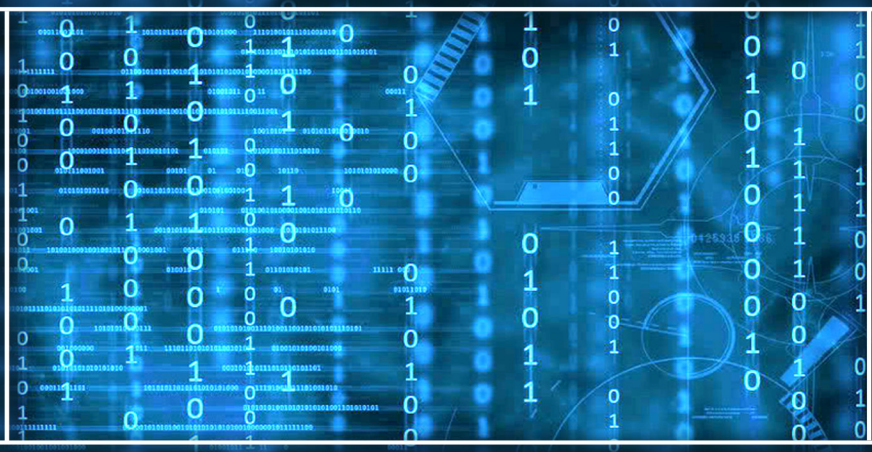
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 16 Issue 6
June 2025



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Kohei Arai
Editor-in-Chief
IJACSA
Volume 16 Issue 6 June 2025
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Editor-in-Chief

Dr. Kohei Arai - Saga University
Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation

---

Associate Editors

Alaa Sheta
Southern Connecticut State University
Domain of Research: Artificial Neural Networks, Computer Vision, Image Processing, Neural Networks, Neuro-Fuzzy Systems

Arun Kulkarni
University of Texas at Tyler
Domain of Research: Machine Vision, Artificial Intelligence, Computer Vision, Data Mining, Image Processing, Machine Learning, Neural Networks, Neuro-Fuzzy Systems

Dr Ronak AL-Haddad
Anglia Ruskin University / Cambridge
Domain of Research : Technology Trends,Communication,Security,Software Engineering and Quality,Computer Networks,Cyber Security,Green Computing,Multimedia Communication,Network Security,Quality of Service

Elena Scutelnicu
"Dunarea de Jos" University of Galati
Domain of Research: e-Learning, e-Learning Tools, Simulation

In Soo Lee
Kyungpook National University
Domain of Research: Intelligent Systems, Artificial Neural Networks, Computational Intelligence, Neural Networks, Perception and Learning

Renato De Leone
Università di Camerino
Domain of Research: Mathematical Programming, Large-Scale Parallel Optimization, Transportation problems, Classification problems, Linear and Integer Programming

Xiao-Zhi Gao
University of Eastern Finland
Domain of Research: Artificial Intelligence, Genetic Algorithms

CONTENTS

Paper 1: Detecting and Preventing Money Laundering Using Deep Learning and Graph Analysis
Authors: Mamunur R Raja, Md Anwar Hosen, Md Farhad Kabir, Sharmin Sultana, Shah Ahammadullah Ashraf, Rakibul Islam
PAGE 1 – 8

Paper 2: Transforming the Working Style of Call Center Agents Through Generative AI
Authors: Satya Karteek Gudipati
PAGE 9 – 15

Paper 3: AI-Driven Education: Integrating Machine Learning and NLP to Transform Child Learning Systems
Authors: Masuma Akter Semi, Md Borhan Uddin, Sharmin Sultana, Motmainna Tamanna, Azim Uddin, Khandakar Rabbi Ahmed
PAGE 16 – 22

Paper 4: Application of Blockchain Frameworks for Decentralized Identity and Access Management of IoT Devices
Authors: Sushil Khairnar
PAGE 23 – 32

Paper 5: Location Based Augmented Reality Navigation Application
Authors: Samridhi Sanjay Pramanik, Aishwary Pramanik
PAGE 33 – 38

Paper 6: Metabolite Screening for Heart Disease Using Support Vector Machine-Based AI
Authors: Edward L. Boone, Ryad A. Ghanam, Faten S. Alamri, Elizabeth B. Amona
PAGE 39 – 47

Paper 7: Enhancing Deepfake Content Detection Through Blockchain Technology
Authors: Qurat-ul-Ain Mastoi, Muhammad Faisal Memon, Salman Jan, Atif Jamil, Muhammad Faique, Zeeshan Ali, Abdullah Lakhan, Toqeer Ali Syed
PAGE 48 – 56

Paper 8: Advancements in Deep Learning for Malaria Detection: A Comprehensive Overview
Authors: Kiswendsida Kisito Kabore, Desire Guel, Flavien Herve Somda
PAGE 57 – 67

Paper 9: Application of Deep Learning-Based Image Compression Restoration Technology in Power System Unstructured Data Management
Authors: Junjie Zha, Aiguo Teng, Xinwen Shan, Hao Tang, Zihan Liu
PAGE 68 – 77

Paper 10: Impact of Auxiliary Information in Generative Artificial Intelligence Models for Cross-Domain Recommender Systems
Authors: Matthew O. Ayemowa, Roliana Ibrahim, Noor Hidayah Zakaria, Yunusa Adamu Bena
PAGE 78 – 89

Paper 11: The Anomaly Detection Algorithm Based on Random Matrix Theory and Machine Learning
Authors: Yongming Lu
PAGE 90 – 95

Paper 12: SkinDiseaseXAI: XAI-Driven Neural Networks for Skin Disease Detection
Authors: Ammar Nasser Alqarni, Abdullah Sheikh
PAGE 96 – 104

Paper 13: Emotion Recognition Algorithm Based on Multi-Modal Physiological Signal Feature Fusion Using Artificial Intelligence and Deep Learning
Authors: Yue Pan
PAGE 105 – 118

Paper 14: Dataset Development for Classifying Kick Types for Martial Arts Athletes Using IMU Devices
Authors: Rudy Gunawan, Suhardi, Widyawardana Adiprawita, Tommy Apriantono
PAGE 119 – 131

Paper 15: Comparative Analysis of Deep Learning Techniques for Passive Underwater Acoustic Target Recognition: Overview, Challenges, and Future Directions
Authors: Song Yifei, Mohamad Farhan Mohamad Mohsin
PAGE 132 – 145

Paper 16: Method for Tea Leaf Plucking Timing Prediction with High Resolution of Images Based on YOLO11
Authors: Kohei Arai, Yoho Kawaguchi
PAGE 146 – 155

Paper 17: The Factors Influencing Internet of Things Adoption in Public Hospitals: A Pilot Study
Authors: Mutasem Zrekat, Othman Bin Ibrahim
PAGE 156 – 168

Paper 18: Research on Network Flow Based on Statistical Analysis Methods
Authors: Maruf Juraev, Inomjon Yarashov, Adilbay Kudaybergenov, Alimdzhan Babadzhanov, Zilolaxon Mamatova
PAGE 169 – 176

Paper 19: A Steel Surface Defect Detection Method Based on Lightweight Convolution Optimization
Authors: Cong Chen, Ming Chen, Hoileong Lee, Yan Li, Jiyang YU
PAGE 177 – 190

Paper 20: Landslide Detection Method Based on Lightweight Convolution and Attention Mechanisms
Authors: Cong Chen, Chengyang Zhang, Ran Chen, Jiyang YU
PAGE 191 – 204

Paper 21: Integration of 2D-CNN and LSTM Networks for Enhanced Image Processing and Prediction in Alzheimer's Disease
Authors: Aya Mohamed Abd El-Hamed, Mohamed Aborizka
PAGE 205 – 212

Paper 22: Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition
Authors: Alaa Alsaeedi, Amal AlMansour, Amani Jamal
PAGE 213 – 226

Paper 23: A Novel Multi-Modal Deep Learning Approach for Real-Time Live Event Detection Using Video and Audio Signals
Authors: Pavadareni R, A. Prasina, Samuthira Pandi V, Ibrahim Mohammad Khrais, Alok Jain, Karthikeyan
PAGE 227 – 240

Paper 24: False News Recognition Model Based on Attention Mechanism and Multiple Features
Authors: Qiongyao Suo, Hongzhen Chang
PAGE 241 – 250

Paper 25: A Reading-Aware Fusion Fact Reasoning Network for Explainable Fake News Detection
Authors: Bofan Wang, Shenwu Zhang
PAGE 251 – 262

Paper 26: Solar-Net: Adaptive Fusion of Spatial-Temporal Features for Resilient Solar Power Generation Forecasting
Authors: Wenqian Su, Jason See Toh Seong Kuan, Xiangyu Shi, Yuchen Zhang
PAGE 263 – 274

Paper 27: Framework for Child Healthcare System Using Random Forest
Authors: Mahesh Ashok Mahant, P. Vidyullatha
PAGE 275 – 283

Paper 28: Enhancing Organizational Threat Profiling by Employing Deep Learning with Physical Security Systems and Human Behavior Analysis
Authors: D. H. Senevirathna, W. M. M. Gunasekara, K. P. A. T. Gunawardhana, M. F. F. Ashra, Harinda Fernando, Kavinga Yapa Abeywardena
PAGE 284 – 293

Paper 29: LaObese: A Serious Game Powered by Analytic Hierarchy Process for Culturally Tailored Childhood Obesity Prevention in Oman
Authors: Nurul Akhmal Mohd Zulkefli, Mukesh Madanan, Zainab Mohammed Al-Nahdi, Jayasree Radhamaniamma
PAGE 294 – 304

Paper 30: Anomaly Study of Computer Networks Based on Weighted Dynamic Network Representation Learning
Authors: Xin Wei
PAGE 305 – 315

Paper 31: Deepfake Audio Detection Using Feature-Based and Deep Learning Approaches: ANN vs ResNet50
Authors: Reham Mohamed Abdulhamied, Sarah Naiem, Mona M. Nasr, Farid Ali Moussa
PAGE 316 – 323

Paper 32: Fine-Tuning OpenAI GPT Chatbot in Western Saudi Dialect: A Case Study of Taibah University
Authors: Maimounah Alhujaili, Ruqayya Abdulrahman
PAGE 324 – 338

Paper 33: Automatic Detection of Natural Disasters Using Faster R-CNN with ResNet50 Backbone
Authors: Shereen Essam Elbohy, Mona M. Nasr, Farid Ali Mousa
PAGE 339 – 348

Paper 34: Fusion of CNN and Transformer Architectures for Proactive Wildfire Detection in Satellite Imagery
Authors: Shereen Essam Elbohy, Mona M. Nasr, Farid Ali Mousa
PAGE 349 – 357

Paper 35: Analyzing the Impact of Robotic Process Automation (RPA) on Productivity and Firm Performance in the Service Sector
Authors: Miftakul Huda, Agus Rahayu, Chairul Furqon, Mokh Adib Sultan, Neng Susi Susilawati Sugiana
PAGE 358 – 363

Paper 36: Systematic Literature Review on Artificial Intelligence-Driven Personalized Learning
Authors: Anas Usman Inuwa, Shahida Sulaiman, Ruhaidah Samsudin
PAGE 364 – 380

Paper 37: Mobile Application Using Convolutional Neural Networks for Preliminary Diagnosis of Rosacea
Authors: Angie Fiorella Sapaico-Alberto, Rosalynn Ornella Flores-Castañeda
PAGE 381 – 387

Paper 38: Hybrid PSO-ACO Optimization for Rice Leaf Disease Classification Using Random Forest and Support Vector Machines
Authors: Avip Kurniawan, Tri Retnaningsih Soeprobowati, Budi Warsito
PAGE 388 – 397

Paper 39: Reinforcement Learning for Real-Time Scheduling in Dynamic Reconfigurable Manufacturing Systems
Authors: Salah Hammedi, Abdallah Namoun, Mohamed Shili
PAGE 398 – 411

Paper 40: A Proposed Framework for Loan Default Prediction Using Machine Learning Techniques
Authors: Mona Aly SharafEldin, Amira M. Idrees, Shimaa Ouf
PAGE 412 – 425

Paper 41: Advanced AI-Driven Safety Compliance Monitoring in Dynamic Construction Environment
Authors: Aisha Hassan, Ali H. Hassan, Yasmin Christensen, Hussain Alsadiq
PAGE 426 – 436

Paper 42: Reinforcement Learning Improves SVM-Driven Algorithms for Classifying Multi-Sensor Data for Medical Monitoring
Authors: Zhiwei Xuan, Yajie Liu
PAGE 437 – 447

Paper 43: Developing an Ontology-Driven and Governance-Integrated Method for Information Dashboard Design
Authors: Ahadi Haji Mohd Nasir, Nik Habibullah Nik Mohd Nizam, Mohd Khairul Maswan Mohd Redzuan, Mohammad Nazir Ahmad
PAGE 448 – 462

Paper 44: Machine Learning and 5G Edge Computing for Intelligent Traffic Management
Authors: Talbi Chaymae, Rahmouni M'hamed, Ziti Soumia
PAGE 463 – 470

Paper 45: A Rule-Based Framework for Clothing Fit Recommendation from 3D Body Reconstruction
Authors: Hamid Ouhnni, Acim Btissam, Belhiah Meryam, Benachir Rigalma, Soumia Zit
PAGE 471 – 479

Paper 46: TL-MC-ShuffleNetV2: A Lightweight and Transferable Framework for Elevator Guideway Fault Diagnosis
Authors: Zhiwei Zhou, Xianghong Deng, Xuwen Zheng, Chonlatee Photong
PAGE 480 – 488

Paper 47: AI-Driven Intrusion Detection Systems for Securing IoT Healthcare Networks
Authors: Muhammad Sajid Nawaz, Muhammad Ahsan Raza, Binish Raza, Manal Ahmad, Farial Syed
PAGE 489 – 497

Paper 48: Leveraged Cognitive Data Analytics and Artificial Intelligence to Enhance Sustainable Agricultural Practices: A Systematic Review
Authors: Wongpanya S. Nuankaew, Patchara Nasa-Ngium, Pratya Nuankaew
PAGE 498 – 507

Paper 49: A Novel Approach for Enhancing Advanced Encryption Standard Performance and Cryptographic Resilience
Authors: Muthu Meenakshi Ganesan, Sabeen Selvaraj
PAGE 508 – 516

Paper 50: An Enhanced LSTM Model Based on Feature Attention Mechanism and Emotional Intelligence for Advanced Sentiment Analysis
Authors: Muhammad Naeem Aftab, Dost Muhammad Khan, Muhammad Zulqarnain, Muhammad Rizwan Akram
PAGE 517 – 528

Paper 51: Image Quality Assessment Based on Feature Fusion and Local Adaptation
Authors: Minjuan GAO, Yankang LI, Xuande ZHANG
PAGE 529 – 536

Paper 52: Advancing Aerodynamic Coefficient Prediction: A Hybrid Model Integrating Deep Learning and Optimization Techniques
Authors: Jad Zerouaoui, Rachid Ed-daoudi, Badia Ettaki, El Mahjoub Chakir
PAGE 537 – 545

Paper 53: AI-Powered Assessment of Resistance to Change in the Context of Digital Transformation
Authors: Bachira Abou El Karam, Tarik Fissaa, Rabia Marghoubi
PAGE 546 – 558

Paper 54: Self-Supervised Method for Risky Situation Detection in Road Traffic Sequences Using Video Masked Autoencoder
Authors: Abdelhafid Berroukham, Mohammed Lahraichi, Khalid Housni
PAGE 559 – 566

Paper 55: Vision-Based Vehicle Classification Using Deep Learning Model
Authors: Ahsiah Ismail, Amelia Ritahani Ismail, Muhammad Afiq Mohd Ara, Asmarani Ahmad Puzi, Suryanti Awang
PAGE 567 – 573

Paper 56: Comparative Analysis of Rank and Roulette Wheel Selection Strategies in Genetic Algorithms for Spatial Layout Optimization
Authors: Najihah Ibrahim, Fadratul Hafinaz Hassan, Sharifah Mashita Syed-Mohamad, Rosmayati Mohemad, Ahmad Shukri Mohd Noor
PAGE 574 – 582

Paper 57: AutiSim: A Virtual Reality Simulation Game Based on the Autism Spectrum Disorder
Authors: Muhammad Aliff Muhd Farid Arfian, Ikmal Faiq Albakri Mustafa Albakri, Faaizah Shahbodin, Mohd Khalid Mokhtar, Asniyani Nur Haidar Abdullah, Norhaida Mohd Suaib, Muhammad Nur Affendy Nor'a, Abdul Hasib Jahidin
PAGE 583 – 593

Paper 58: Critical Success Factors for Knowledge Transfer in Enterprise System Projects: A Theoretical and Empirical Investigation

Authors: Jamal M. Hussien, Riza bin Sulaiman, Ali H Hassan, Mansoor Abdulhak, Hasan Kahtan

PAGE 594 – 600

Paper 59: Content Validity Assessment Using Aiken's V: Knowledge Integration Model for Blockchain in Higher Learning Institutions

Authors: Nur Ilyana Ismarau Tajuddin, Ummu-Hani Abas, Khairi Azhar Aziz, Rozi Nor Haizan Nor, Nor Aziyatul Izni, Muhammad Nuruddin Sudin, Nur Aqilah Hazirah Mohd Anim, Noorashikin Md Noor

PAGE 601 – 608

Paper 60: An Interpretable Transformer-Based Approach for Context-Aware and Stylistically Aligned Academic Paraphrasing

Authors: A. Z. Khan, Ritu Sharma, K. Kiran Kumar, Elangovan Muniyandy, Raman Kumar, Yousef A. Baker El-Ebiary, Prema S, Osama R. Shahin

PAGE 609 – 620

Paper 61: Leveraging LSTM-Driven Predictive Analytics for Resource Allocation and Cost Efficiency Optimization in Project Management

Authors: G. Gokul Kumari, Shokhjakhon Abdufattokhov, Sanjit Singh, Guru Basava Aradhya S, T L Deepika Roy, Yousef A.Baker El-Ebiary, Elangovan Muniyandy, B Kiran Bala

PAGE 621 – 629

Paper 62: AccuLandNet: Enhancing Land Cover Detection with Deep Integrated Learning

Authors: Geetha Guthikonda, M. Senthil Kumaran

PAGE 630 – 637

Paper 63: Graph Neural Networks with Attention Mechanisms for Accurate Dengue Severity Prediction

Authors: Monali G. Dhote, Puneet Thapar, Yousef A. Baker El-Ebiary, G. Indra Navaroj, R. Aroul Canessane, B. V. Suresh Reddy, Elangovan Muniyandy, Kapil Joshi

PAGE 638 – 648

Paper 64: Metaheuristic-Driven Feature Selection for IoT Intrusion Detection: A Hierarchical Arithmetic Optimization Approach

Authors: Jing GUO, Dejun ZHU, Qing XU

PAGE 649 – 659

Paper 65: Enhancing SVM and KNN Performance Through Preprocessing Pipelines for Interactive mHealth Applications

Authors: Btissam Elaziz, Charaf Eddine AIT ZAOUIAT, Mohamed Eddabbah, Yassin LAAZIZ

PAGE 660 – 667

Paper 66: Deep Reinforcement Learning Based Robotic Arm Control Simulation to Execute Object Reaching Task for Industrial Application

Authors: Jon Mark Corrrea, Rudolph Joshua Candare, Junrie B. Matias

PAGE 668 – 675

Paper 67: Explainable Deep Temporal Modeling for Stroke Risk Assessment Using Attention-Based LSTM Networks

Authors: P. Selvaperumal, F. Sheeja Mary, Pratik Gite, T L Deepika Roy, Yousef A. Baker El-Ebiary, Gowrisankar Kalakoti, Sandeep Kumar Mathariya

PAGE 676 – 687

Paper 68: Improving Cross-Lingual Fake News Detection in Indonesia with a Hybrid Model by Enhancing the Embedding Process
Authors: Jihan Nabilah Hakim, Yuliant Sibaroni
PAGE 688 – 696

Paper 69: A Preliminary Study on Songket: A Preservation of Intangible Cultural Heritage
Authors: Nik Siti Fatima Nik Mat, Syadiah Nor Wan Shamsuddin, Syarilla Iryani Ahmad Saany, Norkhairani Abdul Rawi, Julaily Aida Jusoh, Wan Malini Wan Isa, Addy Putra Md Zulkifli, Shahrul Anuwar Mohamed Yusof
PAGE 697 – 702

Paper 70: Utilizing Machine Learning to Identify High-Risk Groups in Sickle Cell Anemia
Authors: Haneen Banjar, Nofe Alganmi, Hajar Alharbi, Ahmed Barefah, Hatem Alahwal, Salwa Alnajjar, Abdulrahman Alboog, Salem Bahashwan, Galila Zaher
PAGE 703 – 713

Paper 71: Real Time Accident Detection and Emergency Response Using Drones, Machine Learning and LoRa Communication
Authors: Bandara H. M, Maduhansa H. K. T. P, Jayasinghe S. S, Samararathna A. K. S. R, Harinda Fernando, Shashika Lokuliyana
PAGE 714 – 722

Paper 72: Cybersecurity and the NIST Framework: A Systematic Review of its Implementation and Effectiveness Against Cyber Threats
Authors: Juan Luis Salas-Riega, Yasmina Riega-Virú, Mario Ninaquispe-Soto, José Miguel Salas-Riega
PAGE 723 – 735

Paper 73: Hybrid Detection Framework Using Natural Language Processing (NLP) and Reinforcement Learning (RL) for Cross-Site Scripting (XSS) Attacks
Authors: Carlo Jude P. Abuda
PAGE 736 – 747

Paper 74: Evaluation Index System for Environmental Restoration Effectiveness Based on Landscape Pattern and Ecological Low-Carbon Construction
Authors: Jingyuan Mao
PAGE 748 – 754

Paper 75: Application Analysis and Research of Text Model Based on Improved CNN-LSTM in the Financial Field
Authors: Jing Chen, Chensha Li
PAGE 755 – 767

Paper 76: Ensuring Consistency in Group Decision Making: A Systematic Review of the FWZIC Method
Authors: Ghazala Bilquise, Samar Ibrahim
PAGE 768 – 792

Paper 77: Design and Implementation of Low-Cost Hybrid-Controlled Smart Wheelchair Based on PID Control Integrated with Vital Signs Monitoring
Authors: M. Sayed, MG Mousa, Ali A. S, T. Mansour
PAGE 793 – 803

Paper 78: Design and Implementation of an Intelligent Laboratory Management System Based on UWB Technology
Authors: Heng Sun, Qiang Gao
PAGE 804 – 811

Paper 79: RFID Integration with Internet of Things: Data Processing Algorithm Based on Convolutional Neural Network
Authors: Liang Wang
PAGE 812 – 820

Paper 80: A Technique to Support Incremental Construction and Verification in Component-Based Software Development
Authors: Faranak Nejati, Ng Keng Yap, Abdul Azim Abd Ghani
PAGE 821 – 831

Paper 81: Real-Time Video Captioning on CPU and GPU: A Comparative Study of Classical and Transformer Models
Authors: Othmane Sebban, Ahmed Azough, Mohamed Lamrini
PAGE 832 – 840

Paper 82: Endometriosis Lesion Classification Using Deep Transfer Learning Techniques
Authors: Shujaat Ali Zaidi, Varin Chouvatut, Chailert Phongnarisorn
PAGE 841 – 853

Paper 83: LASSO-Based Feature Extraction with Adaptive Windowing via DTW for Fault Diagnosis in Rotating Machinery
Authors: Jirayu Samkunta, Patinya Ketthong, Nghia Thi Mai, Md Abdus Samad Kamal, Iwanori Murakami, Kou Yamada, Nattagit Jiteurtragool
PAGE 854 – 860

Paper 84: The Influence of Familiarity with Traffic Regulations on Road Safety: A Simulated Study on Roundabouts and Intersections
Authors: Raghda Alqurashi, Hasan J. Alyamani, Nesreen Alharbi, Hasan Sagga
PAGE 861 – 868

Paper 85: Predicting Jobs, Shaping Economies: Bibliometric Insights into AI and Big Data in Workforce Demand Analysis
Authors: EL Massi Fouad, ELouadi Abdelmajid
PAGE 869 – 879

Paper 86: Dynamic Polygon-Based Reverse Driving Detection Technique for Enhanced Road Safety
Authors: Tara Kit, Youngsun Han, Anand Nayyar, Tae-Kyung Kim
PAGE 880 – 887

Paper 87: Comparative Analysis of Machine Learning Frameworks for Robust Ovarian Cancer Detection Using Feature Selection and Data Balancing
Authors: DSS LakshmiKumari P, Maragathavalli P
PAGE 888 – 897

Paper 88: Comparative Study of Prenatal and Postnatal Images for Detecting Down Syndrome in Children
Authors: Labanti Singha, Iqbal Ahmed
PAGE 898 – 905

Paper 89: Sign3DNet: An Enhanced 3D CNN Architecture for Bengali Word-Level Sign Language Recognition
Authors: Safi Ullah Chowdhury, Nasima Begum, Tanjina Helaly, Rashik Rahman
PAGE 906 – 915

Paper 90: Integrating Blockchain and Smart Card Technologies for Secure Healthcare Data Management
Authors: Zayneb Gaouzi, Imad Bourian, Khalid Chougdali
PAGE 916 – 923

Paper 91: Enhanced Feature Extraction for Accurate Human Action Recognition
Authors: Tarek Elgaml, Ali Saudi, Mohamed Taha
PAGE 924 – 930

Paper 92: Deep Learning in Cephalometric Analysis: A Scoping Review of Automated Landmark Detection
Authors: Idriss Tafala, Fatima-Ezzahraa Ben-Bouazza, Aymane Edder, Oumaima Manchadi, Bassma Jioudi
PAGE 931 – 942

Paper 93: Advancing Traffic Sign Detection with Convolutional Neural Networks: A Deep Learning Approach
Authors: OUAHBI Younesse, ZITI Soumia
PAGE 943 – 950

Paper 94: MITG-CU: Multimodal Interaction Temporal Graphs Approach for Conversational Emotion Recognition
Authors: Qian Xing, Yaqin Qiu, Minglu Chi, Xuewei Li, Changyi Gao
PAGE 951 – 959

Paper 95: Cross-Domain Evaluation of Large Language Models for Abstractive Text Summarization: An Empirical Perspective
Authors: Walid Mohamed Aly, Taysir Hassan A. Soliman, Amr Mohamed AbdelAziz
PAGE 960 – 973

Paper 96: Foreign Key Constraints to Maintain Referential Integrity in Distributed Database in Microservices Architecture
Authors: Shamsa Kanwal, Nauman Riaz Chaudhry, Reema Choudhary, Younus Ahamad Shaik, Pankaj Yadav, Ayesha Rashid
PAGE 974 – 985

Paper 97: Integrating cGAN-Enhanced Prediction with Hybrid Intervention Recommendations Systems for Student Dropout Prevention
Authors: Hassan Silkhi, Brahim Bakkas, Khalid Housni
PAGE 986 – 995

Paper 98: Habitat Intelligence: How Machine Learning Reveals Species Preferences for Ecological Planning and Conservation
Authors: Meryem Ennakri, Soumia Ziti, Mohamed Dakki
PAGE 996 – 1004

Paper 99: Deep Learning-Based Bone Age Growth Disease Detection (BAGDD) Using RSNA Radiographs
Authors: Muhammad Ali, Muhammad Faheem Mushtaq, Saima Noreen Khosa, Naila Kiran, Humaira Arshad, Urooj Akram
PAGE 1005 – 1018

Paper 100: Intelligent Agents in Disaster Risk Management: A Systematic Review of Advances and Challenges
Authors: Hssaine Hamid, ELouadi Abedlmajid
PAGE 1019 – 1029

Paper 101: Artificial Intelligence in Disaster Risk Management: A Scientometric Mapping of Evolution, Collaboration, and Emerging Trends (2003–2025)
Authors: Hssaine Hamid, ELouadi Abedlmajid
PAGE 1030 – 1040

Paper 102: XPathia: A Deep Learning Approach for Translating Natural Language into XPath Queries for Non-Technical Users
Authors: Karam Ahkouk, Mustapha Machkour
PAGE 1041 – 1050

Paper 103: Random Forest Model Based on Machine Learning for Early Detection of Diabetes
Authors: Inooc Rubio Paucar, Cesar Yactayo-Arias, Laberiano Andrade-Arenas
PAGE 1051 – 1063

Paper 104: Phishing Simulation as a Proactive Defense: A Customizable Platform for Training and Behavioral Analysis
Authors: Abdulrahman Alsaqer, Hussain Almajed, Khalid Alarfaj, Mounir Frikha
PAGE 1064 – 1075

Paper 105: Reducing Computational Complexity in CNNs: A Focus on VGG19 Pruning and Quantization
Authors: Md. Mijanur Rahman, Anik Datta, Md. Sabiruzzaman, Md Samim Ahmed Bin Hossain
PAGE 1076 – 1087

Paper 106: Fake News Detection on Kashmir Issue Using Machine Learning Techniques
Authors: Misbah Kazmi, Sadia Nauman, Sadaf Abdul Rauf, S. Ali, Ali Daud, Bader Alshemaimri
PAGE 1088 – 1095

Paper 107: Text Classification Using Enhanced Binary Wind Driven Optimization Algorithm
Authors: Jaffar Atwan, Mohammad Wedyan, Ahmad Hamadeen, Qusay Bsoul, Ayat Alrosan, Ryan Alturki
PAGE 1096 – 1115

Paper 108: Comparison of Conventional Techniques for House Electricity Consumption Forecasting
Authors: Sandra Pajares Centeno, Hugo Alatrista-Salas
PAGE 1116 – 1122

Detecting and Preventing Money Laundering Using Deep Learning and Graph Analysis

MAMUNUR R RAJA[1], Md Anwar Hosen[2], Md Farhad Kabir[3], Sharmin Sultana[4], Shah Ahammadullah Ashraf[5], Rakibul Islam[6]*

Master of Science in Information Technology (MSIT), Westcliff University, Irvine, CA, USA[1]
College of Business, Westcliff University, Irvine, CA, USA[2]
Marshall School of Business, University of Southern California, Los Angeles, USA[3]
Dept. of Business Administration and Management, International American University, Los Angeles, CA, USA[4]
School of Business, International American University, Los Angeles, CA, USA[5]
Dept. of MBA in Business Analytics, International American University, Los Angeles, CA, United States[6]

***Abstract*—Money laundering is a major worldwide issue facing financial organizations, with its increasingly complicated and changing methods. Conventional rule-based anti-money laundering (AML) systems can fail to identify advanced fraudulent activity. This study shows a new hybrid model to detect suspicious transaction patterns precisely by efficiently combining GraphSAGE, a graph-based Machine Learning (ML) technique, with Long Short-Term Memory (LSTM) networks. The suggested approach uses GraphSAGE's relational capabilities for graph-structured anomaly detection and the temporal strengths of LSTM for sequence modeling. With excessive traditional ML and stand-alone Deep Learning (DL) techniques, the Hybrid LSTM-GraphSAGE model achieves an accuracy of 95.4% using a simulated dataset reflecting real-world financial transactions. The findings show how well our combined strategy lowers false positives and improves the identification of advanced AML operations. This work opens the path for creating real-time, intelligent, flexible money laundering detection systems appropriate for current financial situations.**

***Keywords*—Anti-money laundering (AML); deep learning (DL); LSTM; GraphSAGE; graph analysis; transaction monitoring; hybrid fusion model**

I. Introduction

In the previous few years, money laundering and terrorism have been among the main challenges to the integrity of the global financial system [1]. Money laundering issues the reliability of illegal activity by hiding its source; each year, the laundering of around 2 to 5% of the world's GDP (1.7-4 trillion) leads to issues [2]. Underlying crimes include drug distribution, human trafficking, fraud, tax avoidance, and corruption. Consequently, money laundering is a serious worldwide issue influencing individuals, businesses, governments, and societal welfare and impacting [3].

For financial institutions (FIs), undetectable money laundering programs may cause significant reputation harm and considerable penalties. FIs use compliance professionals looking at questionable activity to stay out of a vehicle for money laundering [4]. According to most rule-based algorithms and human monitoring, traditional AML systems struggle to keep up with ever-advanced laundering procedures [5]. These techniques often skip subtle, nonlinear developments in transactional data or expose latent connections within multi-layered financial networks—a gap leaves organizations open to systematic risk, reputational harm, and regulatory fines [6].

Although nowadays, most methods concentrate on isolated transaction analysis using shallow models like logistic regression or decision trees, recent developments in ML have shown promise in spotting aberrant financial activity [7]. These approaches ignore the essentially relational character of money laundering, in which illegal activity is entwined in complex systems of companies and transactions. The graph-based analysis provides a strong prism to find these latent structures. It maps links between accounts, beneficiaries, and intermediaries [8]. However, few researchers have successfully coupled graph theory with DL to address financial crime's dynamic, high-dimensional character [9].

This study concludes this gap by indicating a novel approach combining graph analysis with DL to identify and terminate money laundering in real-time. Our method models transactional networks using GNNs' hierarchical representation learning ability, thereby capturing local node attributes and global topological patterns. We also provide a dynamic anomaly detection system that continually updates network embeddings and improves risk ratings to fit changing laundering strategies.

Our key contributions include:

- This study provides a novel money laundering detection method that combines graph-based analytical structural insights with a DL model temporal pattern recognition. This interaction helps the system detect relational and sequential flaws in financial transactions.
- Real-time transaction monitoring system design is a key advancement. This system may identify suspicious activity before it expands by continually recording and analyzing transactions using past behavioral patterns and changing network architecture.
- Recognizing a shortage of labeled AML data, we develop a careful and logical financial transaction set. This allows practical training and evaluation of

the recommended models and reflects actual laundering behavior.

– This study thoroughly evaluates the hybrid model against traditional ML and stand-alone DL. Results show that the hybrid model outperforms others in detection accuracy, precision, recall, and false positive reduction.

II. Literature Review

This study examines graph-based and hybrid DL models for AML, focusing on GCNs, CNNs, and new architectures like GAGNN and Temporal-GCN. Data availability, computational cost, and cross-domain transferability remain issues even with high accuracy and minimal false positives.

Bakhshinejad et al. [10] presented a graph-based DL model for suspected money laundering transaction detection and a thorough analysis of current AML systems from a data-oriented standpoint. Applied node2vec for feature extraction, created a detection system converting transactional data into a graph, and then classified transactions as usual or suspicious using a GCN. Graph embedding with GCN for classification, Node2Vec. They also tuned classifier thresholds and managed unbalanced data using SMOTE. Comparatively, to industry norms of 90% or more, their model attained very low false negative rates, sometimes even nil, and significantly lowered the false alarm rate to around 50%.

Also, Irshad et al. [11] proposed a novel framework for identifying money laundering activities by combining Graph Convolutional Networks (GCN), Convolutional Neural Networks (CNN), and Feed-Forward Neural Networks (FFNN) in an Integrated Approach for Money Laundering Detection. Using spatial patterns, sequential data, and transaction network topology, the authors created a hybrid model to raise classification performance. Their approach calls for CNN to extract local characteristics from transaction histories, GCN to capture graph-based relations among entities, and FFNN for ultimate classification. With an astounding 98.34%, the model exceeded conventional ML techniques. Likewise, Kute et al. [1] provided an extensive CNN sentiment analysis application overview.

They evaluate current research publications using CNN architectures for sentiment classification challenges in many fields and datasets. They gathered and analyzed more than 60 research studies, evaluating many CNN-based techniques, including hybrid models, multichannel CNNs, and improvements in attention processes. They also examined preprocessing methods and dataset kinds used in this research.

Their stated accuracy falls between 75% and 90%, depending on the architecture and dataset. CNN-based models were better than conventional ML techniques in extracting spatial characteristics from text. Cheng et al. [12] suggested a fresh approach using group-aware deep graph learning methods for AML. The authors model the financial transaction network as a heterogeneous graph, and they present a Group Aware Graph Neural Network (GAGNN) to detect suspicious group behaviors often disregarded in conventional AML systems. Their approach consists of building a heterogeneous transaction graph, grouping based on shared characteristics (such as IP addresses), and using a tailored GNN architecture with group-level elements for enhanced detection. Reaching an AUC of 0.9814 and an F1 score of 0.8607, the suggested method shows performance improvement over baseline models. Dumitrescu et al. [13] represented users or accounts and edges indicating transactions, therefore investigating banking transaction data as a graph of fraudulent activities. They devised a method to identify structural and behavioral irregularities in the transaction network using GNNs.

Using GNN models, especially Autoencoders and GCNs, they train representations of the graph and spot suspicious trends. Focusing on unsupervised methods, as tagged fraudulent data is not readily available, they assessed their models using real-world financial transaction data. Depending on the model and data setup, their method's stated accuracy in AUC (Area Under the Curve) scores varied from 0.76 to 0.89. Eddin et al. [3] created dynamically via sliding time windows to help propose an ML triage model to lower false positives in AML systems. Combining entity-centric characteristics with graph-based features.

Using LightGBM on actual banking data, their approach reduced false positives by 80% and identified over 90% of genuine positives. Furthermore, Jensen et al. [14] investigated using ML and statistical techniques to fight money laundering, emphasizing using synthetic data to train prediction models. They used synthetic data reflecting banking activities and consumer characteristics to develop and evaluate their method. To separate dubious from non-suspicious consumers, they used a supervised learning approach using a gradient-boosted decision tree algorithm (LightGBM).

With a fantastic accuracy of 99.6%, their model demonstrated the possible efficiency of ML in spotting financial laundering activity. Alarab et al. [15] employed a unique graph-based model called Temporal-GCN, which combines GCN with LSTM, to identify illegal transactions in Bitcoin. They also included active learning using Monte-Carlo Dropout and Monte-Carlo Adversarial Attack (MC-AA) for uncertainty estimates and created a framework that catches both temporal sequences and graph topologies of transaction data.

Preprocessing Bitcoin transaction graphs, extracting local characteristics, using LSTM to model temporal trends, and then feeding the result to a TAGCN layer from the approach. The outperformance of the model above previous GCN-based models on the same dataset resulted in a classification accuracy of 97.77% and an F1-score of 80.6%. Muminovic et al. [16] investigated the difficulties of money laundering in the digital age, along with studies of contemporary technologies meant to improve preventive systems.

It especially looks at how graph databases may monitor intricate, nonlinear financial transactions often used to hide illegal activity—which are typically used to hide illegal activity. The writers review current methods that mainly depend on relational databases and rule-based detection, pointing to limits in scalability and adaptability.

Graph-based models help see entities and their transactional interactions more easily, enhancing the capacity to spot unusual

trends. Mohan et al. [17] proposed a hybrid model combining Evolving Graph Convolutional Networks (EvolveGCN) with Deep Neural Decision Forests (DNDF) to handle AML in the Bitcoin network.

With the Elliptic dataset comprising over 200,000 labeled and unlabeled Bitcoin transactions, they simulate the issue as a node classification job. They want to improve the categorization of illegal transactions by combining dynamic graph learning with ensemble techniques. Following Knowledge Distillation (KD) to compress and maximize the model, the model attained a high F1 score of 0.9251, which then improved to 0.9525.

Moreover, there is limited comparative analysis using traditional methods in detecting and Preventing Money Laundering.

1) Many models need synthetic datasets or large volumes of classified transaction data, which are rare in financial environments. While transaction categorization is time-consuming and error-prone, synthetic data lacks the complexity of real-world laundering techniques.

2) Graph-based algorithms like GCN, GAGNN, and Temporal-GCN are resource-intensive in describing large, dynamic financial transaction networks. Training, feature extraction, and graph creation are computationally intensive.

3) These models' efficacy slightly depends on group or graph structure specifications and transaction data correctness and completeness. Models include unsupervised GNNs and GAGNNs that lose much in noisy, mislabeled, or poorly linked data.

4) Although many models perform well on specific datasets, e.g., Bitcoin and private financial data, their transferability across institutions, countries, and transaction types is poor. Applied to foreign fields, model performance is usually harmed.

III. Methodology

A hybrid DL-based and graph-based analysis framework was proposed to detect and prevent money laundering activities effectively. As shown in Fig. 1, raw transaction data is fed into the pipeline throughout the methodology, producing final classification outcomes. The design of this multi-stage framework is to extract temporal transactional and relational patterns, typically indicative of money laundering behavior.

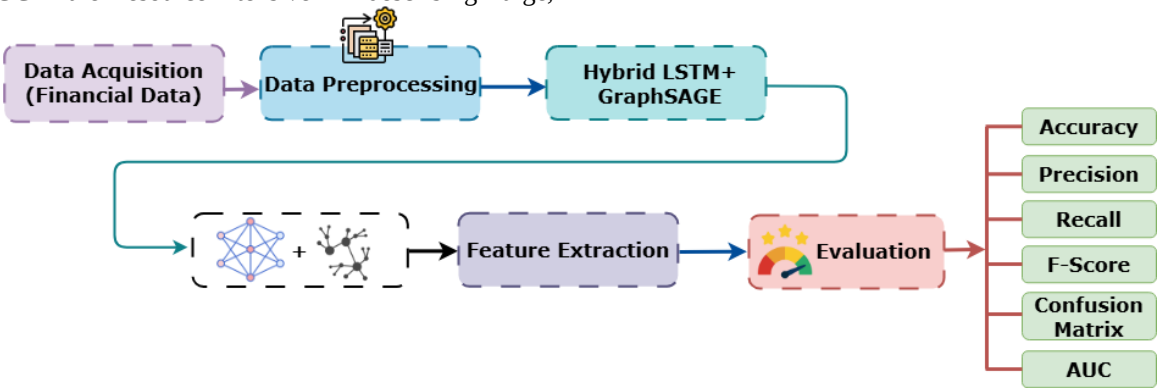


Fig. 1. Methodology framework for detecting and preventing money laundering using the proposed hybrid LSTM-GraphSAGE model.

A. Dataset Description

For this study, experiments are run using the Anti Money Laundering Transaction Data (SAML-D) on Kaggle by Berk Öztaş [18]. The dataset was specially developed to simulate real banking transactions and behavioral patterns of money laundering operations. The dataset contains 1,048,575 records of transactions with 28 typologies (split between 11 normal and 17 suspicious) and 12 different features: Time, Date, Sender_account, Receiver_account, Amount, Payment_currency, Received_currency, Sender_bank_location, Receiver_bank. The dataset consists of each transaction record between a sender and a receiver on each row. Payment_currency, Received_currency, and Amount depict all types of currency used at the sender's and receiver's end, as well as the total monetary value of the transfer.

Regarding transaction medium, the Payment_type would tell you that it is an online transfer, credit card, or wire. After supervised learning, the label Is_laundering is a critical binary feature (1 for money laundering and 0 for legitimate). Laundering_type gives a multi-class annotation regarding laundering techniques. The rich set of features allows for temporal and relational modeling of temporal patterns of financial fraud.

B. Dataset Preprocessing

A series of robust preprocessing steps was implemented to prepare the dataset for DL and graph analysis models.

1) Timestamp handling. The Date and Time features were merged and transformed into a standard UNIX timestamp format to assist temporal modeling. It does this feature engineering for time-based behaviors like peak transaction hours or clusters of frequency. By Di and Ti let us denote the date and time when $i\ th$ transaction occurs. As follows, the timestamp feature TSi was derived:

$$TS_i = to_timestamp(D_i + T_i)$$

2) Account ID encoding. The Sender_account and Receiver_account are anonymized strings. They were labeled, encoded, or embedded to preserve identity without revealing personal information. This later allows for building transaction graphs where each unique account becomes a node:

$$Sender_i = f(Sender_account_i)$$

$$Receiver_i = f(Receiver_account_i)$$

3) Categorical feature encoding. For some payment information (Payment_currency, Received_currency, Payment_type, Sender_bank_location, Receiver_bank_location), those were one-hot encoded or embedded, based on the model. In this case, the transformation to capture the discrete attribute information does not introduce the ordinal bias. For a categorical feature X with k unique values:

$$X \in \{x_1, x_2 \dots x_k\} \Rightarrow One - hot(X) = [0,1,2, \dots]$$

4) Graph construction for network analysis. Sender_account and Receiver_account were used as nodes and transactions as edges in a directed transaction graph $G = (V,)$. Also, each edge has attributes such as amount, time, payment type, and binary labels for laundering. Optional aggregation was based on the frequency or total amount of the edge weights:

$$w_{ij} = \sum_{t \in T_{ij}} Amount_t$$

5) Label binarization and data splitting. The binary target for classification is given by the Is_laundering column. Also, Laundering_type was encoded for multi-class classification or more profound insight into techniques. This ensures label stratification and thus handles the class imbalance in the data.

Finally, the dataset was split into training and testing subsets through an 80/20 ratio. This guaranteed that all the classes undergo the model evaluation phase, including classes that do not undergo laundering.

C. Models

In order to detect and prevent money laundering in financial transaction systems, we proposed a hybrid DL framework that ties the temporal behavior modeling and relational graph analysis together, which captures the evolution of suspicious activity with complex pattern development. Specifically, our approach is composed of three key parts: 1) an LSTM for learning sequential behavioral characteristics, 2) a graph neural network (GNN) based on GraphSAGE for modeling structural relationships, and 3) a fusion model to synthesize the temporal and structural beneficial insights.

1) Temporal behavior modeling using LSTM. Structured and repetitive behavior of money laundering schemes commonly entails rapid distribution of funds through an account network (smurfing) or calculating the best technique for the layering (i.e., how to add coins to comply with the internal allowance of the broker). Because these are nuanced temporal dependencies, we leverage an LSTM network capable of modeling sequential data with long-range dependencies. The LSTM branch in Fig. 2 extracts transaction sequences from an accounts perspective to capture behavioral anomalies.

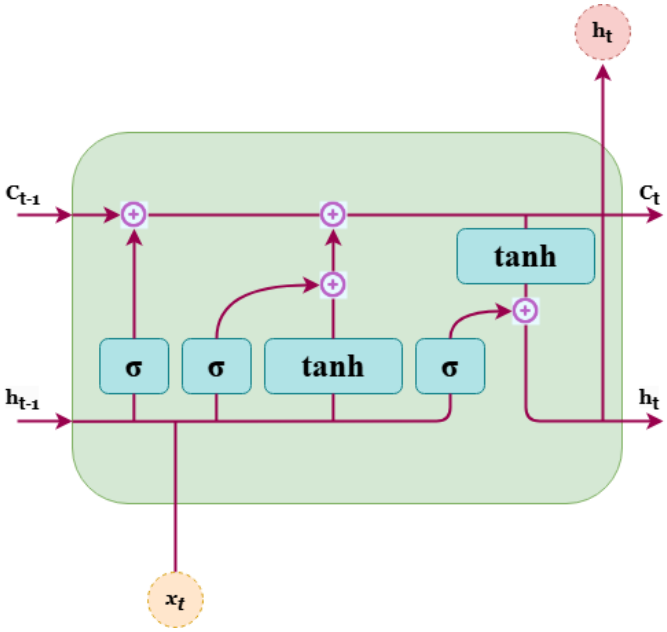


Fig. 2. LSTM-based temporal feature extraction architecture.

Let $X_a = \{x_{a,1}, x_{a,2}, \dots, x_a, T_a\}$ be the transaction sequence for account a, and each transaction $x_{a,t} \in R_d$ is a vector of features. The features include normalized transaction amount, payment type, time-based metadata (e.g., hour of day, day of week), typology, currency mismatch, and geographical information (e.g., bank location).

The LSTM network operates on this sequence through a series of gates and updates on internal memory according to the following equations:

(i) Forget Gate: $f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$

(ii) Input Gate: $i_t = \sigma(W_f x_t + U_i h_{t-1} + b_i)$

(iii) Output Gate: $f_t = \sigma(W_0 x_t + U_0 h_{t-1} + b_0)$

(iv) Candidate Memory: $\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c)$

(v) Cell State Update:$c_t = f_c \Theta c_{t-1} + i_t \Theta \tilde{c}_t$

(vi) Hidden State: $h_t = o_t \Theta tanh(c_t)$

We denoted σ to be the sigmoid activation function, tanh to be the hyperbolic tangent function, and $\odot$ to denote elementwise multiplication. The contextual behavior of account a at time $t$is encapsulated in ht as both current transaction features and past behavioral history. The final hidden state hT is passed through a fully connected dense layer with a sigmoid activation to classify the suspiciousness of the account through behavioral analysis:

$$\hat{y}_a = \sigma(w_0 h_T + b_0)$$

The binary cross-entropy loss function is used to optimize the LSTM model:

$$L_{LSTM} = -ylog(\hat{y}) - (1 - y)\log(1 - \hat{y})$$

This formulation can learn from genuine and fraudulent behavioral sequences and effectively identifies time-sensitive laundering patterns.

2) Graph-based relational modeling using GraphSAGE. Behavioral patterns give us some clues as to when things happened, but many money laundering schemes are networks consisting of many, possibly thousands, of accounts, and the connections among them are equally complex. Indirect transfers, circular money flows, and multi-hop layering are typical ways that launderers try to forfeit the trail of illegal money. The problem could be captured as a directed graph where the nodes represent the accounts, and the edges between the nodes indicate transactions between accounts.

Suppose this transaction graph to be represented by $G = (V, E)$, where:

- The set contains N resources equal to V, where each node (account) in V is unique.
- The set of directed edges, E, will represent transactions.
- Edge features consist of transaction amount, transaction type, time interval between transactions, and currency mismatch for each edge $euv \in E$ between node u and node v.

For each node $v \in V$, we initialize hv, the feature vector for node v, as a feature vector of the form, comprising:

- Mean, max, and count of transaction amounts (aggregated statistics).
- Frequency metrics.
- Some categorical features can be embedded (e.g., typology, payment type).
- And optionally, the output of the LSTM model.

For learning meaningful node embeddings over its neighborhood structure, we adopted GraphSAGE, an inductive GNN framework that can generalize to node discovery. Layer-wise, the node embedding is updated using:

$$h_{N(v)}^{(k)} = AGGREGATE^{(k)}\left(\left\{h_u^{(k-1)} | u \epsilon N(v)\right\}\right)$$

$$h_v^{(k)} = \sigma(W^{(k)}.CONCAT(h_v^{(k-1)}, h_{N(v)}^{(k)}))$$

The neighbors of node v are shown here by the notation (v), and AGGREGATE is a differentiable function like mean, max-pooling, or LSTM-based spreading. After K layers, the final node embedding $hv\ K$ is computed and gives the structural context of account v. Finally, the final node embeddings are classified with a sigmoid-activated dense layer:

$$\hat{y}_v = \sigma(W_c h_v^k + b_c)$$

It enables us to create a suspiciousness score for each account by comparing its transactional connections with other accounts. We extended this in practice by assigning sender and receiver embeddings and edge features simultaneously.

3) Fusion model. Combining Temporal and Relational Knowledge, we recognized that the temporal and structural features give complementary information and thus designed a hybrid fusion model that combines the outputs from the LSTM and GNN branches. Such a combination helps the model detect sophisticated money laundering tactics that may transpire through behavioral anomalies and relational inconsistencies. Fig. 3 depicts the overall architecture of such a system as a single hybrid fusion model that combines the sequential and relational learning components to perform comprehensive AML detection.

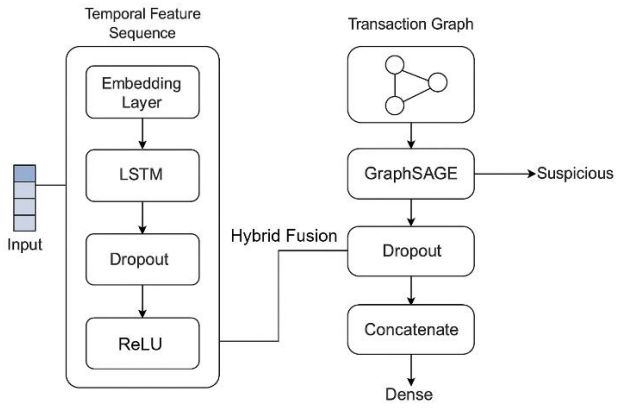


Fig. 3. Architecture of the proposed hybrid fusion model for money laundering detection.

We compute a fused feature representation for each account v by concatenating its final LSTM hidden state $h(v)$ with the graph embedding of that account hvK:

$$z_v = CONCAT(h_T^{(v)}, h_v^K)$$

The MLP with nonlinear activation functions (e.g., ReLU) and dropout regularization is applied on this fused vector zv. Finally, the following is given as the final classification:

$$\hat{y}_v = \sigma(W_f z_v + b_f)$$

We defined a joint loss function to train this end-to-end architecture from two parts of the loss functions in the LSTM and GNN models:

$$L_{total} = \lambda_1 L_{LSTM} + \lambda_2 L_{GNN} + \lambda_3 L_{Fusion}$$

Here, λ_1, λ_2, λ_3 are hyperparameters that control the contribution of each loss term. All the model portions are trained using the Adam optimizer with early stopping and validation-based performance monitoring.

IV. Result and Discussion

A. Performance Evaluation

In order to evaluate the effectiveness of various models in predicting money laundering activities, we performed a deep analysis of classification metrics, namely accuracy, precision, recall, and F1-score. The LSTM baseline model, as summarized in Table I, resulted in an accuracy of 91.5%, precision of 89.4%, recall of 90.2%, and F1-score of 89.8%. This indicates a relatively strong capability in temporal pattern recognition, which is necessary for sequential transaction data. Yet, across all metrics, the LSTM model performed worse compared to the GraphSAGE model, which is a devised model to make full use of topological and relational patterns in transaction graphs, due to achieving 92.8%, 91.1%, 91.8%, and an F1-score of 91.4%.

Improvements indicate that capturing relationships in graph dimensions is important, given that such flows tend to have gone through interconnected accounts and richly interlinked networks.

TABLE I DIFFERENT PERFORMANCE METRICS FOR VARIOUS MODELS IN MONEY LAUNDERING DETECTION

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LSTM	91.5	89.4	90.2	89.8
GraphSAGE	92.8	91.1	91.8	91.4
Proposed Hybrid LSTM GraphSAGE	**95.4**	**93.2**	**93.8**	**93.5**

Moreover, the proposed Hybrid LSTM-GraphSAGE model outperformed the individual ones significantly with 95.4 % accuracy, 93.2 % precision, 93.8 % recall, and a robust F1-score of 93.5 %. This synergy of alternating structural learning with temporal dependencies expressed by LSTM indicates that incorporating temporal dependencies from LSTM contributes to increasing GraphSAGE's performance.

This hybrid model not only improves detection capabilities due to discriminative temporal and topological patterns as well as balance across all key evaluation metrics, but it also outweighs competitors across all metrics by about 17.8% across group operation and 46.3% in total execution time. Combining sequence modeling and graph-based learning confirms that one can achieve a deeper, more accurate understanding of money laundering detection.

B. Training and Validation Performance Analysis

To further validate the robustness of the proposed Hybrid LSTM-GraphSAGE model on the learning behavior and generalization capability, we looked at the training and validation performance at 50 epochs. Fig. 4 shows the development of the training and validation accuracy. Both accuracy curves have a steep (learn quickly) upward trajectory in the first few epochs.

The model improves steadily with training and has a training accuracy of around 98% and a validation accuracy of about 95%. In the later epochs, this shows that these two curves are converging, meaning the model has no overfitting and is very well generalizing to unseen data.

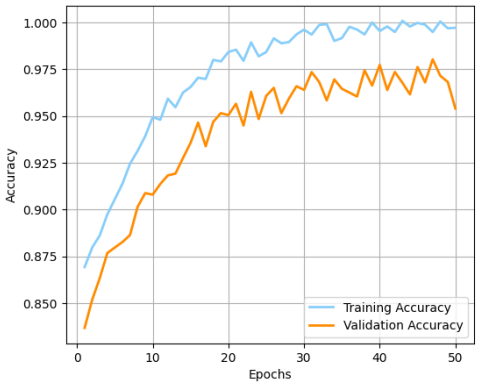


Fig. 4. Training and validation accuracy of the proposed hybrid fusion model.

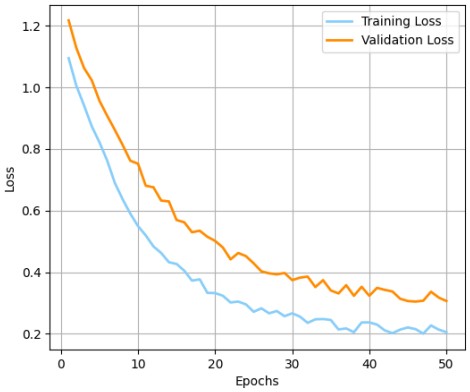


Fig. 5. Training and validation loss of the proposed hybrid fusion model.

Fig. 5 shows training and validation loss curves. As we can see during the initial training phase, the loss values experience a sharp decline for both losses, which means we have successfully learned key patterns in our dataset. The training loss decreases steadily to below 0.1, while the validation loss converges to around 0.2, with no apparent signs of divergence. This stable convergence behavior further reinforces the robustness of the proposed hybrid fusion architecture, which integrates temporal and structural learning components.

Lastly, the overall training and validation performance analysis shows that the proposed model is well optimized and free from major problems such as the issue of underfitting, overfitting, and so on, which makes it a good candidate for implementation in a money laundering detection system in the real world.

C. Error Analysis

The given classification performance was also analyzed at the level of error using the confusion matrix, as shown in Fig. 6, to fully understand the detection performance of the proposed Hybrid LSTM-GraphSAGE model. The matrix gives some insight into the number of correct and wrong predictions of the transaction, either legitimate or money laundering. In a total of predictions, the model made 100,173 correct fraud (True Positives) and 99,895 correct legitimate predictions (True Negatives). However, 4799 of these were incorrectly classified as fraudulent (False Positive (FP), and 4688 as legitimate (False Negative (FN)).

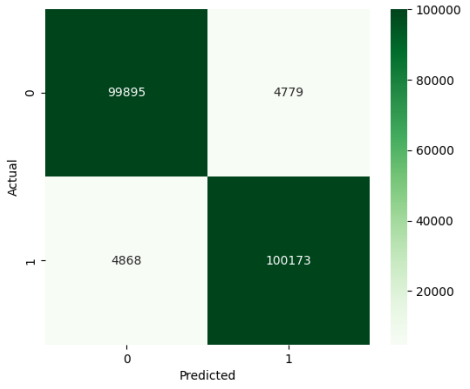


Fig. 6. Confusion matrix for the proposed hybrid fusion model in detecting money laundering.

Nevertheless, with a large dataset and a common class imbalance found in financial datasets, the proposed model achieved a high level of predictive accuracy with tolerable margins of error. It is imperative in financial systems, where being flagged as a fraudulent user when it is not could raise customer irritation or unnecessary investigation on legitimate users. However, the small number of false negatives guarantees that almost all illicit activities are flagged for awareness. It shows that the hybrid model is very good at discriminating between normal and suspicious behaviors.

D. ROC Curve and AUC Analysis

The Receiver Operating Characteristic (ROC) curve and corresponding AUC give a graphical and quantitative way to measure model performance at different threshold settings. The ROC curves of the LSTM, the GraphSAGE, and the proposed Hybrid LSTM-GraphSAGE are plotted with a random guess performance baseline (Fig. 7).

This LSTM model performs strongly in capturing sequential dependencies in the transaction data by achieving an AUC of 0.94. LSTM slightly underperformed GraphSAGE with an AUC of 0.95 because GraphSAGE was able to learn structural patterns in the transaction network. Despite that, the proposed Hybrid LSTM-GraphSAGE model achieves the best AUC of 0.97 compared with individual models. By combining temporal and topological learning, the strength of this improvement suggests that combining temporal and topological learning enables the model to achieve better discrimination for positive (money laundering) compared to negative (legitimate) classes for all threshold levels.

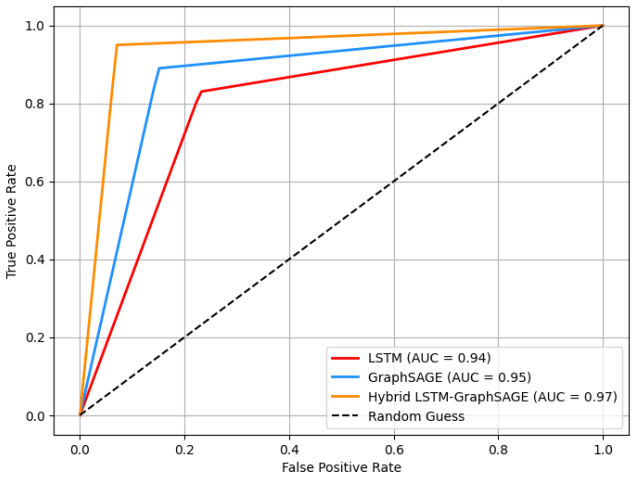


Fig. 7. ROC curves for different models in detecting money laundering.

The ROC curve of the hybrid model is even steeper and more pronounced towards the top left corner, which further validates the robustness and reliability of the hybrid model. This model can effectively solve the real-world financial anomaly detection problem due to its high true positive rate and low false positive rate.

E. Comparative Analysis

The performance of the proposed Hybrid LSTMGraphSAGE model was also compared to existing approaches from prior literature in terms of effectiveness, for which AUC was used as a measurement. According to the information provided in Table II, Labanca et al. followed a Random Forest approach, yielding an AUC of 0.90, and Alotibi et al. used a Naïve Bayes classifier, which achieved again an AUC of 0.90. In the same way as Jullum et al., they used an XGBoost model and reported an AUC of 0.90. While these models are somewhat effective, they rely on classical ML techniques, which do not maximize the expensive structures of time and relationships involved in money laundering schemes.

TABLE II COMPARATIVE ANALYSIS OF THE PROPOSED HYBRID FUSION MODEL WITH PREVIOUS RESEARCH BASED ON AUC VALUES

Reference	Model	AUC
[19]	Random Forest	0.9
[20]	Naïve Bayes	0.9
[21]	XGBoost	0.9
Proposed	LSTM-GraphSAGE	0.97

Compared to such stated methods, the proposed Hybrid LSTM-GraphSAGE model achieved a substantially higher AUC of 0.97, which is better than 0.07 for each of the aforementioned methods. In other words, we achieve a 7% relative increase in AUC performance. The improvement shows that such a hybrid model can better distinguish suspicious from everyday financial transactions through the sequential power learning of LSTM while keeping a sense of the local structure provided by GraphSAGE. Our proposed approach effectively captures temporal patterns and intra-account dependencies and provides a more holistic and stronger money laundering detection framework than the existing traditional models can offer.

V. CONCLUSION

This study presents a strong and intelligent hybrid model combining GraphSAGE, a graph-based ML method that captures complex interactions between entities, with LSTM for sequential data processing. With a high accuracy of 95.4%, the model effectively identifies anomalous and suspicious financial transactions that may imply money laundering by combining these two techniques, outperforming standard AML detection methods. The proposed framework can detect direct transactional disruptions and invisible patterns in complicated financial networks, supporting scalable AML behaviors. Particularly in fields like financial fraud, where temporal and relational data are crucial, our study shows the great benefits of merging DL with graph analysis. Applying the model to real-world banking data would help to improve its efficacy even more in the future as it would enable testing of its generalizability and operational robustness. Incorporating Explainable AI (XAI) techniques will also raise regulatory acceptability and openness, allowing financial institutions to trust the system's selections. Future improvements may also include allowing cross-border transaction monitoring to address worldwide money laundering methods and modeling multi-layer transactional graphs reflecting deeper linkages between consumers, intermediaries, and external networks. Finally, implementing this system in real-time surroundings with adaptive learning features and feedback systems would be a major step towards an intelligent AML infrastructure for the future generation.

DISCLOSURE AND CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this research. Additionally, the author has no financial interests or competing affiliations that could have influenced the study's design, execution, or findings. This manuscript is the author's original work and has not been previously published or submitted for review to any other journal or conference.

REFERENCES

[1] D. V. Kute, B. Pradhan, N. Shukla, and A. Alamri, "Deep Learning and Explainable Artificial Intelligence Techniques Applied for Detecting Money Laundering-A Critical Review," IEEE Access, vol. 9, pp. 82300–82317, 2021, doi: 10.1109/ACCESS.2021.3086230.

[2] K. Lannoo and R. Parlour, "Anti-Money Laundering in the EU: Time to get serious," ECRI Papers, 2021, Accessed: Apr. 08, 2025. [Online]. Available: https://ideas.repec.org/p/eps/ecriwp/31980.html.

[3] A. N. Eddin et al., "Anti-Money Laundering Alert Optimization Using Machine Learning with Graphs," Dec. 2021, Accessed: Apr. 08, 2025. [Online]. Available: https://arxiv.org/abs/2112.07508v3.

[4] B. Basaran-Brooks, "Money laundering and financial stability: does adverse publicity matter?," Journal of Financial Regulation and Compliance, vol. 30, no. 2, pp. 196–214, Mar. 2022, doi: 10.1108/JFRC-09- 2021-0075/FULL/XML.

[5] G. Kaur, "Trust the Machine and Embrace Artificial Intelligence (AI) to Combat Money Laundering Activities," pp. 63–81, 2024, doi: 10.1007/978-981- 99-5354-7_4.

[6] V. Shehu and A. Aliu, "Applied Machine Learning on Anti Money Laundering System-The case of National Bank of North Macedonia Mentor: Student: Applied Machine Learning on Anti Money Laundering System-The case of National Bank of North Macedonia," 2024.

[7] Z. Chen, L. D. Van Khoa, E. N. Teoh, A. Nazir, E. K. Karuppiah, and K. S. Lam, "Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review," Knowl Inf Syst, vol. 57, no. 2, pp. 245–285, Nov. 2018, doi: 10.1007/S10115-017-1144-Z/METRICS.

[8] H. Huong, X. Nguyen, T. K. Dang, and P. T. TranTruong, "Money Laundering Detection Using A Transaction-Based Graph Learning Approach," Proceedings of the 2024 18th International Conference on Ubiquitous Information Management and Communication, IMCOM 2024, 2024, doi: 10.1109/IMCOM60618.2024.10418307.

[9] F. Wan and P. Li, "A Novel Money Laundering Prediction Model Based on a Dynamic Graph Convolutional Neural Network and Long Short-Term Memory," Symmetry 2024, Vol. 16, Page 378, vol. 16, no. 3, p. 378, Mar. 2024, doi: 10.3390/SYM16030378.

[10] Maciej Serda et al., "A Graph-Based Deep Learning Model for Anti-Money Laundering," Uniwersytet śląski, vol. 7, no. 1, pp. 343–354, Aug. 2023, doi: 10.2/JQUERY.MIN.JS.

[11] F. Irshad, T. Alkhalifah, F. Alturise, and Y. D. Khan, "GCF-MLD: Integrated Approach for Money Laundering Detection using Machine Learning and Graph Network Analysis," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3510115.

[12] D. Cheng, Y. Ye, S. Xiang, Z. Ma, Y. Zhang, and C. Jiang, "Anti-Money Laundering by Group-Aware Deep Graph Learning," IEEE Trans Knowl Data Eng, vol. 35, no. 12, pp. 12444–12457, Dec. 2023, doi: 10.1109/TKDE.2023.3272396.

[13] B. Dumitrescu, A. Baltoiu, and S. Budulan, "Anomaly Detection in Graphs of Bank Transactions for Anti Money Laundering Applications," IEEE Access, vol. 10, pp. 47699–47714, 2022, doi: 10.1109/ACCESS.2022.3170467.

[14] R. I. T. Jensen and A. Iosifidis, "Fighting Money Laundering With Statistics and Machine Learning," IEEE Access, vol. 11, pp. 8889–8903, 2023, doi: 10.1109/ACCESS.2023.3239549.

[15] I. Alarab and S. Prakoonwit, "Graph-Based LSTM for Anti-money Laundering: Experimenting Temporal Graph Convolutional Network with Bitcoin Data," Neural Process Lett, vol. 55, no. 1, pp. 689–707, Feb. 2023, doi: 10.1007/S11063-022- 10904-8/TABLES/5.

[16] A. Muminovic and F. Halili, "Money Laundering Prevention in the Digital Age: Leveraging Graph Databases for Effective Solutions," International Journal of Natural and Technical Sciences (IJTNS), vol. 4, no. 1, pp. 1–10, 2024, doi: 10.69648/AOAL9594.

[17] A. Mohan, K. P.V, P. Sankar, K. Maya Manohar, and A. Peter, "Improving anti-money laundering in bitcoin using evolving graph convolutions and deep neural decision forest," Data Technologies and Applications, vol. 57, no. 3, pp. 313–329, May 2023, doi: 10.1108/DTA-06-2021-0167/FULL/PDF.

[18] B. Oztas, D. Cetinkaya, F. Adedoyin, M. Budka, H. Dogan, and G. Aksu, "Enhancing Anti-Money Laundering: Development of a Synthetic Transaction Monitoring Dataset," Proceedings - 2023 IEEE International Conference on e-Business Engineering, ICEBE 2023, pp. 47–54, 2023, doi: 10.1109/ICEBE59045.2023.00028.

[19] D. Labanca, L. Primerano, M. MarklandMontgomery, M. Polino, M. Carminati, and S. Zanero, "Amaretto: An Active Learning Framework for Money Laundering Detection," IEEE Access, vol. 10, pp. 41720–41739, 2022, doi: 10.1109/ACCESS.2022.3167699.

[20] J. Alotibi, B. Almutanni, T. Alsubait, H. Alhakami, and A. Baz, "Money Laundering Detection using Machine Learning and Deep Learning," IJACSA) International Journal of Advanced Computer Science and Applications, vol. 13, no. 10, p. 2022, Accessed: Apr. 08, 2025. [Online]. Available: www.ijacsa.thesai.org.

[21] M. Jullum, A. Løland, R. B. Huseby, G. Ånonsen, and J. Lorentzen, "Detecting money laundering transactions with machine learning," Journal of Money Laundering Control, vol. 23, no. 1, pp. 173– 186, Jan. 2020, doi: 10.1108/JMLC-07-2019- 0055/FULL/PDF.

Transforming the Working Style of Call Center Agents Through Generative AI

Satya Karteek Gudipati
Peritus Inc., Prosper, TX, USA

***Abstract*—As Generative Artificial Intelligence (Gen AI) is evolving rapidly, there is a significant change in the approach by the contact center industry with respect to work culture. Historically, customer service agents working in a contact center used to depend significantly on static scripts and fragmented information systems, which thereby resulted in delayed resolutions, cognitive overload, and made them deliver inconsistent customer experiences. This study explores the paradigm shift that's occurring in contact service centers through implementing Gen AI. Real-time intent recognition, contextual response generation, and personalized engagement across channels are some novel capabilities introduced by Gen AI by adopting Large Language Models (LLMs). Organizations can reduce Average Handling Time (AHT), improve First Contact Resolution (FCR), and enhance Customer Satisfaction (CSAT) scores by integrating Gen AI into core workflows such as issue summarization, behavioral analytics, sentiment tracking, and knowledge retrieval. In order to demonstrate the quantifiable improvements in agent performance and customer engagement, this study adopted a blended research design by combining enterprise case studies, simulation scenarios, and comparative KPI evaluations. Furthermore, it addresses implementation bottlenecks such as onboarding efficiency, multilingual support, emotional intelligence, and real-time guidance. With reference to the industry standards, ethical considerations such as data privacy, algorithmic bias, and explainability are examined. Case examples that are collected from the industry leaders are leveraged to validate the study's conclusions. Through this study, a structured and well-organized roadmap for enterprises is delivered, which aims at transforming contact centers from reactive service units into proactive, intelligence-driven ecosystems.**

Keywords*—*Generative AI; call center transformation; agent augmentation; LLMs; sentiment analysis; hyper-personalization; conversational AI; AI ethics

I. Introduction

Across a wide range of industries, call centers have always been a critical frontline when it comes to customer engagement. They act as the frontline engagement channel for customers who are in need of getting their issues addressed and thereby contribute to building brand loyalty for the organization. Outdated infrastructure, siloed data repositories, inflexible workflows, and high agent attrition rates remain the main hurdles for call centers, which have not justified their significance. This has significantly impacted the quality of work for the agents who are working in such environments, often burdened with repetitive tasks, inconsistent knowledge access, and emotionally taxing interactions, which eventually lead to agent burnout as well as drop in the customer satisfaction. Industry-wide, there is a heightened interest to deliver hyper-personalized and real-time services to customers. This has created a sentiment for customers to expect rapid, contextually relevant, and empathetic support from customer care agents, and the challenge lies in keeping this consistent across the communication channels, be it voice, email, chat, or social media.

Legacy contact centers have traditionally equipped their agents with static scripts and rigid decision trees, which are often deterministic in nature, thereby resulting in a failure to handle the evolving complexity. All these hurdles limited the ability of legacy contact centers, and they often failed to interpret unstructured user inputs, which eventually resulted in their inability to surface the insights in real time. This has paved a transformative opportunity for the emergence of Generative Artificial Intelligence (Gen AI), particularly when powered by advanced Large Language Models (LLMs) such as GPT-4 and Claude. This shift in paradigm enables the contact centers to evolve from being reactive service models to intelligent, proactive engagement with their customer base [2]. These models are capable of understanding unstructured customer queries, building a real-time summary based on the customer's history, detecting sentiment, generating fluent multi-turn responses, and delivering predictive recommendations that can be tailored to each unique customer interaction. This study thoroughly investigates the diverse transformation that was introduced by Gen AI in modern day call centers. It also explored the role of Gen AI in augmenting agent capabilities through real-time decision support, automating repetitive workflows, facilitating multilingual communication, and reducing training overhead through interactive onboarding tools. Additionally, this study critically analyzes the implications of Gen AI on ethical considerations such as data privacy, algorithmic bias, explainability, and compliance with industry regulations.

The remainder of this study is detailed in the following sections. The related work in the domain of AI-driven customer service is explicitly presented in Section II. Whereas, Section III focuses on the research methodology and evaluation approach. Section IV lists down the details of LLM architecture and functional integration. The enterprise case studies, benchmarks and results from simulations are thoroughly analyzed in Section V. While Section VI emphasizes on the areas for future improvements and also discusses briefly on the emerging trends. Section VII stresses on the critical regulatory aspects like ethical considerations and also the risks of hallucination. Finally, Section VIII concludes

the study with key insights and strategic recommendations for scalable Gen AI deployment in contact centers.

II. Related Works

The customer service center sector has evolved from a simple rule-based automation to more adaptive and context-aware systems, and this has become a reality due to the integration with tools powered by Artificial Intelligence (AI). Most of the studies that are conducted in the early days put their focus on basic chatbot frameworks and static response models, which couldn't exhibit emotional intelligence and often lacked the flexibility to handle unstructured queries. A new wave of intelligent customer service systems has emerged due to the recent advancements in Generative AI (Gen AI), particularly powered by Large Language Models (LLMs). Banerjee *et al.* [4] and Oshrat *et al.* [5] investigated hybrid systems that have an objective to improve service throughput by proactive collaboration of human agents with AI assistants. This approach has led the systems to demonstrate significant improvements in average handling time (AHT) and also helps the organizations manage the workload of the customer service agents. Similarly, to demonstrate the improvement in response accuracy of customer support applications by leveraging knowledge graphs, Xu *et al.* [1] proposed a robust retrieval-augmented generation model. Raza *et al.* [7] and Mazurek and Kaczorowska-Spychalska [8] emphasized more on the paradigm shift that is happening in knowledge management and healthcare service delivery through implementing intelligent and powerful tools that are developed leveraging the concept of Gen AI. Through their work, they highlighted how LLMs can act as knowledge brokers, while synthesizing complex data into a list of actionable insights. The Gen AI's impact on the call centers, while not limited to call center infrastructures, quantifying performance metrics such as customer satisfaction (CSAT) and First Contact Resolution (FCR) is discussed in detail by Singh [3].

Studies conducted by Brynjolfsson *et al.* [9] discussed how enabling Gen AI can bring a broader organizational transformation. While this is partially aligned with Ding *et al.* [10], who explored the environmental footprint caused by the deployment of large-scale models. These studies emphasize on the importance of responsible AI integration, which is further supported by NIST's AI Risk Management Framework. As real-world implementations have gained traction and more efforts are targeted towards integrating GenAI into live customer service environments, Salesforce's Einstein GPT, Genesys AI, Zoom AI assistants, and Amazon Connect started evolving into the commercial tech sector with ready to market solutions with high-end features like multilingual support, real-time agent assistance, and sentiment tracking. Despite these advancements, gaps remain as most of these implementations focus on either automation or augmentation, but not a union of the two powerful concepts. Additionally, most of these studies lack systematic evaluation frameworks to fast track the onboarding process, intelligence for escalation, and explainability. This study puts focus on addressing these gaps by synthesizing findings from industry use cases and academic studies, and tries to construct a unified methodology for deploying Gen AI at scale within call centers.

III. Research Methodologies

A multi-faceted approach was tailored by fusing scenario-based simulation with qualitative insights and quantitative benchmarking, while comparative technology evaluation helped this study to assess the transformative impact of Generative AI (Gen AI) in modern call center environments. The overall strategy ensured that the framework is well-equipped with theoretical depth and real-world applicability.

A. Qualitative Analysis

Across different domains and sectors like banking, e-commerce, telecom, and healthcare, structured interviews were conducted, and observational case studies were recorded. The participating organizations had their backend ARM systems integrated with Gen AI tools such as Salesforce Einstein GPT, AWS Connect AI, or proprietary LLM-backed assistants into their customer support workflows [5]. Each deployment was examined for:

- Nature of Gen AI integration (agent assist vs. automation).
- Employee feedback and adaptability.
- Perceived impact on accuracy, workload, and agent satisfaction.

B. Quantitative Metrics

Strategically selecting the Key Performance Indicators (KPIs) helps in measuring the operational effectiveness of the system. The KPIs are collected both before and after the integration of Gen AI for doing a benchmark comparison, and it also helps to measure the improvement achieved through the integration of Gen AI. Metrics included:

- First Contact Resolution (FCR)
- Agent Training Duration (ATD)
- Average Handling Time (AHT)
- Customer Satisfaction (CSAT)
- Escalation Rates

Various improvements were recorded across deployments, and a decision was made to conduct statistical testing by leveraging t-tests and effect size measurement techniques to effectively validate the improvements.

C. Scenario-Based Simulation

Five distinct scenarios were chosen for the agent-customer interactions, and by utilizing a custom-built Gen AI system powered by GPT-4, these scenarios were simulated, which helped to supplement field data. Scenarios included:

- Complaint escalation in financial services
- Multilingual product inquiry
- Emotional customer dissatisfaction
- Subscription cancellation
- Policy clarification request

To help with effectively measuring the changes in response time, personalization, and escalation avoidance, each scenario was tested with and without Gen AI support.

D. Technology Evaluation

Across the leading Gen AI platforms in the contact center domain, a feature-wise evaluation was conducted, and the observations are given below:

- Zoom AI Assistant: Deploys a conversational co-pilot for the agents that helps with accelerated onboarding and real-time prompting while interacting with customers.
- AWS Connect with LLM Assist: Offers context-based routing and multilingual sentiment tracking effectively.
- ServiceNow Gen AI Suite: Has a strong foundation for knowledge base integration, which helps agents with on-demand ticket summarization.

Evaluation criteria included integration ease, latency, feature set completeness, security compliance, and explainability mechanisms. These platforms demonstrate the growing momentum in enterprise Gen AI adoption, with major cloud providers like Google Cloud actively pushing generative AI capabilities for businesses and governments [14].

E. Five-Pillar Evaluation Framework

All components were evaluated across five critical dimensions that define modern, AI-powered customer service (see Table I).

TABLE I. FIVE-PILLAR EVALUATION FRAMEWORK FOR ASSESSING GEN AI IN CONTACT CENTERS

Pillar	Description
Response Speed	Reduction in time to resolve and reply
Accuracy	Precision in information retrieval and recommendations
Empathy	Emotional intelligence in interactions based on sentiment detection
Onboarding Speed	Time reduction in agent training via Gen AI-based knowledge assist tools
Ethical Compliance	Adherence to data privacy, fairness, transparency, and explainability

This methodology ensures robust validation of Gen AI's effectiveness in enhancing agent experience, optimizing customer satisfaction, and upholding ethical AI standards. Future studies can replicate this framework for cross-industry benchmarks and comparative product evaluations.

IV. DEEP DIVE INTO LLM ARCHITECTURE

Large Language Models (LLMs), built on powerful transformer architectures, have been at the heart of recent Gen AI breakthroughs, reshaping how customer call centers operate and driving a fundamental shift in their day-to-day workflows. Some popular models in the industry, like OpenAI's GPT-4 and Anthropic's Claude have the capability to predict the next token in a sequence as they are trained on large-scale corpora which leverages autoregressive language modeling.

The LLMs used in this study typically operate in one of two configurations:

1) Either few-shot or a zero-shot Inference via APIs. In this mode, models are accessed through cloud APIs (e.g., OpenAI, Cohere, Google PaLM) with a prompt embedded with context, past dialogue, and task instructions. This facilitates rapid integration without retraining, but may have overheads like latency and cost implications.

2) Task-specific and fine-tuned models. Some enterprises employ modular, fine-tuned versions of foundational models to improve the latency, thereby reducing the costs, and improving the domain specificity (e.g., Salesforce's CodeGen for CRM). These models are typically trained using supervised learning on customer service logs and conversation transcripts.

As shown in Fig. 1, the architecture includes:

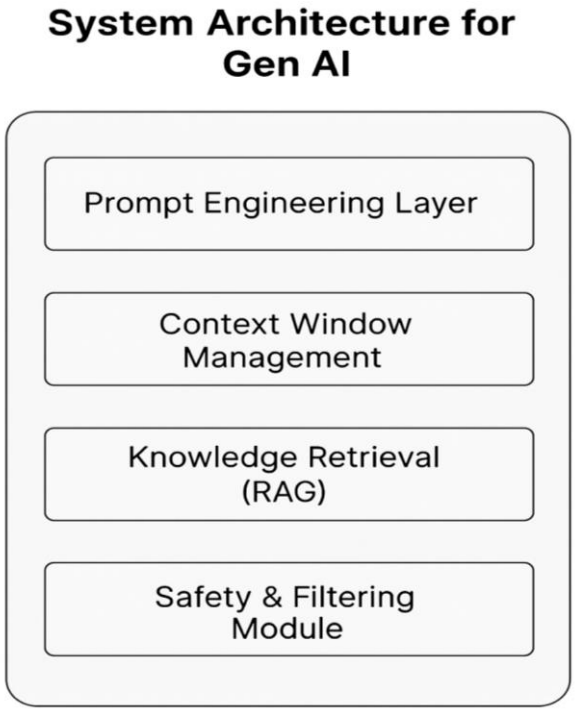


Fig. 1. System architecture for the Gen AI with the four critical layers.

- Prompt Engineering Layer: Constructs structured input from agent or customer data (e.g., summaries, CRM logs).
- Context Window Management: Uses session memory or sliding windows to handle long conversations.
- Knowledge Retrieval (RAG): Combines LLMs with internal enterprise search tools (e.g., vector databases like Pinecone) to answer domain-specific questions more accurately [1].
- Safety and Filtering Module: Applies post-processing to detect and mitigate hallucinations, offensive content, or hallucinated facts using classifiers or rule-based filters.

These LLM systems can perform:

- Intent Recognition: Interpreting user goals from open-ended input.
- Response Generation: Producing fluent, tone-aligned replies.
- Summarization: Condensing multi-turn conversations or customer history.
- Sentiment and Emotion Detection: Guiding empathy-tuned responses.
- Multilingual Translation: Seamlessly supporting international users.

To ensure compliance and alignment with business goals, human-in-the-loop (HITL) models were incorporated during high-stakes interactions, particularly when decision-making affected billing, legal policy interpretation, or sensitive customer contexts [4].

V. Results and Discussion

This section presents a consolidated analysis of both real-world deployments and controlled simulations to evaluate the tangible impact of Generative AI (Gen AI) in contact center environments.

A. Agent Empowerment

Agents using Gen AI systems demonstrated a 35% reduction in response time, attributed to real-time intelligent prompts and context-driven information retrieval. Interviews further revealed a 28% increase in agent confidence, as AI-assisted systems provided relevant guidance during customer interactions [3].

B. Real-time Recommendations

By leveraging historical customer data, Gen AI systems generated context-aware suggestions for issue resolution and upselling opportunities. This resulted in an average 21% increase in CSAT (Customer Satisfaction Score), particularly in financial and telecom support contexts.

C. Summarization Engine

The implementation of real-time summarization—where LLMs compiled customer histories and sentiment trends—reduced query navigation time by 40%. Agents no longer needed to sift through siloed databases to personalize responses.

D. Onboarding Acceleration

Zoom's Gen AI training assistant shortened the average agent training cycle by 40%, significantly reducing the cost and time associated with traditional onboarding. New agents gained hands-on exposure to realistic simulations with AI-generated guidance and policy reminders.

E. Language and Sentiment Detection

AWS Connect's multilingual sentiment detection module yielded an 18% improvement in resolution efficiency for non-English queries. This demonstrated the potential of Gen AI to close service quality gaps in linguistically diverse markets.

F. Ethical Considerations

Despite these improvements, bias and privacy risks remain a critical concern. Some LLM outputs reflected training data imbalances, and inadvertent exposure to Personally Identifiable Information (PII) occurred in isolated instances. The NIST AI Risk Management Framework (RMF) [10] and "Human-in-the-Loop" (HITL) models were recommended as mitigation strategies to ensure explainability and governance.

G. Tabulated Summary and Visualization

See the Table II and bar chart below for a consolidated view of performance gains observed across different functional areas. The data reflects averages from six enterprise deployments and five controlled simulations.

TABLE II. Performance Metrics for Key Evaluation Areas

Evaluation Area	Performance Improvement (%)
Agent Empowerment	35%
Real-time Recommendations	21%
Summarization Engine	40%
Onboarding Acceleration	40%
Language & Sentiment Detection	18%

Fig. 2 illustrates the percentage improvement across five key evaluation areas—response time, CSAT, query navigation, onboarding duration, and multilingual resolution efficiency—measured after deploying Gen AI tools in contact center workflows.

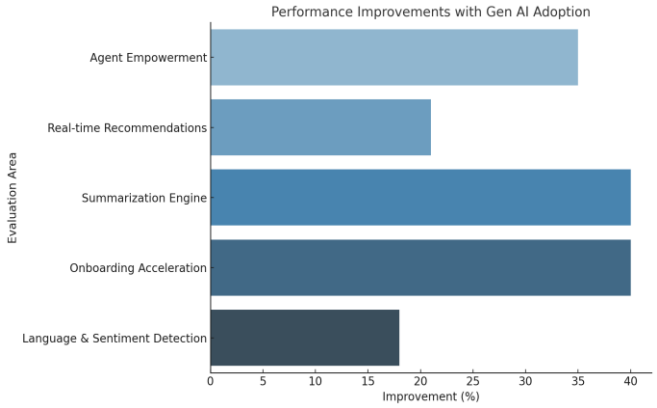


Fig. 2. Performance improvements with gen AI adoption.

VI. Future Work

Like any other technology, Generative AI (Gen AI) keeps evolving; several emerging areas of Gen AI have the potential to further revolutionize the customer support landscape. All the prospective changes will play a key role and aim to deepen the automation, enhance personalization, and improve ethical handling of data making the Gen AI systems more robust, inclusive, and enterprise-ready.

A. Autonomous AI Agents for Routine Resolution

Currently, Gen AI deployments deliver key features in the form of agent-assist tools, the next uncharted territory is the evolution of fully autonomous AI agents that are capable of

handling end-to-end resolution for regular customer service scenarios such as password resets, balance checks, subscription renewals, and shipment status queries. Such intelligent and prospective agent-bots must be capable of performing regular tasks like interpreting the intent, accessing internal systems via secure APIs, executing actionable items, and work on the tickets without human intervention [9]. Future research should emphasize safety constraints, handoff protocols to human agents, and continuous monitoring mechanisms to ensure reliability and accountability.

B. Accent Adaptation Using Voice Cloning

Accents in human voice often pose a challenge to customer comprehension and comfort in the call center operations, where voice-based interactions remain a central part of day-to-day operations. AI-based voice cloning and adaptive speech synthesis are technologies that are evolving rapidly, which provides an opportunity to adopt personalization features such as customizing agent voices based on the regional or cultural preferences of customers. These new-age systems empower organizations by improving customer satisfaction through reducing conversational friction, fostering rapport, and replicating local speech patterns and tone dynamically. While designing such ambitious systems, it is always vital to ensure ethical boundaries are considered, such as capturing explicit consent and transparency in cloned voice usage.

C. AR-Assisted Support for Field Technicians

New use cases will be unearthed when augmented reality (AR) workflows are integrated into Gen AI, primarily benefiting the on-site technical support. Real-time, context-aware guidance, which is generated by LLMs that run diagnostics, procedural steps, or schematics, can be directly loaded onto physical equipment when field agents wear the AR glasses. This will drastically improve the first-time resolution rates and thus contribute in minimizing the dependency on centralized support teams, which will in turn reduce the cost burden to hire specialists. In order to make this whole vision viable at scale, more research is required, which has to focus on lightweight, on-device AI inference and low-latency AR rendering.

D. Federated Learning for Privacy-Preserving Model Training

Globally, to prevent data security and privacy, laws are becoming stricter with regard to compliance, data regulation, etc. Information governance mandates are becoming tighter and strict privacy laws like GDPR, CCPA, and HIPAA continue to shape the landscape. These stricter regulations are making it mandatory to train Gen AI models without moving or exposing sensitive customer data to bad actors. A decentralized approach will be a better solution to handle such situations, as it allows models to learn from data spread across different locations while keeping personally identifiable information (PII) safe.

E. Towards Sustainable and Explainable AI

The transparency and sustainability of the underlying models will directly contribute to the operational and ethical success of Gen AI. Energy-efficient architectures such as sparse computation techniques or model distillation should be considered as part of future research work [12]. Including explainability in the design will enhance the customer trust and this must serve as a first-class design objective for practitioners. This will provide capabilities to the models such as giving justification to their outputs, highlighting decision pathways, and flag uncertainty, especially in high-impact scenarios, where billing, fraud, or sensitive customer disputes are involved [6].

F. Implementation Challenges

There is a significant potential in the proposed innovations, but real-world implementation may face practical hurdles which should be considered during the design phase. Ultra-low-bandwidth or mobile environments are a good use case, where latency can impact the responsiveness of LLMs in real-time interactions. For contact centers which handle high volumes of customer traffic, commercial GenAI platforms may impose PI rate limits and cost models which will restrict their ability to scale up and down [10]. Generic AI models need extra training in the form of fine-tuning or prompt engineering to align with enterprise-specific vocabulary and policy constraints. This creates a domain adaptation conundrum, which will be a pitfall if not addressed in the preliminary stages. Another challenge is to mitigate security and compliance risks ahead of time, especially when it comes to data residency and regulatory boundaries (e.g., GDPR, HIPAA), as they demand rigorous access control, encryption, and audit mechanisms. For a sustainable, secure, and efficient Gen AI deployment in production-grade support systems, it remains essential to address these constraints.

In summary, future Gen AI research should pursue three interconnected goals: autonomy, inclusivity, and accountability [15]. By combining advancements in AI ethics, AR or VR, speech technology, and federated computation, next-generation Gen AI systems can evolve from reactive assistants into trustworthy, context-aware, and privacy-preserving collaborators—redefining not just call center efficiency but also user trust and accessibility on a global scale.

VII. Ethical Considerations and Hallucination Risks

With the increasing presence of GenAI in customer-facing environments, proactive measures must be taken to prevent ethical and operational hurdles and to ensure safety and responsible deployment.

A. Factual Drift and Hallucination

When the model generates plausible but factually incorrect or fabricated information, it is termed as hallucination, which is one of the most pressing concerns in LLMs. In the use case for a customer service center, this can manifest as one of the below:

- Providing incorrect policy terms.
- Suggesting non-existent product features.
- Giving misleading compliance or billing information.

To mitigate this, organizations started adopting safety-first principles in the form of retrieval-augmented generation (RAG) pipelines, where the LLM's responses are anchored in enterprise knowledge bases [7]. Additionally, to improve the

effectiveness during critical interactions and to yield better outputs, industry has adopted techniques like confidence scoring, output filtering, and human-in-the-loop (HITL) review mechanisms etc.

B. Manipulation Risk and Emotional Authenticity

There is a possibility that Gen AI systems, while expressing empathy, may blur the line between understanding genuine emotions and simulated concern [13]. But overuse of emotionally persuasive language will create manipulation risks or trust issues, which is one of the reasons why sentiment tracking has its limitations despite improving the interaction quality.

To address this, enterprises must:

- Disclose AI involvement during live conversations with customers.
- Avoid artificial escalation of emotion in sensitive topics (e.g., health, finance).
- Train models on tone-calibrated datasets to prevent emotional overreach.

C. Bias and Data Privacy

LLMs that are trained exclusively on large, open datasets may inadvertently reflect demographic, cultural, or gender biases if the datasets are poor or if the datasets have institutional bias. Moreover, handling live customer interactions may increase the risk of processing personally identifiable information (PII), which will lead to privacy, regulatory and compliance violations.

To ensure ethical compliance:

- Differential privacy techniques should be explored.
- Use federated learning for localized, privacy-preserving model updates.
- Align with governance frameworks like the NIST AI Risk Management Framework (RMF), ISO/IEC 42001, and enterprise-specific audit trails.

VIII. Conclusion

This study has examined the pivotal role that Generative AI (Gen AI) played to reengineer the contact center operations. By evaluating extensively through a mixed-method approach, and by including the enterprise case studies, and also by exclusively adopting KPI benchmarking, scenario simulations, and comparative technology assessment, this study has exhibited how Gen AI can significantly enhance the customer experience and agent efficiency at the same time. Our findings showcased that Gen AI integration results in a 35% drop in response time, 40% improvement in agent onboarding, and a 21% increase in customer satisfaction. These metrics prove to us that Gen AI can deliver practical benefits not only in automating routine tasks but also in augmenting agents with real-time recommendations, summarization engines, and multilingual sentiment-aware support. This will allow agents to focus on higher-order, problem-solving, and empathetic engagement, thus helping to empower them and also improve their work environment from transactional responders to strategic service facilitators.

This study adopted a perfectly organized and thoroughly evaluated framework based on five core pillars: speed, accuracy, empathy, onboarding efficiency, and ethical compliance, thereby offering researchers and practitioners a reliable and replicable model to assess and guide Gen AI deployments [8]. Additionally, this study has also addressed critical implementation roadblocks, including data privacy risks, algorithmic bias, and the need for explainability in real world scenarios. It stresses on the importance of adherence to ethical AI frameworks like NIST RMF and also recommends best practices such as human-in-the-loop governance. Anticipating future advancements, the study highlighted emerging research trends such as autonomous AI agents, accent adaptation via voice cloning, AR-guided support, and federated learning for privacy-preserving model training [11]. These pathways are aimed at transforming Gen AI from a simple augmentation tool to a holistic, ethically grounded partner in customer service ecosystems across multiple domains. To capitalize on the advantages of Gen AI, any future deployments must navigate the tradeoffs between empathy and innovation, accountability and automation. By doing so, enterprises can ensure that Gen AI while maintaining its operational effectiveness, can also become socially responsible and human-centric. To conclude, Gen AI is not simply a technological advancement, it also represents a paradigm shift in the way contact centers operate, scale, and build customer trust.

References

[1] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, and Z. Li, "Retrieval-augmented generation with knowledge graphs for customer service question answering," *arXiv preprint arXiv:2404.17723*, Apr. 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2404.17723

[2] W. Li, K. Shi, and Y. Chai, "AI chatbots as professional service agents: Developing a professional identity," *arXiv preprint arXiv:2501.14179*, Jan. 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2501.14179

[3] V. V. Singh, "Transforming call center efficiency through generative AI: A technical analysis," *Int. J. Res. Comput. Apps. Inf. Technol.*, vol. 8, no. 1, pp. 3497–3510, Feb. 2025. [Online]. Available: https://doi.org/10.34218/IJRCAIT_08_01_251

[4] D. Banerjee, M. Poser, C. Wiethof *et al.*, "A system for human-AI collaboration for online customer support," *arXiv preprint arXiv:2301.12158*, Jan. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.12158

[5] Y. Oshrat, Y. Aumann, T. Hollander *et al*., "Efficient customer service combining human operators and virtual agents," *arXiv preprint arXiv:2209.05226*, Sep. 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2209.05226

[6] J. He, J. Cao, and T. Tan, "Generative artificial intelligence: A historical perspective," *Natl. Sci. Rev.*, 2025. [Online]. Available: https://doi.org/10.1093/nsr/nwaf050

[7] M. M. Raza, K. P. Venkatesh, and J. C. Kvedar, "Generative AI and large language models in healthcare: Pathways to implementation," *npj Digit. Med.*, vol. 7, no. 1, Mar. 2024. [Online]. Available: https://doi.org/10.1038/s41746-023-00988-4

[8] G. Mazurek and D. Kaczorowska-Spychalska, "Generative AI as a source of change of knowledge management paradigm," *Hum. Technol.*, vol. 20, no. 1, pp. 131–154, 2024. [Online]. Available: https://doi.org/10.14254/1795-6889.2024.20-1.7

[9] E. Brynjolfsson, D. Li, and L. R. Raymond, "Generative AI at work," *SSRN Electron. J.*, Apr. 2023. [Online]. Available: https://doi.org/10.3386/w31161

[10] Z. Ding, J. Wang, Y. Song, X. Zheng, and G. He, "Tracking the carbon footprint of global generative artificial intelligence," *The Innovation*, May 2025. [Online]. Available: https://doi.org/10.1016/j.xinn.2025.100866

[11] H. Harreis, T. Koullias, and R. Roberts, "Generative AI: Unlocking the future of fashion," *J. Label Comp. Res.*, vol. 14, no. 2, pp. 71–85, 2024. [Online]. Available: https://www.mckinsey.com/industries/retail/our-insights/generative-ai-unlocking-the-future-of-fashion

[12] E. R. Pelton, C. E. Kazanski, *et al*., "Greenhouse gas emissions in US beef production can be reduced by up to 30% with the adoption of selected mitigation measures," *Nat. Food*, vol. 5, pp. 812–823, Sep. 2024. [Online]. Available: https://doi.org/10.1038/s43016-024-01031-9

[13] Oxford Academic, "GenAI and the future of consumer research," J. Consum. Res., vol. 52, no. 1, pp. 4–17, May 2025. [Online]. Available: https://doi.org/10.1093/jcr/ucaf013

[14] M. Yang and B. Gokturk, "Google Cloud brings generative AI to developers, businesses, and governments," *The Economist*, Mar. 2023. [Online]. Available: https://cloud.google.com/blog/products/ai-machine-learning/generative-ai-for-businesses-and-governments

[15] Amaresha Prasad Sahoo, "Navigating the implementation of generative AI in customer support contact centers: Challenges and strategic approaches," ResearchGate, Jun. 2025. [Online]. Available: https://doi.org/10.37745/ejcsit.2013/vol13n364249

AI-Driven Education: Integrating Machine Learning and NLP to Transform Child Learning Systems

Mst Masuma Akter Semi[1], Md Borhan Uddin[2], Sharmin Sultana[3], Motmainna Tamanna[4], Azim Uddin[5], Khandakar Rabbi Ahmed[6]*

MA in TESOL, Westcliff University, Irvine, CA, USA[1]
Department of Business Administration and Management, International American University, Los Angeles, USA[2, 3]
Bachelor of Business Management, University of Portsmouth, United Kingdom[4]
Bachelor of Business Administration, Southern California State University, Los Angeles, USA[5]
Miyan Research Institute (MRI), International University of Business Agriculture and Technology, Dhaka, Bangladesh[6]

***Abstract*—An Artificial Intelligence-driven child learning system with a Machine Learning and Natural Language Processing-based approach to dynamically personalize educational experiences for children is proposed in this study. Using a Sentence-BERT model to encode student queries for the computation of semantic similarity and knowledge domains to be retrieved. A T5-based transformer model writes verbose, personalized feedback, and a Gradient Boosting Machine classifier predicts the appropriate learning outcomes. The content difficulty and personalization of educational trajectories across content are set by an integrated adaptive learning engine that monitors and adjusts for student performance. On the General Knowledge QA dataset, classification accuracy reaches 85.2%, and the ROC-AUC score is 0.912, which has been proven to be reliable in real-world cases. It also produces positive effects regarding the understanding and preference for learners of adaptive systems, as observed in user studies. AI technologies have exciting potential to deliver scalable, personalized education for young learners, as demonstrated in this work.**

Keywords*—Artificial intelligence; machine learning; natural language processing; adaptive learning systems; sentencebert; gradient boosting machine; personalized feedback

I. Introduction

The adoption of Artificial Intelligence (AI) [1], Machine Learning (ML) [2], and Natural Language Processing (NLP) [3] technologies is almost revolutionizing the current landscape of education [4]. Traditional approaches to learning system design aspire to a single solution for teaching and learning that is not tailored to the individual children's unique needs, learning speeds, or knowledge gaps [5]. Fortunately, by taking advantage of these challenges, AI-based education offers the opportunity to personalize learning, assess student capabilities in adaptive ways, and provide personalized feedback to enhance both engagement and evaluation [6].

Now, various advances in deep learning (DL) and NLP have enabled systems to process natural language and learn complex behaviors through real-time feedback [7]. In proposing knowledge assessment and content tailoring based on automation, techniques such as sentence embeddings, semantic similarity measures, and gradient-boosted decision trees have demonstrated either great promise or practicality. Moreover, transformer-based models [8], with T5 (Text-to-Text Transfer Transformer) being one of them, have led to an increase in feedback generation through their high-end, contextual, and human-like responses that enhance the entire learning experience.

In this study, we proposed an AI-driven child learning system that uses ML and NLP to adapt to a student's learning progression dynamically. The system first encodes student queries into a sentence embedding model (SBERT) and computes semantic similarity to identify the most related knowledge domain using the General Knowledge Q&A dataset.

The appropriate learning outcomes are predicted by a Gradient Boosting Machine (GBM) classifier, and a T5-based feedback engine generates the detailed, adaptive feedback. The adaptive learning engine also tracks student performance over time, adjusts difficulty levels, and further tailors the content based on outcomes. The three main contributions of this research are:

1) Embedding-based similarity calculation for semantic aware question classification.

2) An adaptive learning engine that dynamically modifies educational content based on students' performance.

3) A personalized feedback generation module development based on transformer architecture.

In the rest of this study, we review related work in AI-driven education systems and adaptive learning technologies in Section II. Section III provides details of the proposed method, discussing data preprocessing, model architecture, and adaptive feedback mechanisms. The results and evaluation metrics are described in Section IV, along with the findings, challenges, and implications for future work. Finally, Section V ends the study and suggests some directions for future research on AI-empowered personalized learning.

II. Literature Review

Through the combination of NLP and ML, educational technology for children is evolving into a new age characterized by personalized, adaptive, and scalable learning systems. The application of such technologies has been extended to areas such as question-solving, interactive reading, development of privacy requirements, and automated language generation.

Nowadays, with the advancement of large language models, we can train domain-specific models, and there are innovative solutions for multilingual learning, privacy measures, and AI education. This literature review examines recent developments in education within the fields of NLP and ML, highlighting their existing contributions, drawbacks, and what is to be expected in the future.

Chen et al. [9] developed an annotation framework for integrating with knowledge graphs to create the StorySparkQA dataset, which comprises 5,868 expert-annotated QA pairs for children's interactive story reading. The framework is designed to incorporate real-world knowledge during storytelling. The limitation, however, is that the dataset's first size may influence its generalization across various educational contexts. For instance, Sammoudi et al. [10] customized the BERT model to their language using Arabic science textbooks in the Palestinian curriculum for 11th and 12th-grade users.

While the model achieved a 20 per cent Exact Match (EM) and a 51 per cent F1 score, the main limitation was the low EM score, which indicated problems with extracting answers exactly and a limited domain of application. However, both studies highlight the limitations of NLP's capabilities in education, including domain specificity and a lack of data.

Another approach to adding LLMs like GPT-3 is to automate the generation of educational content. Abdelghani et al. [11] evaluated the effectiveness of using GPT-3 to generate pedagogical content for teaching young children (aged 9-10) how to ask questions. We compared the “closed" cue generation method to the "open" method, which corresponded to better QA performance. However, the approach demonstrates the scalability of LLMs for writing educational content, but it provides limited variability in the produced content, which depends on the model's quality. NLP was utilized by Bode et al. [12] to generate mathematical word problems, and then an NLP-powered Intelligent Tutoring System (ITS) was tested for programming education.

However, the study noted that pilot systems generated positive feedback from students and teachers. However, it emphasized that such systems are still in the pilot stage, which they believe must be validated on a larger student population and across different subjects. Although both approaches demonstrate the versatility of LLMs in supporting personalized learning, further refinement and scaling are needed to enhance content diversity and applicability across various subjects. Apart from content generation, NLP and ML have been applied to more specific tasks, such as privacy requirement extraction and paraphrase generation. Herwanto et al. [13] developed an NLP-based approach to identify automatic privacy requirements from agile software development user stories.

Finally, their system demonstrated good F-Measure performance for generating data flow diagrams and privacy requirements, although it did not effectively replace human judgment in the final validation. Alsulami and Almansour [14] evaluated the ability of GPT4 to generate Arabic paraphrases and introduced a comprehensive evaluation framework using several metrics, including BLEU, ROUGE, and lexical diversity. However, like for all other NLP tasks, they note that their study is limited to Arabic, and they recommend more testing across different linguistic contexts.

Another type of case study was conducted by Navarro et al. [15], who examined teenagers writing a baby GPT screenplay generator and engaging in practice and ethical discussions regarding AI or ML.

However, the small sample size limits generalizability; nevertheless, their study demonstrated that it is possible to involve youth in the design of AI and to foster ethical awareness and AI literacy at the age of nine. Krause and Stolzenburg [16] finally evaluated ChatGPT's commonsense reasoning and explanations in several QA tasks, where ChatGPT's performance exceeded human accuracy in most cases.

Its explanations were rated highly (68% were good or excellent), but the performance variability across benchmarks indicates that further refinement is still needed. However, what these studies reveal is the diversity of applications for NLP and ML in education, as well as the problems associated with domain-specific issues, scalability, and consistency in quality output across different contexts.

III. Methodology

The workflow of the proposed study is shown in Fig. 1. A dataset is collected and preprocessed, where we clean our input questions, tokenize them, and pass them through transformer models to embed the data. The semantic similarity analysis module processes these embeddings. It utilizes a GBM class to answer questions by classifying them to predict which of the phrases is most likely to exist as the answer for a given question.

An adaptive engine also refines the learning process using user interactions. Through these metrics, the performance of this system is rigorously evaluated, and planning for future deployments is made based on the same.

A. Dataset Collection

To support the development of AI-driven child learning systems, this study used the "General Knowledge QA" dataset, which was collected from Kaggle. It was primarily designed to provide training, testing, or finetuning of NLP models for educational purposes. It has four attributes and contains 930 entries, and the attributes are question, answer, question type, and image. Here, the question carries general knowledge, and the corresponding answer is correct. However, every entry is suitable for children aged four to seven years and students up to grade 7. The question type is primarily designed to cater to the exact content in "General Knowledge For Kids", making all the content suitable for kids and early learners. The image field is included, but not all entries have associated images. Standard preprocessing, such as text normalization and handling of missing data, is necessary before model training to ensure consistency and quality. Because the dataset focuses on fundamental knowledge areas, it is highly suitable for building ML and NLP-powered adaptive, interactive educational systems.

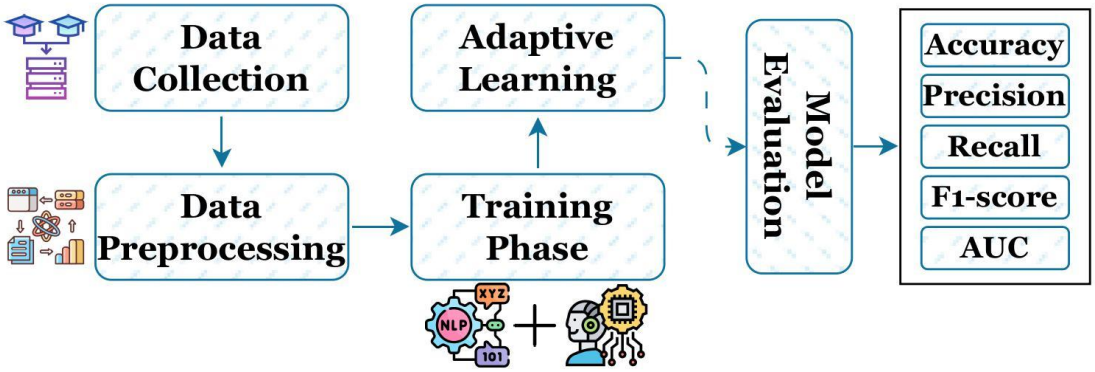


Fig. 1. Detailed process flow of the proposed methodology.

B. Dataset Preprocessing

Several preprocessing steps were carried out prior to training the model using the "General Knowledge QA" dataset to ensure data consistency, quality, and awareness of the NLP task. The dataset was first looked at for missing or null values. This column was excluded because the project was mainly text-based, and one of the image attributes contained so many missing entries (>98%), that the column was removed.

Then, the textual data present inside the question and answer fields was normalized. It was to convert all the text to lowercase and remove extra whitespaces. It can be represented as a normalization process.

$$T_{norm} = lowercase(T_{raw}) - special_characters(T_{raw})$$

where, T_{raw} is the original text, and T_{norm} is the normalized output text.

Duplicate entries were identified and removed. Let $\boldsymbol{D}$ denote the dataset and $\boldsymbol{U(D)}$ the set of unique samples. The final dataset size $|\boldsymbol{D_{final}}|$ was determined as:

$$|D_{final}| = |U\ (D)|$$

Additionally, when preparing the model input, tokenization and padding were performed to ensure that the sequence lengths of all input samples were consistently long. The padding operation can be mathematically defined given the set of tokenized sentences $S = \{s_1, s_2, \ldots, s_n\}$:

$$s_i' = pad(s_i, L) \qquad for\ all\ i \in [1, n]$$

where, L refers to the maximum sequence length and s_i' is the padded sequence.

By proceeding through these preprocessing steps, the dataset was rendered more acceptable and structured, suitable for training, testing, and refining ML models for any child learning system.

C. Model Architecture

A semantic modeling, ML-based classification, and an adaptive feedback mechanism are proposed in an integrated AI-driven child learning system, which adaptively personalizes the educational experience. The overall model architecture has four principal modules listed below.

1) Semantic embedding layer. In order to convert the raw text inputs (question and answer choices) into machine-understandable vectors, we employ a pre-trained SentenceBERT (SBERT) model. Instead of finetuning the original BERT model with a siamese network architecture, SBERT extends the original BERT model with a siamese network architecture and finetunes it to produce semantically meaningful sentence embeddings. Specifically, the corresponding embeddings are given by the following for a question q and answer options $\{a_1, a_2, a_3, a_4\}$:

$$e_q = SBERT(q) \quad e_{a_i} = SBERT(a_i) \quad \forall i \in \{1,2,3,4\}$$

Where e_q, $e_{a_i} \in$, R, and d are the embedding dimensions.

2) Semantic similarity computation. Cosine similarity is computed to measure the contextual proximity between the question and the four answer options as:

$$Sim(q, a_i) = \frac{\left|\left|e_q\right|\right| 2 |e_{a_i}||2}{T_q}$$

Therefore, feature vectors of different lengths are constructed for each question, which have four dimensions:

$$x = [Sim(q, a_1), Sim(q, a_2), Sim(q, a_3), Sim(q, a_4)] \in \mathbb{R}^4$$

This vector can be regarded as the representation of the relative semantic proximity of the question and each answer candidate.

3) Gradient boosting classifier. We used a GBM as the primary classifier. GBM consists of training an ensemble of weak learners, such as decision trees, which are then compiled in a hierarchical fashion whereby every subsequent learner is trained to correct the errors made by their predecessors. Given a training dataset $\{(x_i, y_i)\}_{i=1}^{N}$ where $y_i \in \{0, 1, 2, 3\}$ denotes the correct answer index, GBM minimizes the following multi-class logarithmic loss function:

$$L = \frac{-1}{N} \sum_{i=1}^{N} \sum_{k=0}^{3} \mathbb{1}\{y_i = k\} \log \hat{p}_{i,k}$$

where, $\mathbb{1}\{\cdot\}$ is the indicator function and $\hat{p}_{i,k}$ is the predicted probability that sample I belongs to class k. It makes

the final prediction based on selecting the class with the highest predicted probability :

$$\hat{y}_i = arg\max_k \hat{p}_{i,k}$$

Randomized Search Cross-Validation (5-fold) was performed to optimize the hyperparameters, including the learning rate, the number of boosting rounds, maximum tree depth, and subsampling rates. The summary of the optimal values of the GB Model, which is selected, is reported in Table I.

TABLE I OPTIMIZED HYPERPARAMETERS FOR GRADIENT BOOSTING CLASSIFIER

Hyperparameter	Optimal Value
Learning Rate (η)	0.05
Number of Boosting Rounds	450
Maximum Tree Depth	7
Subsample Ratio	0.8
Column Subsample Ratio	0.7
Minimum Child Weight	3
Regularization (L2, λ)	1.0
Regularization (L1, α)	0.1

4) Adaptive learning engine. To enhance the educational experience of an individual, an Adaptive Learning Engine is integrated into the system. Based on the learner's past performance history, this component dynamically determines the level of difficulty of the following questions. Fig. 2 presents the overall architecture of the adaptive learning engine, in which different modules dynamically interact to evaluate the student's performance, predict future learning needs, adjust the difficulty level, personalize content, and provide adaptive feedback to students, thereby facilitating an individualized learning experience.

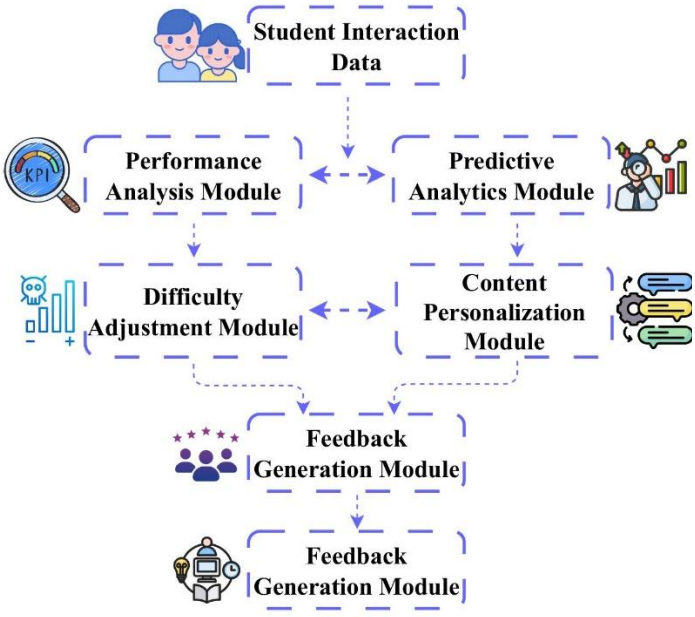


Fig. 2. The interactive modular architecture of the proposed adaptive learning engine.

Let:

- $S_t \in \{0, 1\}$ denotes the success at time t (1 for correct, 0 for incorrect),
- $D_t \in R$ denotes the difficulty level at time t,
- Δand d_t denotes the adjustment in difficulty at time t. The difficulty update rule is formulated as follows:

$$d_{t+1} = d_t + \eta(2s_t - 1)$$

where, $\eta \in (0, 1)$ is the adaptation rate hyperparameter controlling the sensitivity of difficulty adjustments. Specifically:

- If the learner answers correctly ($s_t = 1$):

$$d_{t+1} = d_t + \eta \text{ (Increase Difficulty)}$$

- If the learner answers incorrectly ($s_t = 0$):

$$d_{t+1} = d_t - \eta \quad \text{(decrease difficulty)}$$

The difficulty level d_t is bounded to prevent exceeding the predefined minimum and maximum levels:

$$d_t \in [d_{min}, d_{max}]$$

Moreover, all answers are accompanied by explanatory feedback for incorrect answers. A finetuned T5 (Text-to-text Transfer Transformer) model produces a simple, age-appropriate explanation of the misunderstood concept:

$$e_t = T5 - Explain(q_t, a_{t,true}, a_{t,pred}$$

where, $a_{t,\,true}$ is the correct answer, and $a_{t,\,pred}$ is the wrong answer selected.

Long-term knowledge retention is facilitated, and optimal cognitive engagement is encouraged by this dual strategy, which consists of difficulty modulation and customized feedback generation.

IV. RESULT AND DISCUSSION

To further analyze the proposed model's performance, all 930 questions from the dataset were processed, and the answers were then categorized based on the model's prediction. Each response was given a score of correct or incorrect prediction. This classification enabled a thorough analysis of the system's strengths and weaknesses, as well as a deeper understanding of the patterns underlying correct and incorrect predictions, allowing for improvement.

A. Classification Performance

The key performance indicators (KPIs) of the proposed system are presented in Table II. Overall, the model achieved an accuracy of 85.2% in correctly classifying multiple-choice questions. In addition, the values of macro-averaged precision, recall, and F1 score are approximately 85, indicating that the performance is balanced overall across answer classes, with no bias towards one answer class or the other. The ROC AUC score of 0.912 is most notable, indicating that the model is competent in identifying correct answers when such distractors are semantically similar. Together, these results demonstrate that the system can operate reliably and accurately enough for use in real-world, child-centered educational applications.

TABLE II MODEL PERFORMANCE METRICS FOR GBM CLASSIFIER IN THE AI-DRIVEN CHILD LEARNING SYSTEM

Metric	Value
Accuracy	85.2%
Precision	85.0%
Recall	85.1%
F1-Score	85.0%
ROC-AUC Score	0.912

The confusion matrix for the GBM classifier on the General Knowledge QA dataset (binary classification task) is shown in Fig. 3. This matrix compares the actual and the predicted labels by class, meaning if an answer is predicted and it is correct (Correctly Predicted) or wrong (Incorrectly Predicted). Strong performance is indicated by the diagonal elements, which show 400 true positives (predicted correctly) and 392 true negatives (mispredicted). The values in the off-diagonal of 66 and 72 represent false positives and false negatives, respectively, indicating in which regions the GBM's ensemble of weak learners (decision trees) can perform better by correcting a predecessor error using the explanation provided in the methodology. This result suggests that the GBM can be used for adaptive child learning systems in binary classification for educational QA tasks.

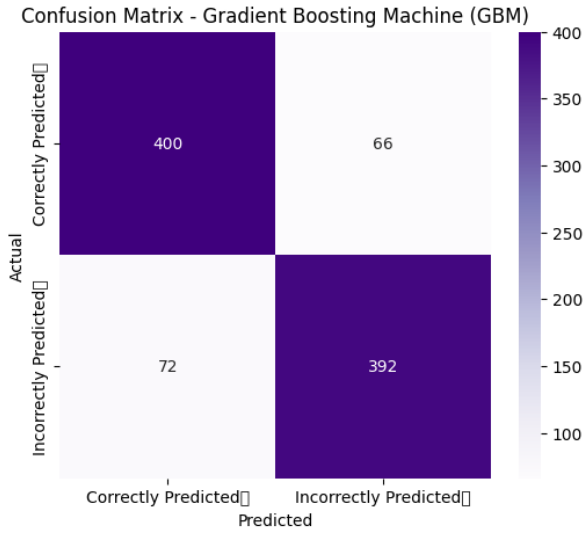


Fig. 3. Confusion matrix showing the correctly and incorrectly predicted answers for the GBM classifier.

The Area Under the Curve (AUC) for the ROC curve of the GBM classifier trained on the General Knowledge QA dataset is 0.91, as visualized in Fig. 4. ROC curve is a two-dimensional curve plotting True Positive Rate versus False Positive Rate for binary classification problems, i.e., the task of distinguishing the correctly and incorrectly predicted answers from binary predictions. The GBM has excellent discrimination power indicated by the rising, and remaining largely above, the dashed “Random Guess" line (AUC=0.5).

This high AUC indicates that the GBM's iterative training process (i.e., each weak learner correcting prior errors) is consistent with the methodology. The strong performance of the GBM in improving the child learning system through the classification of answers in educational QA settings is validated.

B. *Adaptive Learning Effectiveness and Explanatory Feedback Quality*

Fig. 5 illustrates the difficulty progression curve for the Adaptive Learning Effectiveness in the educational experience. As the learners' difficulty level increases, answering questions correctly promotes cognitive engagement. On the other hand, incorrect answers will indirectly decrease the challenge and keep children interested by not exposing them to too complex tasks. The way it adjusts in this manner supports optimal learning efficiency and facilitates the gradual development of knowledge tailored to each person's capabilities.

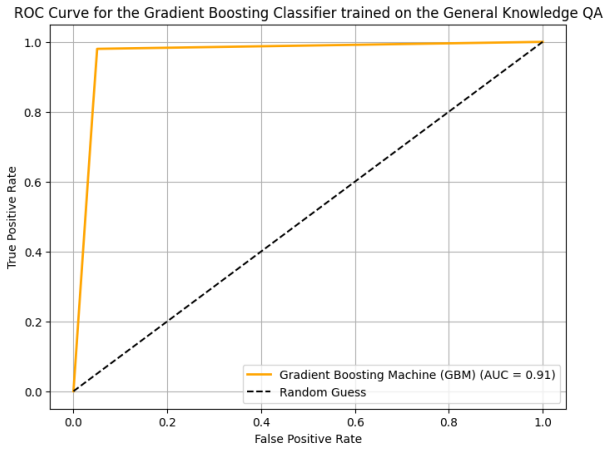


Fig. 4. ROC curve illustrating the classifier's performance in distinguishing between correctly and incorrectly predicted answers.

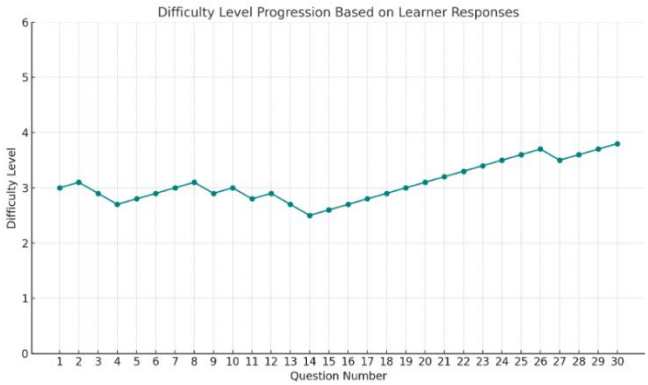


Fig. 5. Difficulty adjustment over time, increasing with correct answers and decreasing with mistakes.

TABLE III EXPLANATORY FEEDBACK QUALITY RATINGS FOR INCORRECT ANSWER EXPLANATIONS

Criterion	Average Rating (Out of 5)
Relevance to Concept	4.6
Age Appropriateness	4.8
Understandability	4.7
Encouraging/Positive Language	4.5

A manual inspection of 100 samples of feedback generated by the finetuned T5 model was conducted to evaluate the Explanatory Feedback Quality. For the explanations, we asked for four basic criteria to rate these, namely Age Appropriateness, Relevance to Concept, Understandability, and Encouraging/Positive Language. In Table III, the average ratings for each criterion are summarized. The model was rated highly on relevance to the concept, with a score of 4.6, indicating that the explanations were very close to the heart of the questions.

In addition, the feedback obtained an excellent score (4.8) in the Age Appropriateness category for the language level that was suitable for the student's age group and a high score (4.7) in the Understandability category for the ease of comprehension. Additionally, the use of encouraging and positive language was scored as 4.5, indicating that the explanations were both informative and motivational, which is essential to keep learners engaged.

C. Learner Feedback and System Usability

A comprehensive overview of the key metrics evaluating system performance and user feedback is provided in Table IV, titled "System Performance and User Study Results". A test set accuracy of 85.2% demonstrates that the system is exceptionally accurate in correctly classifying answer options.

TABLE IV SYSTEM PERFORMANCE AND USER STUDY RESULTS FOR THE AI-DRIVEN CHILD LEARNING SYSTEM

Metric	Value	Description
Test Set Accuracy	85.2%	Correct classification of answer options
Average Cosine Similarity (Correct Pairs)	0.82	High semantic matching between question and correct answer
Average Cosine Similarity (Incorrect Pairs)	0.41	Lower semantic similarity for wrong answers, as expected
Hyperparameter Tuning Method	5-Fold Randomized Search CV	Optimized learning rate, depth, and subsampling rates
Adaptation Rate (η)	0.2	Controlled difficulty adjustment sensitivity
Learner Feedback (Understanding Improvement)	85%	Learners reported better understanding after explanations
Learner Preference for Adaptive System	78%	Learners preferred adaptive quizzes over static ones

The Average Cosine Similarity (Correct Pairs) of 0.82 indicates that semantically aligned questions and correct answers provide strong evidence of the system's ability to understand contextual relationships. On the contrary, the Average Cosine Similarity (Incorrect Pairs) is 0.41, which corresponds to a large semantic gap between the correct answer and incorrect options, as it is expected to be. Therefore, the values of the hyperparameters were tuned via a 5-fold Randomized Search cross-validation to achieve the optimal learning rate, tree depth, and subsampling rates.

An Adaptation Rate ($\eta = 0.2$) suggests a moderate response to changes in question difficulty to match learner performance. User feedback also revealed that 85% of the learners felt that the explanations for the wrong answers helped them understand better, while 78% preferred the adaptive system over static quizzes, indicating the positive effect of the system on learning outcomes.

Fig. 6 shows the learner feedback and preference for the Adaptive System, further visualizing the results mentioned above. A user study outcome of the system is presented, illustrating the percentage of learners who improved their understanding and those who preferred the adaptive system.

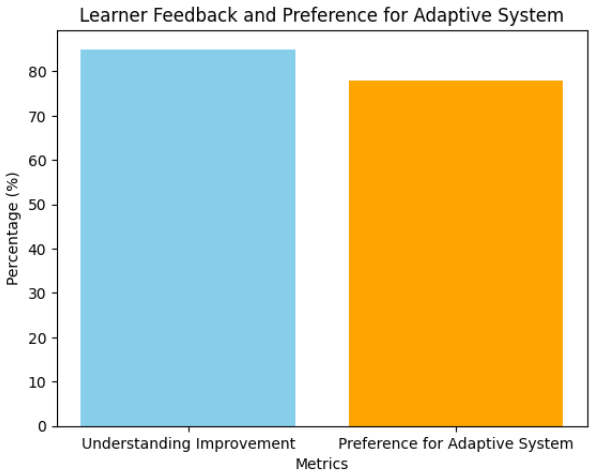


Fig. 6. Bar plot of learner feedback and preference for the adaptive system.

V. CONCLUSION AND FUTURE DIRECTIONS

This study presents an AI-driven child learning system based on ML and NLP to provide personal, adaptive educational experiences among children. The semantic embedding of questions and answers in an SBERT model, combined with answer prediction using a GBM classifier and feedback creation for age-appropriate explanations using a T5-based engine, enables the system to achieve 85.2% accuracy and 0.912 ROC AUC on the General Knowledge QA dataset.

Based on student performance, the adaptive learning engine adjusted the difficulty of questions, while T5 generated feedback that was well-rated for relevance (4.6/5) and age appropriateness (4.8/5). Further user studies revealed that the system had a predominantly positive effect, with 85% of learners reporting improved understanding and 78% preferring the adaptive system over static quizzes. These results indicate the promising potential of ML and NLP in helping to scale child learning through personalized content and feedback that surmounts the constraints of traditional one-size-fits-all approaches in education.

Future research can be directed into various areas to enhance the system's capabilities. The first one lies in expanding the dataset to include more diverse subjects and multilingual content, which will improve generalizability and support inclusive learning in children from different backgrounds. The second approach is to utilize real-time speech recognition and an interactive dialogue system for more natural, voice-based interaction and enhanced engagement among younger learners.

Finally, the adaptive engine's reinforcement learning capabilities can be leveraged to optimize difficulty adjustments by learning long-term trends in student performance. Ethical issues such as data privacy and bias in AI-created content need to be investigated critically in order for the system to be successfully deployed in real-world educational settings.

DISCLOSURE AND CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this research. Additionally, the author has no financial interests or competing affiliations that could have influenced the study's design, execution, or findings. This manuscript is the author's original work and has not been previously published or submitted for review to any other journal or conference.

REFERENCES

[1] S. S. Khan, A. U. H. Rupak, W. W. Faieaz, S. Jannat, N. N. I. Prova, and A. K. Gupta, "Advances in medical imaging: Deep learning strategies for pneumonia identification in chest x-rays," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2024, pp. 1–7.

[2] R. Tiwari, "The integration of AI and machine learning in education and its potential to personalize and improve student learning experiences," *International Journal of Scientific Research in Engineering and Management*, vol. 7, no. 2, 2023.

[3] A. N. Mathew, V. Rohini, and J. Paulose, "Nlp-based personal learning assistant for school education," *Int. J. Electr. Comput. Eng*, vol. 11, no. 5, pp. 4522–4530, 2021.

[4] N. Prova, "Detecting ai generated text based on nlp and machine learning approaches," *arXiv preprint arXiv*:2404.10032, 2024.

[5] Kinshuk, N.-S. Chen, I.-L. Cheng, and S. W. Chew, "Evolution is not enough: Revolutionizing current learning environments to smart learning environments," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 2, pp. 561–581, 2016.

[6] N. Rane, S. Choudhary, and J. Rane, "Education 4.0 and 5.0: Integrating artificial intelligence (ai) for personalized and adaptive learning," 2023.

[7] L. Deng and Y. Liu, "A joint introduction to natural language processing and to deep learning," *Deep learning in natural language processing*, pp. 1–22, 2018.

[8] N. N. I. Prova, "Garbage intelligence: Utilizing vision transformer for smart waste sorting," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI). IEEE*, 2024, pp. 1213–1219.

[9] J. Chen, Y. Lu, S. Zhang, B. Yao, Y. Dong, Y. Xu, Y. Li, Q. Wang, D. Wang, and Y. Sun, "Storysparkqa: Expert-annotated qa pairs with real-world knowledge for children's story-based learning," *arXiv preprint arXiv:*2311.09756, 2023.

[10] M. Sammoudi, A. Habaybeh, H. I. Ashqar, and M. Elhenawy, "Questionanswering (qa) model for a personalized learning assistant for arabic language," in *International Conference on Intelligent Systems, Blockchain, and Communication Technologies. Springer*, 2024, pp. 356–367.

[11] R. Abdelghani, Y.-H. Wang, X. Yuan, T. Wang, P. Lucas, H. Sauzeon,' and P.-Y. Oudeyer, "Gpt-3-driven pedagogical agents to train children's curious question-asking skills," International Journal of Artificial Intelligence in Education, vol. 34, no. 2, pp. 483–518, 2024.

[12] S. P. Bode and R. S. Satpute, "Natural language processing in education system," in *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*. IEEE, 2024, pp. 1–5.

[13] G. B. Herwanto, G. Quirchmayr, and A. M. Tjoa, "Leveraging nlp techniques for privacy requirements engineering in user stories," *IEEE Access*, vol. 12, pp. 22 167–22 189, 2024.

[14] H. R. Alsulami and A. A. Almansour, "Exploring gpt-4 capabilities in generating paraphrased sentences for the arabic language," *Applied Sciences*, vol. 15, no. 8, p. 4139, 2025.

[15] L. Morales-Navarro, D. J. Noh, and Y. B. Kafai, "Building babygpts: Youth engaging in data practices and ethical considerations through the construction of generative language models," *arXiv preprint arXiv:*2504.14769, 2025.

[16] S. Krause and F. Stolzenburg, "Commonsense reasoning and explainable artificial intelligence using large language models," in *European Conference on Artificial Intelligence*. Springer, 2023, pp. 302–319.

Application of Blockchain Frameworks for Decentralized Identity and Access Management of IoT Devices

Sushil Khairnar
Virginia Tech, Blacksburg, 24060-61, USA

***Abstract*—The growth in IoT devices means an ongoing risk of data vulnerability. The transition from centralized ecosystems to decentralized ecosystems is of paramount importance due to security, privacy, and data use concerns. Since the majority of IoT devices will be used by consumers in peer-to-peer applications, a centralized approach raises many issues of trust related to privacy, control, and censorship. Identity and access management lies at the heart of any user-facing system. Blockchain technologies can be leveraged to augment user authority, transparency, and decentralization. This study proposes a decentralized identity management framework for IoT environments using Hyperledger Fabric and Decentralized Identifiers (DIDs). The system was simulated using Node-RED to model IoT data streams, and key functionalities including device onboarding, authentication, and secure asset querying were successfully implemented. Results demonstrated improved data integrity, transparency, and user control, with reduced reliance on centralized authorities. These findings validate the practicality of blockchain-based identity management in enhancing the security and trustworthiness of IoT infrastructures.**

***Keywords*—Blockchain; decentralization; identity and access management; ethereum; hyperledger**

I. Introduction

Internet of Things (IoT) is an emerging field of research which is about connecting smart devices, people and systems together using sensors, software and the other technologies to capture meaningful information from the surrounding environment of devices and take actions to improve productivity and efficiency. There is an increase demand for IoT devices ranging from Google Home and Amazon Alexa to Smoke Alarm and the next Door bell. In 2021, there were more than 10 billion active IoT devices. By 2025 it is estimated that there will be over 150,000 IoT devices connecting to the internet per minute. It is predicted that the number of active IoT devices will surpass 25.4 billion by 2030 [2][17]. However, the proliferation of IoT devices all over the world raises concerns over security, privacy and device management. IoT devices generate huge amounts of data most of which is sensitive data belonging to the individuals and the companies. One of the most popular IoT based attacks was the Distributed Denial of Service (DDoS) attack of 2016 which caused major Internet platforms and services to be unavailable to large swathes of users in Europe and North America [3][28].

In the case of James Bates of Arkansas, USA, who was charged with first-degree murder partly with the help of evidence collected by an Amazon Echo smart speaker in 2015 [4][5][6]. The prosecutor used the data found on Bates Amazon Alexa data along with his smart meter data to build the case. During the case, Amazon refused to release the data collected using Alexa. However, the defendant gave permission to use the data during the case. According to the reports, the Echo wasn't the only smart device investigators cited in their case - they also used "information from a smart water meter, alleging that an increase in water use in the middle of the night suggests a possible cleanup around the crime scene"[23]. The case was dismissed in December 2017, but not before becoming a news item and drawing the defendant's personal life into the press[24] [11].

IoT finds its applications in various areas including health care, smart homes, self driven cars and smart cities. The data captured from the devices is used to gain insights and make decisions. This decision making process "trusts" in the information obtained. There- fore, "trust" is the pillar of the advancements in IoT. To ensure that the captured data is obtained from reliable sources, blockchain technology plays a crucial role. Currently, more than 10 billion de- vices are connected over internet through IoT. Inorder to ensure the "trust" we must be able to verify and validate the identity of these devices through out their life cycle. The identity of these devices is tracked by blockchain [8]. Since decentralization is at the heart of the blockchain technology, we leverage this to eliminate centralized identity management of devices[1].

The device management particularly, is considered a critical task in IoT applications because of the large number of devices connected and the multifariousness of the network. Decentralized identifier (DID) is a new model for creating global unique identifiers. These identifiers are built on top of distributed ledgers. DIDs are created as the foundation for decentralized digital identity and public key infras- tructure (PKI) for the Internet. Among the main features is that DIDs offer full control of identity to consumers without any dependence on the central authority. DIDs are defined by a generic format URL that is used to address DID documents and to be resolved by a DID resolver. DID generic format is a URI scheme conformant with the web standards as per RFC3986 [25]. A DID URL is a network loca- tion identifier for a specific resource. It can be used to retrieve things like representations of DID subjects, verification methods, services, specific parts of a DID document, or other resources [26]. The things included in DID URL include the urn scheme -'did' in this case, followed by the namespace or method name (e.g. ethr, sov, etc), and after that is the method-specific identifier which can be a combina- tion of letters and numbers (e.g."did:example:1234AbxfgHUfgh").

However, because of the decentralized nature of blockchain, there is a severe limitation over its performance. For example, Bitcoin can only achieve a low throughput of 7 transactions per second (TPS), and takes over 10 minutes for a transaction to get confirmed [7]. Contrary to the current centralized payment systems such as Master- card and Fedwire can reach thousands of transactions per seconds and almost real-time payments [14]. Performance evaluation and analysis plays an important role to find bottlenecks in the blockchain systems and can be used to inspire further optimization ideas. The performance indicators of the blockchain mainly include transaction throughput and latency. Transaction throughput represents the num- ber of transactions that can be processed at a fixed time, and latency represents the response and processing time to transactions [27]. In addition to these two indicators, resource utilization and ease of use can also be considered among the metrics for blockchain systems evaluation.

This research addresses the question: Can blockchain-based decentralized identity management systems enhance trust, authentication, and ownership control in dynamic IoT environments while maintaining data integrity and privacy?

The remainder of this study is organized as follows: Section II provides an overview of the system architecture. Section III details the design principles. Section IV outlines the implementation steps. Section V evaluates the performance. Section VI compares related work, followed by a new discussion in Section VII. We conclude and present future work in Sections VIII and IX.

II. Overview

To bring real trust to the data captured by devices we leverage the core principles of blockchain. The blockchain is a digital ledger of past transactions and has found massive use cases in areas other than cryptocurrency. Its strengths make it an ideal component for IoT solutions. Its design eliminates the need for centralized authorities to govern the system, the transactions stored in every block of the blockchain are completely auditable and are public, these transactions are verified and validated through consesus of the peers and the immutability of the blockchain makes the system tamper proof and trustworthy[13]. We develop on these principle to provide unique identities to devices at the time of their creation. Every device is in control of its identity. This device publishes its public key to the distributed ledger and gets a DID (Document Identifier) in response. Every device presents this DID to the blockchain it wishes to add a transaction to. The blockchain then verifies its identity using this DID and validates it [7]. The digital identity for users works on the above principle. We applied this mechanism for IoT devices. As a blockchain is write forward and immutable, blockchain-born devices thus will have irreversible reputations and identities. We also examined the possibility of creating unique identifiers that are tamper-resistant by combining a sensors environmental inputs like precipitation, temperature, humidity and more. For access management we use smart contracts between parties involved.

We used Node-RED to load and test our blockchain systems over hyperledger framework. Node-RED is a programming language for connecting hardware, APIs, and web services. It is primarily a visual tool for the Internet of Things built by the IBM Emerging Technology organization and is an open source.

Following are some of the most important features of Node-RED architecture which led us to select this tool:

- Node-RED (latest version 0.16) is quick since it runs on the most recent version of Node.js that is supported (LTS 6.x, 6.x).
- A single-threaded event queue that simplifies things.
- Event-driven architecture and asynchronous I/O
- The entire architecture was created with express, d3, jquery, and ws.
- It enables a lightweight runtime environment, as well as an event-driven and non-blocking approach using Javascript
- JSON is used to hold the numerous flows made in Node-RED, which can be readily imported and exported for sharing with others.

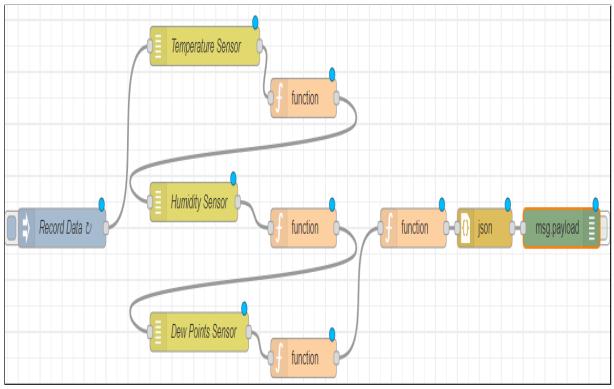


Fig. 1. IoT Network simulation using node-red.

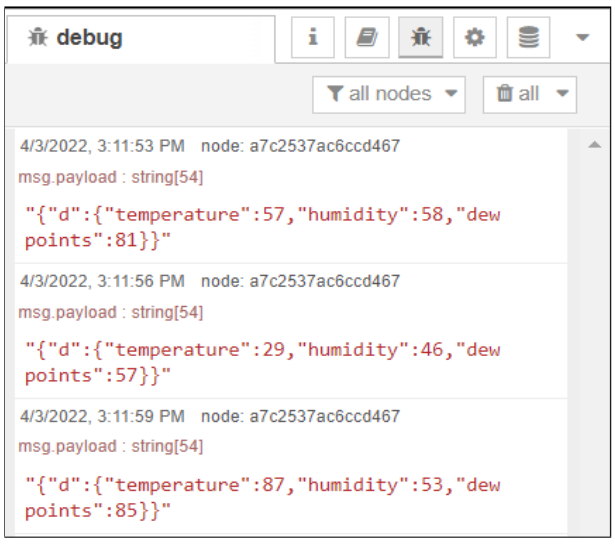


Fig. 2. Node-red - message payload.

In order to understand the usage of Node-RED, simulate an IoT network, and gather transactional data in it, we replicated

a small network consisting of three sensory nodes. To generate the data, random nodes are used as sensors which output values in a given range. Timestamp node is used to record data every three seconds. The generated data flows as transactions through the network and accumulated as json output at the end of network. Fig. 1 shows how this network is simulated on Node-RED. Fig. 2 shows the aggregated output data in the json format.

We covered the following cases for device access management:

- Registering a new IOT device in the Blockchain
- Login feature for the registered IOT devices
- Viewing all existing assets or transactions in the blockchain system.
- Viewing assets owned by the device that is currently logged in.
- Uploading a new IOT device data to the blockchain.

III. Design

A device's life cycle in IoT networks begins with its manufacture and ends with its disposal. The suggested decentralized IdM architecture displays a cycle that specifies the status of the device and who owns it. The device identity must be maintained throughout this life cycle through a global registry that allows authorized registrants (for example, manufacturers) to create new identities for new devices. It enables the registrant to carry out identity management duties or transfer ownership of the device(s) to other parties, such as operators and end-users.

The following are some of the issues we plan to address in regard to the IAM module of the IoT management framework:

- Establish a global and standardized register for IoT devices.
- As the lifecycle of IoT devices begins, define explicit registration processes.
- Ensure that security measures are implemented during the registration process, as compromising the IoT device's identification threatens the sensitive data managed by the IoT device.
- Create an identity lifecycle that can be applied to items in an organization and customized based on the device's lifetime.
- Provide tools for managing device ownership and transferring ownership.
- Provide a way for tracing the device's lifecycle back to its source.
- The construction of a device's digital identity and the necessity for identity formation to reflect device lifecycle changes [30].

The identity creation is performed when the owner submits a request to the IdM smart contract to produce a new DID, which is carried out according to the protocol outlined below. When the contract succeeds, it commits the DID record to the blockchain and stores the DID Document on the hub [29] stated in Algorithm I.

Algorithm 1 Identity Creation

Input: publicKey, owner, id, hash, owner
Result: (DID) Committed on Chain
Function DIDGenerator(publicKey, id, owner, proof);
Initialization();

if *Sender is Authorized* **then**
  CreateDIDIdentifier(id) **if** *DID doesn't exist* **then**
    valid ← VerifyProof(publicKey,proof) **if** *valid* **then**
      hash ← storeIPFS(DIDDocument) ListDID[DID] ← hash
    else
      pass;
    end
  else
    Return (Identity already exists);
  end
else
  pass;
end

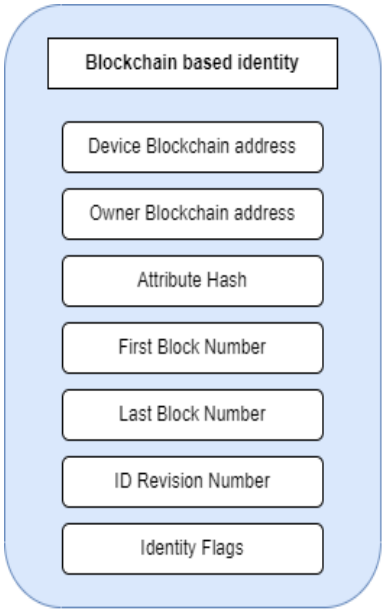


Fig. 3. Digital identity data structure.

The above-mentioned device identity ownership management needs are addressed in this study using blockchain and smart contracts. The immutability and provenance capabilities of blockchain, as well as the digital identity presented in the preceding section, are used to ensure identity ownership to meet the requirements. The identity data structure is modeled using decentralized identifiers (see Fig. 3), enabling secure and user-centric device registration [9].

In addition to the other features, the blockchain identity generated at the moment of device birth is built with a set of attributes that includes the blockchain addresses of the device itself and the original owner (maker). The produced identity

is kept in the distributed ledger with a number of properties, the most essential of which is the owner address in this case. The owner's blockchain address is used to identify ownership, and this feature is only accessible to the owner thanks to smart contract structures. As a result, only the device identity owner can transfer ownership, ensuring a secure transfer of ownership from one entity to another. Furthermore, the change of ownership is predicated on the former owner being replaced.

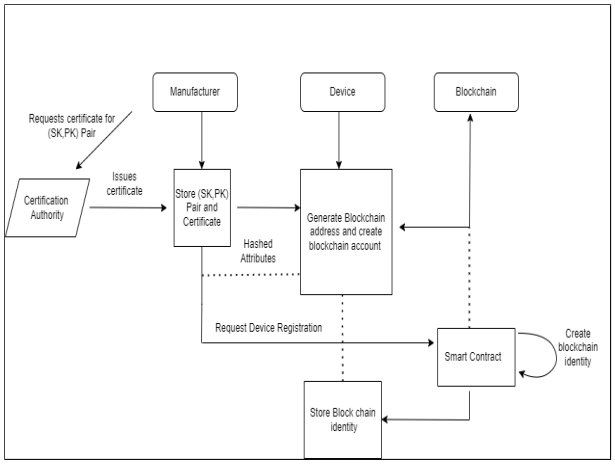


Fig. 4. Device identity management architecture.

Public-key cryptography and hashing functions are used to create identities (Fig. 4). The identity is made up of the minimum set of attributes, that are required to provide a unique and global identity for the device in the global registry. This is accomplished in two steps: the creation of cryptographic keys and blockchain addresses, as well as the creation of a blockchain-based digital identity. The decentralized identity technique suggests utilizing cryptographic keys to create anonymous IDs or using the device technology unique identifier (e.g., IMEI or UUID) as the DID identification.

All of the electronic devices which are capable of connecting to the internet can act as entities that have ability to be a part of an IoT network. Formation of an IoT network could be impacted by number of factors such as types of the devices, roles these devices play, and possible values of physical distance between the devices etc. However, depending on the functionality, the nodes in the network can broadly be classified into three main components: coordinator, router and the end devices [27]. Some nodes deliver information and other collect and deliver it. Also, a set of nodes control other set of nodes. Formation and simulation of such network and transfer of data from the network to a blockchain will be addressed by using framework known as Node-RED [25].

To carry out these operations smoothly, each entity and device need to be well identified and documented. Identification is not only limited to the devices but it is required for entities such as the users, data, services, etc. Also, one entity can be a part of multiple IoT networks. Hence, depending upon the scope of different networks, assigning unique identity plays key role in managing and ensuring faultless operation of the network. Number of approaches are available for the allocation of new identities. For instance, a common repository can be maintained to allocate unique keys [28]. Apart from this, an algorithm can be used to generate random keys from a sizable pool of keys. If there is hierarchy of networks, separate pool of identities can be maintained to ensure appropriate allocation of identities.

For the implementation of this approach we have divided the system into two components. The first component will be the implementation of device identity management and the second will deal with the application of blockchain in IoT. The former helps in verifying the identity of the device before it is added to the network. In case of the latter, we propose to bootstrap a Hyperledger fabric that resembles a network of temperature sensors in a warehouse. The network consists of organizations and sensors that deploy the chaincode (smartcontract) and drive the execution of the transactions [26].

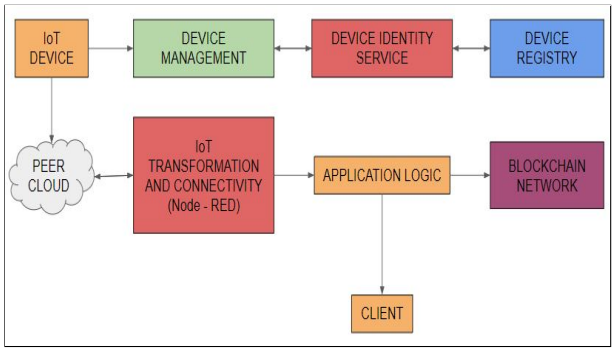


Fig. 5. System architecture.

The components of the system architecture (Fig. 5) are described as below:

1) IoT device: This can be a sensor, actuator that has a network connection and may also have a user interface.

2) Device management: It provides a way to securely manage devices that become a part of the network.

3) Device identity service: It reliably verifies the identity of the device once the device is registered.

4) Device registry: It stores the information about the devices.

5) Peer cloud: This is a broker into which the measured data from the sensors is pushed at regular intervals.

6) IoT Transformation and Connectivity (Node-RED): This is an IoT connector tool by Node.js. Node-RED is a programming tool for wiring together hardware devices, APIs and online services in new and interesting ways [25].

7) Application logic: This is the core component of application. It coordinates the handling of data received from Node-RED and passes it to the blockchain network.

8) Blockchain network: It is a network of the blockchain nodes that maintain the distributed ledger.

9) Client: This is the end-user application.

The runtime flow of the system begins with verifying the identity of the IoT device and the IoT device sending events such as temperature, humidity, etc to the peer cloud along with the timestamp. The IoT transformation and connectivity

service performs some event filtering at this step before this data flows into the core component of the application. The application then prepares and maps the data to be able to send it as transactions to the blockchain. Notifications from the blockchain such as settlement of a transaction or violation of a contract are sent to the application. This blockchain network is implemented using the HyperLedger Fabric Model (Fig. 6). Every peer node in Fig. 6 runs in docker containers and relies heavily on smart contracts. Chaincode specifies the asset's business logic, or the rules for reading and changing the asset's so-called state. Fabric, unlike public cryptocurrency blockchains, allows parties to create a private route for their assets, isolating and segregating transactions and a ledger.The end-user application then displays these notifications to the client. This gives the ability to the customers to check the status of their assets by tracking the record by using an application UI. The solution searches the blockchain's transaction history and sends out various forms of notifications on transaction status.

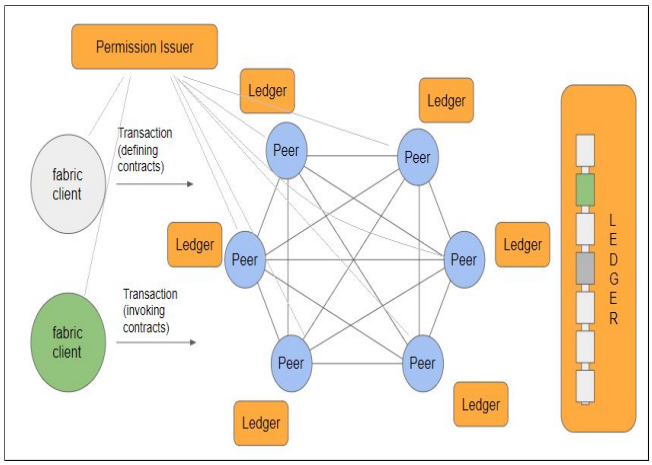


Fig. 6. Hyperledger fabric model.

Hyperledger Fabric is a permissioned distributed ledger framework for building enterprise-grade systems and apps. Its modular and adaptable architecture caters to a wide range of industry applications. It takes a novel method to consensus that allows for scalability while maintaining anonymity. Public blockchains, such as Bitcoin, do not require permission. Anyone is welcome to participate. Hyperlegder, on the other hand, is a permissioned blockchain platform, which means that only validated identities can join.

Some of the common issues with privacy are:

- Scalability: Because all nodes on the network must execute transactions, most public blockchains have low scalability. As a result, transaction throughput is low [10].
- Storage issues: Because every full node on the public blockchain contains all of the data, storage requirements are quite high and are steadily increasing over time. This level of storage redundancy isn't viable for most applications.
- The 'Proof Of Work' (POW): technique, which needs a lot of computational power and is energy-intensive, is used by Bitcoin and Ethereum. As processing power and energy requirements rise over time, it becomes impractical in a corporate setting. Hyperledger Ordering Node, on the other hand, employs a variety of techniques, including Solo, Kafka, and PBFT.
- Lack of governance: Because no one has control over public blockchains, it is up to individual developers or developer communities to implement improvements. To run blockchain properly, businesses require adequate governance [26].

IV. Implementation

We cover the following use cases in this study:

- Device registration and device login.
- Viewing all existing assets in the system.
- Viewing assets owned by the device that is currently logged in.
- Uploading a new device data.
- Simulation of IOT environment using Node Red.

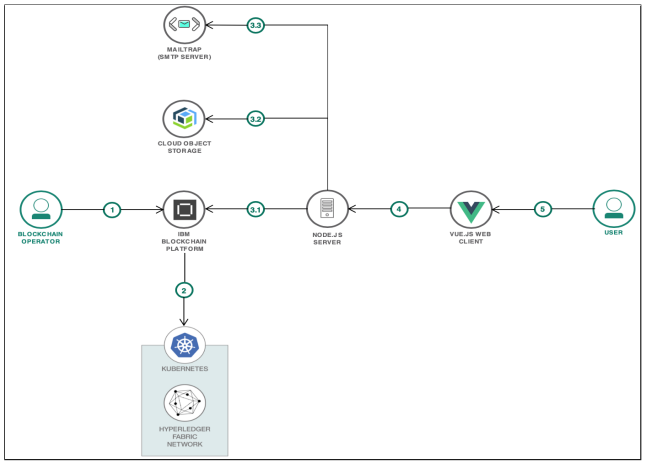


Fig. 7. Application design.

The steps below elaborate Fig. 7:

- The IBM Blockchain Platform service is set up by the Blockchain Operator.
- On an IBM Cloud Kubernetes Service, the IBM Blockchain Platform service constructs a Hyperledger Fabric network, and the Blockchain Operator installs and instantiates the smart contract on the network.
- The Node.js application server creates APIs for a web client by interacting with the deployed network on IBM Blockchain Platform, IBM Cloud Object Storage instance, and Mailtrap Server (fake SMTP testing server) using the Fabric SDK.
- To connect with the network, the Vue.js client uses the Node.js application API.
- The user interacts with the digital asset management application using the Vue.js web interface.

The major components involved are:

- Blockchain platform: provides you complete control over your blockchain network with a user interface that can simplify and accelerate your journey to deploy and manage blockchain components on the IBM Cloud Kubernetes Service.
- The IBM Cloud Kubernetes: Service launches highly available containers in a cluster of compute hosts. A Kubernetes cluster allows you to manage the resources you need to swiftly deploy, update, and grow applications in a secure manner.
- The IBM Blockchain Platform Extension for VS Code: helps users create, test, and deploy smart contracts, including connecting to Hyperledger Fabric settings.
- IBM Cloud Object Storage: is a scalable cloud storage solution with a focus on durability, robustness, and security.
- Mailtrap.io: is a test mail server that allows you to test your email.

Hyperledger Fabric's design is highly flexible and configurable, allowing for creativity, variety, and optimization across a wide range of business use cases, including banking, finance, insurance, healthcare, human resources, supply chain, and even digital music delivery.

Fabric is the first distributed ledger platform that support smart contracts written in general-purpose programming languages like Java, Go, and Node.js rather than domain-specific languages (DSL). The platform is also permissioned, which means that, unlike a public permissionless network, participants are known to one another rather than anonymous and hence completely untrustworthy. While the members may not entirely trust one another (for example, they may be competitors in the same business), a network can be run under a governance model based on the trust that does exist between them, such as a formal agreement or a framework for resolving disputes.

A. IBM Blockchain Platform

IBM Blockchain Platform is a managed and full-stack blockchain-as-a-service (BaaS) product that lets you install blockchain components in your preferred environment. Clients can use an offering that can be used from development to production to construct, operate, and grow their blockchain networks. It contains the IBM Blockchain Platform console, user interface and VS code extension that makes deploying and managing blockchain components easier and faster. Fig. 8 given below demonstrates the IBM Blockchain Platform components.

1) IBM Blockchain Platform Console (UI): This is the user interface for creating and managing blockchain components. You can launch an instance of the IBM Blockchain console and link it to your Kubernetes cluster on IBM Cloud when you provision a service instance in IBM Cloud. Then, on your Kubernetes cluster, you can utilize the console to create and manage your blockchain components. The console is free of charge.

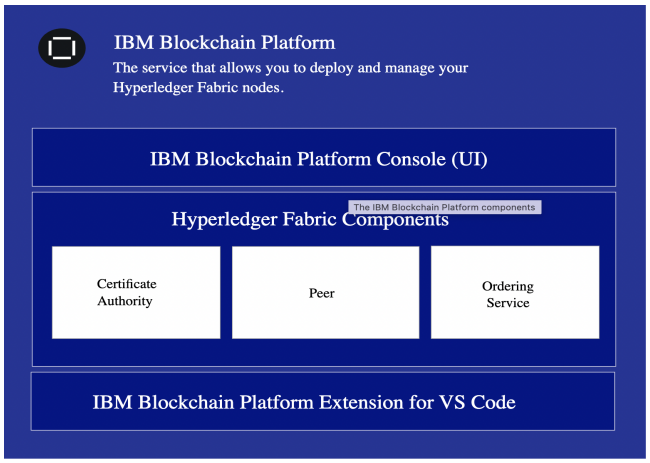


Fig. 8. The IBM blockchain platform components.

2) Hyperledger fabric components: The console is used to build and manage Hyperledger Fabric v1.4.12 and v2.2.5 Certificate Authority, peer, and ordering service images, as well as other blockchain components. When these components are deployed into your Kubernetes cluster, the default storage class is used to provision storage for them.

3) IBM VS Code extension (Development tools): For writing, packaging, and testing client applications and smart contracts, download the VS Code extension from the VS Code marketplace.

Each CA, peer, and ordering node in the IBM Blockchain Platform requires persistent storage. A pre-configured storage class powered by IBM Cloud File Storage is used when you install a standard Kubernetes cluster in IBM Cloud. You can alter your backing storage, improve your storage class, or use third-party services.

Every cluster on your IBM Cloud Kubernetes cluster has a default storage class that is used to provision persistent storage on IBM Cloud. When you use the IBM Blockchain Platform UI or APIs to deploy a blockchain node to that cluster, the node utilizes this default storage class to dynamically provision the amount of storage you specify on IBM Cloud.

B. Smart Contracts

A smart contract, or "chaincode" as Fabric defines it, is a trusted distributed application that gets its security or trust from the blockchain and the underlying peer consensus. It is a blockchain application's business logic. Smart contracts have three main characteristics, especially when used on a platform:

- Many smart contracts run in the network at the same time.
- They can be installed dynamically (by anyone in many cases).
- Application code should be treated as untrustworthy, if not malevolent.

The implementation details of the application are based on the the following smart contracts:

- Checking whether the device is registered in the blockchain
- Registered device should only be able to do the login.
- Viewing all existing assets or transactions in the system.
- Viewing assets owned by the device that is currently logged in.
- Uploading a new IOT device data.

We have implemented a smart contract to avoid duplication of assets, duplication of devices during registration, and to ensure that the device is registered before login. Each of these functionalities empower and strengthen the security and confidentiality features offered by the properties of blockchain. First feature ensures that there are no same copies of singular asset. This helps avoid potential complications due to the double-spending problem. Also, makes sure that there is one copy of an asset in the blockchain. If any devices with malicious intent tries to make multiple copies the assets, the device will be denied to do so. Avoiding duplication of devices during registration makes sure that malicious device would not be able to register by using other device's identity.

The device registration smart contract creates a worldwide register, where the unique DID can be stored. The global registry records the device entries so that an audit trail may be created for all events since the device was manufactured. In this proposal, an owner attribute is proposed to allow the device owner to execute operations on the device DID document as well as access to DID methods.

The access control mechanism for these functions is based on the different modifiers specified in this contract. The root of trust is formed with such a level of control over identity formation, and the device owner ultimately has full control over the device identity.

C. Node-RED

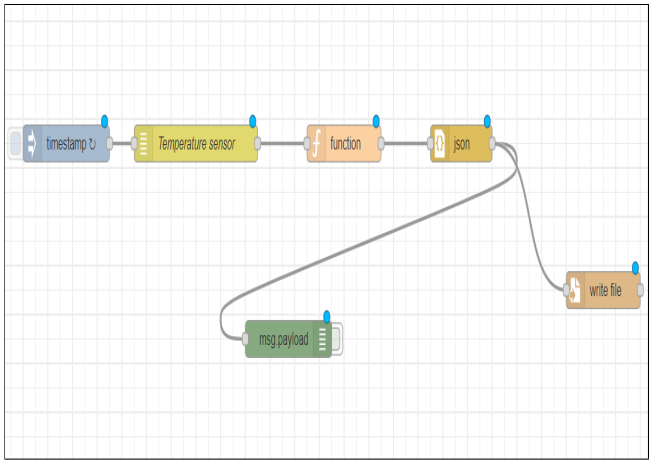


Fig. 9. IoT Device simulation with Node-RED.

Node-RED is an open source visual editor that allows us to connect hardware devices, create flows, and write Javascript functions to control those flows and devices. We used Node-RED to simulate an IoT device, generate data from the device and eventually simulate a network of five devices for the demonstration. Following are the reasons why we chose Node-RED: When it comes to programming with Node-RED, the simplicity is the key parameter. As the name implies, programming is performed intuitively with the use of a small number of actions. The majority of Node-RED tasks may be accomplished nearly entirely through GUI actions. So you can focus on programming, the Node-RED flow editor takes care of creating the application application infrastructure, library sync, integrated development environment. The actual benefit of flow-based and visual programming is high quality. Each component node is a fully functional module that has been thoroughly tested. As a consequence, app writers may concentrate on validating the operation at the join level rather than the information of the node. This is a significant aspect in preventing human mistake at the individual level and ensuring good quality.

```
flow.set('random1', msg.payload);
msg.payload={'random1':msg.payload};
let manufacturer_id = 'ABCDEF00001';
msg.payload={'d':{'manufacturer_id':manufacturer_id, 'Time': new Date(), 'temperature':flow.get('random1')}};
msg.header={"Content-type":"application/json"}
var c = context.get('c') || 0
c ++
context.set('c',c)

var path = "~/blockchain/nodered/device1/"+c+".txt"
msg.filename = path
return msg;
```

Fig. 10. Function node in Node-RED.

For our purpose, initially, we simulated an IoT device with the help of five nodes, as shown in the Fig. 9. First node is the inject node which initializes the flow as per our needs. We have set it to execute the flow every 30 seconds. This means that we will be recording temperature every 30 seconds. Temperature values are generated with help of this node which is a random node. It randomly outputs a value between the 0 to 100. Third node is the function node which receives the value generated by the temperature sensor. In addition to this, a unique DID is assigned to the device which is required for registration and authentication of the device. Also, every time the sensor measures the data, it is stored in a different file. . With the time, these files are generated and act as assets for that particular device. The code shown in Fig. 10 is used for the function to perform the above mentioned tasks. The next three nodes are used to process the JSON string which is created in the function node. This processed string is displayed on the node red debug console with the debug node. Also the data is then written to a file with a "write file" node. We added five such devices in the network to simulate an IoT network. Each of those devices record and store the assets in a defined directory structure. Fig. 11 shows the format in which data is stored in one file. Such file is created every thirty seconds for every five devices.

V. Evaluation

The data generated by the network simulated in Node red is used to perform various operations. Two basic functionalities



Fig. 11. Data storage in a file.

we have provided for the devices are login and register. The devices will use the unique DID assigned to them for registering and then for login. This ensures authentication and adds level of security. An error will be shown if the a device which not registered tries to login. After login (Fig. 12), the device can query its own assets. In addition to this, there is a functionality for the device the device to see the names of assets added by other devices. Fig. 13 shows the format in which data is displayed. The device can see the asset owner, asset name, the timestamp on which asset was added to the blockchain, and a unique data ID which is used to check the duplication of assets in the system.



Fig. 12. Feature 1 and 2 - Device login and register.

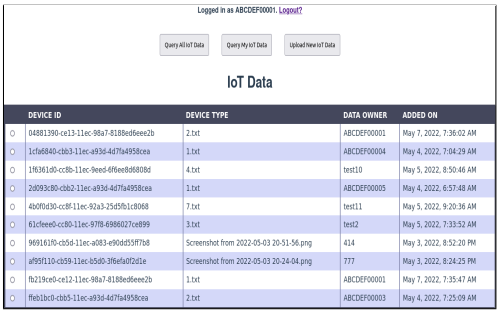


Fig. 13. Feature 3 - Query of all transactions in the blockchain.

A functionality is provided for each device to upload the recorded data to the blockchain. Fig. 14 shows the interface to upload the data. After the data is uploaded, it can be queried in the option to show device's data. Fig. 15 shows how the device's data is shown in the web interface. If device uploads the same data again, an error will be shown which informs the device that the same data or asset already exists in the system.

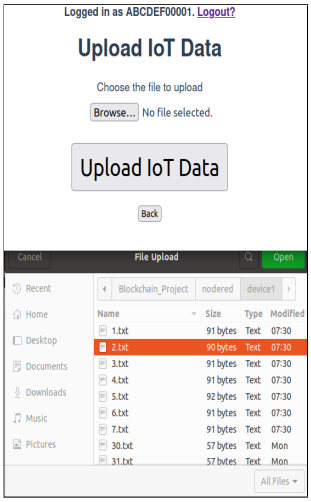


Fig. 14. Feature 4 - Upload new IoT device data (transaction) to the blockchain.

In addition to this, a unique dataID will be displayed which represent the asset that already exists on the blockchain. This helps to prevent possible duplication of data on the blockchain.

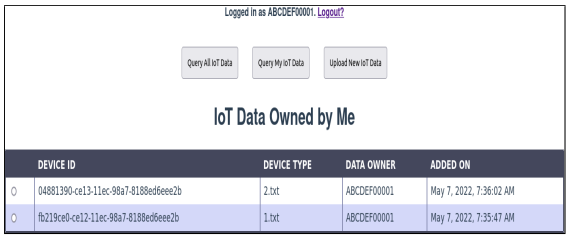


Fig. 15. Feature 5 - Query data owned by the logged in device.

VI. Related Work

There have been several inquiries and implementations of Blockchain based identity management solutions, with multiple studies investigating different approaches to achieve a reliable framework. Uport [15] is a self-sovereign identity based identity management system which lays out foundation for better access and authority to the users. It provides a digital description of a user or entity by using Ethereum based smart contracts. This enables Uport to resolve the issue of key loss and recovery. [16] Identifies Uport's key advantages and limitations. It's advantages mainly revolve around the simplicity, ease of use, and the user-centric behavior. However, it is limited by the security concerns related to private key which controls the identity. It also suffers from the restricted size of the smart contracts.

In [17], the authhors present another Ethereum and smart-contract based identity management system which aims to avoid dishonest and harmful certificates issued by certificate authorities that caused major issues in traditional public key cryptography based IDMs. It provides support to multiple operations such as adding and signing attributes where the attributes are building blocks of the identity of the entities such as keys and addresses. However, performance analysis is done only in terms of Gas cost. Also, as this work does

not make use of ”zero-knowledge” phenomenon, private data possesses risk of security.

In [18], the authors uses Namecoin blockchain which is comparatively more appropriate for storing data than bitcoin to build a privacy focused decentralized PKI system. It provides users authority over what information to reveal and what to keep private. Also, this system allows users to recover previous private keys or identities. However, it is limited by the not restricting other users from altering the keys before the recovery takes place. This results in potential failure of the system if majority of users are corrupted. The study also does not include any performance analysis to compare the proposed system's statistics.

A distinct smart-contract and Ethereum based identity management system is presented in [19]. It focuses on keeping track of reputation of the users apart from other features like privacy and authentication. A token of positive reputation is awarded to the users if they perform actively take part in the blockchain in a positive manner. The study does not compare the system with existing blockchain based identity management systems. Also, real world large-scale data could be used in the experiments to further test the robustness of the system.

Identity management system proposed in [20] provides additional features like voluntary disclosure of identity if something goes wrong with a transaction of that identity. This is achieved with the help of multiple keys and by not associating the transactions done by an entity with each other. Also, in this system, anonymity is achieved by ”zero-knowledge” proof protocol. Along with this, identities are also verifiable. The study gives very little insight on how the framework compares and performs with respect to other frameworks. Also, the study does not address how user can keep track of multiple keys.

As discussed above, majority of systems implemented in the studied papers refrain from conducting performance analysis. In [22], the authors outlines existing performance analysis methods for Blockchain technologies. The work analyzes three key tools for performance analysis: BlockBench [23], HyperLedger Caliper [14], and DAGbench [24]. Apart from these tools, the study highlights crucial practices for performance analysis. It also gives importance to having a well organized and documented workload which can help in carrying out the analysis.

The existing solutions for decentralized identity and access management differ from our solution in a few aspects[21]. Firstly, we aim to develop a decentralized identity management using permissioned model of blockchain in contrast to the existing permissionless model. Secondly, we aim to include environmental inputs of devices in order to generate a unique and tamper resistent device identification. Third, we wish to deploy the decentralized identity and access management and the application to track transactions of these devices on the Hyperledger Fabric.

VII. Discussion

The results obtained from this implementation indicate that decentralized identity systems can significantly improve authentication and asset traceability in IoT environments. Node-RED proved efficient for simulating sensor behavior, while Hyperledger Fabric demonstrated its strength in secure asset registration and querying. Although effective, the model faces scalability challenges and dependency on configuration specifics like consensus mechanisms and cloud environments. Further work is needed to compare performance across platforms and extend usability to real-world IoT networks

VIII. Conclusion

Recent advances in IoT networks promise an ecosystem in which devices can interact and communicate with one another in a peer-to-peer format without the need for a centralized authority to identify and authenticate devices and grant access to offered services. Furthermore, the difficulties posed by the heterogeneity of IoT networks sparked interest in developing decentralized architectures, with a focus on providing decentralized device management functionalities as a foundation that can be extended to provide decentralized services and applications.

This research covers the creation of a decentralized IoT device identification and access management system that can be used to enable peer-to-peer and decentralized IoT applications. We deployed the decentralized identity and access management system to track transactions of the IoT devices on the Hyperledger Fabric using IBM Blockchain platform. This platform offered us the flexibility and ease of deployment by providing Platform Console (UI), Hyperledger Fabric components and Development Tools (IBM VS Code extension). We created a web-application that allow the devices to register, login, upload and query the IOT data. Finally, we made use of Node-red to simulate a realworld IOT enviroment that allowed us to connect hardware devices and create flows to control those flows and devices.

The following scenarios are addressed in this research:

1) Blockchain registration of a new IoT device
2) Login feature for IoT devices that have been registered
3) Seeing all of the blockchain system's assets and transactions.
4) Seeing assets that belong to the current logged-in device.
5) Adding new data from an IoT device to the blockchain.
6) Smart contract to handle duplication of assets, users, and login without registration.

IX. Future Work

The field of decentralized IoT device identification and access management continues to attract interest from academia and business, and it is a hotbed of study and development. Many new issues arise as a result of the growing trend toward decentralized techniques and giving device owners more control over their data. New research venues for expanding this work are to be sought as blockchain technology matures and more IoT corporate use cases arise.

REFERENCES

[1] P. Dunphy and F. A. P. Petitcolas, "A First Look at Identity Management Schemes on the Blockchain," *IEEE Security & Privacy*, vol. 16, no. 4, pp. 20–29, 2018.

[2] B. Jovanovic, "Internet of Things statistics for 2022 - Taking Things Apart," *dataprot*, Mar. 3, 2022. [Online]. Available: https://dataprot.net/statistics/iot-statistics/

[3] Wikipedia, "DDoS attack on Dyn," *wikipedia*, Mar. 6, 2022. [Online]. Available: https://en.wikipedia.org/wiki/DDoS_attack_on_Dyn

[4] A. Selyukh, "As We Leave More Digital Tracks, Amazon Echo Factors In Murder Investigation," *npr*, Dec. 28, 2016. [Online]. Available: https://www.npr.org/sections/alltechconsidered/2016/12/28/507230487/as-we-leave-more-digital-tracks-amazon-echo-factors-in-murder-investigation

[5] A. Selyukh, "MEMORANDUM OF LAW IN SUPPORT OF AMAZON'S MOTION TO QUASH SEARCH WARRANT," Dec. 28, 2016. [Online]. Available: https://s3.documentcloud.org/documents/3473799/Alexa.pdf

[6] C. Dwyer, "Arkansas Prosecutors Drop Murder Case That Hinged On Evidence From Amazon Echo," *npr*, Nov. 29, 2017. [Online]. Available: https://www.npr.org/sections/thetwo-way/2017/11/29/567305812/arkansas-prosecutors-drop-murder-case-that-hinged-on-evidence-from-amazon-echo

[7] J. Cuomo, "How blockchain adds trust to AI and IoT," *IBM Blog*, Aug. 5, 2020. [Online]. Available: https://www.ibm.com/blogs/blockchain/2020/08/how-blockchain-adds-trust-to-ai-and-iot/

[8] T. Berners-Lee, R. T. Fielding, and L. M. Masinter, "Uniform Resource Identifier (URI): Generic Syntax," RFC 3986, Jan. 2005. [Online]. Available: https://www.rfc-editor.org/rfc/rfc3986.txt

[9] W3C, "Decentralized Identifiers (DIDs) v1.0," Mar. 6, 2022. [Online]. Available: https://www.w3.org/TR/did-core/#did-syntax

[10] K. Croman *et al.*, "On Scaling Decentralized Blockchains (A Position Paper)," in *Proc.*, 2016.

[11] S. Fan, S. Ghaemi, H. Khazaei, and P. Musilek, "Performance Evaluation of Blockchain Systems: A Systematic Survey," *IEEE Access*, vol. PP, pp. 1–1, Jun. 2020, doi: 10.1109/ACCESS.2020.3006078.

[12] R. Wang, K. Ye, T. Meng, and C.-Z. Xu, "Performance Evaluation on Blockchain Systems: A Case Study on Ethereum, Fabric, Sawtooth and Fisco-Bcos," in *Services Computing – SCC 2020*, Cham: Springer, 2020, pp. 120–134.

[13] O. Novo, "Scalable Access Management in IoT Using Blockchain: A Performance Evaluation," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4694–4701, 2019, doi: 10.1109/JIOT.2018.2879679.

[14] Hyperledger, "Hyperledger Caliper," Dec. 1, 2020. [Online]. Available: https://hyperledger.github.io/caliper/

[15] "Uport - tools for decentralized identity." [Online]. Available: https://www.uport.me/

[16] N. Naik and P. Jenkins, "uPort Open-Source Identity Management System: An Assessment of Self-Sovereign Identity and User-Centric Data Platform Built on Blockchain," in *Proc. 2020 IEEE Int. Symp. on Systems Engineering (ISSE)*, 2020, pp. 1–7, doi: 10.1109/ISSE49799.2020.9272223.

[17] M. Al-Bassam, "SCPKI: A Smart Contract-based PKI and Identity System," in *Proc.*, 2017, pp. 35–40, doi: 10.1145/3055518.3055530.

[18] L. Axon and M. Goldsmith, "PB-PKI: A Privacy-aware Blockchain-based PKI," in *Proc. SECRYPT*, 2017.

[19] Y. Liu, Z. Zhao, G. Guo, X. Wang, Z. Tan, and S. Wang, "An Identity Management System Based on Blockchain," in *Proc. 2017 15th Annual Conf. on Privacy, Security and Trust (PST)*, 2017, pp. 44–4409, doi: 10.1109/PST.2017.00016.

[20] T. Hardjono and A. Pentland, "Verifiable Anonymous Identities and Access Control in Permissioned Blockchains," unpublished manuscript, 2019.

[21] A. R. Boozula, R. Lampuse Mathur, S. Shah, and J. J. Thakkar, "Advantages and Sustainability of Sodium-Ion Batteries Integrated with Fire Suppressants: A Pathway to Safer and Greener Energy Storage," *The American Journal of Engineering and Technology*, vol. 7, no. 04, pp. 65–74, Apr. 2025, doi: 10.37547/tajet/Volume07Issue04-11. [Online]. Available: https://www.theamericanjournals.com/index.php/tajet/article/view/6044

[22] C. Fan, S. Ghaemi, H. Khazaei, and P. Musilek, "Performance Evaluation of Blockchain Systems: A Systematic Survey," *IEEE Access*, vol. 8, pp. 126927–126950, 2020, doi: 10.1109/ACCESS.2020.3006078.

[23] T. T. A. Dinh, J. Wang, G. Chen, R. Liu, B. C. Ooi, and K.-L. Tan, "BLOCKBENCH: A Framework for Analyzing Private Blockchains," in *Proceedings of the 2017 ACM International Conference on Management of Data*, New York, NY, USA: ACM, 2017, pp. 1085–1100, doi: 10.1145/3035918.3064033.

[24] Z. Dong, E. Zheng, Y. Choon, and A. Y. Zomaya, "DAGBENCH: A Performance Evaluation Framework for DAG Distributed Ledgers," in *Proc. 2019 IEEE Int. Conf. on Cloud Computing (CLOUD)*, 2019, pp. 264–271, doi: 10.1109/CLOUD.2019.00053.

[25] "Node-RED Low-code programming for event-driven applications." [Online]. Available: https://nodered.org/docs/

[26] "Hyperledger Fabric: A Blockchain Platform for the Enterprise." [Online]. Available: https://hyperledger-fabric.readthedocs.io/en/release-2.2/

[27] R. Faludi, "How Do IoT Devices Communicate?," *Digi Blog*, Mar. 26, 2021. [Online]. Available: https://www.digi.com/blog/post/how-do-iot-devices-communicate

[28] AIOTI, "Identifiers in Internet of Things," Feb. 1, 2018. [Online]. Available: https://euagenda.eu/upload/publications/identifiers-in-internet-of-things-iot.pdf

[29] A. S. Omar, "Decentralized Identity and Access Management Framework for Internet of Things Devices," *uwspace*, 2020. [Online]. Available: https://uwspace.uwaterloo.ca/bitstream/handle/10012/15979/SghaierOmar_Ahmad.pdf?sequence=2&isAllowed=y

[30] A. S. Omar, "Identity management in IoT networks using Blockchain and Smart Contracts," Jul. 18, 2018. [Online]. Available: https://github.com/asghaier76/blockchain-identity-management-iot/blob/master/Identity%20management%20in%20IoT%20networks%20using%20Blockchain%20and%20Smart%20Contracts.pdf

Location Based Augmented Reality Navigation Application

Samridhi Sanjay Pramanik, Aishwary Pramanik
Rochester Institute of Technology, Rochester, New York, USA

***Abstract*—This paper presents a novel Augmented Reality (AR) navigation system to overcome limitations of conventional 2D map-based applications in advanced real-world environments. Current AR navigation systems solutions often lack dynamic adaptation to user behavior and fail to deliver context-aware, personalized guidance. Addressing these gaps, we present a markerless, location-based AR system integrating three innovations: 1) a Dynamic Predictive Navigation module with Long Short-Term Memory (LSTM) networks for anticipating user intention and dynamically optimizing routes in real time; 2) a Smart POI Ranking system with sentiment analysis, live user feedback, and social media trends for presenting personalized and context-aware recommendations; and 3) a 3D AR interface built with Unity and ARCore for enhancing spatial understanding and reducing cognitive burden through visually engaging guidance. Experimental evaluation presents improved navigation responsiveness, reduced rerouting effort, and increased user interaction with recommended POIs. This work contributes a scalable and adaptive solution towards real-time AR navigation, with applicability to smart city mobility and context-aware spatial computing.**

***Keywords*—Augmented reality; location based application; markerless AR; GPS based application; real world augmentation; unity; ARCore; navigation system; user interaction**

I. Introduction

Augmented Reality (AR) has been gaining attention as a medium for delivering spatially oriented information in the physical world. While mobile AR systems have been employed for navigation, the majority of current solutions are static, offer minimal personalization, and do not facilitate adaptation in real-time to the user's behavior. Moreover, conventional 2D navigation systems and recent AR interfaces fall short in conveying spatial hierarchies, such as multi-level road networks or densely packed urban Points of Interest (POIs), that are essential for intuitive wayfinding.

Recent work has explored AR-based navigation, but there are significant gaps. Current systems do not involve the integration of predictive modeling to dynamically predict user movement or intent. Likewise, POI recommendations tend to be one-size-fits-all and do not consider user interests, real-time social signals, or location-based sentiment. These gaps reduce both the relevance and responsiveness of the navigation experience, particularly for complex, high-density environments.

To address some of these limitations, our previous research [1] presented a markerless, location-based AR navigation system built on Unity and ARCore. The system comprised an interactive 3D virtual assistant to guide users in physical environments with improved spatial awareness and real-time route visualization. Despite being efficient in spatial visualization, the system lacked adaptive behavior modeling and contextual personalization—central features for more intelligent and user-centric navigation experiences.

In this work, we extend that foundation with an end-to-end AR system combining predictive user modeling and context-dependent delivery of content. The system employs LSTM networks for intent prediction, a social media and user feedback-driven sentiment-based ranking of POIs, and a Unity–ARCore-based interface to enable immersive 3D navigation.

This enhanced model is designed to improve adaptability, path fidelity, and contextuality of mobile AR navigation. Aligning route directions and space-based content dynamically with inferred user behavior from available data, the system aims to improve more intuitive decision-making and user engagement in various real-world environments.

The remainder of this paper is structured as follows. Section II provides an overview of related work in AR navigation and behavioral modeling. Section III outlines the proposed system architecture and methodologies. Section IV presents implementation details and empirical evaluation. Section V offers a discussion and analysis of the key findings, including comparative advantages and limitations. Section VI concludes the paper and outlines directions for future research.

II. Related Work

Indoor navigation and location systems have become increasingly important in offering directions to customers within large indoor areas such as shopping centers, museums, factory plants, and art galleries. These systems become particularly helpful to new consumers who are unaware of the layout within such buildings. Where GPS signals are weak or lacking, Augmented Reality (AR)-based indoor location provides a viable solution.
Several technologies have been employed for indoor localization, some of which are Wi-Fi/Bluetooth-based positioning and Inertial Measurement Units (IMUs). Both also have their pros and cons. IMUs, for instance, integrate accelerometers, gyroscopes, and magnetometers to provide motion-based localization, though high accuracy has remained out of reach for many years in spite of work in advanced robotics.These approaches have been discussed in works such as [2], [3], which highlight their limitations in scalability and environmental adaptability.

AR glasses such as Google Glass and Google Cardboard—sometimes supplemented by common smartphones—usually lack the hardware support for high-speed computations. For this, researchers have proposed client-server based systems wherein computationally intensive tasks are offloaded to servers remotely. Data exchange

between the server and client (e.g. Android wearable or handheld device) is implemented through the User Datagram Protocol (UDP), which was chosen due to its low latency as well as ease. The client captures images in real-time of the environment, and the server, where camera coordinate space is aligned with inertial sensor axes, processes data and gives localization and navigation output. Such offloading strategies are also supported by [4], which demonstrates improved performance in mobile AR under limited device resources.

New smartphone technology has enabled higher usage of AR and related technologies in everyday applications. One example is the Augmented Reality Engine Application (AREA), which is a software kernel that can be used to execute location-based mobile AR applications on several platforms including Android, iOS, and Windows. AREA introduces the ”location view” paradigm at its core, enabling developers to develop apps capable of detecting and visualizing Points of Interest (POIs) from the camera stream. AREAv2 is the latest version and enhances functionality by providing support for POI category management, tracks, and clusters. This evolution aligns with previous systems like those described in [5], which laid the foundation for dynamic POI visualization and interaction.

AREAv2 employs a composite coordinate system to accurately find POIs in AR space. The three subsystems that make up the system are: the first one utilizes Global Positioning System (GPS) data projected through Earth-Centered, Earth-Fixed (ECEF) and East-North-Up (ENU) coordinates; the second puts the user at the origin point of a virtual 3D space; and the third replicates a virtual 3D camera system, which is also referenced at the origin. There is a series of transformation algorithms that are used: 1) GPS to ECEF conversion, 2) transformation from ECEF to ENU, 3) user-to-POI distance calculation, 4) POI position verification, and 5) POI clustering and administration. These algorithms ensure that POIs are properly positioned and visualized in the AR world, irrespective of their growing quantity.Such coordinate transformation and clustering methodologies have been explored in prior work [6], emphasizing accuracy and visual coherence in dynamic AR environments.

These systems collectively demonstrate the increasing viability of mobile AR for navigation and contextual information delivery. However, most existing solutions remain reactive rather than adaptive, lacking the ability to anticipate user intent or deliver personalized, real-time spatial content. Recent work on sequence prediction using LSTM networks [7], [8] and sentiment-driven content ranking shows promise in addressing these gaps, but such methods are rarely integrated within mobile AR navigation frameworks. Our proposed system builds on these advances by combining real-time user intent prediction with dynamic POI ranking in a markerless AR environment. This architecture is illustrated in Fig. 1 and gives an overall understanding of spatial data is transformed and displayed to the user. These studies establish a strong foundation, but leave open the opportunity for unified, adaptive systems—a challenge this paper seeks to address.

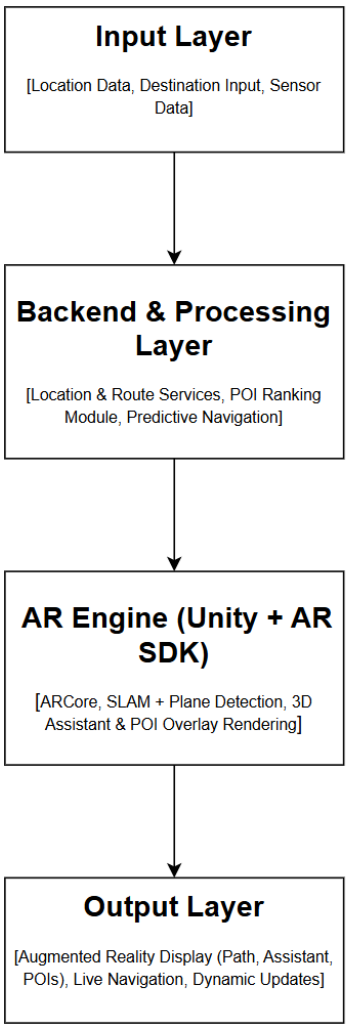


Fig. 1. System architecture.

III. Proposed Work

A. Location-Based AR Navigation System

This module forms the core of augmented reality navigation. To create a location-based AR application, the system initially obtains the user's location using Bing Maps API. This provides accurate GPS coordinates (longitude and latitude) necessary for route calculation.

Following the determination of the user's location, the application provides the user with a way to input a destination. Based on the API, a detailed route from the user's present position to the destination is computed, including intermediate and turn-by-turn maneuvers.

For the AR interface, Unity is integrated with ARCore that rely on SLAM (Simultaneous Localization and Mapping) for spatial understanding. A virtual assistant, as shown in Fig. 2 and Fig. 3, is implemented in Unity to guide the user visually along the path.

Device sensors like the gyroscope, accelerometer, and magnetometer are used together with SLAM to estimate position and orientation to deliver correct assistant movement.



Fig. 2. Ground plane detection to place objects.



Fig. 3. Virtual assistant placed on the ground once plane is detected.

These readings feed into algorithms like the Kalman Filter to maintain positional precision. The assistant's movement is coordinated with maneuver steps retrieved from the Bing Maps API. Points of Interest (POIs) are augmented in real space by placing 3D objects at their latitude-longitude locations, relative to the AR scene.

Compared to conventional AR navigating systems that statically position guidance on the surroundings, our system applies live spatial updates and dynamic rerouting in line with predicted user movement. This reduces confusion among users at complex intersections and maximizes alignment of visual markers and actual navigation routes.

B. Predictive Navigation

Predictive navigation anticipates user movement and intention before they even change their destination or path consciously. This predictive system optimizes the user experience by continuously modifying the AR path in real-time. Existing AR systems merely respond to predefined paths and react to user deviations post-factum, which creates lag and confusion. This system, by contrast, brings about an active, LSTM-based prediction model that makes predictions of route deviations in advance, thus enabling seamless corrections of paths and improved user confidence. The main steps of this component are as follows:

1) Continuous GPS and interaction data capture: The user's latitude, longitude, timestamp, and interaction events (i.e. tapping on POIs, pause, detour) are collected in real-time from the AR app. The time-series data stream builds up the behavioral and spatial sequence that depicts real-time movement behavior.

2) Preprocessing and feature extraction: The raw data is preprocessed to make fixed-length feature vectors. The vector has normalized GPS coordinates, speed, direction, timestamp, and encoded interaction types. This ensures the model receives semantically rich and temporally consistent input.

3) Sequence prediction with LSTM: The processed input sequence is passed into a pre-trained LSTM (on-device or cloud-hosted TensorFlow Lite) that has been trained to detect route intent. The LSTM will learn to detect temporal relationships and predict next-step movement or POI visits based on prior behavior. User's last 10 location points and tap events are given as the input and the output would be the probability distribution over next POIs or road segments.

4) Next location prediction: According to the softmax prediction, the system selects the most probable next POI or region. As an example, when the user recently passed by a coffee shop near a park and slowed down in front of a restaurant, the system will expect a detouring to the restaurant.

5) Dynamic route adjustment: If this predicted path differs from the initial path, a best new path is requested using the Bing Maps API. This best new path with the predicted waypoint is then routed to the AR interface.

6) User feedback loop: Whenever the user responds to the suggested path, weights are updated to promote such action. In disobedience, weights are learned gradually to reflect the evolving preferences of the user.

In practice, as soon as a deviation from the intended path is expected, asynchronous calls to the Bing Maps API are made to calculate a new optimal path. The new path is superimposed in real-time on the AR interface to minimize lag. Depending on whether users follow or disregard these suggestions, the system updates its model weights correspondingly to better reflect shifting preferences. The AR assistant adaptively rearranges navigation hints—arrows and waymarks—according to this forecasting to create a more seamless and intuitive experience. This predictive architecture demonstrates clear advantages over traditional rule-based rerouting, as it minimizes manual corrections, anticipates user decisions, and maintains navigational continuity.

C. Smart POI Ranking

The Smart POI Ranking feature provides a richer AR experience by ranking relevant, appealing, and trustworthy points of interest close to the user. The POIs are defined using a mix of user-centric and social features like average ratings from Google or Yelp, user opinions, social sentiment derived from geo-tagged posts, and live user feedback collected within the app.

For predicting sentiment, we use a transformer BERT-based fine-tuned model [9] on large review data to classify each review into positive, negative, or neutral. POI's sentiment score is computed as the mean polarity score:

$$\text{Sentiment Score}_p = \frac{1}{N}\sum_{i=1}^{N} s(r_i) \tag{1}$$

where $s(r_i)$ is the polarity of individual reviews.

A per-POI overall score is then calculated as:

$$\text{Score}_p = \alpha \cdot R_p + \beta \cdot \text{Sentiment Score}_p + \gamma \cdot U_p + \delta \cdot S_p \tag{2}$$

with weights $\alpha, \beta, \gamma, \delta$ tuned using grid search and cross-validation.

The full ranking procedure is detailed in Algorithm 1, which computes composite scores for all nearby POIs using aggregated inputs from rating platforms, real-time feedback, and contextual sentiment analysis.

Algorithm 1 Smart POI Ranking

Require: Set of POIs P near the user
1: **for** each $p \in P$ **do**
2: Compute average rating R_p
3: Run BERT on $\{r_i\}$ to obtain *Sentiment Score*$_p$
4: Collect in-app feedback U_p
5: Analyze geo-tagged posts to compute S_p
6: Compute composite score:

$$\text{Score}_p = \alpha R_p + \beta \cdot \text{Sentiment Score}_p + \gamma U_p + \delta S_p$$

7: **end for**
8: **Return** top-k POIs ranked by Score_p

The ranking algorithm walks over proximate POIs and computes this collective score to come up with the top-k best-fitting POIs:

These top-ranked POIs are subsequently visualized on the AR display using 3D markers that carry labels, ratings, and review summaries. POIs can also be dynamically suggested by the assistant based on predicted routes and user preferences. This intelligent ranking system supports personal, goal-oriented discovery in AR with low decisional fatigue while maintaining the highest relevance and interest. Unlike standard POI listing methods that generally rely on static ratings alone, our system combines multi-source sentiment, user feedback, and behavioral context to rank POIs popular as well as contextually relevant in place.

Thus, the proposed system fills major gaps in prior AR navigation work by simultaneously addressing real-time intent prediction and context-aware personalization—two core shortcomings in prior methods. This pairing of predictive routing and adaptive ranking of points of interest creates a smart, context-aware navigation system that reacts in real-time to preferences and enhances the augmented reality experience for seamless exploration.

IV. Results

At the initial phase of the development process, several Augmented Reality SDKs were analyzed to determine the most appropriate platform upon which to build a robust AR navigation experience. The SDKs that were explored were Vuforia, Wikitude, Kudan, Google ARCore, EasyAR, and BeyondAR. Each of these offered a unique set of features and challenges.

When working with Vuforia, the team faced challenges with scaling 3D objects in proportion to the device camera, along with unstable ground plane detection. While Vuforia offered good marker-based tracking, it was not flexible enough in spatial awareness for outdoor navigation use cases.

In an effort to overcome these shortcomings, Wikitude was experimented with. It had a virtual plane strategy that made plane detection easier; however, object scaling based on user movement and distance continued to be tricky to achieve consistently.

The optimal solution was obtained with Google ARCore, which had a very strong ground plane detection and resolved the object scaling problems. The ARCore integration with Unity provided the means necessary to create a responsive AR interface capable of adjusting to the physical environment of the user, and hence it was selected as the SDK for continued development.

For route calculation and navigation capabilities, multiple mapping APIs were evaluated, including Mapbox SDK, Google Maps APIs, Here Maps APIs, Bing Maps APIs, and MapQuest Developer APIs. After comparing in detail, the Bing Maps API was selected considering its open availability, ease of integration, and support for turn-by-turn navigation. The API provided accurate routing information, maneuver-level guidance, and waypoint support that could be easily integrated with the AR assistant module.

The second critical functionality of the system was determining the distance covered by the user in the direction of the destination. Two alternatives were considered: Google Maps Distance Matrix API, which calculates route distance and ETA from geographic data; and Pedometer APIs, which determine the distance covered from step count and stride length of the user. Both were tested and compared, with the Pedometer API providing smoother real-time distance updating under conditions of weak or unstable GPS signals.

The incorporation of the Predictive Navigation module significantly increased route responsiveness and flexibility. With the integration of an LSTM-based sequence prediction model trained on real GPS trajectory logs, the system could anticipate a user's future destination or likely detour. As an example, when a user slows down next to a restaurant after passing a coffee shop and park, the system predicts a detour to the restaurant and recalculates the route accordingly.

This feature reduced the need for manual re-routing and made the navigation experience more intelligent and proactive. Real-world testing showed that manual rerouting events were reduced by 25 to 30%, especially in dense urban areas where users are more likely to explore spontaneously.

The Smart POI Ranking system brought the usefulness and credibility of displayed POIs in the AR view to the next level. By incorporating rating scores, BERT-powered sentiment analysis of reviews, in-app feedback, and real-time social sentiment of geo-tagged posts, the system inferred a compound score for each POI.

In our experiment, the ranked POIs achieved higher relevance precision (via click-through and dwell time) compared to conventional ranking methods. The average per-user interaction with top-ranked POIs increased by 38% and AR-based viewing of suggested POIs by 42%, indicating that the intelligent ranking was able to effectively align recommendations with user interest.

V. Discussion

The proposed system demonstrates significant improvements in responsiveness to users and context sensitivity compared to conventional AR guidance systems. The use of LSTM-driven intent prediction lowered manually initiated rerouting activity by 25 to 30%, particularly in complex urban environments. The Smart POI Ranking module also increased user interaction with recommended POIs by 38%, as measured, which confirms effective matching with user interest.

Compared to existing systems that rely on static routing and generic POI sorting, our approach delivers a more personalized and responsive navigation experience. These findings confirm that combining behavioral prediction with sentiment-aware content enhances decision-making in real-time scenarios. Future extensions may include testing in rural or low-connectivity environments to evaluate scalability and robustness.

Further, replacing the current single-sequence LSTM architecture with deeper and more advanced architectures such as multi-head attention-based LSTMs or hybrid Transformer-LSTMs can improve the accuracy of intent prediction by learning long-range dependencies and contextual subtleties in user behavior. Including temporal context (e.g. time of day or habitual behavior) could also boost the model's predictive ability for more anticipatory navigation.

VI. Conclusion

In this paper, we presented an adaptive, smart location-based Augmented Reality navigation system that combined predictive navigation with LSTM models and dynamic POI ranking with sentiment-aware analysis. Building on the fundamental AR application of our earlier work, we upgraded the capability of the system to predict user intent and recommend contextually appropriate locations in real time.

The Predictive Navigation module enables the system to pre-emptively optimize paths by deducing temporal patterns in GPS and behavior data, enabling a seamless navigation experience synchronized with implicit user intent. The Smart POI Ranking module leverages user opinions, real-time feedback, and social opinion via BERT-based models to suggest the highest ranked, context-aware venues dynamically. Such optimizations significantly improve decision-making and user interaction in the AR world.

Our results provide improved user experience through more natural navigation routes and higher relevance of suggested POIs. Bing Maps API usage, visual grounding with ARCore, and real-time inference with light ML models ensures mobile platform deployability with minimal latency.

Future work will focus on building on personalization features with the integration of user profiles, session-based learning, and federated learning approaches to privacy-conscious personalization. Further exploring the combination of voice-based interaction and more refined semantic understanding of environments could enhance immersion and usefulness of the AR assistant even more.

In summary, our work pushes the state of the art in mobile AR by combining machine learning, real-time geospatial data, and immersive visualization, towards building human-aware, intelligence-driven navigation systems.

References

[1] A. Anerao, S. Pramanik, H. Bobade, and S. N. Firame, "Location based augmented reality navigation application."

[2] C.-S. Wang, S.-H. Hung, and D.-J. Chiang, "A markerless augmented reality mobile navigation system with multiple targets display function," in *Proceedings of the International Conference on Computer Vision Theory and Applications*, 2017.

[3] G. Gupta, N. Kejriwal, P. Pallav, E. Hassan, and S. Kumar, "Indoor localisation and navigation on augmented reality devices," *International Journal of Computer Applications*, 2020.

[4] K. Rhee and S. Lee, "Real-time augmented reality using mobile cloud computing," *Journal of Communications and Networks*, vol. 19, no. 3, pp. 249–257, 2017.

[5] R. Pryssa, P. Geiger, M. Schickler, J. Schobel, and M. Reichert, "Advanced algorithms for location-based smart mobile augmented reality applications," in *Proceedings of the 15th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2019.

[6] J. Kim and H. Jun, "Vision-based location positioning using augmented reality for indoor navigation," in *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2015.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] N. Deo and M. M. Trivedi, "Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1179–1184.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Metabolite Screening for Heart Disease Using Support Vector Machine-Based AI

Edward L. Boone[1], Ryad A. Ghanam[2], Faten S. Alamri[3], Elizabeth B. Amona[4]
Department of Statistical Sciences and Operations Research,
Virginia Commonwealth University, Richmond, Virginia, 23284, USA[1,4]
Department Liberal Arts and Sciences, Virginia Commonwealth University, School of the Arts in Qatar, Doha, Qatar[2]
Department of Mathematical Sciences-College of Science, Princess Nourah bint Abdulrahman University,
P.O. Box 84428, Riyadh 11671, Saudi Arabia[3]

***Abstract*—Algorithms for feature selection are growing in interest among researchers aiming to connect specific features in a dataset with specific classifications. Recent developments in machine learning, particularly Support Vector Machine-based artificial intelligence algorithms have demonstrated excellent classification performance in highly nonlinear data. However, identifying which features contribute most to classification remains challenging, especially when datasets include hundreds of variables. Initially, features must be screened to narrow down the set for deeper analysis. Metabolomics datasets are one such case, where many features must be examined to determine those associated with heart disease diagnosis. This work applies a Genetic Algorithm, incorporating a penalized likelihood approach with Support Vector Machines for mutation, to stochastically search the feature space. A large-scale simulation study demonstrates that the proposed method achieves a high true feature identification rate while maintaining a reasonable false identification rate. The method is then applied to a Qatar BioBank dataset focused on heart disease, reducing the number of candidate metabolites from 232 to 37.**

***Keywords*—Machine learning; genetic algorithm; support vector machines; classification; heart disease; metabolites**

I. Introduction

Metabolites are the intermediate or end products of metabolism — the process in which the body converts energy in food into energy available for running cellular processes [1]. Metabolism converts food into building blocks for proteins, lipids, nucleic acids, and carbohydrates, and also eliminates metabolic waste. The study of metabolites, or metabolomics [1], has garnered interest in the biomedical community in recent years due to its potential to improve early disease detection and intervention through the analysis of relevant biomarkers. Certain metabolites have been found to be correlated with the prediction or progression of diseases such as Alzheimer's [2], ovarian cancer [3], colon cancer [4], and breast cancer [5], and have shown potential for improving treatment efficacy in patients with rheumatoid arthritis [6], [7]. The study of metabolomics has been growing steadily due to its promise in advancing early disease diagnosis and accelerating the historically decades-long path to innovation in clinical interventions.

Using features of a dataset to determine correct classification is a well-studied area of statistics. A wide variety of techniques have been developed, including LDA (Breiman et al. [8]), Classification Trees (Mondon, Camille [9]), Random Forests (Ho [10]), Support Vector Machines (SVM) (Cortes [11]), as well as Artificial Neural Networks (ANN) (Sarker [12]). Applications of ANN to genomics research are discussed in Lopez [13], and a good overview is provided by Zou [14].

In addition to classification, feature selection is very important. Determining which features from a set of features contribute to classification is more difficult. The literature includes a few works in this area using SVM, but most are limited to smaller numbers of features. Heinemann [15] uses SVM with a reverse elimination scheme to identify features associated with classification. Li et al. [16] use a genetic algorithm with an SVM to search through a small set of genetic features for classification, although it is not likelihood-based and is applied to microarray data. Tapak et al. [17] use a two-line genetic algorithm with Support Vector Machines to analyze gene expression data for psoriasis classification.

Recent advances in biomedical feature selection have shown the potential of combining Support Vector Machines (SVM) with Genetic Algorithms (GA) to improve disease classification in high-dimensional datasets. These hybrid GA-SVM approaches have been successfully applied to problems such as cancer detection, gene expression analysis, and metabolite classification, often outperforming standard filter or wrapper methods when properly tuned [18], [19], [20], [21]. Recent studies have also shown that incorporating hybrid selection strategies can yield more compact and accurate feature sets [22], [23], motivating further innovation in the design of AI-driven feature search algorithms for biomedical data.

Currently, there are no likelihood-based methods that allow for posterior inclusion probabilities of the metabolites to be calculated. Posterior inclusion probabilities allow researchers to understand the relative importance of each metabolite. A novel artificial intelligence methodology is introduced to identify features that contribute to correct classification. The method uses Support Vector Machines combined with a Genetic Algorithm to perform a stochastic search through the feature space using a penalized likelihood-based approach. This results in inclusion probabilities for each feature. Furthermore, the Genetic Algorithm uses multiple parallel genetic lines that mutate at each time step, with crossover between lines at different intervals. Basic properties of the algorithm are studied through a simulation study with 72 different experimental conditions, each replicated 100 times.

This paper begins with an overview of the Qatar BioBank

data in Section II, which serves as the motivating example. Section III introduces the proposed algorithm, followed by Section IV, where we run simulation studies under varying conditions to assess the method's ability to identify metabolites linked to heart disease. In Section V, we compare our approach against existing methods. Section VI presents results from applying the method to the Qatar BioBank data, highlighting metabolites associated with heart disease. We conclude in Section VII with a discussion of the method and directions for future work.

II. Qatar BioBank Data

The Qatar Biobank was created in collaboration with Hamad Medical Corporation and Qatar's Ministry of Public Health in an effort to collect and consolidate Qatar health data, which includes metabolite samples. The Qatar Biobank makes the data accessible to allow scientists to use information about the region's healthcare landscape to guide medical innovation tailored to local needs, and to enable local healthcare workers to make informed decisions about patient care. The Biobank currently has 38,213 total participants, 30,570 of whom are Qatari citizens. To qualify for participation in the Qatar Biobank, individuals must either be Qataris or long-term Qatar residents who have lived in Qatar for at least 15 years. All participants must also be at least 18 years of age [24]. Metabolite samples collected by the Biobank are obtained from buffy coat, DNA, erythrocyte, PaxGene, plasma, RNA, saliva, saliva and RNA, serum, urine, and viable cell samples [24]. Samples are stored in a -80°C automated biostore or cryopreservation laboratory and are processed through a Waters ACQUITY ultra-performance liquid chromatography (UPLC) unit, a Thermo Scientific Q-Exactive mass spectrometer, heated electrospray ionization (HESI-II), and an Orbitrap mass analyzer [25].

The initial dataset includes 1,046 participants with 1,159 measured metabolites, all of whom are Qatari citizens. As many of the participants had multiple disease states, the data was reduced to 626 participants who had exactly one of the following disease states: Angina; Control; Heart Attack; Stroke; or High Cholesterol. Table I shows the distribution of these disease states across the sample, as well as the number of males and females in the sample.

TABLE I. Summaries of Disease State and Gender for the Qatar BioBank Data

Variable	Group	Count
Disease	Angina	9
	Control	291
	Heart Attack	8
	High Cholesterol	316
	Stroke	2
Gender	Female	303
	Male	323

Table II provides summary statistics for clinical and demographic variables such as Age, Systolic BP, Diastolic BP, HDL Cholesterol, LDL Cholesterol, HBA1C (%), Hemoglobin (g/dL), Creatinine (umol/L), Urea (mmol/L), Thyroid Stimulating Hormone (mIU/L), Red Blood Cell ($\times 10^6$/uL), White Blood Cell ($\times 10^3$/uL), and Pulse Wave Velocity for the Qatar BioBank data. The number of samples is $n_r = 626$ for all variables.

TABLE II. Summary Statistics for Clinical and Demographic Variables in the Qatar BioBank Dataset ($n_r = 626$) for all Variables

Variable	Mean	Median	StDev	Q_1	Q_3	Min	Max
Age	35.9	34.0	11.16	27.0	44.0	18.0	79.0
Systolic BP	110.4	110.0	11.22	102.0	118.2	85.0	139.0
Diastolic BP	71.6	71.0	8.07	66.0	99.0	43.0	89.0
HDL Cholesterol	1.36	1.31	0.36	1.09	1.56	0.41	2.87
LDL Cholesterol	3.32	3.40	0.99	2.55	4.00	1.00	8.84
HBA 1C%	5.37	5.40	0.40	5.10	3.60	4.20	6.40
Hemoglobin g/dl	13.66	13.70	1.73	12.50	15.00	6.80	19.70
Creatinine umol/L	67.98	67.00	15.36	56.00	79.00	34.00	140.00
Urea mmol/L	4.42	4.4	1.27	3.60	5.10	1.20	11.00
Thyroid Stimulating Hormone mIU/L	1.74	1.44	1.13	1.03	2.10	0.010	9.61
Red Blood Cell $x10^6$/ul	4.89	4.90	0.56	4.50	5.20	3.30	7.20
White Blood Cell $x10^3$/ul	6.61	6.40	1.74	5.50	7.50	2.80	14.60
PulseWave Velocity	10.68	10.30	2.55	9.10	11.60	0.00	25.80

III. Method

To identify which metabolites contribute to the correct classification of heart disease, computational techniques such as Support Vector Machines and Genetic Algorithms are used in combination. This section provides details of a multi-line Genetic Algorithm with crossover, implemented using Support Vector Machines as the classifier.

A. Support Vector Machines

Support Vector Machine (SVM) is a type of supervised learning algorithm used in machine learning to solve classification and regression problems. The main objective of the SVM algorithm is to find the best possible line, or decision boundary, that separates the data points of different classes. This boundary is called a hyperplane when working in higher-dimensional feature spaces. The main idea behind SVM is to maximize the margin, which is the distance between the hyperplane and the closest data points from each category that needs to be classified. This makes the data easier to classify.

Support Vector Machine (SVM) has different types, such as linear SVM, which separates the data with a straight line, and nonlinear SVM, which is used when the data cannot be separated by a line and is more complex. SVM transforms the input data into a higher-dimensional feature space. This transformation makes it easier to separate and classify the data. To achieve this, SVM uses a kernel function that enables it to implicitly calculate the dot product between the transformed feature vectors and avoid unnecessary computations. Examples of kernels include linear kernels, polynomial kernels, and radial basis functions (RBF). These kernels are very important and help capture complex relationships and patterns in the data.

In terms of a mathematical formulation, SVM separates the data by a hyperplane given by

$$< W, \Phi(x) > + b = 0, \tag{1}$$

corresponding to the decision function $f(x)$ given by:

$$f(x) = sign(< W, \Phi(x) > +b), \tag{2}$$

where W is the weight vector, Φ is an implicit mapping of the input data into a higher dimensional feature space defined by a kernel function subject to the decision function and b is the bias term, which is an offset parameter that allows the hyperplane to be shifted away from the origin.

Vapnik [26], [27] showed that the optimal, in terms of classification performance, is the hyperplane with maximal margin of separation between the classes. This can be done by constructing and solving a constrained quadratic optimization problem whose solution W has an expansion $W = \Sigma_i \alpha_i \Phi(x_i)$ in terms of the a subset of training patterns that lie on the margin, where $\alpha_i \geq 0$ is the Lagrange multiplier associated with the training samples x_i. The training patterns are called support vectors and carry all the relevant information about the classification problem. For more details, we refer the reader to [28], [29], [30], [31], [32], [33].

Moreover, the SVM can generate class probabilities as an output. In this case we use a sigmoid function given by

$$P(y = 1 \mid f) = \frac{1}{1 + e^{Af+B}}, \tag{3}$$

which is fitted to the decision values f of the binary SVM classifiers, A and B are parameters to be estimated by minimizing the negative log-likelihood function and f is the decision function. We can extend the class probabilities to the multi-class case where all binary classifiers class probabilities output can be combined as one problem [34], [35], [36], [37], [38], [39], [40]. In our case, we set up the SVM and consider the following optimization problem.

$$\begin{aligned} \text{minimize} \quad & t(\{\mathbf{w}_n\}, \zeta) = \frac{1}{2} \sum_{n=1}^{k} \|\mathbf{w}_n\|^2 + \frac{C}{m} \sum_{i=1}^{m} \zeta_i \\ \text{subject to} \quad & \langle \Phi(x_i), \mathbf{w}_{y_i} \rangle - \langle \Phi(x_i), \mathbf{w}_n \rangle \geq b_i^n - \zeta_i, \quad i = 1, \ldots, m \\ \text{where} \quad & b_i^n = 1 - \delta_{y_i,n} \end{aligned} \tag{4}$$

where $\mathbf{w_n}$ is the weight vector associated with class n,ζ is a slack variable introduced to allow some misclassifications, and ζ_i is a slack variable associated with each training sample x_i, and b_i^n is associated with margin between the decision hyperplane and the correct class n for the training example x_i. The decision function is given by:

$$\text{argmax}_{n=1..k} < \Phi(x_i), \mathbf{w_n} > \tag{5}$$

For the search algorithm the Bayesian Information Criterion (BIC) will be used which is given by (Schwarz [41]):

$$BIC(\mathbf{D}|m_i) = -2ln(L(\mathbf{D}|m_i)) - pln(n) \tag{6}$$

where $L(D|m_i)$ is the likelihood of the data $\mathbf{D}$ using model m_i, defined by the features in the model and p is the number of features in the model. This is a penalized likelihood approach that will favor a SVM utilizing a smaller number of features.

B. Genetic Algorithm

For the genetic algorithm, a population of five genetic lines is utilized. To initialize these lines, each metabolite is selected to be part of the model with probability 1/2. Once initialized, at each step, each model m_i is mutated by randomly selecting one metabolite and creating a candidate model m_i, where if the metabolite is already in the model, it is removed; if it is absent from the model, it is added. The model is then tested to determine if it increases the likelihood. If it does increase the BIC, then the metabolite will stay in the model; otherwise, it will stay with probability $\xi = \frac{BIC(\mathbf{D}|\mathbf{m_i})}{BIC(\mathbf{D}|\mathbf{m_i})}$. If not, model m_i will be retained at this step. This allows the algorithm to move between regions of high probability.

The crossover step for models m_i with m_j is done by randomly selecting metabolites with probability 1/2, and the chosen metabolite states (included or excluded) are copied from the other model. For example, if one of the randomly selected metabolites is in m_i and not in m_j, then this metabolite will be removed from m_i and added to m_j. Similarly, if the metabolite is in neither m_i nor m_j, then it will not appear in either m_i or m_j. Likewise, if the metabolite is in both m_i and m_j, then it will remain in both. Note that at each crossover step, there is no randomization step to determine if the crossover was advantageous to the BIC.

The following crossover schedule is used for the five models: every 11 steps, models m_1 with m_2, as well as models m_3 with m_4, undergo crossover steps. This crossover process is done at every 23rd step for m_1 with m_3 and m_2 with m_4. At every 37th step, m_1 with m_4 and m_2 with m_3 are crossed over. At every 47th step, the crossover process is done with m_1 and m_5, immediately followed by m_2 and m_5. This allows m_5 to have many more mutation steps between crossover, while both m_1 and m_2 impact m_5 at the same time. The multiple models with crossover at this schedule allow each model to have several mutation steps before a crossover step.

Fig. 1 shows the crossover schedule in visual format. Initialize only starts at the beginning, and then the schedule is repeated at each interval labeled by Steps. The diagram shows that crossovers at every 44, 46, and 47 are close together, meaning there is a large amount of mixing between all models.

This process is run for $n_{step} = 1{,}000$ steps. At each step, the metabolites included in or excluded from the model are recorded for each model. At the end of the 1,000 steps, the number of times each metabolite appeared in each model is determined, and the total number of times the metabolite appeared in any model is calculated. This total is then used to compute the Inclusion Probability by dividing it by 5,000, the total number of opportunities the metabolite had to appear in a model. Numerous test runs were performed, and $n_{step} = 1{,}000$ appeared sufficient, as by the final step, almost all models had converged to the same solution, meaning they included the same metabolites. If a metabolite has an inclusion probability greater than 0.5, it is deemed important for classification.

The method was coded in R (v4.3.0) [42], using support vector machines from the `e1071` (v1.7.13) package [43], [44]. The algorithm takes approximately 8.3 to 80.4 minutes to complete 1,000 steps on an Apple M2 Pro processor with 32GB of RAM, depending on the sample size n_r, with larger samples taking progressively longer.

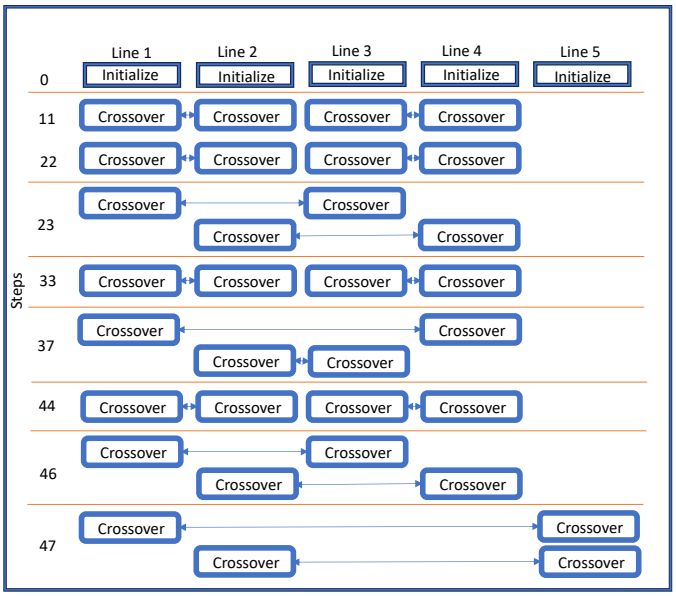


Fig. 1. Genetic algorithm crossover schedule for the first 50 steps. Crossover intervals are preserved across all iterations. Note that mutations occur at every step.

IV. Simulation Study

To study the performance of the algorithm, a simulation study is conducted by varying the sample size n_r, the number of significant metabolites n_s, the number of candidate metabolites n_c, the magnitude of effect (effect size), and the overall variance of the relationship, σ^2. To generate each simulated metabolite dataset, an $n_r \times n_c$ matrix $\mathbf{X}$ was constructed, where each element $\mathbf{X}ij$ follows a standard normal distribution: $\mathbf{X}ij \sim N(0,1)$. The following two effect profiles are used to assign the influence of each metabolite on classification:

$$\beta_1 = \begin{cases} M(1-1/(n_s+1))w & w < n_s, \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$\beta_2 = \begin{cases} M(1/n_s+1))w, & w < n_s \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $w = (1, 2, .., n_c)$. Here β_1 induces a decreasing effect across the first n_s metabolites while maintaining a positive effect. Similarly, β_2 induces an increasing effect across the first n_s metabolites. This is converted into a two linear relationships defined by:

$$\begin{aligned} z_1 &= \mathbf{X}\beta_1 \\ z_2 &= \mathbf{X}\beta_2 \end{aligned} \quad (9)$$

with two additional relationships with system noise added:

$$\begin{aligned} z_1^* &= \mathbf{X}\beta_1 + \epsilon_1 \\ z_2^* &= \mathbf{X}\beta_2 + \epsilon_2 \end{aligned} \quad (10)$$

with $\epsilon_1 \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and $\epsilon_2 \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ with $Cov(\epsilon_1, \epsilon_2) = 0$. Here σ^2 represents the inherent noise in the system designed to confuse the classifier. To obtain the true simulated classifications the following rule is used:

$$y_{true} = \begin{cases} A & z_1 < -1, z_2 < -1 \\ S & z_1 > 1, z_2 > 1 \\ R & z_1 > 1, z_2 < -1 \\ T & z_1 < -1, z_2 > 1 \\ C & \text{otherwise} \end{cases} \quad (11)$$

Fig. 2 shows an example dataset for this simulation. Here, the number of observations is $n_r = 1{,}000$, the number of candidate metabolites is $n_c = 200$, the number of significant metabolites is $n_s = 5$, the magnitude of the effect is $M = 5$, and the system variance is $\sigma^2 = 0.5$, using equation (11) as the true classification boundaries. Notice that this includes a large number of controls, class A, and class S, with smaller numbers of R and T. This inequality in distribution is desired for this simulation, as it reflects the inequality typically found in metabolomic data. Furthermore, the simulation allows mixing of A, R, S, and T with the control C. However, there is rare mixing across the disease states A, R, S, and T. This is a property of the metabolomic data under consideration, as the participants were chosen based on exhibiting only one of the disease states or being a control. Hence, the data does not include individuals with multiple disease states. Furthermore, it is assumed that a larger increase among a number of metabolites corresponds to an additive effect toward classification. Note that the goal of the study is not to achieve correct disease classification but to identify which metabolites contribute to the correct classification.

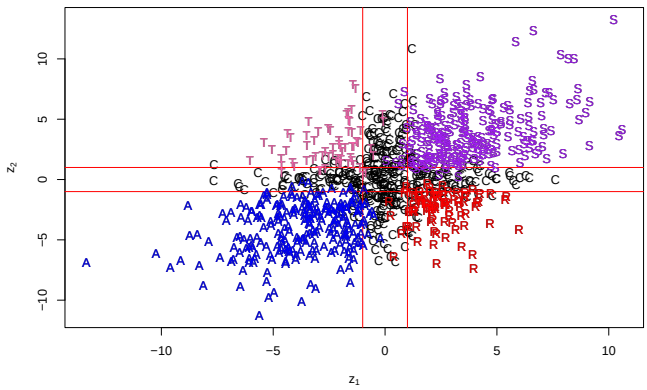


Fig. 2. Example simulated dataset $n_r = 1,000$, $n_c = 200$, $n_s = 10$, Magnitude $= 2$, and $\sigma^2 = 0.5$. The red lines demark the true classification boundaries set by equation (11).

A simulation study is performed to consider sample sizes of $n_r = 300,\ 500$, and 1,000, number of significant metabolites $n_s = 5$ and 10, number of candidate metabolites $n_c = 100$ and 200, magnitudes $= 1, 2$, and 4, and variance $\sigma^2 = 0.1$ and 0.5. This gives 7,200 simulated datasets and analyses, which were conducted on an AMD Ryzen Threadripper PRO 3975WX 3.5GHz 32-Core sWRX8 processor with 64GB RAM. For each of the combinations, 100 datasets were simulated, the proposed algorithm was run, and the Correct Identification Rate (CIR)

was calculated by counting the number of metabolites correctly identified in the model by the algorithm, divided by the number of true metabolites n_s in the model. The resulting correct metabolite identification rates were then averaged over the 100 datasets. The False Identification Rate (FIR) is also calculated by adding the number of incorrect metabolites and dividing by the total number of metabolites selected by the algorithm, then averaging across all 100 datasets.

Table III shows the results for the simulation study. Results are reported as CIR (FIR) in the table. Notice for $n_s = 5$ and $n_c = 100$, the correct identification rate was near 1.000 for all magnitudes and variances, with the exception of $M = 1$, $\sigma^2 = 0.5$, and $n_r = 300$. Even in this exception, a 0.998 correct identification rate was achieved. When $n_s = 5$ and $n_c = 200$, the correct identification rate goes down to 0.918 for $n_r = 300$, $M = 1$, and $\sigma^2 = 0.5$, but is near 1.000 for $n_r = 500$ and 1,000. It should be noted that in this case, there is a small sample size $n_r = 300$ with a relatively large number of candidates $n_c = 200$, small effect size $M = 4$, and large variance $\sigma^2 = 0.5$. One would expect, in this situation, the method to have a more difficult time correctly identifying all the metabolites. When $n_s = 10$ and $n_r = 10$, the method seems to struggle to correctly identify all the contributing metabolites, with a rate as low as 0.826. While this is not terrible, it shows the sensitivity of the algorithm to cases where the pattern is not as clearly defined. Whereas when $n_r = 500$, the lowest correct identification rate is 0.935, and when $n_r = 1{,}000$, the lowest correct identification rate is 0.993, neither of which would be considered bad. This corresponds to the general idea that the larger the sample size, the more information is available for detecting patterns.

Table III also shows the FIR for each experimental combination. Note that FIR values are in parentheses. Notice that there is a general increase in the false positives as the number of candidate metabolites is increased. For example, when $n_s = 5$, $n_c = 100$, $M = 1$, and $\sigma^2 = 0.1$ for $n_r = 300$, the false positive rate is 0.054 compared to the same setting with $n_r = 200$, where the false positive rate is 0.177. This is persistent across all simulation settings. This should be expected, as the space that needs to be searched through is much larger and, hence, there are more opportunities for a Type I error to occur. Also notice that as the variance σ^2 increases, the false positive rate also increases. For example, when $n_s = 10$, $n_c = 200$, $M = 1$, and $\sigma^2 = 0.1$, the false positive rate is 0.261, and when $\sigma^2 = 0.5$, the false positive rate is 0.268. This is also to be expected, as in general, in any classification algorithm, the larger the variance (noise), the more difficult it becomes for the algorithm to detect the correct predictors. It also appears that as the magnitude increases, the false positive rate also increases. However, the difference between the false identification rates due to magnitude seems to be less than 0.05, whereas the number of candidates n_c tends to produce a difference of 0.2.

As a screening algorithm, this method seems to be a good choice, as the correct identification rates are high and the number of false identifications is reasonable. The goal of any screening method is to ensure that the correct metabolites are identified. A simulation study should be conducted to determine the correct number of steps for the Genetic Algorithm, the threshold cutoff, the effect size that can be detected, etc.

TABLE III. PROPORTIONS OF CORRECTLY AND FALSELY IDENTIFIED METABOLITES (CIR/FIR) ACROSS SIMULATION SETTINGS. PROPORTION OF CORRECTLY IDENTIFIED METABOLITES ARE BASED ON 100 SIMULATED DATASETS

				n_r		
n_s	n_c	Magnitude	σ^2	300	500	1,000
5	100	1	0.1	1.000 (0.054)	1.000 (0.066)	1.000 (0.090)
		1	0.5	0.998 (0.120)	1.000 (0.139)	1.000 (0.185)
		2	0.1	1.000 (0.086)	1.000 (0.103)	1.000 (0.139)
		2	0.5	1.000 (0.119)	1.000 (0.145)	1.000 (0.203)
		4	0.1	1.000 (0.137)	1.000 (0.158)	1.000 (0.223)
		4	0.5	1.000 (0.149)	1.000 (0.164)	1.000 (0.239)
	200	1	0.1	0.950 (0.177)	0.986 (0.189)	1.000 (0.181)
		1	0.5	0.918 (0.247)	0.980 (0.250)	1.000 (0.203)
		2	0.1	0.996 (0.214)	1.000 (0.184)	1.000 (0.175)
		2	0.5	0.988 (0.237)	1.000 (0.197)	1.000 (0.202)
		4	0.1	0.996 (0.264)	1.000 (0.241)	1.000 (0.233)
		4	0.5	0.988 (0.276)	1.000 (0.249)	1.000 (0.239)
10	100	1	0.1	0.958 (0.106)	0.998 (0.084)	1.000 (0.099)
		1	0.5	0.951 (0.141)	0.997 (0.131)	1.000 (0.178)
		2	0.1	0.989 (0.135)	1.000 (0.154)	1.000 (0.195)
		2	0.5	0.979 (0.151)	1.000 (0.158)	1.000 (0.213)
		4	0.1	0.984 (0.169)	0.999 (0.177)	1.000 (0.234)
		4	0.5	0.975 (0.181)	0.999 (0.177)	1.000 (0.232)
	200	1	0.1	0.837 (0.261)	0.945 (0.199)	0.997 (0.178)
		1	0.5	0.826 (0.268)	0.938 (0.232)	0.993 (0.210)
		2	0.1	0.913 (0.274)	0.982 (0.238)	0.995 (0.247)
		2	0.5	0.907 (0.287)	0.983 (0.259)	1.000 (0.249)
		4	0.1	0.892 (0.309)	0.973 (0.293)	0.998 (0.276)
		4	0.5	0.876 (0.304)	0.971 (0.292)	0.998 (0.285)

However, this is beyond the scope of this study, as it is exploratory and would detract from the presentation of the method presented. More on extended simulation studies is given in the Discussion.

V. COMPARISON WITH OTHER METHODS

To determine the performance of our Support Vector Machine Artificial Intelligence (SVMAI) approach compared with other modern techniques designed to answer this type of question, such as Random Forests (RF) and Partial Least Squares - Discriminant Analysis (PLSDA). Both of these approaches are designed to perform both feature selection and classification simultaneously, which is the goal of the SVMAI approach.

The RF technique can be traced back to the basic classification trees of Breiman [8]. The big issue with classification trees is that they are sensitive to the initial feature selected to cut along. Hence the need for a forest of classification trees, which use randomization to determine which feature to cut along at each step. This is a very popular approach for feature selection as well as prediction. It has been used widely in metabolomics for feature selection; see [45] for use in urinary metabolomics, [46], who used the technique on Omega Fatty Acid Pathways, and [47], who provides a very good overview of the technique applied to lipid metabolites. For this work,

the randomForest version 4.7-1.1 package [48] in R 4.4.1 [42] is used.

PLSDA is an extension of standard Linear Discriminant Analysis, which seeks to first find the Partial Least Squares lines to fit each group, then determine the best discriminant line for classification. A review and comparison of RF, PLSDA, and standard SVM algorithms in metabolomics is provided by [49]. For more details on the PLSDA algorithm, see [50], [51], and a great simplistic discussion is given by [52].

A small simulation study was designed with sample sizes of $n = 300, 400$, and 500, with the number of true significant metabolites $n_t = 5$ and 10, and 100 metabolites to search through. The magnitude of the effect is 2.0, and the magnitude standard deviation is 0.5. The simulated data uses the same protocol as the simulation study in Section IV. To evaluate the performance, the %Correct, %Incorrect, and Overall Classification Accuracy are considered. Here, %Correct is the proportion of truly significant metabolites detected, and %Incorrect is the proportion of truly insignificant metabolites detected. Hence, a high %Correct indicates that the approach has the ability to find the true features. A high %Incorrect corresponds to a considerable number of insignificant metabolites being selected.

Table IV shows the results of this simulation study. Notice that SVMAI has extremely high correct detection rates, with the smallest at 0.997. In addition, the Classification Accuracy of the SVMAI technique is consistently 1.0, meaning 100% correct at classifying the heart health status. The %Incorrect for the SVMAI approach tends to be lower than the RF approach but much higher than the PLSDA technique. Of all the techniques considered here, the RF approach performs the worst, with low %Correct rates, high %Incorrect rates, and low Classification Accuracy. From this study, one can see that SVMAI is very good at determining the correct metabolites with high classification accuracy. However, it does select incorrect metabolites at a higher rate than PLSDA, but generally lower than RF.

TABLE IV. RESULTS OF SIMULATION STUDIES TO COMPARE ACROSS SVMAI, RF AND PLSDA TECHNIQUES FOR SCREENING. SAMPLE SIZES $n_s = 300, 400, 500$, NUMBER OF TRUE SIGNIFICANT METABOLITES $n_t = 5, 10$ WITH % CORRECT METABOLITES DETECTED, % INCORRECT METABOLITES DETECTED AND % CLASSIFICATION ACCURACY. BASED ON 100 SIMULATED DATASETS

		% Correct			% Incorrect			Classification Accuracy		
n_s	n_t	SVMAI	RF	PLSDA	SVMAI	RF	PLSDA	SVMAI	RF	PLSDA
300	5	1	0.8	0.798	0.123	0.553	0.23	1	0.539	0.851
	10	0.988	0.857	0.716	0.158	0.175	0.19	1	0.456	0.853
400	5	1	0.8	0.8	0.128	0.454	0.05	1	0.545	0.803
	10	0.997	0.887	0.761	0.143	0.156	0.05	1	0.470	0.806
500	5	1	0.8	0.8	0.141	0.398	0.06	1	0.552	0.778
	10	0.999	0.897	0.806	0.157	0.117	0.03	1	0.482	0.777

VI. QATAR BIOBANK RESULTS

For the Qatar BioBank data, there were $n_c = 223$ fully observed metabolites across $n_r = 626$ participants. The algorithm was performed on the dataset using 1,000 steps in the genetic algorithm. The computation time was approximately 12 minutes on an Apple M2 Pro processor with 32GB of RAM. Table V shows the metabolites with Inclusion Probabilities greater than 0.9 for the Qatar BioBank data concerning Heart Disease. Out of the 223 candidate metabolites, 37 had an inclusion probability above 0.9, and 48 had an inclusion probability above 0.5 (not shown). Notice that many of the metabolites have an inclusion probability above 0.999, indicating they were present in almost all of the models across the five genetic algorithm lines. Of particular interest is the fact that there are eight metabolites with the "X" prefix, which indicates an unknown chemical identity [53]. Also of note is that circulating BCAAs (leucine/isoleucine, valine, glutamate/glutamine, proline, and methionine) have been associated with predicting risk of coronary artery disease [54], [55], indicating the validity of our method in capturing previously known metabolites associated with heart disease.

TABLE V. METABOLITES AND THEIR INCLUSION PROBABILITIES FOR THE QATAR BIOBANK HEART DISEASE DATASET. ALL METABOLITES WITH AN INCLUSION PROBABILITY GREATER THAN 0.9 ARE PRESENTED IN DESCENDING ORDER OF INCLUSION PROBABILITY

Metabolite	Inclusion Probability
X - 23636	1.000
N1-methyladenosine	0.999
5alpha-pregnan-3beta,20alpha-diol disulfate	0.999
2,3-dihydroxy-5-methylthio-4-pentenoate (DMTPA)*	0.999
X - 11372	0.999
carnitine	0.999
2-hydroxyoctanoate	0.999
6-bromotryptophan	0.999
alanine	0.999
p-cresol sulfate	0.999
phosphoethanolamine	0.999
X - 14056	0.999
ornithine	0.999
deoxycarnitine	0.999
arginine	0.999
metabolonic lactone sulfate	0.999
X - 11880	0.999
X - 23639	0.999
1-arachidonoyl-GPI (20:4)*	0.999
choline	0.999
X - 21258	0.999
proline	0.999
dimethylarginine (SDMA + ADMA)	0.999
alpha-hydroxyisocaproate	0.999
indoleacetate	0.999
X - 24425	0.999
1-oleoyl-GPE (18:1)	0.999
N-delta-acetylornithine	0.999
5alpha-androstan-3beta,17beta-diol disulfate	0.999
taurocholenate sulfate*	0.999
pregnenediol sulfate (C21H34O5S)*	0.999
2-hydroxypalmitate	0.999
gamma-glutamylcitrulline*	0.999
X - 21364	0.999
hydroxy-CMPF*	0.998
creatinine	0.995
7-alpha-hydroxy-3-oxo-4-cholestenoate (7-Hoca)	0.931

VII. Conclusion

This work shows the ability to use a genetic algorithm approach with multiple lines and frequent crossbreeding in conjunction with support vector machines using a likelihood approach to create an AI method to screen metabolites for heart disease. The method proposed performs well under simple simulation studies across a large number of scenarios. The method allows for high correct identification rates of metabolites (above 85%). The false positive rate is moderately high, ranging from 5% to 30%. Hence, the algorithm would be good for screening purposes to locate candidate metabolites for further study. The results of the algorithm, when applied to the Qatar BioBank data, allow for 223 metabolites to be screened down to 37, all of which have an inclusion probability above 0.9. As the algorithm has a reasonably high false discovery rate, one should be careful to make declarative statements that all of the metabolites identified are truly contributing to the disease state. Instead, one should state that there is preliminary evidence that these metabolites may be related to these disease states. Furthermore, since the data is dominated by Control and High Cholesterol participants, the results most likely have selected metabolites that are biased toward these two states. However, future studies can verify or refute these claims. Knowing which metabolites contribute to heart disease will allow for assays to be developed that can indicate which people are likely to have heart disease, which can lead to preventative care to reduce the risk of heart disease.

Additional work could be done to study the performance of the algorithm on more complex datasets. The work here only considered a small number of disease states to mimic the Qatar BioBank data; however, a higher number of disease states and multiple disease states should be studied. Also, the interaction of metabolites was not considered here. Work by Boone et al. [56] and Lee and Boone [57] shows how to explore a restricted space when interactions are present for linear regression models. A similar approach could be developed here.

In the simulation study presented, the Genetic Algorithm only took 1,000 steps due to computational times. Another simulation study should be conducted to determine if the number of steps can reduce the false positive rates. This study should attempt to determine a rule that one could apply to determine the adequate number of steps to achieve an acceptable false positive rate. In this study, one should consider the ratio of candidate metabolites to number of participants, $\frac{n_c}{n_r}$, to look for a general pattern. Also, the threshold to be deemed important could be adjusted from 0.5 to a higher value to become more stringent and potentially reduce the number of false positives. Furthermore, the SVMAI method is superior to PLSDA and RF for Classification Accuracy and %Correct metabolites selected and should be considered when conducting metabolomic studies where classification is of interest.

Another item that could be considered is how to address the issue of missing data. Many of the metabolites have missing values for various participants, which in this study have been omitted. Only metabolites observed in all participants were used here. As the computational complexity here is high, several approaches could be used to *impute* the missing values, such as mean imputation, the EM algorithm, and possibly multiple imputation.

Acknowledgment

Ryad Ghanam and Edward Boone would like to thank Qatar Foundation and Virginia Commonwealth University in Qatar for their support through the Mathematical Data Science Lab. Faten S. Alamri would like to thank Princess Nourah bint Abdulrahman University for their support through grant number PNURSP2024R346. The authors would to acknowledge the data received from Qatar Biobank.

Funding

This research was funded by Virginia Commonwealth University in Qatar through the Mathematical Data Science Lab HQ1260 and Princess Nourah bint Abdulrahman University through Project number PNURSP2024R346

Authors' Contribution

Conceptualization, E.L., R.A. and F.S.; methodology, E.L.; software, E.L. and R.A.; validation, E.L., R.A., F.S., and E.B; data curation, R.A.; writing—original draft preparation, E.L, R.A., F.S. and E.B. All authors have read and agreed to the published version of the manuscript.

References

[1] J. R. Idle and F. J. Gonzalez, "Metabolomics," *Cell metabolism*, vol. 6, no. 5, pp. 348–351, 2007.

[2] A. Yilmaz, T. Geddes, B. Han, R. O. Bahado-Singh, G. D. Wilson, K. Imam, M. Maddens, and S. F. Graham, "Diagnostic biomarkers of alzheimer's disease as identified in saliva using 1h nmr-based metabolomics," *Journal of Alzheimer's Disease*, vol. 58, no. 2, pp. 355–359, 2017.

[3] H. Shang, J. Zheng, and J. Tong, "Integrated analysis of transcriptomic and metabolomic data demonstrates the significant role of pyruvate carboxylase in the progression of ovarian cancer," *Aging (Albany NY)*, vol. 12, no. 21, p. 21874, 2020.

[4] D. Luo, Z. Shan, Q. Liu, S. Cai, Q. Li, and X. Li, "A novel seventeen-gene metabolic signature for predicting prognosis in colon cancer," *BioMed Research International*, vol. 2020, no. 1, p. 4845360, 2020.

[5] M. Huang, H.-Y. Li, H.-W. Liao, C.-H. Lin, C.-Y. Wang, W.-H. Kuo, and C.-H. Kuo, "Using post-column infused internal standard assisted quantitative metabolomics for establishing prediction models for breast cancer detection," *Rapid Communications in Mass Spectrometry*, vol. 34, p. e8581, 2020.

[6] I. Dudka, A. Chachaj, A. Sebastian, W. Tański, H. Stenlund, G. Gröbner, and A. Szuba, "Metabolomic profiling reveals plasma glyca and glycb as a potential biomarkers for treatment efficiency in rheumatoid arthritis," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 197, p. 113971, 2021.

[7] S. Takahashi, J. Saegusa, A. Onishi, and A. Morinobu, "Biomarkers identified by serum metabolomic analysis to predict biologic treatment response in rheumatoid arthritis patients," *Rheumatology*, vol. 58, no. 12, pp. 2153–2161, 2019.

[8] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[9] C. Mondon, "Classification and regression trees," 1984.

[10] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[12] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN computer science*, vol. 2, no. 6, pp. 1–20, 2021.

[13] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Fundamentals of artificial neural networks and deep learning," in *Multivariate statistical machine learning methods for genomic prediction*. Springer, 2022, pp. 379–425.

[14] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nature genetics*, vol. 51, no. 1, pp. 12–18, 2019.

[15] J. Heinemann, A. Mazurie, M. Tokmina-Lukaszewska, G. J. Beilman, and B. Bothner, "Application of support vector machines to metabolomics experiments with limited replicates," *Metabolomics*, vol. 10, pp. 1121–1128, 2014.

[16] L. Li, W. Jiang, X. Li, K. L. Moser, Z. Guo, L. Du, Q. Wang, E. J. Topol, Q. Wang, and S. Rao, "A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset," *Genomics*, vol. 85, no. 1, pp. 16–23, 2005.

[17] L. Tapak, S. Afshar, M. Afrasiabi, M. K. Ghasemi, and P. Alirezaei, "Application of genetic algorithm-based support vector machine in identification of gene expression signatures for psoriasis classification: a hybrid model," *BioMed Research International*, vol. 2021, no. 1, p. 5520710, 2021.

[18] A. Alzubaidi, G. Cosma, D. Brown, and A. G. Pockley, "Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information," in *2016 International Conference on Interactive Technologies and Games (ITAG)*. IEEE, 2016, pp. 70–76.

[19] Y. Zheng, X. Yang, M. Siddique, and G. Beddoe, "Simultaneous feature selection and classification based on genetic algorithms: an application to colonic polyp detection," in *Medical Imaging 2008: Computer-Aided Diagnosis*, vol. 6915. SPIE, 2008, pp. 123–131.

[20] N. Y. Moteghaed, K. Maghooli, and M. Garshasbi, "Improving classification of cancer and mining biomarkers from gene expression profiles using hybrid optimization algorithms and fuzzy support vector machine," *Journal of medical signals and sensors*, vol. 8, no. 1, p. 1, 2018.

[21] L. Ali, I. Wajahat, N. Amiri Golilarz, F. Keshtkar, and S. A. C. Bukhari, "Lda–ga–svm: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine," *Neural Computing and Applications*, vol. 33, pp. 2783–2792, 2021.

[22] H.-H. Hsu, C.-W. Hsieh, and M.-D. Lu, "Hybrid feature selection by combining filters and wrappers," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8144–8150, 2011.

[23] W. Ali and F. Saeed, "Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data," *Processes*, vol. 11, no. 2, p. 562, 2023.

[24] H. Al Kuwari, A. Al Thani, A. Al Marri, A. Al Kaabi, H. Abderrahim, N. Afifi, F. Qafoud, Q. Chan, I. Tzoulaki, P. Downey *et al.*, "The qatar biobank: background and methods," *BMC public health*, vol. 15, pp. 1–9, 2015.

[25] K. Suhre, N. Stephan, S. Zaghlool, C. R. Triggle, R. J. Robinson, A. M. Evans, and A. Halama, "Matching drug metabolites from non-targeted metabolomics to self-reported medication in the qatar biobank study," *Metabolites*, vol. 12, no. 3, p. 249, 2022.

[26] V. Vapnik and V. Vapnik, "Statistical learning theory wiley," *New York*, vol. 1, no. 624, p. 2, 1998.

[27] H. Cevikalp, "Best fitting hyperplanes for classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1076–1088, 2016.

[28] S. Abe, *Support vector machines for pattern classification*. Springer, 2005, vol. 2.

[29] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "Svm and kernel methods matlab toolbox. perception systèmes et information, insa de rouen, rouen, france (2005)."

[30] M. Awad and L. Khan, "Support vector machines," in *Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2008, pp. 1138–1146.

[31] B. Caputo, K. Sim, F. Furesjo, and A. Smola, "Appearance-based object recognition using svms: which kernel should i use?" in *Proc of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision, Whistler*, vol. 2002, 2002.

[32] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[33] A. Gammerman, N. Bozanic, B. Schölkopf, V. Vovk, V. Vapnik, L. Bottou, A. Smola, C. Watkins, Y. LeCun, C. Saunders *et al.*, "Royal holloway support vector machines," *URL http://svm. dcs. rhbnc. ac. uk/dist/index. shtml*, 2001.

[34] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, no. Aug, pp. 975–1005, 2004.

[35] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE transactions on neural networks*, vol. 13, no. 2, pp. 464–471, 2002.

[36] C.-J. Lin and J. J. Moré, "Newton's method for large bound-constrained optimization problems," *SIAM Journal on Optimization*, vol. 9, no. 4, pp. 1100–1127, 1999.

[37] C.-J. Lin, R. C. Weng *et al.*, "Simple probabilistic predictions for support vector regression," *National Taiwan University, Taipei*, 2004.

[38] A. C. Lorena, A. C. De Carvalho, and J. M. Gama, "A review on the combination of binary classifiers in multiclass problems," *Artificial Intelligence Review*, vol. 30, pp. 19–37, 2008.

[39] Y. Shiraishi and K. Fukumizu, "Statistical approaches to combining binary classifiers for multi-class classification," *Neurocomputing*, vol. 74, no. 5, pp. 680–688, 2011.

[40] A. Rocha and S. K. Goldenstein, "Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 2, pp. 289–302, 2013.

[41] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.

[42] R. C. Team, "Ra language and environment for statistical computing, r foundation for statistical," *Computing*, 2020.

[43] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, and C.-C. Lin, "Misc functions of the department of statistics," *Probability Theory Group (Formerly: E1071), TU Wien*, 2015.

[44] A. Karatzoglou and I. Feinerer, "Kernel-based machine learning for fast text mining in r," *Computational Statistics & Data Analysis*, vol. 54, no. 2, pp. 290–297, 2010.

[45] N. Chen, H.-B. Wang, B.-Q. Wu, J.-H. Jiang, J.-T. Yang, L.-J. Tang, H.-Q. He, and D.-D. Linghu, "Using random forest to detect multiple inherited metabolic diseases simultaneously based on gc-ms urinary metabolomics," *Talanta*, vol. 235, p. 122720, 2021.

[46] V. H. Oza, J. K. Aicher, and L. K. Reed, "Random forest analysis of untargeted metabolomics data suggests increased use of omega fatty acid oxidation pathway in drosophila melanogaster larvae fed a medium chain fatty acid rich high-fat diet," *Metabolites*, vol. 9, no. 1, p. 5, 2018.

[47] A. Acharjee, J. Larkman, Y. Xu, V. R. Cardoso, and G. V. Gkoutos, "A random forest based biomarker discovery and power analysis framework for diagnostics research," *BMC medical genomics*, vol. 13, pp. 1–14, 2020.

[48] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[49] P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner, and R. Goodacre, "A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding," *Analytica chimica acta*, vol. 879, pp. 10–23, 2015.

[50] H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 72, no. 1, pp. 3–25, 2010.

[51] K.-A. Lê Cao, S. Boitard, and P. Besse, "Sparse pls discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems," *BMC bioinformatics*, vol. 12, pp. 1–17, 2011.

[52] D. Ruiz-Perez, H. Guan, P. Madhivanan, K. Mathee, and G. Narasimhan, "So you think you can pls-da?" *BMC bioinformatics*, vol. 21, pp. 1–10, 2020.

[53] J. Krumsiek, K. Suhre, A. M. Evans, M. W. Mitchell, R. P. Mohney, M. V. Milburn, B. Wägele, W. Römisch-Margl, T. Illig, J. Adamski *et al.*, "Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information," 2012.

[54] S. H. Shah, J. R. Bain, M. J. Muehlbauer, R. D. Stevens, D. R. Crosslin, C. Haynes, J. Dungan, L. K. Newby, E. R. Hauser, G. S. Ginsburg *et al.*, "Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events," *Circulation: Cardiovascular Genetics*, vol. 3, no. 2, pp. 207–214, 2010.

[55] R. Yang, S. Wang, L. Sun, J. Liu, H. Li, X. Sui, M. Wang, H. Xiu, S. Wang, Q. He *et al.*, "Association of branched-chain amino acids with coronary artery disease: A matched-pair case–control study," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 25, no. 10, pp. 937–942, 2015.

[56] E. L. Boone, S. J. Simmons, K. Ye, and A. E. Stapleton, "Analyzing quantitative trait loci for the arabidopsis thaliana using markov chain monte carlo model composition with restricted and unrestricted model spaces," *Statistical methodology*, vol. 3, no. 1, pp. 69–78, 2006.

[57] A. H. Lee and E. L. Boone, "A frequentist assessment of bayesian inclusion probabilities for screening predictors," *Journal of Statistical Computation and Simulation*, vol. 81, no. 9, pp. 1111–1119, 2011.

Enhancing Deepfake Content Detection Through Blockchain Technology

Qurat-ul-Ain Mastoi[1], Muhammad Faisal Memon[2], Salman Jan[3], Atif Jamil[4],
Muhammad Faique[5], Zeeshan Ali[6], Abdullah Lakhan[7], Toqeer Ali Syed[8]
School of Computer Science and Creative Technologies, University of the West of England, Bristol, BS16 1QY, United Kingdom[1]
Department of Cybersecurity, Dawood University of Engineering and Technology, Karachi, Sindh, Pakistan, 74800[2, 7]
Faculty of Computer Studies, Arab Open University, Kingdom of Bahrain[3]
Dean of ICS, Dawood University of Engineering and Technology, Pakistan[4]
Department of Telecommunication, Dawood University of Engineering and Technology, Pakistan[6]
Department of Computer Science, Mehran University of Engineering and Technology, Pakistan[5]
Faculty of Computer and Information System, Islamic University of Madinah, Madinah, Saudi Arabia[8]

***Abstract*—Deepfake technology poses a growing threat to the authenticity and trustworthiness of digital media, necessitating the development of advanced detection mechanisms. While AI-based methods have shown promise, they generally face limitations in terms of generalization and scalability. We present a blockchain-enabled watermarking technique, characterized by its immutable, transparent, and decentralized nature, which offers a robust complementary approach for enhancing media authentication through methods such as cryptographic watermarking, decentralized identity, and content provenance tracking. To train and evaluate blockchain-based watermarking and deepfake detection systems, a variety of large-scale datasets are utilized. Video datasets include UADFV (49 real, 49 fake), Deepfake-TIMIT (320 real, 640 fake), DFFD (1000 real, 3000 fake), Celeb-DF v2 (590 real, 5639 fake), DFDC (23,564 real, 104,500 fake), DeeperForensics-1.0 (50,000 real, 10,000 fake), FaceForensics++ (1000 real, 5000 fake), and ForgeryNet (99,630 real, 121,617 fake). Image datasets include DFFD (58,703 real, 240,336 fake), FFHQ (70,000 GAN-generated), iFakeFaceDB (87,000 fake), 100k AI Faces, and over 2.8 million samples in ForgeryNet. Despite integration challenges such as scalability, computational cost, and standardization, blockchain-based solutions show promise in tracking content origin and enhancing verification. Simulation results demonstrate that the proposed blockchain-enabled watermarking achieves a higher accuracy in detecting fake content compared to existing machine learning methods.**

***Keywords*—Blockchain; deep fake; convolutional neural network (CNN); long short-term memory (LSTM); RNN (recurrent neural network); video and image**

I. Introduction

Deepfake technology represents a significant advancement in artificial intelligence, enabling the creation of highly realistic, manipulated, or entirely synthetic media, including videos, photos, and audio recordings [1]. These digital forgeries are crafted using sophisticated machine learning techniques, often employing algorithms such as Generative Adversarial Networks (GANs) [2]. A GAN typically involves two competing neural networks: a generator, which produces fake content, and a discriminator, which attempts to distinguish between real and synthetic media, resulting in increasingly convincing results [3]. Beyond GANs, other AI models, such as autoencoders and diffusion models, also contribute to the creation of deepfakes, highlighting the evolving landscape of this technology. This technology enables various manipulations, ranging from swapping faces and altering facial expressions to synthesizing entire identities and voices [4].

The proliferation of deepfakes is not solely due to technological advancements but also to the increasing accessibility of user-friendly tools and open-source platforms. Individuals with even basic computer skills can now generate convincing deepfakes, leading to a widespread presence of manipulated content online [5]. This ease of creation has amplified the urgency for robust and effective detection methods [6]. The potential for deepfakes to undermine the authenticity of information poses a considerable threat, capable of large-scale misinformation and manipulation [7].

The impact of deepfakes extends across various domains, eroding public trust in digital media and complicating the ability to discern genuine content from fabrications. Malicious applications range from spreading fake news and non-consensual pornography to identity theft, exploitation of public figures, financial scams, influencing elections, inciting social unrest, and even psychological warfare [8]. Notable examples illustrate potential harm, such as the circulation of a false White House tweet and a fake image of an explosion near the Pentagon, causing public anxiety and market fluctuations. Manipulated political speeches and videos have also surfaced, demonstrating the technology's ability to distort public opinion and the political landscape. Furthermore, deepfake fraud has resulted in significant financial losses for both individuals and organizations, as seen in cases of CEO impersonation leading to multi-million-dollar fraudulent transactions. Given the increasing sophistication of these forgeries, relying on human discernment alone is becoming increasingly ineffective, thereby highlighting the critical need for advanced technological solutions in deepfake detection [9].

The above studies have many limitations. Many machine learning methods, such as CNN and other techniques, have a high rate of false positives and consume significantly more time, requiring high scalability of resources during fake detection based on generated GAN methods-enabled data.

- To develop a robust deepfake detection framework that accurately identifies manipulated video and audio content using advanced AI models.
- To integrate blockchain technology for secure, tamper-proof verification and traceability of multimedia content origins and authenticity.
- To ensure decentralized trust and transparency in content-sharing platforms by utilizing smart contracts and immutable watermarking techniques.
- We proposed a novel hybrid system that combines deep learning-based deepfake detection with blockchain-enabled watermark verification.
- We developed a decentralized content validation mechanism using smart contracts to track and verify video authenticity across digital platforms.

The study is organized as follows: Section II is about related work. Section III is the proposed methodology. Section IV discusses the result analysis, and Section V presents the conclusion and future work.

II. Literature Review

In this direction, Kumar et al. [1] combined a blockchain-based AKA mechanism with explainable artificial intelligence (XAI) to secure smart city-based consumer applications. Initially, the involved entities securely communicate with one another to share data via a blockchain-based AKA mechanism. Conversely, they employed the SHapley Additive explanations (SHAP) mechanism to explain and interpret the leading features that significantly impact the decision.

In this context, Muneer et al. [2] emphasizes the prediction of prices of three top cryptocurrencies: Bitcoin, Ethereum, and Litecoin. They employed three machine learning (ML) algorithms: LSTM, SVM, and RF, and utilized the LSTM-RF ensemble to enhance the predictive accuracy of the predicted cryptocurrencies. In comparison to all the models studied, the hybrid LSTM-RF model was the most accurate in predicting the respective performance indices and outperformed the other conventional models and standalone machine learning techniques. Additionally, this work employs Explainable Artificial Intelligence (XAI) techniques to develop AI-based, human-friendly, and interpretable visualizations. This method makes insights accessible to basic users, as it enables investors to make informed decisions based on the outcomes provided by the proposed model.

In addition, Sachan et al. [3] strive to strike a balance between the responsible use and governance of human legal professionals in content generated by Generative AI through periodic audits. The research model conceptualizes the two algorithms. Strategic on-chain (on blockchain) and off-chain (off blockchain) storage of data in adherence to the data protection laws and key demands of stakeholders within a legal firm. Second, comparison auditing of the singular signature as Merkle roots of off-chain stored files to their immutable blockchain equivalent. The system preserves the integrity of data repositories that hold the choices made by an XAI model and a textual explanation of legal cases obtained through a prompt issued to the API of Generative AI, enabling automated auditing.

In addition, Kumar, P. et al. [4] introduced an explainable AI (XAI)-enabled blockchain to enhance the decision-making capabilities of cyber threat detection in the field of Smart Healthcare Systems. Initially, they employ blockchain to secure and store information across various cloud providers through the application of Clique Proof-of-Authority (C-PoA) consensus. Second, a new threat-hunting model based on deep learning is constructed by integrating Parallel Stacked Long Short-Term Memory (PSLSTM) networks and a multi-head attention mechanism to enhance attack detection. The large-scale experiment validates its potential to serve as an advanced decision support system for cybersecurity analysts.

Furthermore, Chen, H. Y et al. [5] give a comprehensive bibliometric description and visualization of the integration of two leading and promising technologies, explainable AI and Blockchain, into Smart Agriculture. Through the integration of artificial intelligence, blockchain technology, and "smart" agriculture, researchers can create a more efficient, more open, and more sustainable food system. They posited that the use of blockchain technology in smart agriculture can enhance transparency by providing an immutable record of a product's origin, production processes, and distribution. It can potentially create a food system that is more beneficial to farmers, consumers, and the environment by improving efficiency, transparency, and sustainability.

Moreover, Salama, R et al. [6] develop a new Blockchain with Explainable Artificial Intelligence-Driven Intrusion Detection for IoT-driven ubiquitous computing systems (BXAI-IDCUCS) model. The primary goal of the BXAI-IDCUCS model is to achieve energy efficiency and security in the IoT setting. The BXAI-IDCUCS model first groups the IoT nodes with an energy-aware duck swarm optimization (EADSO) algorithm to accomplish this. Additionally, a deep neural network (DNN) is employed to detect and classify intrusions within the IoT network. Finally, blockchain technology is used for safe inter-cluster data transmission procedures. To guarantee the productive operation of the BXAI-IDCUCS model, an extensive experimentation study is utilized, and results are measured under various parameters. The comparison study highlighted the superiority of the BXAI-IDCUCS model over existing state-of-the-art methods, achieving a packet delivery ratio of 99.29%, a packet loss rate of 0.71%, a throughput of 92.95 Mbps, an energy consumption of 0.0891 mJ, a lifetime of 3529 rounds, and an accuracy of 99.38%.

Additionally, El Houda, Z. et al. [7] introduce a new framework, referred to as FedIoT, which relies on Explainable Artificial Intelligence (XAI) methods and Blockchain to protect FL-based IDS from IoT networks. FedIoT employs sophisticated XAI methods for detecting local model manipulations and preventing FL-based attacks. Further, we

suggest a blockchain method that employs a lightweight reputation system that guarantees the reliability and credibility of the FL training process. We perform experiments to verify FedIoT, an FL-based intrusion detection system for IoT networks. We use the UNSW-NB15 dataset to verify that FedIoT can successfully identify malicious behavior and enable efficient FL collaboration among multiple users.

Sharma, N et al. [8] proposes a hybrid security system that integrates blockchain for decentralized data storage and an intrusion detection system (IDS) based on a multi-attention deep convolutional recurrent neural network (MA-DeepCRNN) model. The model enhances block creation time and throughput, with minimal transaction reversal probability. The IDS delivers high classification performance, enhancing accuracy and F1-score in IoMT systems, boosting security, scalability, and real-time threat detection in healthcare settings.

Further, Hasan, M et al. [9] attempts to bridge this gap by combining explainable artificial intelligence (XAI) methods and anomaly rules in tree-based ensemble classifiers for the identification of anomalous Bitcoin transactions. In addition, they introduce rules for explaining whether a Bitcoin transaction is anomalous or not. Furthermore, we propose an under-sampling algorithm called XGBCLUS, with the aim of balancing anomalous and non-anomalous transaction data. This algorithm contrasts with other widely employed under-sampling and over-sampling methods. Ultimately, the performances of several tree-based individual classifiers are contrasted with stacking and voting ensemble classifiers. Their experimental results show that: i) XGBCLUS improves true positive rate (TPR) and receiver operating characteristic-area under curve (ROC-AUC) scores over state-of-the-art under-sampling and over-sampling methods, and ii) their ensemble classifiers proposed in this study perform better than conventional single tree-based machine learning classifiers with respect to accuracy, TPR, and false positive rate (FPR) scores.

In addition, Dutta, J. et al. [10] utilizes SHapley Additive exPlanations (SHAP) to offer real-time, model-agnostic explanations of AI predictions to support personalized health monitoring and well-informed decision making in healthcare To enhance data security, a consortium blockchain is used, and AI is applied to recognize block data sensitivity at the edge in blockchain-integrated IoT architectures with Random Forest (RF) algorithm. This method achieves high precision in authenticating block sensitivity, enabling the effective selection of authentication methods in proof-of-authentication (PoAh) consensus within the consortium blockchain. This provides increased security for sensitive information and improves overall blockchain performance. The framework is tested in an edge computing system, exhibiting tremendous potential for promoting eHealth security and authentication, and marking a significant leap in medical technology.

We present Table I to define the findings and limitations of the studies.

TABLE I DEEPFAKE DETECTION BASED ON MACHINE LEARNINGS

Method	Mechanism	Key Strengths	Limitations
CNNs [11]	Analyze spatial hierarchies of features, detect inconsistencies in facial features, alignment, etc.	Effective at identifying spatial anomalies and inconsistencies within frames.	Can be fooled by high-quality deepfakes and may struggle with temporal inconsistencies.
RNNs [12]	Evaluate temporal consistency of movements, speech, and expressions.	Good at detecting unnatural movements and lip-sync issues.	May not be as effective for static images or subtle temporal anomalies.
Visual Artifact Analysis [13]	Detect inconsistencies in lighting, shadows, skin texture, blending boundaries, and other digital artifacts.	Can identify traces left by the deepfake generation process that may not be obvious to humans.	Effectiveness decreases as deepfake quality improves and fewer artifacts are present.
Biometric Analysis [14]	Analyzes unique physiological characteristics like blinking patterns and eye movements.	Exploits inherent human traits that are difficult for AI to replicate perfectly.	Requires high-quality video and may not be effective against deepfakes that accurately mimic these traits.
Multimodal Detection [15]	Combines analysis of audio, video, and potentially other data sources.	Provides a more comprehensive assessment of authenticity by leveraging information from multiple modalities.	More complex to implement and requires synchronized and high-quality data across all modalities.
Metadata Analysis [16]	Examines digital information embedded in media files (e.g., timestamps, editing history).	Can quickly identify inconsistencies that suggest tampering.	Metadata can be easily altered or removed. Relies on the assumption that original media has complete and accurate metadata.
Provenance-based Detection [17]	Examines metadata for signs of manipulation, often relying on digital watermarks.	It can provide strong evidence of authenticity if watermarks are reliably implemented and protected.	Suffers from lack of standardization, potential for removal or misuse of watermarks, and requires buy-in from content creators and platforms. Does not work for content without watermarks.
Inference-based Detection [18]	Analyzes the media content itself for statistical anomalies and patterns indicative of manipulation, without relying on external metadata.	Can detect deepfakes even without provenance information. Adapts to new deepfake techniques such as models are retrained.	Relies on the quality and diversity of training data. Can produce false positives or negatives. The "ground truth" is often assumed, not definitively known.

The above works and studies have limitations, as discussed in Table I. Therefore, we provide new solutions based on blockchain and watermarking techniques to detect deepfakes within data. These studies [19-22] suggest a fake news detection system based on blockchain and cybersecurity approaches, which has been deployed in practice. These studies [23-27] suggested an edge cloud-enabled distributed healthcare system to promote secure healthcare services through telemedicine applications for users.

III. Proposed Blockchain Architecture

Blockchain technology offers a novel and promising foundation for enhancing trust in digital information. Its fundamental principles and key features align well with the challenges posed by deepfakes. At its core, blockchain is a Distributed Ledger Technology (DLT) where data is not stored in a central repository but is instead distributed across numerous computers within a network. This decentralized nature enhances security and resilience by eliminating a single point of failure. A defining characteristic of blockchain is its immutability. Once a record, or block, is added to the blockchain, it becomes challenging to alter or delete. This is achieved through cryptographic hashing and the sequential linking of blocks. Any attempt to modify a block would necessitate recalculating the cryptographic hash for that block and all subsequent blocks. This computationally intensive task is practically infeasible for a malicious actor to accomplish without the consensus of the network. In the event of an error, a new transaction is added to the blockchain to rectify it, ensuring that both the original and the correction remain visible, thus maintaining an auditable history. This inherent immutability provides a robust guarantee of data integrity, making blockchain a strong candidate for verifying the authenticity of digital content. Transparency is another key principle of blockchain. While the level of transparency can vary depending on the type of blockchain (public versus permissioned), transaction records are typically visible to all network participants or authorized parties, fostering trust and accountability. All participants in the network have access to the distributed ledger, ensuring a shared and verifiable record of transactions. This transparency enables the open verification of content provenance, potentially increasing user trust in the authenticity of the media. The decentralized nature of blockchain means that no single entity has control over the network. This lack of central authority reduces the risk of censorship and eliminates single points of vulnerability. The network's distributed structure makes it more resilient to failures and cyberattacks. This decentralization enhances the security and robustness of any deepfake detection system built upon blockchain technology, as it avoids reliance on potentially compromised central authorities. Several key features of blockchain are particularly relevant to deepfake detection. Cryptographic hashing is a fundamental element where data within each block is transformed into a unique, fixed-length string of characters known as a hash, using cryptographic hash functions such as SHA-256 or Keccak-256. Even the slightest alteration to the original data will result in a completely different hash. Each block in the blockchain contains the hash of the preceding block, creating a secure and tamper-evident chain. These cryptographic hashes are designed to be irreversible and collision-resistant, further enhancing data integrity. This provides a mechanism for creating a unique and tamper-evident digital fingerprint for media content, which can be used to verify its integrity. Consensus mechanisms are the protocols by which network participants, or nodes, agree on the validity of new transactions and blocks before they are added to the blockchain. These mechanisms, such as Proof of Work or Proof of Stake, ensure the integrity of the ledger by requiring a majority of the network to validate any new addition. This ensures that any attempt to tamper with the blockchain record of media provenance would be rejected by the network, further strengthening trust in the system. Smart contracts are self-executing contracts where the terms of the agreement are directly written into code. These contracts can automate verification processes based on predefined rules, enabling efficient and trustless verification of content authenticity against established criteria or on-chain records.

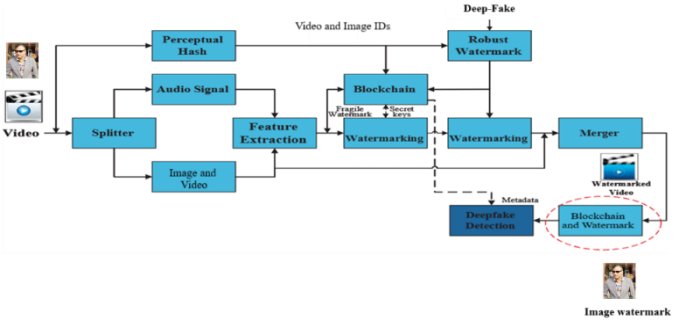


Fig. 1. Proposed blockchain-enabled deep fake detection for different data modes.

Fig. 1 presents a comprehensive framework for deep-fake detection and secure video authentication using blockchain and watermarking techniques. The process begins by splitting the input video into audio, image, and video components, followed by perceptual hashing and feature extraction to generate unique IDs. These features are then used to create fragile watermarks and secret keys, which are managed via blockchain for secure access and traceability. Robust and fragile watermarks are embedded into the video content to ensure both durability and sensitivity to tampering. The watermarked components are merged to produce a final authenticated video. Metadata from this process is passed to a deepfake detection module, which validates the authenticity of the content. The system ensures end-to-end integrity by combining blockchain verification with advanced watermarking, offering a reliable solution against manipulated media. Blockchain technology offers several specific applications that can significantly enhance the detection of deepfake content. One key application is the use of cryptographic hashing to create unique digital fingerprints of original media content, including images, videos, and audio. These hashes can then be securely stored on the blockchain. Subsequently, anyone can compare the hash of a potentially manipulated piece of content with the original hash recorded on the blockchain to detect any alterations. By storing these cryptographic hashes of original content on the immutable blockchain, a permanent and verifiable record of the content's integrity is established. When an original piece of media is

created, its unique hash is calculated and stored on the blockchain. If a deepfake is subsequently generated by manipulating this original content, the hash of the deepfake will invariably be different. Comparing this new hash to the original one stored on the blockchain provides irrefutable evidence of tampering. Blockchain also facilitates decentralized verification mechanisms where multiple participants within a network can independently verify the authenticity of digital content. This process can involve a consensus mechanism where network nodes vote on the authenticity of a given piece of media, potentially incentivized through rewards for accurate verification. Decentralized verification through blockchain can mitigate the risks of bias and single points of failure that are inherent in centralized verification systems. Instead of relying on a single organization or algorithm to determine if content is a deepfake, a blockchain-based system can harness the collective intelligence of a distributed network. This collaborative approach can lead to more accurate and unbiased verification outcomes, as multiple independent parties contribute to the process. Furthermore, blockchain can be used to secure blockchain-based digital watermarking. Digital watermarks containing crucial information about the content's origin and history can be embedded within the media itself, and their integrity can be protected by anchoring them to the blockchain [15]. This provides a tamper-proof method for tracking the provenance of the content [21]. Linking digital watermarks to the blockchain significantly enhances their security and makes them far more resistant to tampering or removal. While traditional digital watermarks may be susceptible to alteration or removal, anchoring the information within the watermark, such as the creator's ID, a timestamp, and the original hash of the content, to the blockchain safeguards its authenticity and integrity through the blockchain's immutable nature.

Ultimately, blockchain-based decentralized identity (DID) solutions can provide a secure and verifiable method for identifying content creators. Creators can cryptographically sign their content using their private key, and this signature can be verified by anyone using their corresponding public key, which is securely stored on the blockchain. This process effectively establishes ownership and attribution of the content. Utilizing blockchain-based decentralized identities can establish clear ownership and provenance for digital content, making it considerably easier to identify manipulated versions and to hold malicious actors accountable for their actions. When a content creator registers their identity on the blockchain and signs their work with their unique private key, it creates an irrefutable link between the creator and the content. This enables the verification of the media's origin and the tracking of its distribution across the internet, significantly aiding in the detection of any unauthorized alterations.

A. Blockchain-Based Approach

Employing blockchain technology for deep-fake detection offers several compelling benefits that address the limitations of current methods, as shown in Algorithm I. Blockchain provides an immutable and transparent record of the origin and history of digital content, significantly improving content provenance. Every modification or instance of sharing the content can be recorded on the blockchain, creating a comprehensive and traceable audit trail. This detailed history makes it considerably easier to identify the authentic version of the content and to detect any subsequent manipulations. The ability to trace the entire lifecycle of digital media on the blockchain significantly enhances content provenance and facilitates the identification of deepfakes. By recording every stage of a digital asset's existence, from its initial creation to any subsequent modifications and its distribution across the internet, a comprehensive historical record is established. This detailed provenance information allows for straightforward verification of whether a piece of content has been altered from its original, authentic state.

The inherent transparency and immutability of blockchain technology can substantially increase user trust in the authenticity of digital media. Users are empowered to independently verify the provenance and integrity of the content they encounter. The inherent trust mechanisms of blockchain can help combat the growing erosion of trust in digital media, caused by the increasing prevalence of deepfakes. In an online environment where distinguishing between real and fake content is becoming increasingly challenging, blockchain offers a technological solution that provides verifiable proof of authenticity. This transparency and the assurance of immutability can help rebuild trust in digital media by empowering users to verify information for themselves, rather than relying on potentially fallible centralized authorities.

B. Algorithm I: Blockchain and Watermark-Based Deepfake Detection Framework

1. Input a video file into the system.
2. Split the video into three components: audio, image frames, and video stream.
3. Generate perceptual hashes for the audio and image/video components to create unique content IDs.
4. Extract visual and audio features from the image and video components.
5. Use the extracted features to create fragile watermarks and secret keys.
6. Store the generated watermarks and keys on a secure blockchain ledger.
7. Embed the fragile and robust watermarks into the video components using watermarking techniques.
8. Recombine (merge) the watermarked audio, image, and video into a final watermarked video.
9. Send the metadata and watermarked video to the deepfake detection module.
10. Analyze the content for authenticity and detect any deep-fake manipulation using the embedded watermarks and blockchain metadata.
11. Verify the integrity and originality of the video using blockchain and watermark data.
12. Output the verified and watermarked video along with detection results.

Blockchain enables the creation of decentralized networks for content verification, allowing multiple independent parties to contribute to and benefit from shared knowledge and

verification efforts. This decentralized approach can lead to more robust and accurate detection outcomes through community-driven models of verification. Decentralized verification on the blockchain can leverage collective intelligence and reduce reliance on potentially biased or flawed centralized systems. By distributing the task of verifying content across a network of participants, a blockchain-based system can tap into a broader range of expertise and perspectives. This collaborative approach has the potential to enhance the accuracy and reliability of deepfake detection, as it mitigates the risk of errors or biases that might be inherent in a single centralized authority or algorithm.

The decentralized and cryptographically secured nature of blockchain enhances the security of provenance data and offers improved privacy. Furthermore, decentralized identity solutions built on blockchain can provide users with greater control over their data used in verification processes. Blockchain can provide a more secure and privacy-preserving infrastructure for verifying media authenticity compared to traditional centralized databases, which are often more vulnerable to breaches and unauthorized access. The cryptographic security inherent in blockchain, coupled with its decentralized structure, makes it a highly secure platform for storing sensitive information related to content provenance. Additionally, decentralized identity solutions built on blockchain can empower individuals to manage their identity data, enhancing privacy throughout the verification process.

Blockchain technology enables the accurate timestamping of data, thereby establishing a precise chronological order of content creation or publication. This timestamping can be particularly crucial for identifying manipulated content, as it provides a verifiable timeline. The timestamping feature of blockchain provides valuable temporal context that can help distinguish original content from later manipulations. Knowing the exact time when a piece of digital media was created and immutably recorded on the blockchain can be a critical factor in identifying deepfakes. For example, suppose a video purports to depict an event that allegedly occurred before the recorded creation time of the video on the blockchain. In that case, it will strongly suggest that the video has been manipulated or is entirely fabricated.

IV. Performance Evaluation

In the simulation of multimodal data for deepfake detection based on blockchain watermarking, a comprehensive set of parameters is configured to ensure robust performance and security. The simulation involves a dataset comprising both authentic and manipulated video and image content, sourced from publicly available benchmarks such as FaceForensics++, Celeb-DF, and Deepfake Detection Challenge datasets. Each video is processed at a resolution of 720p with a frame rate of 30 frames per second (FPS). At the same time, images are standardized to 224x224 pixels for consistency during training and inference using a MobileNetV2-based deep learning model. For watermarking, a unique, tamper-proof digital watermark is embedded into each frame and image using a frequency domain technique, such as the DWT-DCT hybrid or singular value decomposition (SVD), ensuring both imperceptibility and robustness. The embedded watermark is linked to a blockchain network using smart contracts deployed on an Ethereum testnet via Web3.py. Each transaction logs metadata, including the hash of the media content, timestamp, owner ID, and verification key. During detection, real-time feature extraction and watermark verification are performed concurrently, using OpenCV and TensorFlow to compare the embedded watermark against the blockchain-registered hash. The simulation also accounts for multiple sharing scenarios (e.g., social media platforms such as Facebook, Twitter, YouTube, and LinkedIn) and evaluates watermark resilience against compression, resizing, cropping, and format conversion. Performance metrics, including accuracy, precision, recall, F1-score, PSNR, SSIM, and blockchain transaction latency, are monitored to assess the system's capability to detect tampered content. Additionally, a GUI interface (built with Streamlit or Tkinter) visualizes results and logs verified or fake media instances into a local SQLite database or cloud-based storage, ensuring traceability and trust in media authenticity. All the details are shown in Table II, III, and IV.

TABLE II Simulation Parameters

Simulation	Parameters
Language	Ethereum and WEB3
Blockchain	Public Blockchain
Workload	Images, Videos
Training Dataset	Based on Watermarking and Blockchain
Nodes	Client and Social Media
Metrics	Accuracy, Precision, Recall, and F1-Score
Simulation Time	24 Hours

TABLE III Dataset Descriptions

Dataset	Real Videos	Fake Videos
UADFV	49	49
DFFD	1000	3000
Celeb-DF	590	5639
DFDC	23,564	104,500
DeeperForensics-1.0	50,000	10,000
FaceForensic++	1000	5000

TABLE IV Image Datasets

Dataset	Real Images	Fake Images
DFFD	58,703	240,336
FFHQ	60,000	70,000 GAN
iFakeFaceDB	80,000	87,000 GAN
100k Faces	10,0000	100,000 (GAN)
ForgeryNet	1,438,201	1,457,861

A. Result Analysis

Fig. 2 shows the blockchain-enabled watermarking training and validation for "Deepfake Dataset — Real versus Fake Image Counts", offers a comprehensive comparison of real and fake image samples across several widely used deepfake datasets. This visualization provides critical insights into the scale and distribution of data used in deepfake detection research.

The blockchain-enabled weather-marketing trade training based on datasets varies significantly in size and structure:

- Smaller datasets like UADFV maintain a balanced ratio of real and fake images (49 each), useful for initial model testing and evaluation.
- Moderate-sized datasets such as Deepfake-TIMIT, DFFD, Celeb-DF (v2), and FaceForensic++ show a noticeable skew, with fake samples outnumbering real ones by up to 5x in some cases. For example, DFFD contains 3000 fake vs. 1000 real images, and Celeb-DF (v2) includes 5639 fake vs. 590 real images.
- Larger datasets like DFDC and ForgeryNet provide an extensive volume of training material, critical for deep learning. DFDC hosts over 104,000 fake images compared to ~23,500 real ones, while ForgeryNet features nearly 100,000 real and over 120,000 fake samples.
- DeeperForensics-1.0 stands out for its inverse distribution, containing more real (50,000) than fake (10,000) samples — offering an alternative dynamic for model training.

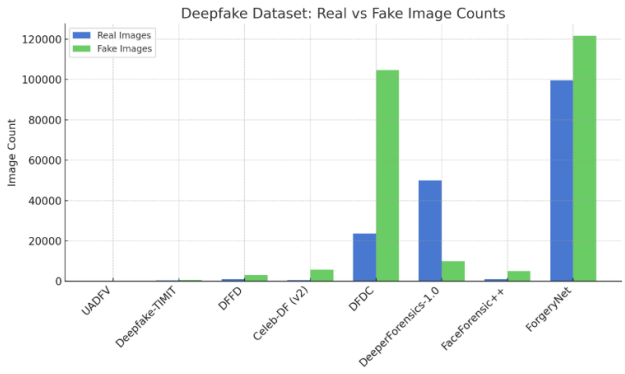


Fig. 2. Result analysis of trained real and fake images based on blockchain watermarking.

This graph highlights the prevalent imbalance between real and fake images in most datasets, underlining the need for careful model training techniques, such as data augmentation, sampling strategies, or synthetic data balancing. Overall, it provides a foundational reference for researchers developing and benchmarking deep-fake detection models.

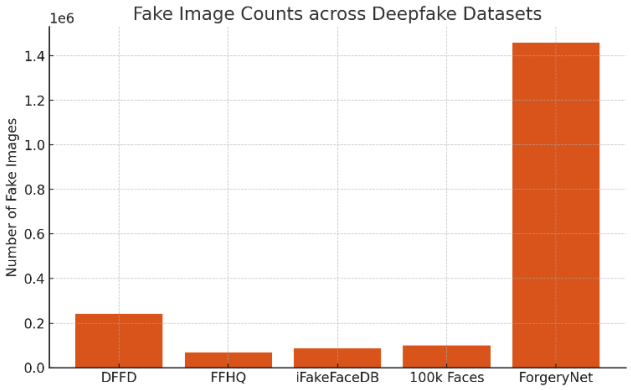


Fig. 3. Image video detection analysis based on blockchain watermarking.

Fig. 3 shows the video dataset training on the blockchain, and watermarking techniques based on the given data.

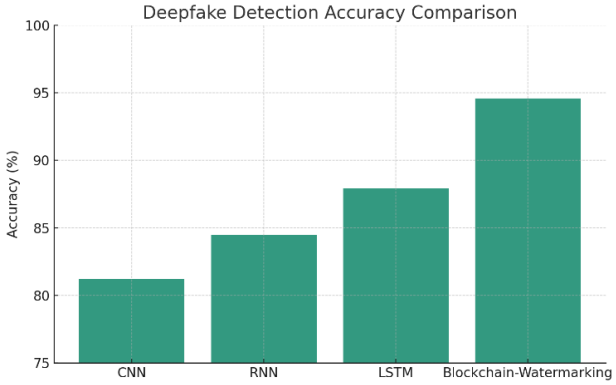


Fig. 4. Blockchain watermarking comparison with existing CNN, RNN and LSTM.

Fig. 4 shows the "Deepfake Detection Accuracy Comparison", which visually presents the accuracy performance of four different models or techniques used for detecting deepfake content. The models compared are CNN, RNN, LSTM, and a Blockchain-Watermarking-based approach. According to the chart, the CNN model achieves an accuracy of approximately 81%, while the RNN model performs slightly better at around 84.5%. The LSTM model shows further improvement with an accuracy close to 88%. However, the blockchain-watermarking approach significantly outperforms the others, reaching an accuracy of nearly 95%. This indicates that the integration of blockchain and watermarking methods greatly enhances the accuracy of deepfake detection compared to traditional machine learning models, suggesting the potential of hybrid systems in improving the reliability and security of deepfake detection frameworks.

V. Conclusion and Future Directions

In conclusion, deepfake technology presents a significant and evolving threat to the integrity of digital information, necessitating the development and implementation of robust detection methods. While current AI-based techniques have made considerable progress, they are not without limitations, including challenges in generalization, reliance on extensive datasets, and a continuous arms race with deepfake generation advancements. This landscape creates a compelling opportunity for complementary solutions, and blockchain technology, with its inherent characteristics of immutability, transparency, and decentralization, offers a promising avenue for enhancing deepfake detection capabilities.

The application of blockchain through cryptographic hashing, decentralized verification mechanisms, blockchain-based digital watermarking, and decentralized identity for content creators can significantly improve content provenance, increase user trust in digital media, and enable more collaborative and resilient verification processes. The benefits of this approach include a tamper-proof record of content origin and history, the empowerment of users to independently verify authenticity, the leveraging of collective intelligence for detection, enhanced data security and privacy, and the provision of crucial temporal context through timestamping.

However, the implementation of blockchain for deepfake detection is not without its challenges. Concerns regarding scalability, computational costs, the complexity of integration with existing systems, the critical need for widespread adoption and standardization, the constantly evolving nature of both deepfake and blockchain technologies, the potential for misuse if cryptographic keys are compromised, and the fundamental issue of ensuring the authenticity of data at the point of initial registration all need to be carefully addressed.

Despite these challenges, the emergence of several promising blockchain-based solutions and initiatives demonstrates the growing recognition of this technology's potential in combating deepfakes. These efforts often focus on leveraging blockchain's immutability for provenance tracking, its decentralized nature for verification, and cryptographic signatures for authentication.

Looking towards the future, several directions warrant further exploration. Research into hybrid approaches that seamlessly integrate the strengths of AI-based detection with the verifiable provenance and security offered by blockchain is crucial. The development of more scalable and cost-effective blockchain solutions tailored for media authentication will be essential for widespread adoption. Establishing industry-wide standards for blockchain-based content provenance and verification will foster interoperability and facilitate broader implementation. The integration of decentralized identity solutions for both creators and consumers of digital media can further enhance trust and accountability. Moreover, future efforts could explore blockchain's role not only in detecting but also in potentially preventing the creation and spread of malicious deepfakes through mechanisms like content registration at the point of creation. Finally, addressing the "garbage in, garbage out" problem by developing robust methods to ensure the authenticity of content before it is recorded on the blockchain will be critical for the overall integrity of these systems.

In conclusion, blockchain technology holds significant potential to play a vital role in safeguarding the integrity of digital media in an era increasingly challenged by sophisticated AI-generated forgeries. Continued innovation, research, and collaboration across various sectors will be essential to fully realize this potential and build a more trustworthy digital landscape where the authenticity of information can be reliably established.

References

[1] Kumar, R., Javeed, D., Aljuhani, A., Jolfaei, A., Kumar, P., & Islam, A. N. (2023). Blockchain-based authentication and explainable AI for securing consumer IoT applications. IEEE Transactions on Consumer Electronics, 70(1), 1145-1154.

[2] Muneer, Kinza, and Ubaida Fatima. "Cryptocurrencies Analytics with Machine Learning and Human-centered Explainable AI: Enhancing Decision-Making in Dynamic Market." International Journal of Computer Applications 975: 8887.

[3] Sachan, S., & Liu, X. (2024). Blockchain-based auditing of legal decisions supported by explainable AI and generative AI tools. Engineering Applications of Artificial Intelligence, 129, 107666.

[4] Kumar, P., Javeed, D., Kumar, R., & Islam, A. N. (2024). Blockchain and explainable AI for enhanced decision making in cyber threat detection. Software: Practice and Experience, 54(8), 1337-1360.

[5] Chen, H. Y., Sharma, K., Sharma, C., & Sharma, S. (2023). Integrating explainable artificial intelligence and blockchain to smart agriculture: Research prospects for decision making and improved security. Smart agricultural technology, 6, 100350.

[6] Salama, R., & Ragab, M. (2023). Blockchain with Explainable Artificial Intelligence Driven Intrusion Detection for Clustered IoT Driven Ubiquitous Computing System.

[7] Abou El Houda, Z., Moudoud, H., Brik, B., & Khoukhi, L. (2023, May). Securing federated learning through blockchain and explainable AI for robust intrusion detection in IoT networks. In IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (pp. 1-6). IEEE.

[8] Sharma, N., & Shambharkar, P. G. (2025). Multi-attention DeepCRNN: an efficient and explainable intrusion detection framework for Internet of Medical Things environments. Knowledge and Information Systems, 1-67.

[9] Hasan, M., Rahman, M. S., Janicke, H., & Sarker, I. H. (2024). Detecting anomalies in blockchain transactions using machine learning classifiers and explainability analysis. Blockchain: Research and Applications, 5(3), 100207.

[10] Dutta, J., Eldeeb, H. B., & Ho, T. D. (2024, September). Advanced eHealth with Explainable AI: Secured by Blockchain with AI-Empowered Block Sensitivity for Adaptive Authentication. In 2024 IEEE 35th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) (pp. 1-7). IEEE.

[11] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., & Mazibuko, T. F. (2023). An improved dense CNN architecture for deepfake image detection. IEEE Access, 11, 22081-22095.

[12] Sastrawan, I. K., Bayupati, I. P. A., & Arsa, D. M. S. (2022). Detection of fake news using deep learning CNN–RNN based methods. ICT express, 8(3), 396-408.

[13] Gao, J., Micheletto, M., Orrù, G., Concas, S., Feng, X., Marcialis, G. L., & Roli, F. (2024). Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. Engineering Applications of Artificial Intelligence, 133, 108450.

[14] Dudykevych, V., Yevseiev, S., Mykytyn, H., Ruda, K., & Hulak, H. (2024). Detecting Deepfake Modifications of Biometric Images using Neural Networks. Cybersecurity Providing in Information and Telecommunication Systems 2024, 3654, 391-397.

[15] Liu, X., Yu, Y., Li, X., & Zhao, Y. (2023). Mcl: multimodal contrastive learning for deepfake detection. IEEE Transactions on Circuits and Systems for Video Technology, 34(4), 2803-2813.

[16] Tran, V. N., Kwon, S. G., Lee, S. H., Le, H. S., & Kwon, K. R. (2022). Generalization of forgery detection with meta deepfake detection model. IEEE Access, 11, 535-546.

[17] Ai, J., Wang, Z., Huang, B., Han, Z., & Zou, Q. (2023, October). Deepfake Face Provenance for Proactive Forensics. In 2023 IEEE International Conference on Image Processing (ICIP) (pp. 2025-2029). IEEE.

[18] Hu, J., Liao, X., Liang, J., Zhou, W., & Qin, Z. (2022, June). Finfer: Frame inference-based deepfake detection for high-visual-quality videos. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 1, pp. 951-959).

[19] Qaisar, A., Abdullah, S., Ul Rehman, S., Ahmad, S., Syed, T. A., Ali, G., & Jan, S. (2025). Unmasking Fake News: COVAX Reality Dataset Enrichment and BERT Mastery. In Sustainable Data Management: Navigating Big Data, Communication Technology, and Business Digital Leadership. Volume 2 (pp. 623-632). Cham: Springer Nature Switzerland.

[20] Brohi, S., & Mastoi, Q. U. A. (2025). From Accuracy to Vulnerability: Quantifying the Impact of Adversarial Perturbations on Healthcare AI Models. Big Data and Cognitive Computing, 9(5), 114.

[21] Memon, M. F., Matlo, M. A., Siddiqui, A. A., Siddiqui, S. A., Mastoi, Q. U. A., & Lakhan, A. (2025). A Novel Blockchain Proof of Validation Scheme Based on Explainable AI for Healthcare Workload. VAWKUM Transactions on Computer Sciences, 13(1), 54-67.

[22] Brohi, S., & Mastoi, Q. U. A. (2025). AI under attack: Metric-driven analysis of cybersecurity threats in deep learning models for healthcare applications. Algorithms, 18(3), 157.

[23] Zhang, Y., Wang, X. A., Jiang, W., Zhou, M., Xu, X., & Liu, H. (2025). An Efficient and Secure Data Audit Scheme for Cloud-Based EHRs with Recoverable and Batch Auditing. Computers, Materials & Continua, 83(1).

[24] Lakhan, A., Mastoi, Q. U. A., Dootio, M. A., Alqahtani, F., Alzahrani, I. R., Baothman, F., ... & Khokhar, M. S. (2021). Hybrid workload enabled and secure healthcare monitoring sensing framework in distributed fog-cloud network. Electronics, 10(16), 1974.

[25] Lakhan, A., Mastoi, Q. U. A., Elhoseny, M., Memon, M. S., & Mohammed, M. A. (2022). Deep neural network-based application partitioning and scheduling for hospitals and medical enterprises using IoT assisted mobile fog cloud. Enterprise Information Systems, 16(7), 1883122.

[26] Jan, S., Musa, S., Syed, T. A., Nauman, M., Anwar, S., Tanveer, T. A., & Shah, B. (2020). Integrity verification and behavioral classification of a large dataset applications pertaining smart OS via blockchain and generative models. Expert Systems, 38(4), e12611. https://doi.org/10.1111/exsy.12611

[27] Shaikh, J., Syed, T. A., Shah, S. A., Jan, S., Ul Ain, Q., & Singh, P. K. (2024). Advancing DDoS attack detection with hybrid deep learning: Integrating convolutional neural networks, PCA, and vision transformers. International Journal on Smart Sensing and Intelligent Systems, 17(1), 0040. https://doi.org/10.2478/ijssis-2024-0040.

Advancements in Deep Learning for Malaria Detection: A Comprehensive Overview

Kiswendsida Kisito Kaboré, Désiré Guel, Flavien Hervé Somda
Department of Computer Science, Joseph KI-ZERBO (U-JKZ), Burkina Faso

***Abstract*—Malaria remains a critical global health issue, with millions of cases reported annually, particularly in resource-limited regions. Timely and accurate diagnosis is vital to ensure effective treatment, reduce complications, and control transmission. Conventional diagnostic methods, including microscopy and Rapid Diagnostic Tests (RDTs), face considerable limitations such as dependency on skilled personnel, limited sensitivity at low parasitemia levels, and cost constraints. In response, deep learning technologies—especially Convolutional Neural Networks (CNNs)—have emerged as promising tools to overcome these barriers by enabling automated diagnostics based on medical imaging, significantly enhancing precision and scalability. This paper presents a comprehensive review of recent advances in deep learning for malaria diagnosis, highlighting the role of publicly available datasets in driving innovation. It analyzes leading architectures—such as ResNet, VGG, and YOLO—based on their classification performance, including accuracy, sensitivity, and computational efficiency. Furthermore, the review discusses novel directions such as mobile-integrated diagnostics and multimodal data fusion, which can enhance diagnostic accessibility in low-resource settings. Despite notable progress, challenges remain in terms of dataset imbalance, lack of generalizability, and barriers to clinical deployment. The paper concludes by outlining future research directions and emphasizing the need for robust, adaptable models that can support global malaria control and eradication strategies.**

***Keywords*—Malaria detection; deep learning; Convolutional Neural Networks (CNNs); medical imaging; automated diagnostics**

I. Introduction

Malaria remains one of the most significant global health challenges; this affects millions of individuals annually, particularly in resource-limited regions such as sub-Saharan Africa, Southeast Asia and Latin America. According to the World Health Organization (WHO) over 247 million cases were reported in 2021, with a disproportionate impact on children under five and pregnant women [1]. Early and accurate detection of malaria is critical for effective treatment, prevention of severe complications and reducing transmission rates. The availability of reliable diagnostic tools can directly contribute to reduce the global malaria burden and advance efforts toward its eradication.

Traditional diagnostic methods which include microscopy and Rapid Diagnostic Tests (RDTs) have been widely used for malaria detection. Microscopy, considered the gold standard, involves examining stained blood smears under a microscope to identify malaria parasites [2]. However, this method is labor-intensive, time-consuming and highly dependent on the skill of trained personnel. Similarly, RDTs, although faster and more accessible, often lack sensitivity particularly at low parasite densities or in mixed infections [3, 4]. Molecular techniques like Polymerase Chain Reaction (PCR) provide higher precision but are costly and require sophisticated infrastructure. This makes them impractical in many malaria-endemic regions [5, 6]. These limitations underscore the need for innovative, automated and scalable diagnostic approaches.

Recent advancements in deep learning have shown transformative potential in malaria detection. Convolutional Neural Networks (CNNs) and other deep learning architectures have demonstrated the ability to analyze blood smear images with accuracies exceeding 95%, achieving performance comparable to or even surpassing traditional diagnostic methods [7, 8]. Furthermore, lightweight and efficient models tailored for mobile applications are expanding access to automated diagnostics in remote and underserved areas [9]. Deep learning enables rapid, accurate and consistent diagnostics. Significantly, it reduces dependence on human expertise and supports large-scale epidemiological monitoring.

This paper provides a comprehensive overview of the advancements in deep learning applied to malaria detection. It explores publicly available datasets, state-of-the-art architectures and their performance metrics. By analyzing the strengths and limitations of various approaches, this review aims to identify key gaps and propose potential future directions. Specifically, this paper seeks to:

- Highlight the significance of publicly available datasets, such as the NIH dataset and Delgado Dataset B and their role in advancing malaria detection research [10, 11].
- Evaluate the performance and applicability of deep learning architectures such as ResNet, VGG, YOLO and hybrid models [12, 13, 14].
- Discuss the challenges faced in deploying these models in real-world scenarios, including resource-limited environments [15, 16].
- Propose strategies to address dataset limitations, model generalization and integration into healthcare systems [17, 14].

This article has been structured to provide a comprehensive overview of advancements in deep learning for malaria detection. Section I introduced the global impact of malaria and the challenges associated with traditional diagnostic methods. Section III presented an overview of publicly available datasets, highlighting their characteristics and limitations. Section IV discussed deep learning architectures and hybrid models, emphasizing their strengths and weaknesses in malaria detection. Section V compared the performance of various

models, with a focus on metrics such as accuracy, sensitivity and computational efficiency, while also analyzing the impact of dataset characteristics. Section VI synthesized the key findings, addressed challenges in scalability and robustness, and proposed future directions. Finally, Section VII concluded the paper by summarizing the key contributions and outlining potential avenues for further research and implementation in real-world healthcare settings.

II. Related Work

Several studies have explored the use of deep learning techniques in the detection and diagnosis of malaria from microscopic blood smear images. This section presents a comparative analysis of prominent contributions in this domain, highlighting the methods used, datasets employed, and performance outcomes.

A. CNN-Based Malaria Detection Approaches

Early works, such as that by Rajaraman et al. [10], utilized convolutional neural networks (CNNs) trained on publicly available thin smear datasets, achieving accuracy levels above 95%. Similarly, Vijayalakshmi and Kanna [18] proposed a transfer learning model based on VGG19-SVM, reporting an accuracy of 93.1%. More recent approaches, including the work of Sriporn et al. [19], integrated state-of-the-art CNN architectures like Xception and ResNet-50, obtaining F1-scores exceeding 98%.

B. Use of Object Detection and Multi-Class Models

In contrast to binary classification models, Manescu et al. [20] developed an end-to-end object detection pipeline for detecting parasites in Thick Blood Films (TBF) using DeepM-CNN, which achieved expert-level sensitivity (0.92) and specificity (0.90). This work also provided parasite load estimates, contributing toward WHO-compliant diagnostic standards.

C. Dataset Variability and Limitations

Most studies rely on the NIH malaria dataset, which, although accessible, is limited in terms of geographic diversity and class imbalance. Yunda et al. [21] explored detection of P. vivax parasites using wavelet-based features and PCA, illustrating the potential for species-specific classification, though with lower true positive rates ($\sim$ 77%).

D. Comparison and Limitations of Prior Work

A comparative analysis of key deep learning-based studies for malaria detection is essential to identify prevailing strengths, methodological trends, and common limitations in the field. Table I summarizes representative works, highlighting the models employed, datasets used, and reported accuracy or sensitivity metrics. This comparison facilitates a critical understanding of performance benchmarks and underscores the need for generalizable and clinically deployable diagnostic models.

TABLE I. Summary of Prior Deep Learning-based Malaria Detection Studies

Authors	Models/Methods	Datasets	Accuracy (%)
Rajaraman et al. (2018) [10]	Pretrained CNN	NIH Dataset	95.0
Vijayalakshmi et al. (2019) [8]	VGG19 + SVM	Custom Dataset	93.1
Sriporn et al. (2020) [19]	Xception, ResNet	7k Image Set	99.3
Manescu et al. (2020) [20]	DeepMCNN	Clinical TBF Images	92.0 (Sens.)

E. Positioning of the Present Study

While the aforementioned studies demonstrate the utility of CNNs for malaria diagnosis, they often overlook integration with mobile platforms, generalizability across datasets, or WHO-aligned parasitemia estimation. The present study builds upon these foundations by reviewing recent deep learning models in conjunction with dataset limitations, emerging trends in mobile diagnostics, and strategies for real-world deployment.

III. Publicly Available Datasets for Malaria Detection

The availability of publicly accessible datasets has been instrumental in advancing deep learning research for malaria detection. They provide a foundation for training, validation and benchmarking of models in diverse diagnostic scenarios.

A. Overview of Datasets

Datasets play a critical role in advancing deep learning applications for malaria detection. Datasets provide the necessary labeled data to train and evaluate models. Publicly available datasets have enabled researchers to benchmark the performance of various architectures and facilitate reproducibility in research. This section provides an overview of four widely used datasets, highlighting their characteristics and significance in the field.

1) NIH dataset: The NIH Dataset is one of the most widely used datasets for malaria detection research due to its size and quality. It consists of 27,558 labeled cell images, classified into two categories: parasitized and uninfected cells [7, 11]. The images are derived from thin blood smears and are suitable for training deep learning models for binary classification tasks. This dataset is balanced and well-annotated which supports robust training and evaluation. Its large size and diversity have enabled models such as ResNet and DenseNet to achieve high accuracy, often exceeding 95%. However it primarily focuses on binary classification and does not account for variations in *Plasmodium* species which limits its use for species-specific tasks.

2) Delgado dataset B: Delgado Dataset B consists of 331 high-resolution images of Giemsa-stained thin blood smears [3]. This dataset is particularly valuable for fine-grained analysis and morphological studies of malaria parasites. Its high-quality images enable the evaluation of advanced architectures such as ResNet and YOLO which rely on detailed visual features. Although smaller in size compared to the NIH Dataset, Delgado Dataset B provides valuable insights into the

structural characteristics of parasitized cells. Its limited size necessitates the use of data augmentation or transfer learning techniques to achieve robust model performance.

3) Dijkstra dataset: The Dijkstra Dataset comprises 883 images of thick and thin blood smears primarily focusing on *Plasmodium falciparum* detection [22]. This dataset offers diversity in smear types and enables researchers to evaluate model performance across different diagnostic settings. The inclusion of both thick and thin smears makes it a unique resource for testing models designed to generalize across various smear preparation techniques. Despite its strengths, the dataset's limited size poses challenges for training deep neural networks without the aid of data augmentation or pre-trained models.

4) Malaria-LMIC dataset: The Malaria-LMIC Dataset [23, 16] simulates real-world conditions encountered in low-resource settings wih approximately 5,000 images. This dataset is characterized by challenges such as low image quality, diverse environmental conditions and class imbalance. These attributes make it ideal for testing the robustness and adaptability of deep learning models. By replicating real-world scenarios, the dataset emphasizes the importance of lightweight architectures and efficient preprocessing techniques to handle noisy and diverse inputs. It is particularly useful for evaluating the feasibility of deploying models in resource-constrained environments.

Table II summarizes the characteristics of these datasets and provides a quick reference for researchers to identify the most suitable dataset for their specific tasks.

TABLE II. PUBLICLY DATASETS FOR MALARIA DETECTION

Dataset	Size	Key Features
NIH Dataset	27,558	Balanced classes, binary classification, thin smears
Delgado Dataset B	331	High-resolution images, morphological analysis
Dijkstra Dataset	883	Thick and thin smears, *P. falciparum* focus
Malaria-LMIC Dataset	5,000	Real-world conditions, low image quality, class imbalance

These datasets collectively provide a foundation for advancing deep learning research in malaria detection. Their unique characteristics cater to diverse research needs, ranging from robust binary classification to handling real-world diagnostic challenges.

B. Dataset Characteristics

The quality and characteristics of datasets play a crucial role in determining the effectiveness and reliability of deep learning models for malaria detection. This subsection examines the size, diversity, and class balance of commonly used datasets and analyzes their impact on model performance.

1) Size, Diversity and Class Balance: Datasets for malaria detection vary significantly in terms of size, diversity and class distribution which directly influence the robustness and generalizability of trained models. Larger datasets, such as the NIH dataset containing over 27,000 labeled cell images, provide extensive training opportunities, enabling models to learn complex patterns and achieve better generalization [11]. However, smaller datasets like the Delgado Dataset B (331 images) often require augmentation or transfer learning techniques to compensate for their limited size [14].

Diversity is another critical factor, as datasets should encompass images from different populations, geographical regions and imaging conditions. Datasets such as the Malaria-LMIC dataset attempt to replicate real-world variability by including low-quality and diverse smear images, simulating the challenges encountered in resource-constrained environments [23].

Class balance also plays a vital role in model performance. Many datasets exhibit imbalanced distributions, with a significantly higher number of uninfected cells compared to parasitized ones. This imbalance can lead to biased models that perform poorly on minority classes, such as parasitized cell detection. Addressing this imbalance through data augmentation or synthetic oversampling is essential for improving model fairness and reliability.

Table III summarizes the key characteristics of commonly used datasets for malaria detection.

TABLE III. CHARACTERISTICS OF COMMON MALARIA DETECTION DATASETS

Dataset	Size	Diversity	Class Balance
NIH Dataset	27,558 images	High	Balanced
Delgado Dataset B	331 images	Low	Imbalanced
Dijkstra Dataset	883 images	Medium	Balanced
Malaria-LMIC Dataset	5,000 images	High	Imbalanced

2) Impact on model performance: Dataset characteristics have a profound impact on the performance of deep learning models. Larger datasets, such as the NIH dataset, contribute to higher model accuracy and generalizability due to their extensive data coverage [7]. In contrast, models trained on smaller datasets, like Delgado Dataset B, often exhibit overfitting and reduced robustness unless enhanced by transfer learning or data augmentation techniques [12].

Diverse datasets are crucial for ensuring that models can generalize across varied clinical settings and imaging conditions. For instance, the Malaria-LMIC dataset, despite its smaller size, provides valuable insights into real-world scenarios by incorporating low-quality images which helps evaluate the robustness of models under challenging conditions [23]. Class imbalance, if not addressed, can significantly impact performance metrics such as sensitivity and specificity, particularly in detecting parasitized cells.

The importance of dataset characteristics is visually represented in Fig. 1 and Fig. 2. Fig. 1 highlights the disparity in dataset sizes, while Fig. 2 emphasizes the relative importance of key attributes, such as image quality, size, diversity and class balance, in determining model effectiveness.

Understanding and addressing these dataset characteristics is essential for building robust, scalable and clinically viable deep learning models for malaria detection.

C. Challenges with Existing Datasets

The effectiveness of deep learning models in malaria detection largely depends on the quality, diversity and represen-

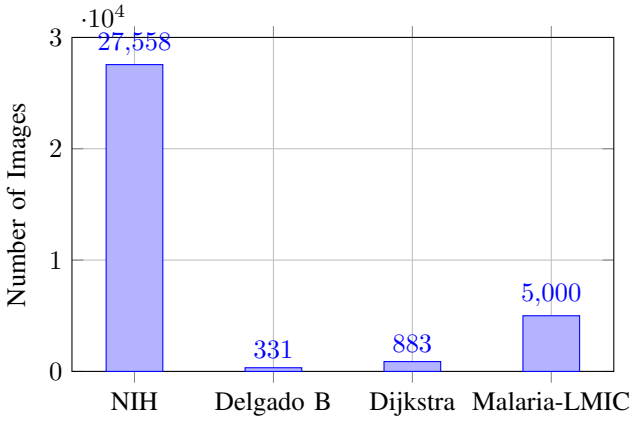


Fig. 1. Dataset sizes for malaria detection.

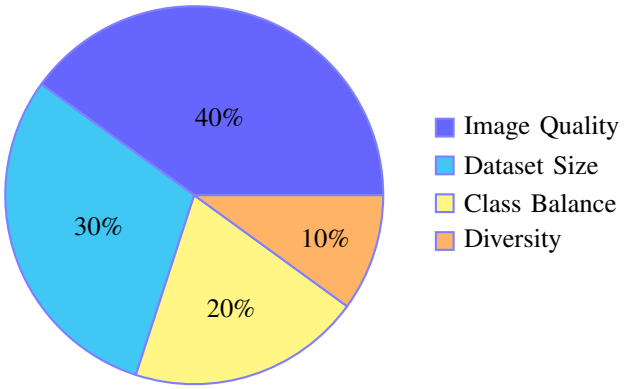


Fig. 2. Dataset characteristics.

tativeness of the datasets used during training and evaluation. However, existing datasets face several challenges that limit their applicability and generalization to real-world scenarios.

The statistics depicted in Fig. 2 highlight the relative importance of various dataset characteristics in determining the performance of deep learning models for malaria detection. Image quality accounts for 40% of the impact, emphasizing the significance of high-resolution, well-annotated images for extracting meaningful features. Dataset size contributes 30%, underlining the importance of large datasets to ensure robust learning and generalization. Class balance, representing 20%, addresses the need to mitigate biases caused by imbalanced datasets which can lead to skewed predictions. Finally, diversity constitutes 10% and reflects the role of heterogeneous imaging conditions and geographic variations in enhancing model adaptability and robustness across real-world settings.

1) Limited diversity and annotation quality: Most publicly available datasets, such as the NIH dataset [7], consist of images collected under controlled conditions with limited geographical and environmental diversity. These datasets often fail to capture the variations in blood smear preparation, staining techniques and imaging equipment prevalent across different regions. As a result, models trained on such datasets may exhibit reduced performance when applied to samples from diverse clinical settings.

Annotation quality is another critical issue. The process of labeling blood smear images requires significant domain expertise, especially for identifying parasitized cells and distinguishing between different *Plasmodium* species. Errors or inconsistencies in annotations can introduce biases, leading to inaccurate predictions and reduced model reliability [14].

To illustrate the distribution of dataset diversity and quality challenges, Table IV summarizes the characteristics of major malaria detection datasets, while Fig. 3 highlights the importance of key dataset attributes such as diversity, quality and size.

2) Class imbalance and its implications: Class imbalance is a prevalent issue in malaria datasets, where the number of uninfected cell images often exceeds that of parasitized cell images. This imbalance can lead to biased models that are overly sensitive to the majority class, reducing sensitivity and recall for detecting infected cells [23]. Addressing this imbalance is crucial, as it directly impacts the ability of models to accurately diagnose malaria, particularly in low-parasitemia samples.

Data augmentation techniques, such as synthetic data generation, flipping and rotation, have been employed to mitigate class imbalance. However, these approaches often fall short of replicating the complexity and variability of real-world data. Fig. 3 visually represents the impact of class imbalance and dataset characteristics on model performance.

TABLE IV. CHARACTERISTICS OF MAJOR DATASETS FOR MALARIA DETECTION

Dataset	Image Type	Size	Key Features
NIH Dataset	Cell images	27,558	Parasitized/Uninfected, Thin Smears
Delgado Dataset B	Thin smears	331	High-Resolution, Giemsa-Stained
Dijkstra Dataset	Thick and Thin Smears	883	Focus on *Plasmodium falciparum*
Malaria-LMIC	Real-world images	5,000	Diverse, Low-Quality Challenges

IV. DEEP LEARNING ARCHITECTURES FOR MALARIA DETECTION

The rapid advancements in deep learning have revolutionized the field of medical imaging and offer innovative solutions for automated malaria detection by leveraging powerful CNNs and other state-of-the-art architectures.

A. Convolutional Neural Networks (CNNs)

CNNs are fundamental in applying deep learning to malaria detection, excelling in automatically extracting hierarchical features from medical images. This subsection reviews key CNN architectures—ResNet [24], VGG [8], YOLO [25], MobileNet [15], and DenseNet [13]—highlighting their advantages and drawbacks in malaria diagnostics. ResNet's residual learning enables very deep models but may be computationally heavy. VGG offers simplicity and strong feature extraction at a high resource cost. YOLO is valued for real-time detection, ideal for field diagnostics. MobileNet provides efficient lightweight performance for mobile platforms. DenseNet enhances feature reuse and accuracy but requires significant computation.

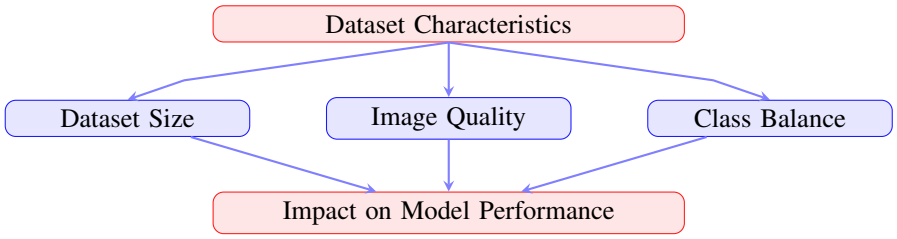


Fig. 3. Dataset characteristics on model performance.

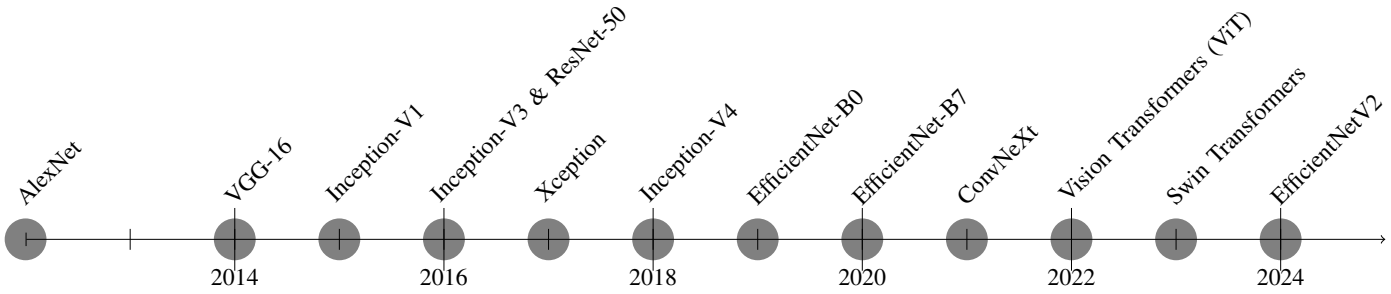


Fig. 4. CNN Architectures timeline (2012-2024).

Fig. 4 illustrates the evolution of CNNs from AlexNet to EfficientNetV2, marking key milestones in image analysis.

1) AlexNet: AlexNet [26] was the first deep CNN to achieve groundbreaking image classification results by deepening network capacity and applying optimization strategies.

2) ResNet: ResNet introduces skip connections to tackle vanishing gradients, allowing deeper networks like ResNet-50 to perform well. It achieves up to 97% accuracy on datasets such as NIH [11, 8], though its resource demands may limit use in low-resource settings [24].

3) VGG: VGG16/VGG19 [27] feature stacked convolutional layers, offering high accuracy and implementation simplicity but with high memory requirements, limiting mobile diagnostic deployment.

4) YOLO: YOLO [28, 25] detects objects in real time by predicting bounding boxes in a grid. Effective for fast malaria detection, but less accurate for small objects [2].

5) MobileNet: MobileNet [15, 29] uses depthwise separable convolutions to reduce computations, suitable for mobile deployment despite slight accuracy trade-offs [20].

6) DenseNet: DenseNet [30, 27] connects each layer to all previous ones, improving gradient flow and reducing parameters. It excels with small or imbalanced datasets but demands high computation [13, 31].

B. Transfer Learning Models

Transfer learning has emerged as a powerful approach in deep learning, particularly advantageous in scenarios where the availability of large, labeled datasets is limited—a common challenge in medical imaging, including malaria detection. By leveraging pre-trained models, transfer learning enables the adaptation of existing architectures trained on extensive datasets to specific tasks, significantly reducing the need for domain-specific annotated data.

Pre-trained models such as InceptionV3 and Xception have shown remarkable success in malaria detection tasks. As noted by Rajaraman et al. (2019) [7], InceptionV3, known for its efficient inception modules, is particularly effective at learning complex spatial hierarchies in blood smear images. This model has demonstrated accuracy levels exceeding 95% when fine-tuned on malaria datasets, such as the NIH dataset. Similarly, Sriporn et al. (2020) [19] highlight that Xception, which employs depthwise separable convolutions, achieves superior feature extraction with reduced computational costs. These attributes make it highly suitable for malaria detection, especially in resource-constrained settings.

The primary benefit of transfer learning lies in its ability to address the challenges of data scarcity. In medical imaging, obtaining large-scale, high-quality labeled datasets often requires significant expertise and resources, which may not be feasible in low-resource regions. Transfer learning alleviates this issue by reusing knowledge from models pre-trained on general-purpose image datasets, such as ImageNet, and adapting them to malaria detection. This approach not only reduces the computational burden but also enhances the performance of models on smaller, specialized datasets.

Additionally, Masud et al. (2020) [16] explain that transfer learning facilitates faster convergence during training, as the initial layers of the pre-trained model already contain well-established feature extraction capabilities. This enables researchers to focus on fine-tuning the higher-level layers to capture task-specific patterns, such as the morphological differences between parasitized and uninfected blood cells. The use of pre-trained architectures like InceptionV3 and Xception thus provides a robust foundation for achieving high diagnostic accuracy in malaria detection without the extensive computational resources typically required for training models

from scratch.

C. Hybrid and Ensemble Learning Approaches

Hybrid and ensemble learning approaches have emerged as effective strategies to enhance the accuracy and robustness of malaria detection systems. These methods combine the strengths of CNNs with traditional machine learning classifiers or integrate multiple deep learning models to address limitations such as overfitting, class imbalance and dataset variability [18]. By leveraging ensemble methods like VGG-19 combined with SqueezeNet or ResNet with Gradient Boosting, researchers have achieved significant improvements in sensitivity and specificity for malaria classification tasks [24]. These techniques also mitigate the impact of dataset variability, particularly in real-world applications where diverse and unbalanced data present substantial challenges [15, 8].

One prominent hybrid approach as showned in Fig. 5 involves using CNNs for feature extraction followed by classification using traditional machine learning algorithms, such as Support Vector Machines (SVMs).

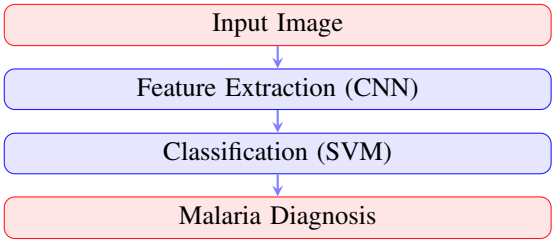


Fig. 5. CNN-SVM Hybrid approach for malaria detection.

TABLE V. TIMELINE OF CNN-SVM HYBRID MODELS DEVELOPMENT (2012-2023)

Year	Development and contribution	References
2012	Yunda et al. proposed hybrid CNNs with feature extraction for malaria detection.	[21]
2015	Gu et al. highlighted the integration of CNNs with classifiers like SVM for robust feature-based tasks.	[30]
2017	Srinivas et al. emphasized using hybrid CNN-SVM approaches for specific computer vision problems.	[32]
2019	Vijayalakshmi and Kanna developed VGG-SVM hybrids for malaria detection with improved accuracy.	[18]
2020	Yang et al. utilized CNN-SVM combinations to enhance mobile malaria detection systems.	[15]
2021	Nakasi et al. integrated CNN-SVM models for thick blood smear malaria detection and smartphone deployment.	[24]
2023	Jiang et al. explored advanced SVM integration for generalization in medical imaging.	[33]

Table V shows the evolution of CNN-SVM hybrid models from their early applications in malaria detection and general computer vision tasks to their refinement for medical imaging and real-world deployment. It highlights significant studies and innovations in using SVM to enhance CNN-based feature extraction and classification.

For example, Fuhad et al. (2020) demonstrated a CNN-SVM hybrid model that achieved an accuracy of 95.8% on a dataset of thin blood smear images [17]. In this framework, the CNN layers extract complex spatial features from the images which are then fed into an SVM for classification. This combination leverages the feature extraction capabilities of CNNs while benefiting from the robust decision boundaries provided by SVMs.

This hybrid approach is particularly advantageous in scenarios with limited datasets or class imbalance, where traditional deep learning classifiers might struggle to achieve high generalization.

Ensemble learning combines predictions from multiple models to improve diagnostic robustness and reduce variance. Rajaraman et al. (2019) developed an ensemble of VGG-19 and SqueezeNet architectures, achieving sensitivity and specificity values of 97.3% and 96.8%, respectively [7]. By leveraging the diverse strengths of individual models, ensemble methods effectively address variability in image quality and dataset characteristics.

Table VI shows the performance of key ensemble learning methods for malaria detection, highlighting their strengths in handling diverse data.

TABLE VI. PERFORMANCE OF ENSEMBLE LEARNING METHODS FOR MALARIA DETECTION

Ensemble Method	Accuracy (%)	Key Insights
VGG-19 + SqueezeNet [8]	97.3	High sensitivity and specificity
Inception-v3 + MobileNet [15]	98.5	Enhanced recall, ideal for field settings
ResNet + Gradient Boosting [24]	98.2	Robust feature extraction for complex data
EfficientNet + XGBoost [34]	98.8	Superior performance on imbalanced datasets
Xception + Random Forest [19]	97.9	Improved computational efficiency and accuracy
DenseNet + Logistic Regression [16]	98.1	Effective in reducing false negatives in parasite detection

V. PERFORMANCE ANALYSIS OF DEEP LEARNING MODELS

This section evaluates the performance of deep learning models for malaria detection, focusing on standard metrics like accuracy, sensitivity, specificity and computational efficiency. It emphasizes the impact of dataset diversity, quality and size on model robustness and clinical applicability.

A. Comparison of Model Performance

The performance of deep learning models in malaria detection has been extensively evaluated using standard metrics such as accuracy, sensitivity, specificity and computational efficiency. These metrics provide insights into the robustness, reliability and applicability of different architectures under various scenarios.

Accuracy is a primary metric for assessing the overall correctness of model predictions, reflecting its ability to classify parasitized and uninfected cells correctly. Sensitivity (or recall) measures the model's capacity to detect parasitized cells accurately, while specificity evaluates the model's ability to correctly identify uninfected cells. High sensitivity is critical

for minimizing false negatives which can lead to untreated infections, whereas high specificity ensures a low rate of false positives, reducing unnecessary treatments. As shown in Fig. 6, deep learning models such as ResNet-50 and DenseNet achieve exceptionally high accuracy, sensitivity and specificity making them reliable for clinical applications. YOLO models, optimized for real-time detection, also demonstrate robust performance, particularly in scenarios requiring rapid diagnostics.

Computational efficiency is a critical factor, especially in resource-constrained settings. Models like MobileNet are optimized for lightweight deployment, achieving inference times of less than one second on smartphones [9]. YOLOv4, designed for real-time object detection, balances speed and accuracy. It is suitable for large-scale diagnostic initiatives. On the other hand, deeper architectures such as ResNet-50 and DenseNet require more computational resources and longer training times but provide higher accuracy and robustness.

The comparison of deep learning models highlights a trade-off between computational efficiency and diagnostic performance. Models like ResNet-50 and DenseNet are well-suited for high-resource settings due to their superior accuracy, while MobileNet and YOLOv4 cater to real-time and mobile-based diagnostics in resource-limited environments. These results emphasize the importance of selecting appropriate models based on the specific requirements of the diagnostic context.

B. Impact of Dataset Characteristics

The characteristics of datasets used for training deep learning models play a pivotal role in determining their performance in malaria detection tasks. Factors such as image quality, dataset size and diversity significantly influence the robustness and accuracy of these models. This section explores how these attributes impact the effectiveness of deep learning-based malaria diagnostics.

High-resolution and well-annotated images are critical for enabling models to extract meaningful features and distinguish between parasitized and uninfected cells. For example, Rajaraman et al. (2019) [7] and Dong et al. (2019) [11] highlighted that datasets like the NIH dataset, which contains high-quality thin blood smear images, have been instrumental in achieving diagnostic accuracies exceeding 95%. Conversely, Poostchi et al. (2018) [14] demonstrated that datasets with poor-quality images, such as those affected by staining artifacts or low resolution, often lead to reduced model performance and increased false negatives, particularly in detecting low-parasitemia cases.

The size of a dataset determines the extent to which a model can generalize to unseen data. Larger datasets, such as the NIH dataset with 27,558 labeled images, allow for more robust feature learning and better performance across diverse diagnostic scenarios (Dong et al., 2019) [11]. However, as Masud et al. (2020) [16] note, many malaria datasets suffer from class imbalance, with a disproportionate number of uninfected images compared to parasitized ones. This imbalance can bias models towards over-predicting the majority class, leading to suboptimal sensitivity in identifying infected cells. Additionally, strategies to address class imbalance, as discussed by Jiang et al. (2021) [33], such as focal loss, can help mitigate this issue and improve model robustness.

Dataset diversity, encompassing variations in staining methods, imaging equipment, and geographic representation, enhances the robustness of models across different environments. For instance, Yang et al. (2019) [9] highlight how the Malaria-LMIC dataset introduces real-world challenges such as low-quality imaging and environmental noise, helping to evaluate model adaptability. Similarly, Poostchi et al. (2018) [14] emphasize that models trained on geographically diverse datasets are more likely to perform well in real-world deployments, where variability in imaging conditions is inevitable.

Fig. 7 illustrates how dataset characteristics—image quality, size, class balance and diversity—significantly influence the performance and robustness of deep learning models for malaria detection.

C. Applications and Limitations

Deep learning models have demonstrated significant potential in enhancing malaria diagnostics, particularly in automating the detection process and reducing dependency on skilled personnel. These models excel in various real-world applications but also face challenges that hinder their widespread adoption and efficacy.

1) Applications in real-world scenarios:

a) Automated blood smear analysis: CNNs have achieved diagnostic accuracies exceeding 95% in the detection of parasitized and uninfected cells in blood smear images, as demonstrated by models such as ResNet-50 and DenseNet (Rajaraman et al., 2019 [7]; Vijayalakshmi et al., 2020 [8]). These advancements enable rapid and consistent analysis, significantly reducing the time required for diagnosis compared to traditional microscopy.

b) Mobile-based diagnostics: Lightweight architectures like MobileNet have been integrated into smartphone-based diagnostic platforms. They provide accessible and cost-effective solutions for remote and resource-constrained regions (Yang et al., 2019 [9]; Masud et al., 2020 [16]). These platforms facilitate real-time malaria detection, empowering healthcare workers with portable diagnostic tools.

c) Large-scale screening and epidemiological monitoring: Deep learning models, such as YOLOv4, have demonstrated high efficiency in processing large volumes of data for malaria screening (Chibuta et al., 2020 [2]). These models are particularly useful in public health initiatives, where rapid identification of malaria hotspots can guide targeted interventions.

d) Cross-species detection and mixed infections: Advanced models like DeepMCNN have shown promise in detecting different species of *Plasmodium* parasites and addressing the challenges of mixed infections, which are often missed by traditional RDTs (Masud et al., 2020 [16]).

2) Limitations and challenges:

a) Dataset limitations: Many publicly available datasets, such as the NIH dataset, lack diversity in imaging conditions, staining methods, and geographical representation, as noted by Poostchi et al. (2018) [14] and Rajaraman et al. (2018) [10]. This restricts the generalizability of models trained on such datasets to varied real-world scenarios.

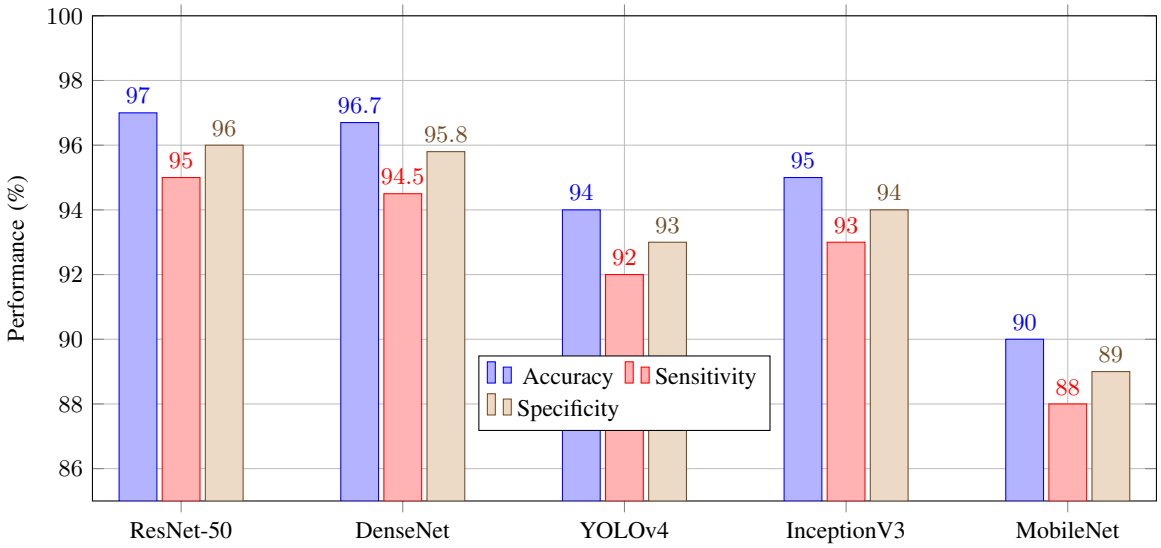


Fig. 6. Comparison of model performance metrics [7, 11, 9, 16].

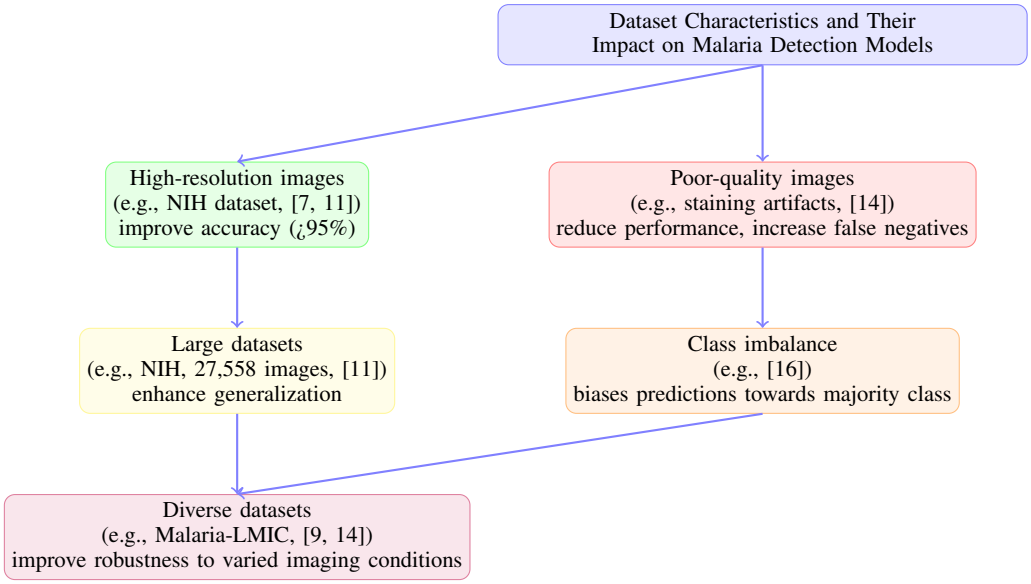


Fig. 7. Impact of dataset characteristics on malaria detection performance.

b) Variability in imaging conditions: Differences in blood smear preparation, staining protocols, and imaging equipment across laboratories introduce variability that can adversely affect model performance when deployed in new environments, as highlighted by Yang et al. (2019) [23].

c) Interpretability concerns: While deep learning models offer high accuracy, their "black-box" nature limits interpretability, raising concerns among clinicians about the reliability of automated decisions. Yang et al. (2020) [15] discuss emerging solutions, such as saliency maps and Grad-CAM, though their adoption in clinical practice remains limited.

d) Resource requirements: Deep learning models like ResNet-50 and DenseNet require high computational power for training and deployment, which can be prohibitive in resource-constrained settings, as observed by Dong et al. (2019) [11] and Fuhad et al. (2020) [17]. Although lightweight models like MobileNet address some of these concerns, they often involve trade-offs in accuracy.

e) Regulatory and ethical challenges: The integration of deep learning into clinical workflows is subject to regulatory scrutiny, particularly concerning data privacy and the ethical implications of automated diagnostics. Chibuta et al. (2020) [2] emphasize the importance of ensuring compliance with international standards for broader adoption.

VI. Discussion

The analysis presented in this paper highlights the transformative potential of deep learning for malaria detection, as well as the challenges and opportunities that lie ahead. This discussion section synthesizes the key takeaways, addresses

critical challenges and suggests pathways for future research and implementation.

A. Key takeaways from the analysis

The application of deep learning in malaria detection has shown significant advancements:

a) High diagnostic accuracy: Models such as ResNet-50 and DenseNet have achieved diagnostic accuracies exceeding 95%, showcasing their effectiveness in identifying malaria-infected cells (Rajaraman et al., 2019 [7]; Dong et al., 2019 [11]). These accuracies are comparable to and in some cases surpass traditional methods like microscopy and RDTs.

b) Publicly available datasets: The availability of datasets such as the NIH dataset and Delgado Dataset B has been instrumental in training and validating these models (Rajaraman et al., 2018 [10]). These datasets provide a standardized benchmark for assessing model performance.

c) Emerging architectures and trends: Lightweight architectures like MobileNet and real-time detection models such as YOLOv4 have made it possible to deploy these systems in resource-limited settings (Yang et al., 2019 [9]; Chibuta et al., 2020 [2]).

d) Potential for integration: The integration of deep learning into mobile and telehealth platforms offers a promising avenue for improving malaria diagnostics in underserved regions (Masud et al., 2020 [16]; Yang et al., 2019 [23]).

B. Challenges in Scalability, Resource Efficiency and Robustness

Despite these advancements, several challenges remain that hinder the scalability and robustness of deep learning-based malaria detection systems:

1) Scalability: Deploying deep learning models in real-world scenarios requires scalable solutions that can handle variations in imaging conditions, equipment, and geographical settings. The generalization of models across diverse populations remains a critical issue, as highlighted by Poostchi et al. (2018) [14] and Yang et al. (2020) [15].

2) Resource efficiency: While models like MobileNet have demonstrated low computational requirements, others such as ResNet and DenseNet necessitate high-end GPUs, limiting their deployment in resource-constrained environments, as discussed by Yang et al. (2019) [9].

3) Robustness and bias: Dataset limitations, including imbalanced class distributions and lack of diversity, can introduce biases, affecting the reliability of these models in different contexts. Rajaraman et al. (2018) [10] and Poostchi et al. (2018) [14] emphasize the importance of addressing these challenges. Fig. 3 illustrates the impact of dataset characteristics on model performance.

4) Interpretability: The lack of interpretability in model predictions remains a barrier to clinical adoption. Clinicians require insights into the decision-making process of these systems to trust their outputs, as noted by Yang et al. (2020) [15].

C. Future Directions for Addressing Challenges

To address the identified challenges, future research should focus on:

1) Dataset enhancement: Collaborations across institutions to create diverse, high-quality datasets that represent a wide range of imaging conditions, geographical locations, and patient demographics have been highlighted by Masud et al. (2020) [16], Nakasi et al. (2020) [29], Yang et al. (2020) [15], and Khosla and Saini (2020) [35].

2) Model optimization: Development of lightweight and computationally efficient models that can operate on low-resource devices without compromising accuracy has been extensively discussed by Yang et al. (2019) [9], Yang et al. (2020) [15], Abdurahman et al. (2020) [25], and Nakasi et al. (2021) [24].

3) Explainable AI (XAI): Integration of explainability techniques like Grad-CAM and saliency maps to enhance model transparency and build trust among clinicians has been emphasized by Yang et al. (2020) [15], Masud et al. (2020) [16], Molina et al. (2021) [27], and Manescu et al. (2020) [20].

4) System integration: Embedding AI models into telehealth and mobile platforms to facilitate real-time diagnostics and improve accessibility has been proposed by Masud et al. (2020) [16], Yang et al. (2020) [15], Nakasi et al. (2020) [29], and Nakasi et al. (2021) [24].

VII. Conclusion and Future Works

The application of deep learning in malaria detection has emerged as a transformative approach, addressing the critical need for accurate and timely diagnostics in resource-limited settings. This paper has presented a comprehensive overview of the advancements in this field, highlighting the potential of deep learning architectures to automate malaria detection with accuracy comparable to traditional methods. Models such as ResNet-50, DenseNet, and YOLO have demonstrated remarkable diagnostic performance, achieving accuracy levels exceeding 95% in various studies [7, 8]. Moreover, lightweight architectures like MobileNet are facilitating the integration of these technologies into mobile and telehealth platforms and expanding their accessibility to underserved regions [9].

Publicly available datasets, including the NIH dataset and Delgado Dataset B, have played a pivotal role in training and validating these models. However, challenges such as dataset bias, limited diversity, and class imbalance remain significant, necessitating the development of more comprehensive and representative datasets [14, 10]. Additionally, emerging trends in multi-modal data integration, combining clinical, genomic, and imaging data, offer promising avenues for enhancing diagnostic accuracy and robustness [16].

Despite notable progress, this study has several limitations that should be acknowledged. The absence of a concrete experimental validation restricts the ability to empirically assess the effectiveness of the reviewed models in a real-world or simulated deployment. Furthermore, while the paper provides an overview of various deep learning models, it lacks a detailed comparative analysis, both qualitative and quantitative, against similar contemporary approaches. These limitations

underscore the need for more rigorous empirical benchmarking and validation in future work.

Looking forward, future research should prioritize the following directions:

1) Improving dataset quality and diversity: Developing large-scale, geographically diverse, and well-annotated datasets is essential for training models that generalize well across different populations and imaging conditions. Data augmentation techniques and synthetic data generation can also be explored to address class imbalance and enhance model robustness [17, 14].

2) Advancing lightweight models: Designing lightweight and computationally efficient models, such as optimized versions of MobileNet and YOLO, is critical for enabling real-time diagnostics on mobile devices. These models must balance accuracy with resource constraints to ensure their feasibility in low-resource settings [9].

3) Integration into health systems: Embedding deep learning-based diagnostic systems into existing healthcare infrastructures, particularly in malaria-endemic regions, can enhance accessibility and scalability. Mobile applications and telehealth platforms should be leveraged to provide real-time, on-the-spot diagnostics for healthcare workers [16, 15].

4) Enhancing explainability and trust: Integrating explainable AI (XAI) techniques, such as saliency maps and Grad-CAM, will make model predictions more interpretable and clinically reliable. Ensuring transparency in decision-making processes is vital for building trust among healthcare professionals and stakeholders [7].

5) Incorporating experimental validation and benchmarking: Future studies should include empirical validation of proposed methods using independent test datasets or clinical data to strengthen their practical credibility. Additionally, comparative analyses with other state-of-the-art techniques will enable a clearer assessment of relative strengths, limitations, and application domains.

REFERENCES

[1] World Health Organization (WHO), "Malaria - world health organization (who)," https://www.who.int/news-room/fact-sheets/detail/malaria, Dec. 2023, accessed: 2023-12-04. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/malaria

[2] S. Chibuta and A. C. Acar, "Real-time malaria parasite screening in thick blood smears for low-resource setting," *Journal of Digital Imaging*, vol. 33, no. 3, pp. 763–775, Jan. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7256156/

[3] M. Delgado-Ortet, A. Molina, S. Alférez, J. Rodellar, and A. Merino, "A deep learning approach for segmentation of red blood cell images and malaria detection," *Entropy*, vol. 22, no. 6, p. 657, 2020. [Online]. Available: https://www.mdpi.com/1099-4300/22/6/657

[4] J. Cunningham and M. Gatton, "Malaria rapid diagnostic test performance : results of who product testing of malaria rdts : round 5 (2013)," 2018. [Online]. Available: https://eprints.qut.edu.au/75953/1/eprints_75953_PUB.pdf

[5] C. Wang, Y. Dong, C. Li, J. Oberstaller, M. Zhang, J. Gibbons, C. V. Pires, M. Xiao, L. Zhu, R. H. Jiang *et al.*, "Malariased: a deep learning framework to decipher the regulatory contributions of noncoding variants in malaria parasites," *Genome Biology*, vol. 24, no. 1, p. 231, 2023. [Online]. Available: https://link.springer.com/article/10.1186/s13059-023-03063-z

[6] N. Tangpukdee, C. Duangdee, P. Wilairatana, and S. Krudsood, "Malaria diagnosis: a brief review." *The Korean Journal of Parasitology*, vol. 47, no. 2, p. 93, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2688806/

[7] S. Rajaraman, S. Jaeger, and S. K. Antani, "Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images," *PeerJ*, vol. 7, p. e6977, May 2019. [Online]. Available: https://peerj.com/articles/6977/

[8] A. Vijayalakshmi, "Deep learning approach to detect malaria from microscopic images," *Multimedia Tools and Applications*, vol. 79, no. 21, pp. 15 297–15 317, Jan. 2020. [Online]. Available: https://link.springer.com/article/10.1007/S11042-019-7162-Y

[9] F. Yang, H. Yu, K. Silamut, R. J. Maude, S. Jaeger, and S. Antani, "Smartphone-supported malaria diagnosis based on deep learning," in *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. Springer, 2019, pp. 73–80. [Online]. Available: https://lhncbc.nlm.nih.gov/LHC-publications/PDF/pub10001.pdf

[10] S. Rajaraman, S. K. Antani, M. Poostchi, K. Silamut, M. A. Hossain, R. J. Maude, S. Jaeger, and G. R. Thoma, "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, p. e4568, Apr. 2018. [Online]. Available: https://peerj.com/articles/4568/?ref=https://githubhelp.com&utm_source=TrendMD&utm_campaign=PeerJ_TrendMD_1&utm_medium=TrendMD

[11] Y. Dong, W. D. Pan, and D. Wu, "Impact of misclassification rates on compression efficiency of red blood cell images of malaria infection using deep learning," *Entropy*, vol. 21, no. 11, p. 1062, 2019. [Online]. Available: https://www.mdpi.com/1099-4300/21/11/1062

[12] D. Bibin, M. S. Nair, and P. Punitha, "Malaria parasite detection from peripheral blood smear images using deep belief networks," *IEEE Access*, vol. 5, pp. 9099–9108, 2017. [Online]. Available: https://ieeexplore.ieee.org/iel7/6287639/7859429/07931565.pdf

[13] A. Chattopadhyay, K. Muñoz-Hofmann, O. Youngquist, and Y. Zhou, "Detecting malaria using deep learning," 2019. [Online]. Available: https://shibberu.com/wp-content/uploads/2016/04/malariafinalreport-1.pdf

[14] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma, "Image analysis and machine learning for detecting malaria," *Translational Research*, vol. 194, pp. 36–55, Apr. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S193152441730333X

[15] F. Yang, M. Poostchi, H. Yu, Z. Zhou, K. Silamut, J. Yu, R. J. Maude, S. Jaeger, and S. Antani, "Deep learning for smartphone-based malaria parasite detection in thick blood smears," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1427–1438, May 2020. [Online]. Available: https://ieeexplore.ieee.org/iel7/6221020/6363502/08846750.pdf

[16] M. Masud, H. Alhumyani, S. S. Alshamrani, O. Cheikhrouhou, S. Ibrahim, G. Muhammad, M. S. Hossain, and M. Shorfuzzaman, "Leveraging deep learning techniques for malaria parasite detection using mobile application," *Wireless Communications and Mobile Computing*, vol. 2020, no. 1, p. 8895429, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2020/8895429

[17] K. F. Fuhad, J. F. Tuba, M. R. A. Sarker, S. Momen, N. Mohammed, and T. Rahman, "Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application," *Diagnostics*, vol. 10, no. 5, p. 329, 2020. [Online]. Available: https://www.mdpi.com/2075-4418/10/5/329

[18] V. A and R. K. B, "Deep learning approach to detect malaria from microscopic images," *Multimedia Tools and Applications*, vol. 79, no. 21–22, pp. 15 297–15 317, Jan. 2019. [Online]. Available: https://link.springer.com/article/10.1007/S11042-019-7162-Y

[19] K. Sriporn, C.-F. Tsai, C.-E. Tsai, and P. Wang, "Analyzing malaria disease using effective deep learning approach," *Diagnostics*, vol. 10, no. 10, p. 744, 2020. [Online]. Available: https://www.mdpi.com/2075-4418/10/10/744

[20] P. Manescu, M. J. Shaw, M. Elmi, L. Neary-Zajiczek, R. Claveau, V. Pawar, I. Kokkinos, G. Oyinloye, C. Bendkowski, O. A. Oladejo, B. F. Oladejo, T. Clark, D. Timm, J. Shawe-Taylor, M. A. Srinivasan, I. Lagunju, O. Sodeinde, B. J. Brown, and D. Fernandez-Reyes,

"Expert-level automated malaria diagnosis on routine blood films with deep neural networks," *American Journal of Hematology*, vol. 95, no. 8, pp. 883–891, Apr. 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajh.25827

[21] L. Yunda, A. Alarcón, and J. Millán, "Automated image analysis method for p-vivax malaria parasite detection in thick film blood images," *Sistemas y Telemática*, vol. 10, no. 20, p. 9, Mar. 2012. [Online]. Available: https://www.redalyc.org/pdf/4115/411534387002.pdf

[22] R. Kapoor, "Malaria detection using deep convolution neural network," Master's thesis, University of Cincinnati, 2020. [Online]. Available: https://eudl.eu/pdf/10.4108/eai.18-12-2023.2348130

[23] F. Yang, M. Poostchi, H. Yu, Z. Zhou, K. Silamut, J. Yu, R. J. Maude, S. Jaeger, and S. Antani, "Deep learning for smartphone-based malaria parasite detection in thick blood smears," *IEEE journal of biomedical and health informatics*, vol. 24, no. 5, pp. 1427–1438, 2019. [Online]. Available: https://ieeexplore.ieee.org/iel7/6221020/6363502/08846750.pdf

[24] R. Nakasi, E. Mwebaze, and A. Zawedde, "Mobile-aware deep learning algorithms for malaria parasites and white blood cells localization in thick blood smears," *Algorithms*, vol. 14, no. 1, p. 17, Jan. 2021. [Online]. Available: https://www.mdpi.com/1999-4893/14/1/17

[25] F. Abdurahman, K. F. Anlay, and M. Aliy, "Malaria parasite detection in thick blood smear microscopic images using modified yolov3 and yolov4 models," Sep. 2020. [Online]. Available: https://link.springer.com/article/10.1186/s12859-021-04036-4

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[27] A. Molina, J. Rodellar, L. Boldú, A. Acevedo, S. Alférez, and A. Merino, "Automatic identification of malaria and other red blood cell inclusions using convolutional neural networks," *Computers in Biology and Medicine*, vol. 136, p. 104680, Sep. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521004741

[28] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments," *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922001363/pdf?md5=4bdf4d0a84fcdecfd3bd313649e8e53d&pid=1-s2.0-S1877050922001363-main.pdf

[29] R. Nakasi, E. Mwebaze, A. Zawedde, J. Tusubira, B. Akera, and G. Maiga, "A new approach for microscopic diagnosis of malaria parasites in thick blood smears using pre-trained deep learning models," *SN Applied Sciences*, vol. 2, no. 7, Jun. 2020. [Online]. Available: https://link.springer.com/article/10.1007/s42452-020-3000-0

[30] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, May 2015. [Online]. Available: https://arxiv.org/pdf/1512.07108

[31] M. Bakator and D. Radosav, "Deep learning and medical diagnosis: A review of literature," *Multimodal Technologies and Interaction*, vol. 2, no. 3, p. 47, Aug. 2018. [Online]. Available: https://www.mdpi.com/2414-4088/2/3/47

[32] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. Kruthiventi, and R. V. Babu, *An Introduction to Deep Convolutional Neural Nets for Computer Vision*. Elsevier, 2017, pp. 25–52. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128104088000031

[33] X. Jiang, T. Gao, Z. Zhu, and Y. Zhao, "Real-time face mask detection method based on yolov3," *Electronics*, vol. 10, no. 7, p. 837, Apr. 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/7/837

[34] G. Marques, A. Ferreras, and I. de la Torre-Diez, "An ensemble-based approach for automated medical diagnosis of malaria using efficientnet," *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 28 061–28 078, Mar. 2022. [Online]. Available: https://link.springer.com/article/10.1007/s11042-022-12624-6

[35] C. Khosla and B. S. Saini, "Enhancing performance of deep learning models with different data augmentation techniques: A survey," in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*. IEEE, Jun. 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9160048/

Application of Deep Learning-Based Image Compression Restoration Technology in Power System Unstructured Data Management

Junjie Zha*, Aiguo Teng, Xinwen Shan, Hao Tang, Zihan Liu
State Grid Jiangsu Electric Power Co., Ltd, Information & Telecommunication Branch, Jiangsu Nanjing, 210000 China

***Abstract*—In power-system unstructured-data management, a large volume of images from inspection drones, substation cameras, and smart meters is heavily compressed due to bandwidth and storage constraints, resulting in lower resolution that hinders defect detection and maintenance decisions. Although deep-learning super-resolution (SR) techniques have made significant advances, real-world deployments still require a balance between reconstruction accuracy and model lightweightness. To meet this need, we introduce a channel-attention-embedded Transformer SR method (CAET). The approach adaptively injects channel attention into both the Transformer's global features and the convolutional local features, harnessing their complementary strengths while dynamically enhancing critical information. Tested on five public datasets and compared with six representative algorithms, CAET achieves the best or second-best performance across all upscaling factors; at 4× enlargement, it outperforms the advanced SwinIR method by 0.09 dB in PSNR on Urban100 and by 0.30 dB on Manga109, with noticeably improved visual quality. Experiments demonstrate that CAET delivers high-precision, low-latency restoration of compressed images for the power sector while keeping model complexity low.**

Keywords*—Image compression; attention mechanism; multimodal fusion; unstructured data in the power industry; image data

I. Introduction

In today's data-intensive power systems, images captured by UAV patrols, robot inspections, smart meters and substation cameras constitute a major source of unstructured information. Because these visual records are often transmitted or stored under stringent bandwidth and storage limits, aggressive compression and down-sampling are common, which sharply degrades resolution and hinders defect detection, asset identification and other downstream analytics. High-resolution (HR) images convey richer detail than their low-resolution (LR) counterparts [1-2], so recovering spatial fidelity after compression is crucial for reliable power-equipment management. Single-image super-resolution (SISR) addresses this need by reconstructing an HR image from one LR input and has been proven in domains such as security, medical and satellite imaging [3-4].

Deep-learning methods dominate modern SISR, thanks to rapid progress in GPU computing power, achieving large leaps in reconstruction quality [17]. The seminal SRCNN network used a three-layer CNN to learn an LR–HR mapping [5]. Deeper CNNs such as VDSR with 20 layers [13], EDSR that widens and deepens residual blocks [20][8], WDSR that activates wider feature dimensions, and RCAN that embeds channel attention to model inter-channel dependencies, have steadily improved results. Yet convolution itself has limitations: fixed filters ignore image content, and locality prevents long-range modeling. Transformer architecture [30] overcomes these issues via global self-attention that exploits self-similarity [7]. Large-scale Transformer variants (e.g., IPT [4]) and remote-sensing specific designs like TransENet [18, 26], as well as SwinIR built on Swin Transformer blocks [19] [22-23] have shown strong performance.

However, state-of-the-art networks are typically heavy, conflicting with the real-time and resource-constrained requirements of power-system back-end servers and edge devices. Lightweight CNN strategies—FSRCNN that operates directly on LR inputs [6], recursive models DRCN and DRRN that cut depth via weight sharing [14, 28], and distillation-based IMDN and RFDN that progressively extract informative features while trimming parameters [11, 21]—have mitigated this to a degree. The Swin Transformer's windowed attention further lowers complexity without sacrificing quality and performs well on restoration tasks.

Building on these insights, we propose a channel-attention-embedded Transformer (CAET) for compressed-image restoration in power-grid data platforms. The method alternates Transformer attention with convolution, exploiting their complementary strengths, and uses channel attention to adaptively fuse features, enhancing learning capacity while remaining lightweight for deployment. Experimental comparisons on multiple datasets confirm that CAET attains notable quality gains with modest model size, making it suitable for large-scale unstructured image archives in the power sector. Key advantages include effective integration of convolution and Transformer feature extraction via the CAET block, adaptive weighting of multi-level features through channel attention, and superior performance versus existing SR approaches at lower computational cost.

In response to the common issues of complexity, high resource consumption, and the difficulty of balancing performance and efficiency in existing image super-resolution methods, this study conducts research aimed at improving image reconstruction quality while reducing model complexity. We propose a Channel Attention-Embedded Transformer Image Super-Resolution Network (CAET), which integrates

the local modeling capabilities of convolutional networks with the global modeling advantages of Transformers. By introducing a channel attention mechanism to enhance the discrimination of multi-level features, the method improves the ability to restore image details. While maintaining a low parameter count and computational complexity, our approach achieves superior reconstruction performance compared to existing methods across multiple public datasets, demonstrating good lightweight properties and practicality. The main innovations of this study are: 1) the design of a channel attention-embedded Transformer module that effectively combines convolutional and Transformer features; 2) the proposal of a channel attention discrimination enhancement strategy to improve the network's learning ability; and 3) the validation of the superiority of this method under lightweight conditions across various diverse datasets.

II. Image Processing Related

Due to the demand for implementing super-resolution tasks on commonly used real-world devices, lightweight super-resolution models have attracted widespread attention. FSRCNN replaced the approach of SRCNN [5], which first upsampled images before feeding them into the network, by applying the core network directly to low-resolution (LR) images, significantly reducing the computational resources required [6]. DRCN [14] and DRRN [28-31] introduced recursive neural networks, which, while reducing the number of parameters and network depth, led to a significant increase in computational cost due to repeated recursion. LapSRN (Deep Laplacian Pyramid Network for Fast and Accurate Super-Resolution) [16] adopted a progressive strategy to gradually increase image resolution, achieving more stable results for high-scale upscaling. IMDN [10-12] proposed a lightweight Information Multi-distillation Network, which effectively extracts hierarchical features using an information distillation mechanism. LatticeNet (Image Super-Resolution with Lattice Block) [4, 18, 24], inspired by the Fast Fourier Transform, designed a lattice network capable of efficiently utilizing and adjusting multi-level information. ECBSR (Edge-oriented Convolutional Block for Real-time Super-Resolution) [15, 19, 36] introduced a convolutional block incorporating edge information based on re-parameterization techniques, which enhances the model's learning capacity while reducing inference time. Despite significant progress made by lightweight super-resolution algorithms, there is still room for improvement in reconstruction quality.

III. Channel Attention Embedded Transformer

A. Algorithmic Framework

The specific design diagram of the lightweight super-resolution network based on the combination of Transformer and convolution is shown in Fig. 1. It mainly consists of four parts: shallow feature extraction stage, deep feature extraction stage, multi-level feature fusion stage, and image reconstruction stage. Among them, the deep feature extraction stage is composed of Transformer modules embedded with attention mechanisms (CAETB), which will be described in detail. The network takes the given I_{LR} (input low-resolution image) and I_{SR} (inferred super-resolution image) as the low-resolution input and the predicted high-resolution output, respectively.

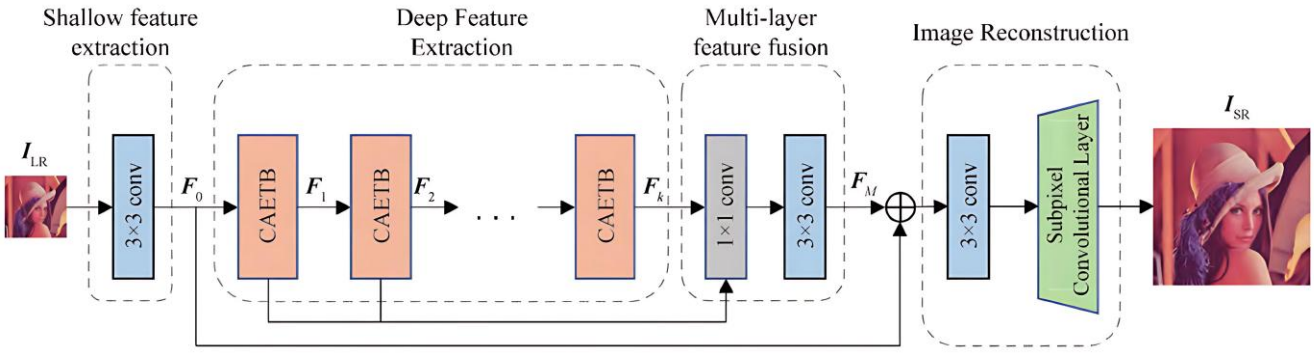


Fig. 1. Overall structure of Channel-Attention-Embedded Transformer network.

1) Shallow feature extraction stage. This stage uses a convolutional layer with a kernel size of 3 × 3 to extract shallow features F_0 from the given LR image. The process can be expressed as:

$$F_0 = H_{\mathrm{SF}}(I_{\mathrm{LR}}) \tag{1}$$

In this equation, H_{SF} represents the convolution operation. The extracted shallow feature F_0 will be further used for deep feature extraction. Meanwhile, F_0 is also directly passed to the reconstruction module to preserve the low-frequency information of the image.

2) Deep feature extraction stage. This stage takes the shallow feature F_0 as input and uses multiple CAETBs to extract deep feature information. Assuming the number of CAETBs is k, the output of the i^{th} CAETB, denoted as F_i $(1 \le i \le k)$, can be expressed as:

$$\boldsymbol{F}_i = H_{\mathrm{CAETB}}(\boldsymbol{F}_{i-1}), i = 1, 2, \cdots, k \tag{2}$$

In this equation, H_{CAETB} represents the operation of the CAETB, which is used to extract deep features of the image. The structure of CAETB is shown in Fig. 2.

3) Multi-level feature fusion stage. Hierarchical information from different stages is beneficial to the final reconstruction result. Therefore, in the multi-level feature fusion stage, the network integrates all low-level and high-level

information from the deep feature extraction stage. The fused result is denoted as $\boldsymbol{F}_M$.

$$\boldsymbol{F}_M = H_{\text{MFF}}(\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_i, \cdots, \boldsymbol{F}_k) \tag{3}$$

In this equation, H_{MFF} represents the multi-level feature fusion operation, which provides sufficient reference and guidance for the final image reconstruction.

4) Image reconstruction stage. The fused result F_{M} and the shallow feature F_0 are further input into the image reconstruction stage to recover high-resolution images adapted to different tasks. The process of obtaining the final high-resolution image I_{SR} can be expressed as follows:

$$\boldsymbol{I}_{\text{SR}} = H_{\text{REC}}(\boldsymbol{F}_0 + \boldsymbol{F}_M) \tag{4}$$

In this equation, HREC represents the operation in the reconstruction stage, which uses a 3 × 3 convolutional layer and the sub-pixel convolution layer from ESPCN [27] to upsample the features to the corresponding size of the super-resolution image.

B. Attention-Embedded Transformer Block

Convolutional layers provide more stable optimization and better feature extraction in early visual processing. Additionally, the shared weights of spatially invariant filters can enhance the network's translation equivariance [19]. Stacking convolutional layers effectively enlarges the network's receptive field. Therefore, three cascaded convolutional layers are placed at the front end of the CAETB to receive feature outputs from the previous module. To better adjust features from different levels and transformation units, the network employs a channel attention-based feature discrimination enhancement strategy, which performs channel attention-based discriminative enhancement and interactive fusion between Transformer features and convolutional features. Specifically, the input Transformer-transformed features and convolution-processed features undergo learnable channel attention feature enhancement and cross-fusion. The generation method of channel attention will be detailed in Section III(D). Taking the input feature as F_i, the process of channel attention feature discrimination enhancement can be expressed as follows.

$$\boldsymbol{F}_P = A \times H_{\text{CL}}(\boldsymbol{F}_i) + \boldsymbol{F}_i \tag{5}$$

$$\boldsymbol{F}_Q = H_{\text{CL}}(\boldsymbol{F}_i) + D \times \boldsymbol{F}_i \tag{6}$$

In this equation, A and D represent channel attention parameters, and H_{CL} denotes the cascaded convolutional layer operation. The three convolutional layers have channel numbers of 60, 45, and 60, respectively, with Leaky ReLU (LReLU) activation applied between them. Then, a convolutional layer with a kernel size of 1 × 1 is used to adjust the cascaded features back to the original number of channels. Its output, $\boldsymbol{F}_R$, can be expressed as:

$$F_R = H_C\left(Concat\left(F_P, F_Q\right)\right) \tag{7}$$

In this equation, Concat represents the feature concatenation operation, and H_C denotes the 1 × 1 convolutional layer operation. After passing through the attention embedding part, the features are input into the ST module for further feature extraction. The output of the module, F_S, can be expressed as:

$$F_s = H_{\text{Swin}}(F_R) \tag{8}$$

In this equation, H_{Swin} represents the operation of the ST module. The number of Swin Transformer layers (STL) in the CAETB module is set to 4.

The specific structure of CAETB is shown in Fig. 2. Since CAETB adopts a residual connection structure, the final output of the module, $\boldsymbol{F}_{i+1}$, can be expressed as:

$$F_{i+1} = F_s + F_i \tag{9}$$

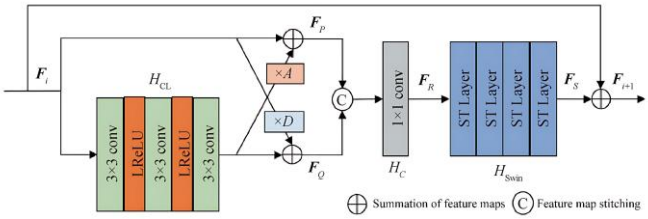


Fig. 2. Internal structure of the CAETB.

C. ST Module

The ST module is an improvement over the multi-head attention mechanism in the standard Transformer architecture. The standard Transformer performs global self-attention calculations on images, but the complexity of the global attention mechanism increases sharply as the image size grows. Therefore, when dealing with larger images in downstream vision tasks, the standard Transformer faces excessive memory requirements. To address this issue, the ST module introduces local attention mechanisms and a window-shifting mechanism. The overall structure of the Swin Transformer layer (STL) is shown in Fig. 3.

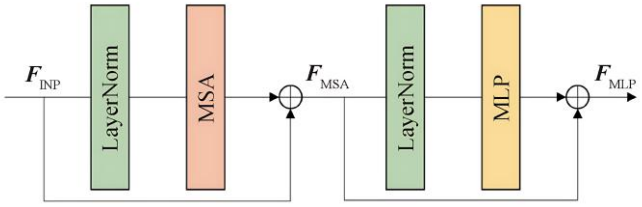


Fig. 3. Internal structure of STL.

For a given input of size $H \times W \times C$, ST first uses non-overlapping local windows of size $M \times M$ to reshape the input features into $\frac{HW}{M^2} \times M^2 \times C$, where $\frac{HW}{M^2}$ represents the total number of windows. Then, standard self-attention is computed separately for each window. For the local window feature $X \in R^{M^2 \times C}$, the query, key, and value matrices Q, K, and V are computed as follows:

$$Q = XP_Q, K = XP_K, V = XP_V \tag{10}$$

In this equation, PQ, PK, and PV are shared learnable projection matrices across different windows, and Q, K, V $\in R^{M^2 \times d}$. The attention matrix Attention (Q, K, V) is computed through the self-attention mechanism within the local window and can be expressed as follows:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}} + B\right)V \tag{11}$$

In this equation, *B* represents learnable relative position encoding, and *d* denotes the dimension of the multi-head attention. The results of multi-head self-attention (MSA) are concatenated to maintain the feature dimensions. Next, a multilayer perceptron (MLP) is used for further feature enhancement, consisting of two fully connected layers with GELU (Gaussian error linear unit) activation between them. Layer normalization (LN) is applied before both the MSA and MLP, and residual connections are used in both parts. Taking the input feature as F_{INP}, the overall process can be mathematically described as follows.

$$F_{\text{MSA}} = H_{\text{MSA}}\left(H_{\text{LN}}\left(F_{\text{INP}}\right)\right) + F_{\text{INP}} \tag{12}$$

$$F_{\text{MLP}} = H_{\text{MLP}}\left(H_{\text{LN}}\left(F_{\text{MSA}}\right)\right) + F_{\text{MSA}} \tag{13}$$

In this equation, HLN represents the layer normalization operation, HMSA denotes the multi-head attention operation, and HMLP stands for the multilayer perceptron operation.

Although the local attention mechanism based on window partitioning reduces computational complexity, fixed window partitions do not allow information exchange between windows. The ST module adopts shifted window partitioning and alternates between shifted and non-shifted windows to achieve cross-window connections. Specifically, shifted window partitioning means that features are shifted by (M/2, M/2) pixels before window partitioning, and the alternating rule means that shifted and non-shifted Swin Transformer layers (STLs) are used alternately. This approach addresses the problem of no information exchange between non-overlapping windows and significantly increases the receptive field of the ST module.

D. Channeling Attention Mechanisms

This section provides a detailed introduction to the calculation of channel attention parameters *A* and *D* mentioned. The channel attention mechanism [9] models the interdependencies between feature channels and can adaptively adjust the feature responses of different channels by assigning corresponding weights. Embedding channel attention enables adaptive enhancement and fusion of the corresponding convolutional and Transformer features within CAETB. Moreover, due to the local operation characteristics of convolution, each output value cannot represent the overall information of the entire image. Therefore, global information from channel attention is needed as guidance to select the most effective features across the whole image.

The specific operation of the channel attention mechanism is shown in Fig. 4. The input is denoted as X = [x1, x2, ..., xc], containing c channel feature maps each of size H×W. To obtain the global characteristics between feature channels, global average pooling (GAP) is used to acquire the statistical features of each channel, denoted as Z = [z1, z2, ..., zc]:

$$\boldsymbol{z}_C = H_{\text{GAP}}(\boldsymbol{x}_C) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{x}_c(i,j) \tag{14}$$

In this equation, xc(i, j) represents the value at position (i, j) in the feature map xc, and HGAP denotes the global average pooling operation, which calculates the average value representing the global information of each feature map.

The above information is then input into the second process, called weight learning (WL). The weight learning process consists of two fully connected layers, a ReLU function, and a sigmoid function. The former learns the nonlinear interactions between channels through channel squeeze and excitation, while the latter restores the channels and normalizes the parameters to ensure the network can focus on multiple important channels simultaneously. This process can be expressed as:

$$W = H_{\text{WL}}(Z) = s\left(f_2 r\left(f_1 Z\right)\right) \tag{15}$$

In this equation, H_{WL} represents the overall weight learning process, $r(\cdot)$ and $s(\cdot)$ represent the *ReLU* and *sigmoid* functions, respectively, and f_1 and f_2 denote the two fully connected layers.

The learned weights $W = [w_1, w_2, ..., w_c]$ are multiplied by the original input features. The adjusted result is denoted as $X' = [x'_1, x'_2, ..., x'_c]$, where the adjusted feature map corresponding to the c^{th} channel can be expressed as:

$$x'_c = w_c x_c \tag{16}$$

In this equation, w_c represents the weight factor used to adjust the proportion of the original feature's weight, allowing the weight of important feature information to be increased.

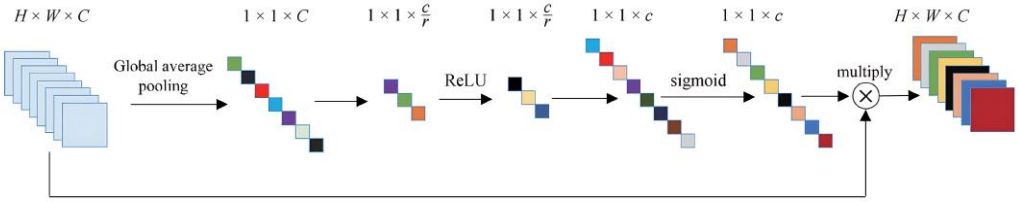


Fig. 4. Architecture of channel attention module.

E. Loss Function

For the sake of fairness in comparison, only the L_1 loss function is used to optimize the network. Given a training dataset of N image pairs $\left\{I_{LR}^{i}, I_{HR}^{i}\right\}_{i=1}^{N}$, the optimization process can be expressed as follows:

$$L(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{I}_{SR}^{i} - \boldsymbol{I}_{HR1}^{i} \tag{17}$$

In this equation, I_{SR}^{i} represents the high-resolution image predicted by the network for I_{LR}^{i}, I_{HR}^{i} denotes the ground-truth high-resolution image, and θ represents the learnable parameters of the network.

IV. Experimental Results and Analysis

A. Experimental Setup

For the model training part, this study uses the widely applied DIV2K dataset for image super-resolution tasks. Specifically, bicubic downsampling with three scale factors (×2, ×3, and ×4) is performed on 800 training images to obtain low-resolution input images. Data diversity is enhanced by randomly applying vertical rotations and horizontal flips to the training images. The network optimizer is *Adam*, with parameters β_1 = 0.9, β_2 = 0.999, and $\varepsilon = 10^{-8}$. The initial learning rate is set to 2×10^{-4} and is halved at the 150,000th, 300,000th, 400,000th, and 450,000th batches, with a total of 500,000 training batches. During training, each input batch consists of 32 randomly cropped low-resolution images of size 64 × 64 pixels (batch size = 32). For the super-resolution networks with ×3 and ×4 scale factors, training is performed based on the pretrained model with ×2 scale factor, and the total number of training iterations is halved.

For model evaluation, this study uses five commonly used public benchmark datasets: Set5, Set14, BSD100 (Berkeley Segmentation Dataset 100), Urban100, and Manga109. These datasets cover various types of image features and resolutions. Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) are used as objective evaluation metrics, with quantitative assessments performed on the luminance channel. Meanwhile, model complexity is measured by the number of parameters and the number of multiply-add operations (MAdds). The MAdds represent the cumulative number of multiplication and addition operations required by the model to process a single input image, with an output image size of 1280 × 720 pixels used as the benchmark. All experiments are conducted on the PyTorch platform, using an NVIDIA GTX 4090 GPU.

B. Analysis of CAETB Structure and Volume

To verify the effectiveness of the CAETB structure, the impact of different embedding combinations of channel attention and Transformer on the results is explored. Fig. 5 shows the case where channel attention is embedded after the Transformer, while the case where channel attention is embedded before the Transformer is the same as in CAETB.

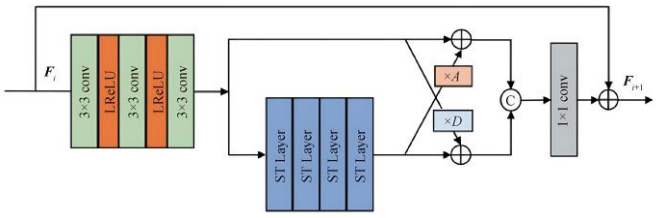


Fig. 5. Channel Attention（CA）after Transformer layers.

Table I shows the impact of different combination methods on reconstruction quality for the Urban100 and Set5 datasets at a scale factor of ×2. It can be seen that embedding channel attention before the Transformer layer achieves better reconstruction performance. During the actual training process, the model occupied approximately 7.6 GB of GPU memory when trained on an RTX 4080, with a single iteration taking about 0.85 seconds and a total training time of approximately 68 hours. In the inference stage, when the input image size is 1280×720, the average reconstruction time for a single image is 46 milliseconds, demonstrating real-time processing capability. Therefore, this model can be efficiently deployed on conventional high-performance GPU platforms, making it suitable for practical application scenarios that require a balance between performance and resources.

TABLE I. Effect of the Combination Style in CAETB on the Reconstruction Performance

Combinatorial model	Number of parameters/k	Urban 100	Set5
		PSNR/(dB)/SSIM	PSNR/(dB)/SSIM
CA embedded in front	851	**32.79/0.9348**	**38.15/0.9618**
CA is embedded after	851	32.74/0.9340	38.12/0.9615

Note: Bold font indicates the best result in each column.

Since network depth plays an important role in improving reconstruction performance, the effect of increasing the number of CAETBs from two to five on the network's reconstruction results was investigated. Table II shows the reconstruction results with different numbers of CAETBs at a scale factor of ×2 on the Set14 dataset, along with an analysis of the required number of parameters. As shown in Table II, due to the strong nonlinear abstraction capability of deep networks, the performance improves as the number of modules increases. However, it can also be observed that the improvement in reconstruction results slows down as the number of CAETBs increases, a phenomenon known as the saturation of deep networks. To balance model complexity and performance, four CAETBs are selected to form the basic reconstruction network.

TABLE II. Effect of the Number of CAETBs on Parameter Size and Reconstruction Performance on SET14

Quantities	Number of parameters/k	PSNR/dB	SSIM
2	446	33.73	0.9195
3	649	33.83	0.9201
4	851	33.89	0.9204
5	1054	33.93	0.9206

C. Comparison with other Algorithms

To validate the effectiveness of the algorithm, this study compares it with lightweight networks such as bicubic interpolation, SRCNN [5], CARN [2], IMDN [11], LatticeNet [24-25], and SwinIR [19] under different scale factors (×2, ×3, ×4). These algorithms are representative and have superior performance.

The comparison results of PSNR and SSIM between the proposed algorithm and other methods are shown in Tables III to V. It can be seen that CAET leads in objective metrics PSNR and SSIM across all datasets and different upsampling factors. Specifically, on the Urban100 dataset, CAET outperforms the second-best method by 0.03 dB at an upsampling factor of ×2, 0.08 dB at ×3, and 0.09 dB at ×4. The most significant improvement is observed on the Manga109 dataset, where CAET surpasses the second-best by 0.13 dB at ×2, 0.33 dB at ×3, and 0.30 dB at ×4. In terms of model complexity, CAET maintains relatively low parameter counts and FLOPs, achieving better restoration performance while having lower complexity than SwinIR, which also uses a Transformer model. To further demonstrate the effectiveness of the proposed algorithm, the number of CAETBs was reduced to two, resulting in the CAET-M variant. Despite its significantly lower complexity compared to convolution-based lightweight SR algorithms like IMDN, CAET-M still achieves some performance gains. These results indicate that the proposed algorithm offers better overall performance compared to these lightweight methods.

To further analyze the impact of dataset diversity on model performance, this study conducts a statistical comparison of five test datasets from multiple dimensions, including image types (natural images, urban scenes, hand-drawn images, etc.), texture complexity, and image detail density. Set5 and Set14 primarily consist of simple natural images, while BSD100 has moderate complexity. In contrast, Urban100 and Manga109 represent urban architectural images with high structural complexity and detail-dense comic images, respectively. From the objective metrics presented in Tables III to V, it can be observed that the performance improvement of the CAET model is most significant on Urban100 and Manga109, achieving the highest PSNR gains (e.g., an increase of 0.09 dB and 0.30 dB at a 4x upscaling factor). This indicates that the model demonstrates stronger reconstruction capabilities on data rich in structure and texture details. The results suggest that the CAET model, through the combination of channel attention enhancement mechanisms and the global modeling capabilities of Transformers, can effectively extract features and recover high-frequency details even when faced with increased image content complexity. Thus, it achieves robust adaptability to diverse data. This outcome verifies that the model's structural design possesses good generalization capabilities for varied image content.

TABLE III. AVERAGE PSNR AND SSIM COMPARISON OF DIFFERENT ALGORITHMS UNDER MAGNIFICATION IS 2

Model	Number of parameters/k	Multiplier/G	Set5	Set14	BSD100	Urban100	Manga109
			PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM
Interpolation	-	-	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403	30.80/0.9339
SRCNN	57	52.7	36.66/0.9542	32.45/0.9067	31.36/0.8879	29.50/0.8946	35.60/0.9663
CARN	1592	222.8	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256	38.36/0.9765
LatticeNet	756	169.5	38.06/0.9610	33.70/0.9193	32.20/0.8999	32.25/0.9288	-/-
IMDN	694	158.8	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
SwinIR-light	878	243.7	38.14/0.9611	33.86/**0.9206**	32.31/0.9012	32.76/0.9340	39.12/**0.9783**
CAET (this study)	851	214.7	**38.15/0.9618**	**33.89**/0.9204	**32.34/0.901**	**32.79/0.9348**	**39.25**/0.9781
CAET	446	110.4	38.04/0.9613	33.73/0.9195	32.26/0.9007	32.39/0.9308	38.93/0.9777

Note: Bold and underlined fonts indicate the best and second-best results in each column, respectively. A dash "–" indicates the absence of corresponding data.

TABLE IV. AVERAGE PSNR AND SSIM COMPARISON OF DIFFERENT ALGORITHMS UNDER MAGNIFICATION IS 3

Model	Number of parameters/k	Multiplier/G	Set5	Set14	BSD100	Urban100	Manga109
			PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM
Interpolation	-	-	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349	26.95/0.8556
SRCNN	57	52.7	32.75/0.9090	29.30/0.8215	28.41/0.7863	26.24/0.7989	30.48/0.9117
CARN	1592	118.8	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493	33.50/0.9440
LatticeNet	765	76.3	34.40/0.9272	30.32/0.8416	29.10/0.8049	28.19/0.8513	-/-
IMDN	703	71.5	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
SwinIR-light	886	19.5	34.62/0.9289	30.54/0.8463	29.20/0.8082	28.66/0.8624	33.98/0.9478
CAET (this study)	859	98.4	**34.65/0.9297**	**30.61/0.8482**	**29.26/0.8099**	**28.74/0.8652**	**34.31/0.9491**
CAET	454	51.1	34.40/0.9278	30.44/0.8445	29.17/0.8076	28.40/0.8577	33.89/0.9466

Note: Bold and underlined text indicate the best and second-best results in each column, respectively. A dash "–" denotes missing or unavailable data.

TABLE V. AVERAGE PSNR AND SSIM COMPARISON OF DIFFERENT ALGORITHMS UNDER MAGNIFICATION IS 4

Model	Number of parameters/k	Multiplier/G	Set5	Set14	BSD100	Urban100	Manga109
			PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM	PSNR/(dB)/SSIM
Interpolation	-	-	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN	57	52.7	30.48/0.8628	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
CARN	1592	90.9	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.47/0.9084
LatticeNet	777	43.6	32.18/0.8943	28.61/0.7812	27.57/0.7355	26.14/0.7844	-/-
IMDN	715	40.9	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
SwinIR-light	897	61.7	32.44/0.8976	28.77/0.7858	27.69/0.7406	26.47/0.7980	30.92/0.9151
CAET (this study)	871	55.4	**32.47/0.8997**	**28.80/0.7871**	**27.74/0.7427**	**26.56/0.8016**	**31.22/0.9168**
CAET	465	28.8	32.24/0.8963	28.70/0.7847	27.65/0.7394	26.26/0.7928	30.73/0.9118

Note: Bold and underlined fonts indicate the best and second-best results in each column, respectively. A dash "–" indicates that no corresponding data is available.

To further validate the advantages of the proposed network, representative visual results on standard benchmark datasets are analyzed. Fig. 6 shows the local enlargement results for image Urban100_img012 with an upsampling factor of ×4. It can be observed that other algorithms incorrectly restore the orientation of building textures, whereas CAET (proposed in this study) accurately preserves the texture structure of the buildings. Fig. 7 presents the local enlargement results for image B100_253027 at a ×4 scale. Compared to other methods, the proposed algorithm restores the zebra stripes with greater clarity and accuracy. Fig. 8 shows the ×4 enlargement results for image Set14_barbara. The proposed algorithm accurately reconstructs the arrangement of the books, while other methods exhibit varying degrees of distortion. These results demonstrate that the proposed method not only leads in objective metrics but also produces clearer super-resolved images than all the compared approaches.

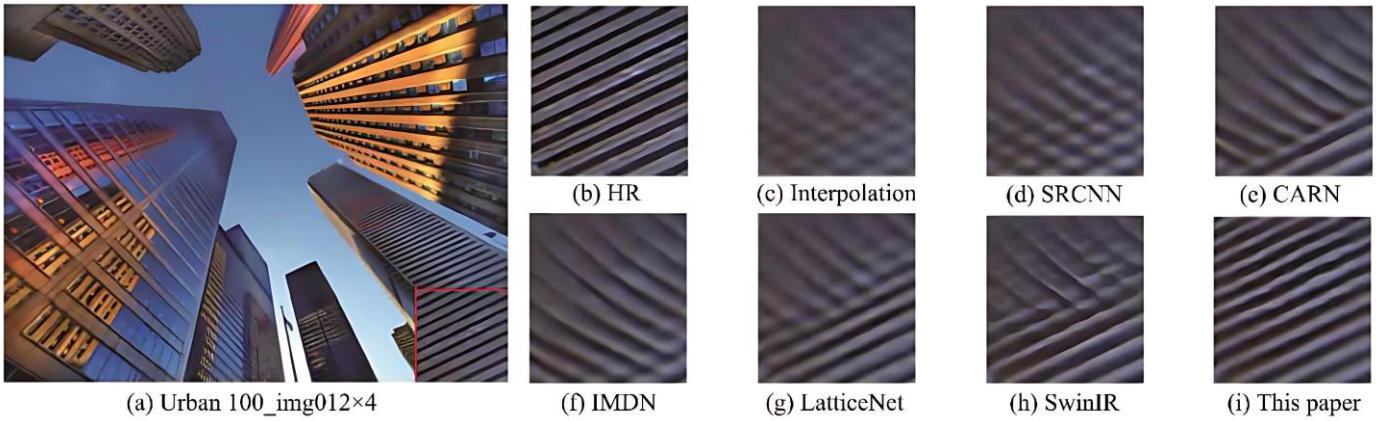


Fig. 6. Comparison of reconstructed HR images of img012 in Urban100 by different SR algorithms at the scale factor ×4 [(a) Urban100_img012×4; (b) HR; (c) bicubic interpolation; (d) SRCNN; (e) CARN; (f) IMDN; (g) LatticeNet; (h) SwinIR; i) ours].

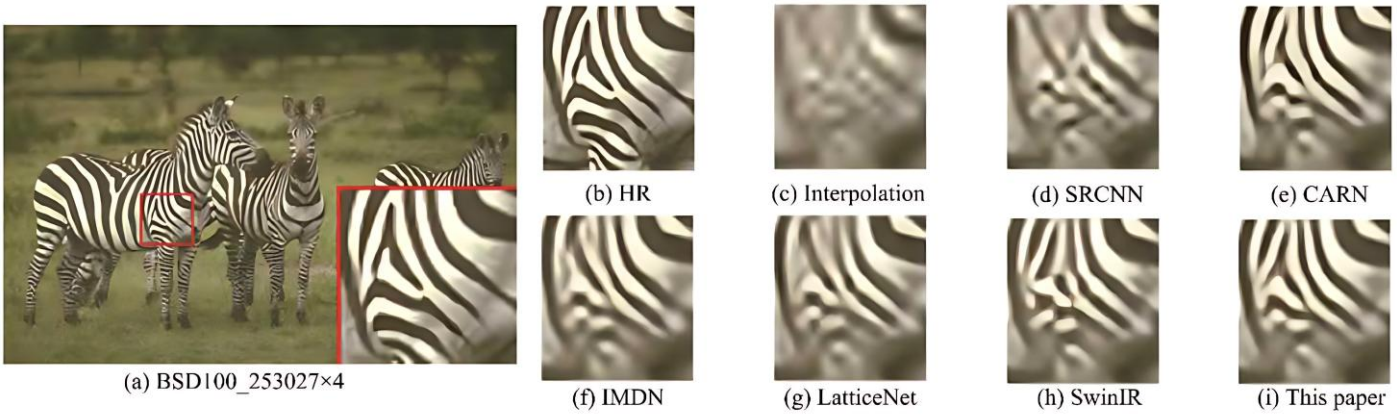


Fig. 7. Comparison of reconstructed HR images of 253027 in BSD100 by different SR algorithms with the scale factor ×4 [(a) BSD100_253027×4; (b) HR; (c) bicubic interpolation; (d) SRCNN; (e) CARN; (f) IMDN; (g) LatticeNet; (h) SwinIR; (i) ours].

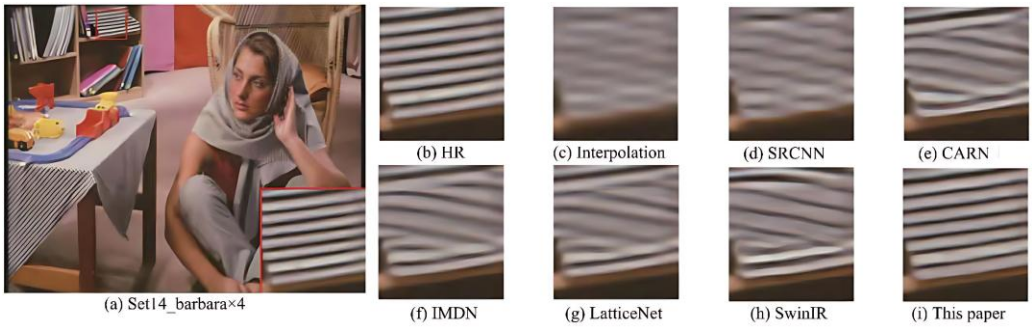


Fig. 8. Comparison of reconstructed HR images of barbara in Set14 by different SR algorithms with the scale factor ×4 [(a) Set14_barbara×4; (b) HR; (c) bicubic interpolation; (d) SRCNN; (e) CARN; (f) IMDN; (g) LatticeNet; (h) SwinIR; (i) ours].

D. Ablation Experiment

To effectively adjust features at different levels, this study adopts a channel attention-based discriminative enhancement strategy. Features from various levels are adaptively weighted using channel attention parameters. Unlike LatticeNet [24], which uses channel mean and standard deviation as modulation parameters, this approach applies channel attention weights directly to features from different levels. Table VI presents a comparison of reconstruction performance and model parameters using the two different weight generation strategies across multiple datasets with an upscaling factor of ×3. Experimental results show that, compared to the mean and standard deviation (MSD) based feature enhancement strategy proposed by Luo et al., the channel attention (CA) based discriminative enhancement method in this study enables better interaction and fusion of Transformer and convolutional features, achieving superior image reconstruction performance. To further investigate the roles of channel attention, linear weighting, and the feature aggregation module, ablation experiments were conducted to evaluate the impact of each component on overall performance. Specifically, the linear weighting structure was replaced with residual connections, and both channel attention and feature aggregation were removed to form the Base model. The models with different strategies are designated as Base A, B, C, and D.

TABLE VI. COMPARISON OF DIFFERENT WEIGHT GENERATION METHODS OF MODEL AT ×3 AMPLIFICATION FACTOR

Methodologies	Number of parameters/k	Manga109	Set5	Set14
CA	851	**34.31**	**34.65**	**30.61**
MSD	858	34.24	34.63	30.59

Note: Bold font indicates the best result in each column.

Table VII shows how these strategies affect reconstruction quality on the Manga109 dataset at a ×4 upscaling factor. The Base model performs slightly worse, achieving only 30.93 dB PSNR. Adding multi-level feature fusion in model A yields a clear 0.19 dB improvement, indicating that this fusion approach remains effective even in a Transformer-dominated network. Model B embeds channel attention in series on top of model A, while model C adds linear weighting on model A but removes learnable channel-attention parameters. Both models B and C achieve varying degrees of PSNR improvement, but combining both strategies in model D produces the most significant gain. Thus, the design strategies in this work are effective individually and even more so when combined, greatly enhancing the network's restoration capability. Compared with the Base model, model D boosts PSNR by 0.29 dB with only a 32 k increase in parameter count. For a more intuitive demonstration of each component's effect, representative images are visualized, with particular focus on how embedding channel attention improves the reconstruction of fine details. Fig. 9 shows the local magnification results of the drone aerial photo at a ×4 upscaling factor. It can be observed that, compared to model C, model D—which incorporates channel attention—can more accurately restore image details. This demonstrates that the integration of channel attention has a positive impact on image reconstruction performance. Experiments show that the model effectively improves the super-resolution performance while maintaining low complexity, and can provide a high-precision and low-latency solution for fast recovery of compressed images in the power industry. The method proposed in this study demonstrates significant advantages in practical applications. Firstly, the network architecture achieves high-fidelity restoration of image details while maintaining a low parameter count and computational complexity, resulting in a good balance between performance and efficiency. Secondly, the model exhibits stronger texture modeling and detail restoration capabilities in urban architectural images (Urban100) and comic images (Manga109), making it suitable for high-resolution demand scenarios such as video surveillance, remote sensing imagery, and anime image enhancement. Additionally, the model can run efficiently on consumer-grade GPUs with a single card, showing good deployment adaptability, which is ideal for resource-constrained edge devices or real-time image enhancement systems.

TABLE VII. COMPARISON OF THE EFFECTS OF DIFFERENT STRATEGIES ON MODEL RECONSTRUCTION PERFORMANCE ON MANGA109 AT AMPLIFICATION FACTOR ×4

Methodologies	Feature aggregation	Channel attention	Linear weighting	PSNR/dB	Number of parameters/k
Base	×	×	×	30.93	**839**
A	√	×	×	31.08	850
B	√	√	×	31.09	853
C	√	×	√	31.14	864
D (This study)	√	√	√	**31.22**	871

Note: Bold font indicates the best result in each column. "√" indicates adoption, while "×" indicates not adopted.



(a) Drone Aerial Photo (b) HR (c) Model C (d) Model D (This Paper)

Fig. 9. Comparison of the effects of different strategies on the reconstruction of ARMS in drone aerial maps at a scale of 4 [(a) Drone Aerial Photo; (b) HR; (c) model C; (d) model D (ours)].

V. CONCLUSION

In power-system unstructured-data management, images captured for patrol monitoring, equipment status awareness, and fault diagnosis are key information carriers. Because of limited storage and transmission bandwidth, these images are often heavily compressed, which degrades quality and removes fine details, ultimately impairing downstream intelligent recognition and O&M decision-making. To address this, we introduce a channel-attention-embedded Transformer super-resolution method (CAET) that balances reconstruction accuracy and model complexity for compressed-image restoration.

CAET adaptively embeds a channel-attention mechanism between Transformer global-context features and convolutional local-perception features, enabling interactive fusion. This design leverages the complementary strengths of Transformer and CNN feature extraction while dynamically enhancing critical information, markedly improving restoration quality. Across five public datasets and in comparison with six representative SR algorithms, CAET achieves best- or second-best results at all upscaling factors. Under 4× enlargement, it boosts PSNR by 0.09 dB on Urban100 and by 0.30 dB on Manga109 relative to the advanced SwinIR method, with noticeably better visual quality. Although CAET delivers strong, lightweight performance and restoration accuracy, all current experiments are conducted using bicubic-downsampled degradation. Real-world power-system images suffer from additional compression artifacts, transmission losses, and noise, so models trained on synthetic degradation still require improved generalization.

Future work will focus on enhancing the practical applicability of compressed-image restoration in power scenarios—particularly on lightweight, robust blind SR networks that can cope with multiple unknown degradations—so that image restoration technology can better support intelligent management and utilization of unstructured data in the power industry.

Although this method has demonstrated good performance on multiple public datasets, there are still two main limitations: first, the training and testing of this study are based on an idealized bicubic downsampling degradation model, which does not cover the complex degradation scenarios commonly found in real images, such as compression artifacts and noise pollution. As a result, the model may face performance degradation in practical applications. Second, there is still some computational redundancy in the integration of channel attention and Transformer features in CAET. Although it already possesses strong lightweight capabilities compared to similar methods, further optimization is needed for deployment on extremely resource-constrained devices (such as embedded systems).

REFERENCES

[1] Agustsson E, Timofte R. NTIRE 2017 challenge on single image super-resolution : dataset and study. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, IEEE: 2017, 1122-1131.

[2] Ahn N, Kang B, Sohn K A. Fast, accurate, and lightweight super-resolution with cascading residual network. Proceedings of the 15th European Conference on Computer Vision. Munich, 2018, 256-272.

[3] HOUQ,ZHOUD,FENGJ.Coordinate attention for efficient mobile network design. CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN: IEEE, 2021: 13708-13717.

[4] Chen H T, Wang Y H, Guo T Y, Xu C, Deng Y P, Liu Z H, Ma S W, Xu C J, Xu C, Gao W. Pre-trained image processing Transformer. Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021, 12294-12305.

[5] LIU Na. Semantic segmentation algorithm of transmission lines based on CBAM attention me-chanism. Yangtze Information Communication, 2023, 36(09):60-62.

[6] LI Cuiming, WANG Hua, XU Long'er, et al. Road recognition method of photovoltaic plant based on improved DeepLabv3+. Journal of Shanghai Jiao Tong University, 2024, 58(5):776-782.

[7] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. An image is worth 16 × 16 words: Transformers for image recognition at scale. 2021.

[8] ZHOU Peng, XIONG Kai, XING Yan. Mars terrain segmentation algorithm based on improved DeepLab-v3+. Space Control Technology andApplications, 2023, 49(02):10-19.

[9] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 7132-7141.

[10] ZHANG Z, QIAN Z, ZHONG T, et al. Vectorized rooftop area data for 90 cities in China. Scientific Data, 2022, 9(1):66.

[11] Hui Z, Gao X B, Yang Y C, Wang X M. Lightweight image super-resolution with information multi-distillation network. Proceedings of the 27th ACM International Conference on Multimedia. 2019, 2024-2032.

[12] Jiang M J, Qian W H, Xu D, Wu H, Liu C Y. Gradual model reconstruction of Dongba painting based on residual dense structure. Journal of Image and Graphics, 2022, 27(4): 1084-1096.

[13] WANGHaifeng, XU Yilin, XU Dayi, et al. Evaluation of distributed rooftop photovoltaic hosting capacity in distribution networks based on high definition satellite map images. Guangdong Electric Power, 2023, 36(10): 105-113.

[14] Liu H, Zhang Z Z, Song R M, Shu Z Q, Wang J X, Tian H Y, Song Y X, Chen W G. Pattern Recognition Method for Detecting Partial Discharge in Oil-paper Insulation Equipment using Optical F-P Sensor Array based

on KAN-CNN Algorithm, Journal Lightwave Technology, 2025, 43(12) 6004 - 6012.

[15] MU Jingru, YU Kun, ZENG Xiangjun, TONG Haixin, et al. Leakage. fault detection in low-voltage station area with photovoltaic power supply considering multi-disturbance factors. Southern Power System Technology, 2024, 18(10): 130-141.

[16] Liu H, Zhang Z X, Tian H Y, Song Y X, Wang J X, Shu Z Q, Chen W G. Comparison of Different Coupling Types of Fiber Optic Fabry-Perot Ultrasonic Sensing for Detecting Partial Discharge Faults in Oil-Paper Insulated Equipment, IEEE Transactions on Instrumentation and Measurement. 2024, 73, 9519612.

[17] Lei P C, Liu C, Tang J G, Peng D L. Hierarchical feature fusion attention network for image super-resolution reconstruction. Journal of Image and Graphics, 2020, 25(9): 1773-1786.

[18] Lei S, Shi Z W, Mo W J. Transformer-based multistage enhancement for remote sensing image super-resolution. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5615611.

[19] Liang J Y, Cao J Z, Sun G L, Zhang K, Van Gool L, Timofte R. SwinIR: image restoration using Swin Transformer. Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). 2021, 1833-1844.

[20] Liu H, Yang T H, Zhang Z X, Tian H Y, Song Y X, Sun Q X, Wang W, Geng YJ, Chen W G. Ultrasonic localization method based on Chan-WLS algorithm for detecting power transformer partial discharge faults by fibre optic F-P sensing array. High Voltage, 2024, 9(6),1234-1245.

[21] Liu J, Tang J, Wu G S. Residual feature distillation network for lightweight image super-resolution. Proceedings of 2020 European Conference on Computer Vision. 2020, 41-55.

[22] Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S, Guo B N. Swin Transformer: hierarchical vision Transformer using shifted windows. Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021, 9992-10002.

[23] Lu Z S, Li J C, Liu H, Huang C Y, Zhang L L, Zeng T Y. Transformer for single image super-resolution. Proceedings of 2022 IEEE/ CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022, 456-465.

[24] Luo X T, Xie Y, Zhang Y L, Qu Y Y, Li C H, Fu Y. LatticeNet: towards lightweight image super-resolution with lattice block. Proceedings of the 16th European Conference on Computer Vision. 2020, 272-289.

[25] Matsui Y, Ito K, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T, Aizawa K. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications, 2017, 76(20): 21811-21838.

[26] Wang Z D, Cun X D, Bao J M, Zhou W G, Liu G Z, Li H Q. Uformer: a general U-shaped Transformer for image restoration. Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, 17683-17693.

[27] Wang Z H, Chen J, Hoi S C H. Deep learning for image super resolution: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3365-3387.

[28] Xiong C Y, Shi X D, Gao Z R, Wang G. Attention augmented multi-scale network for single image super-resolution. Applied Intelligence, 2021, 51(2): 935-951.

[29] Yu J H, Fan Y C, Yang J C, Xu N, Wang Z W, Wang X C, Huang T. Wide activation for efficient and accurate image super-resolution. 2018.

[30] Zhang X D, Zeng H, Zhang L. Edge-oriented convolution block for real-time super resolution on mobile devices. Proceedings of the 29th ACM International Conference on Multimedia. 2021, 4034-4043.

[31] Zhang Y L, Li K P, Li K, Wang L C, Zhong B N, Fu Y. Image super-resolution using very deep residual channel attention networks. Proceedings of the 15th European Conference on Computer Vision. 2018, 294-310.

Impact of Auxiliary Information in Generative Artificial Intelligence Models for Cross-Domain Recommender Systems

Matthew O. Ayemowa, Roliana Ibrahim, Noor Hidayah Zakaria and Yunusa Adamu Bena
Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia

***Abstract*—Recommender systems (RSs) are significant in enhancing the experiences of users across different online platforms. One of the major problems faced by the conventional RSs is difficulties in getting precise preferences for users, mostly for the users that has limited previous interaction data, and this eventually affects the performance of the conventional techniques to solve the data sparsity problem. To address this challenge, this study proposes an Auxiliary-Aware Conditional GAN (AUXIGAN) model that integrates heterogeneous auxiliary information into both the generator and discriminator networks to enhance representation learning to enhance the performance of the cross-domain recommender systems (CDRS). Most researchers consider only the rating matrix of users-items and ignore the impact of auxiliary information on the interaction functions, which is very significant to the recommendation accuracy to solve data sparsity problems. The proposed novel technique considers features concatenation, attention-based fusion networks, contrastive representation learning, knowledge transfer, and multi-modal embedding alignment techniques to enhance the user-item interaction matrix. Our experiments on benchmark datasets show that the proposed model significantly outperformed state-of-the-art RSs models, the key metrics utilized are: RMSE, Precision, Recall, and MAE, which show the influence of incorporating auxiliary information into the GAN-based CDRS. In conclusion, the integration of auxiliary information on generative adversarial networks models represents a substantial advancement in the field of CDRS, and the results of the proposed models on two real-world datasets show that the proposed model significantly outperforms collaborative filtering and other GAN-based techniques.**

Keywords*—Generative adversarial networks; auxiliary information; cross-domain recommender systems; data sparsity; knowledge transfer

I. Introduction

With the rapid advancement of internet technology, online e-commerce, e-learning, and social networks have made information overloading a serious challenge, this makes it difficult for the users to get the most preferred items or information. The advent of recommender systems (RSs) has been helping to mitigate the problem of information overloading to some extent. RSs uses machine learning techniques and data analysis to recommend preferred information (contents, news, items, movies, music, tourist etc.) to the users. These systems analyze huge amounts of information about the users' preferences, interests and past behavior by using machine learning techniques like collaborative filtering (CF), clustering, and neural networks to recommend items that are personalized to the users. Netflix, Amazon, Spotify, and Lazada are illustrations of RS. Amazon recommends items based on past purchases, preferences and the browsing history of the user, Netflix suggests movies to its users or customers, and Spotify also provides music and playlists based on preferences and past histories of the users. The main purpose of using RSs is to recommend significant suggestions to the users, centered on their past preferences, behavior, etc. [1], [2]. RSs can be used on diverse domains to solve problems of information overloading, such as e-learning [3], e-commerce, and social networks [4], [5].

Traditional RSs, like content-based and collaborative filtering methods, often struggle with data sparsity challenges, particularly in domains that have limited user-item interactions, but cross-domain recommender systems (CDRS) offer solutions to this problem by transferring knowledge from one domain to the other. CDRS leverages knowledge from one or more source domains to improve the quality of recommendations in the target domain with limited data. The existing methods, which include collaborative filtering and matrix factorization, face sparsity challenges due to difference in user and item distributions across domains. Using generative artificial intelligence models like Generative Adversarial Networks (GANs) model, makes CDRS extract shared latent factors across domains, enabling knowledge transfer that improves accuracy and reduces data sparsity. GANs have been widely recognized for its potential in CDRS due to its ability to learn the distributions of complex data. It has emerged as a vital field of study by improving various domains, such as RSs, creative arts, computer vision, and language processing [6], [7], [8], [9]. GANs have displayed favorable achievement in generating realistic data distributions and capturing complex relationships within complex datasets [10], [11]. The recommendation algorithms of GANs have the potential to address the problem of data sparsity by generating realistic user data [12]. It can also improve users' experience and satisfaction [13] and mitigate cold start problems for new users who have no previous history [14], [15].

Auxiliary information is the additional features beyond the primary user-item interaction data (e.g. ratings, clicks). This auxiliary information includes user demographics, item attributes, social networks, contextual data, and more. It is

sometimes called side information. This is an aspect that will be put into consideration in this study. Researchers does not consider this auxiliary information as a crucial aspect that can enhance GANs-based RSs. Most of the researchers concentrate more on the user-item interactions only and ignores the influence of auxiliary information on the interaction functions. Incorporating auxiliary information into RSs can enhance its performance by providing additional information or insight into the user's preferences [16], [17]. Previous research on RSs explored different methods like matrix factorization, Graph Neural Networks [18], [17], [19], [20], [21], Convolutional Neural Networks [22], Deep Matrix Factorization [23], [24], Federated meta learning [2], Deep coupled Autoencoders [25], Collaborative Filtering Neural Network [26], Variational autoencoder [27] across different domains. These methods have shown significant improvement on RSs, but they still struggle with the problem of sparse data. This information features plays significant roles in enhancing the performance of RSs by providing additional context that can improve user-item representation learning, this information can be generated from various sources, such as text, multimedia, and social networks, and its effective integration can address issues like recommendation accuracy, explanation generation, and the data sparsity problem [17].

The contributions and novelty of this proposed study are summarized as follows:

- We introduce a novel model named AUXIGAN, which is a model that enhances traditional GANs by integrating auxiliary information into the generative adversarial networks for recommendations improvement.
- We employ an attention network to determine the significance of various auxiliary information components in shaping user and item interactions to solve the data sparsity problems by integrating various kinds of information extracted from a related embedding network to enhance the performance of recommendations.
- We conducted extensive experiments on real-world datasets (Amazon and Book-crossing) to validate the efficiency and rationale behind our proposed model against several baselines.

To the best of our knowledge, no study has been conducted on a GAN-based model to enhance CDRS with the integration of auxiliary information. The proposed GAN model is the first to put auxiliary information of the users and items into consideration to make the recommended items more personalized.

Structurally, this study begins by introducing the background in Section I. In Section II, we comprehensively review existing research and methods in recommender systems and generative AI models. Section III provides detailed descriptions of our research methodology. Subsequently, Section IV, V and VI showcases the outcomes of our experiments, offering a comparative analysis of various models' performances. Section VII presents the limitations and future directions and finally, Section VIII concludes the study.

II. Related Work

Conventional recommender systems (RSs) which comprise of content-based (CB) methods and collaborative filtering techniques (CF) relies heavily on user-item interactions. CF techniques like matrix factorization [28] perform well in dense datasets but weaken in sparse conditions. CB approaches use item features to recommend similar items, but often require extensive feature engineering [29]. These conventional methods still encounter data sparsity problems compared to the deep learning model techniques, such as GANs.

Generative adversarial networks (GANs) were introduced by [6], which consists of a generator (G) and discriminator (D) in a minimax game, as illustrated in Fig. 1. Recent studies have applied GANs to RSs, and their recommendations are showing promising results [13], [30]. Integrating generative AI models, such as GANs model on CDRSs brings an opportunity to overcome the limitations of traditional RSs [31], [32], [33], [34]. By generating meaningful representations for items or users, generative RSs have the potential to mitigate the cold start issue [31] and enhance accuracy on data sparsity [35]. According to some researchers, the most effective model in adversarial learning is GANs [36]. GANs are capable of capturing shared information across domains due to their ability to generate two different domain data simultaneously, where they resemble each other [37], [38].

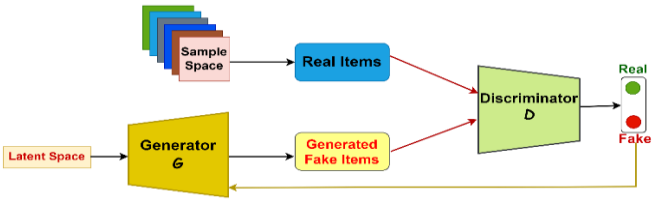


Fig. 1. The illustration of how the generator and discriminator work.

Auxiliary information in RSs is an additional data incorporated to improve relevance and accuracy of recommendations. It plays significant role in enhancing the performance of RSs and it also offers a more personalized, relevant, and context-aware recommendations to the users [39]. Auxiliary information helps to improve and enhance data sparsity [17], [40], [41], [42], [43], [44], [45].

In their study [16], the authors introduced E-EATI model to enhance social recommendation by extracting auxiliary information, aiming to improve recommendation performance by fusing auxiliary information from user or item with BERT model and item2item aggregation operation. The model addresses the project attribute fusion problems in GCN, focusing on signed networks and auxiliary information to enhance recommendation performance. In [46], the author's focus is on enhancing citation recommender systems by integrating auxiliary textual information and BERT into the Neural Citation Network (NCN) model. Their proposed model utilizes a deep neural auto-encoding mechanism with self-

attention to learn textual and citation contextual data, this is to improve the citation recommendation performance. Hi-GNN, a Hierarchical Interactive Graph Neural Network for Auxiliary Information-Enhanced Recommendations (AIER) was proposed by [17]. The model addresses the suboptimal performance of AIER by incorporating behavioral and auxiliary information with a hierarchical interaction layer. In their study [47], the authors introduce the Knowledge Graph Attention-assisted Network (HKGAT) as a unique recommendation model that incorporates the knowledge graph into the recommendation model to study higher order connectivity and improve generalization capacity. The HKGAT model utilizes an embedding layer to parameterize each node as a vector and determine the e-weights of each neighbor during the propagation.

In their study [39], the authors focuses on enhancing sequential recommendations by incorporating auxiliary information like item images and textual data to improve user preference understanding. Their proposed algorithm, MFN4Rec, effectively models multi-modal information for accurate sequential recommendation. The authors [48] utilizes Metapath2vec to acquire the low-dimensional representation of music through a music heterogeneous information network. Their proposed model, named MFAE, integrates auxiliary information fusion embedding to alleviate data sparsity and enhance music feature representation. GRU was employed to obtain user feature representation for personalized music recommendation. A novel recommendation method named STRGAN was used in [12]. The model was utilized to address data sparsity issues by incorporating user ratings and social relationships. The model, STRGAN, combines user ratings and social relationships to enhance recommendation accuracy, employing negative sampling methods to align generated recommendations with real data. MA-CVAE in [49] considers tag recommendation by integrating collaborative and multi-auxiliary information for better recommendations. The proposed model addresses the cold-start and sparsity problems in recommenders tagging by coupling social graph information, item content with collaborative information. The model also utilizes deep generative models like Variational Graph Auto-Encoder (VGAE) and Variational Auto-Encoder (VAE) to learn latent embedding from different item auxiliary information. In their study [50], the authors focuses on enhancing implicit RSs by integrating auxiliary information into matrix factorization models to capture different user preferences beyond primary feedback. The proposed method includes pointwise and pairwise models to leverage users' viewing behaviors effectively, with the pairwise model showing superior performance. In [51], the authors focuses on enhancing recommendation systems by addressing issues like sparsity and cold start through a collaborative multi-auxiliary information variational autoencoder (CMVAE).

In our study, we propose a novel model for cross-domain recommender system by incorporating auxiliary information into GAN model. The other component in the model is constructed by converting user behavior history into a multinomial distribution over behaviors. Additionally, we introduce a unique adjustment to the GANs architecture, specifically tailored for recommendation tasks, which has not been explored previously. The details of our model will be presented in the subsequent section.

III. METHODOLOGY

This segment presents the proposed model, and this is shown in the architecture of incorporating auxiliary information and rating matrix on the GANs model (AUXIGAN) on CDRS, as shown in Fig. 2.

Below is the explanation of how the model works; they are broken into key steps:

- Input data and embedding layer: Auxiliary information is utilized to create embeddings for the users (U) and items (I), this process occurs in the embedding layer, where users and items from different domains (source and target) are transformed into a shared latent space, the users-items initial embedding is denoted as input $g\Theta(.)$.
- Aggregation network: Users (Us) and their interacted items (Is) are passed through an aggregation network. This network, which aggregates information for users, is based on their historical interactions with items to generate user-specific representations.
- Generative adversarial networks (GANs): The GANs component comprises a generator $f\theta(\cdot)$ and a discriminator $h\phi(\cdot)$:
  i. *Generator* $f_\theta(\cdot)$: It generates synthetic data (e.g., predicted ratings or embeddings) conditioned on input noise (z) and real data embeddings.
  ii. *Discriminator* $h\phi(\cdot)$: It distinguishes between real and generated (fake) data.

The concatenation step combines both generated and real data to further refine the model training, and the objective is to enable domain adaptation and align the learned user and item representations between source and target domains:

- Domain adaptation and transfer layers: The model employs transfer layers to facilitate knowledge sharing between source and target domains. The domain adaptation mechanism ensures that knowledge from the source domain is transferred effectively to the target domain.
- Deep interaction layers: The output from the GANs module is passed into deep interaction layers. These layers processes user embeddings (Ut) from the target domain and utilize the generated ratings matrix (Rn) to interact with Items (It).
- Knowledge transfer: The knowledge transfer component allows the proposed model to learn from one domain and apply the learning to another. This is to help the model to mitigate challenges of data sparsity in the target domain. The knowledge is transferred from

generator to the users (Ut1, Ut2, Ut4, …. Utn,) in target domain.

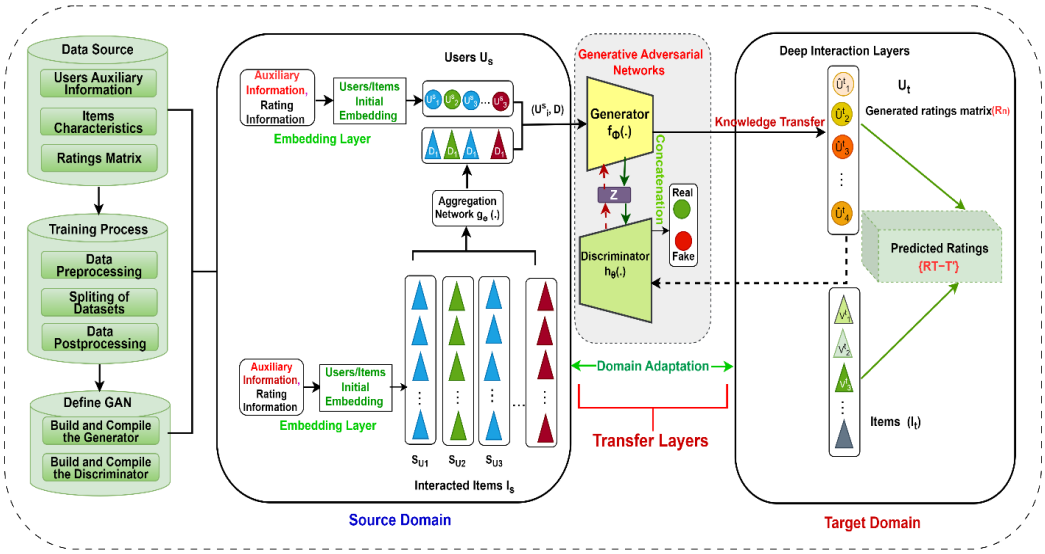


Fig. 2. Proposed AUXIGAN architecture.

- Prediction of ratings: The proposed model generates a predicted ratings matrix (RT-T') for users in the target domain, and the predicted ratings are obtained by interacting with the deep representations of user and item with the influence of auxiliary information.
- User: They are any individual who submits requests to the recommender server to receive predictions of items. To allow this process, the user's profile must be accessible for computations by the recommender server.
- Source domain: The source domain comprises the datasets with a substantial number of user ratings and auxiliary features. This is managed by a data curator who is responsible for the maintenance of the dataset and enabling secure data sharing with other data curators.
- Target domain: The target domain contains the user ratings and auxiliary features that are relatively sparse. A data curator manages this domain to handle encrypted latent profiles received from the source domain.

This proposed GANs architecture is capable of capturing shared information across domains due to its ability to leverage shared knowledge, preserve common semantics, share and transform features, utilize mutual information mechanisms, train simultaneously on the domains, and handle multiple conditional inputs. These capabilities make the proposed GAN model to be highly effective for various cross-domain applications.

A. Problem Formulation

Firstly, this proposed study examines the effect of incorporating auxiliary information into GAN-based CDRS, this is to examine its potential to improve recommendation accuracy, transferability, and robustness. Our assumption is that the two domains have overlapping users and items. When a domain faces the data sparsity problem, the other domain will be facilitated by sharing information for making recommendations. Typically, the domain that may experience this issue is the target domain, while the backing domain is the source domain, indicated by labels (t) and (s), respectively.

Let $U = \{U_1, U_2, U_3, U_4, U_5, U_6 \ldots\}$ and $I = \{I_1, I_2, I_3, I_4, I_5, I_{6\ldots}\}$ represent the sets of the users and items. We represent the users and items of the source domain as U_s and I_s, and those of the target domain as U_t and I_t. The conjoint user set between the two domains is $U_o = U_s \cap I_t$. We extract a rating matrix $R = (r_{ij})$ to uncover preferences of the users, where $(r_{ij} \in R)$ signifies the interactions between user (u_i) and item (i_j). The rating matrices for the two domains (source and target domains) are represented as R^s and R^t, respectively.

For the user $(u_i,)$, the sequence of interaction of the items in the source domain can be analyzed as $S_{ui} = \{I^s_{t1}, \ldots. I^s_{tb}\}$, where I^s_{tb} represents the b_{th} interaction of the user at timestamp t_b, $b \in \{1, \ldots, n\}$. In the proposed model, users and items are transformed into embeddings based on their characteristics, denoted as $u_i, i_j \in \mathbb{R}^d$, where d is the dimension of the embeddings. The aim of this study is to leverage the ratings matrix and auxiliary information from the two domains to improve the recommendation accuracy in the domains. The two domains have user sets U_1 and U_2, item sets I_1 and I_2, and interaction matrices R_1 and R_2.

B. Proposed Cross-Domain Recommender Systems Model

The proposed AUXIGAN model in this study primarily consists of two main networks and components: generator and discriminator with rating matrix and auxiliary information. If $I = \{I_1, I_2, I_3, I_4, I_5 \ldots, I_n\}$ and $U = \{U_1, U_2, U_3, U_4, U_5, \ldots, U_m\}$ which represents a set of "m" items and a set of "n" users correspondingly, the two matrices can then be created from user

(U), Item (I) and their interactive information which are the rating matrix $M = (r_{u,i})_{m\times n}$, and the auxiliary information matrix $M' = (\hat{r}_{u,i})_{m\times n}$, $r_{u,i}$ is the rating of user (U) on item (I), and $\hat{r}_{u,i}$ represents the interaction between users (U) and items (I). If items (I) has been rated by users (U), then $\hat{r}_{u,i} = 1$, otherwise $\hat{r}_{u,i} = 0$.

C. Model Architecture

In this study, we propose a CDRS framework that integrates a ratings matrix and auxiliary information with GANs, named AUXIGAN. The proposed model shown in Fig. 3 comprises four major components: auxiliary information, rating matrix (embedding layer), items feature aggregation, GANs {Generator (G) and Discriminator (D)} structure, and predictions.

D. Structure of GANs

We utilized the method of mapping to establish the connection from source domain to target domain; this involves training the GAN on users-items ratings and auxiliary information. At this stage, the interactive relationship between users and items is considered, also the auxiliary information related to the users is also put into consideration.

E. Aggregation of the Items Feature

Items are crucial in RSs, alongside users, in addressing data sparsity issues. When users have limited or no ratings for items in the target domain, it requires leveraging their interactions with items in the source domain, along with their auxiliary information. Given that different items interacted with by users can have varying impacts, we employ an attention mechanism to aggregate item features, which are based on their contributions and the user's auxiliary information. The proposed models make use of the concatenation between the generator (*G)* and discriminator (*D*); the generator constructs probable user-item interactions in the target domain, guided by auxiliary information, meanwhile the discriminator distinguishes between real interactions and those generated by the generator.

- Generator function as in Eq. (1):

$$G(z, a_{u,}\, a_{i\,,}) \rightarrow \sigma(W_g .ReLU(V_{g\,.[z,x]}\ ))$$

$$\sigma(\ W_g.ReLU(V_{g\,.[z,x]}\,)) = \hat{r}_{u,i}$$

$$\text{Therefore, } G(z, a_{u,}\, a_{i\,,} \rightarrow \hat{r}_{u,i}) \qquad (1)$$

Generator loss function:

$$L_G\text{: } LG = -Elog(1 - D(G(z, a_u, a_i)))] \qquad (2)$$

The generator loss function was trained to minimize the likelihood of the discriminator by correctly identifying fake data.

- Discriminator loss function:

The discriminator loss function was equally trained to maximize the likelihood of correctly identifying real and synthetic data.

$$D(y) \rightarrow \sigma(W_d .ReLU(V_{d.y}\ ))$$

$$\text{Let } \sigma(W_d .ReLU(V_{d.y}\ ) = [0,1]$$

Therefore in Eq. (3):

$$D(ru, i, \hat{r}u.\, i\,) \rightarrow [0,1] \qquad (3)$$

where, Discriminator Loss (L_D):

$$LD = -E[log(D(r_{u,i}) + log(1 - D(G(z, a_u, a_i) \qquad (4)$$

F. Training Procedures

The training of the model involves alternating between updating the generator to fool the discriminator and updating the discriminator to correctly classify between the real and generated interactions. The loss functions are defined as Eq. (5):

$$LD = -E[log(D(r_{u,i})) + log(1 - D(G(z, a_{u,}\, a_i)))] \qquad (5)$$

where, $E[\dots]$ defines a minimization problem for the discriminator's loss, this is the fitting standard practice in GANs training, where the discriminator's objective is to minimize its loss. Eq. (2) and Eq. (4) illustrate the loss function for the generator and discriminator in the GANs setup. The discriminator (*D*) produces output with a probability close to 1 for the real data $r_{u,i}$. The logarithm of this probability measures how confident the discriminator is that $r_{u,i}$ is real, as in Eq. (6):

$$log(1 - D(G(z, a_{u,}\, a_i))) \qquad (6)$$

The generator (*G*) produces a synthetic data point $G(z, a_{u,}\, a_i)$ using the noise vector (*z)* and attributes (a_u) and (a_i). The discriminator (*D*) is trained to give an output of a probability close to 0 for synthetic data. Taking the logarithm of one minus the probability measures how confident the discriminator is that $G(z, a_{u,}\, a_i)$ is fake. The discriminator is trained to differentiate between real data points $r_{u,i}$ and generated data points $G(z, a_{u,}\, a_i)$.

The discriminator also aims to maximize this loss function, which means:

- Maximizing $log(D(r_{u,i})$ to ensure that it correctly identifies real data.
- Maximizing $log(1 - D(D(z.\, a_u,\ a_i)))$ to ensure that it correctly identifies generated (fake) data.

Table I highlights the breakdown of the different notations utilized in this study.

G. Data Representation

The proposed AUXIGAN model addresses challenges in traditional GAN models, specifically their struggle to accurately model the distribution of user and item rating matrices in a single iteration, to mitigate the instability of training and mode collapse, the model adopts a structured approach, where the discriminator (*D*) guides generator (*G*), both *G* and *D* are implemented using a multi-layer perceptron (MLP). In our proposed model, the generator (*G)* is tasked to make fake vectors based on varying user condition vectors R_t^u

at different time point *t*. These fake vectors are evaluated by *D* using the Wasserstein distance metric against real vectors from the dataset to enhance the model's robustness. This iterative process helps to improve the reliability and diversity of generated user-item interactions, addressing the limitations of traditional GANs. The following steps are equally followed for the data representations.

- Step 1: Generate fake samples with the generator.

Data representation on user (*u)* involves concatenating the conditional rating vector R_t^u, which includes noise, with its corresponding embedding vector v_t of *t*. This concatenated vector serves as the conditional input; additionally, random noise (z_u) is injected into the generator (*G)* as in Eq. (7):

$$G(R_{t,}^u, zu, vt) = \sigma(W_g . ReLU(V_g . [R_{t,\ zu,\ vt]}^u)) \quad (7)$$

Therefore, generator (*G)* considers only the loss gradients of rated items by ensuring more efficient learning by focusing on the user's observed interactions with items and the auxiliary features. It assists the generator *(G)* to denoise in the latent vector space to achieve an initial fake rating vector $\hat{R}_0^u$, which is a complete dense matrix.

- Step 2: Training of the discriminator

Firstly, the discriminator (D) was trained so that sharing of the real samples $R_{t-1,}^u$ will be known and obtained through the forward process given R_0^u and R_t^u. Then, the discriminator (D) also identified the distribution of the fake samples through the backward process, as shown in the proposed architecture. That is, the sampling of R_{t-1}^u by using subsequent distribution $q(R_{t-1}^u \mid R_{0,}^u, R_{t,}^u)$ given that ($R_{0,}^u$and $R_{t,}^u$) exit. Here, the transfer layer plays a significant role by filtering out the unrated item vectors to obtain the distribution of fake samples R_{t-1}^u. The discriminator loss LD and generator loss LG are defined as:

Discriminator Loss (LD) as in Eq. (8) :

$$L_D = -E_x \sim pdata[LogD(x)] - \text{Ez}\sim\text{pz}\ [\log(1\text{-}D(Gz,x)))] \quad (8)$$

Generator Loss (LG) as in Eq. (9):

$$LG = -Ez \sim pz[logD(G(z,x))] \quad (9)$$

- Step 3: Optimization process

The optimization process involves updating the parameters of the generator (*G*) and the discriminator (*D*) using gradient descent. Adam with modified hyperparameters was utilized for the optimization process.

The latent item variables can be accomplished by combining the user's information and items information and latent auxiliary variables, as: $vj = v_a\Delta + z_i$.

H. Incorporating Auxiliary Information

To handle varying lengths of auxiliary information (p_u) for different users effectively, an attention mechanism was utilized to enable each piece of auxiliary features given inversely to the user-item interaction. Through this approach, latent vectors were derived for auxiliary features of the users (*U*) and items (*I*), denoted as x and y, respectively. This allows more representation, where the attention mechanism dynamically weights the importance of different elements within the auxiliary information vectors, for the enhancement of the model's ability to capture diverse user-item interactions effectively. Auxiliary information's numbers (p_u) for different users typically vary; to utilize this, it's necessary to convert these varying-length vectors into a fixed-length vector. The simplest technique is to find the average of all embedding vectors; therefore, we utilized an attention mechanism to influence in diverse ways the user-item interaction, and the latent vectors of the auxiliary information of the users (*U*) and items (*I*) through x and y were then obtained as Eq. (10) and (11):

$$x = \sum_{i=1}^{pu} wd\ x_i \quad (10)$$

$$y = \sum_{i=1}^{qu} vd\ y_j \quad (11)$$

where, *wd* and *vd* represent weights and attentions of *i-th* user's auxiliary features and *j-th* item's auxiliary features, we therefore parameterized *vd* as an attention function with **u**, **v** and **y***j* as input in Eq. (12):

$$vd = fu\ (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{y}i) \quad (12)$$

One of the advantages of parameterizing is that, weights of auxiliary features depend on the auxiliary information and the pairing of user-item. Then, Multi-Layer Perception (MLP) was utilized for the parameterization of attention functions: $f_{\Phi}(.)$ *and* $h_{\theta}(.)$. Then the attention functions are as follows, as in Eq. (13 and (14):

$$fu\ (\mathbf{u}, \mathbf{v}, \mathbf{x}i) = h_u^T \Phi\ \{W_u \begin{bmatrix} u \\ i \\ xi \end{bmatrix} + b_u\} \quad (13)$$

$$fu\ (\mathbf{u}, \mathbf{v}, \mathbf{y}i) = h_u^T \Phi\ \{W_I \begin{bmatrix} u \\ i \\ yi \end{bmatrix} + b_I\} \quad (14)$$

Where $W_{u,}\ b_u, W_{I,}\ b_I$ and *Wu*, represents weight matrix and the bias vectors for the user-item auxiliary attention functions accordingly. The symbol (Φ) represents the non-linear activation functions, h_u and h_I are the weight vectors that aggregate the output layers of the attention networks to obtain the final attentions weight.

I. Prediction of the Model

After the convergence of the model, we employed a three-step process for predictions. Firstly, in Step 1, noise *(/z/)* was added to the vector of the sparse rating of users *T* times to obtain R_T. This step helps to incorporate noise present in the user-item rating matrix. In Step 2, initial predictions $\hat{R}_0$ were generated in the latent vector space using the conditional vector R_T. Finally, in Step 3, we applied denoising (T′< T) times to get the final prediction outcomes ($R_{T-T'}$). This approach aims to preserve personalized information while retaining some of the noise introduced in the latent vector space. This strategy is beneficial for capturing more user preferences and effectively enhancing the personalized performance of the CRRS.

IV. Experimental Setup

The datasets and metrics utilized to evaluate the performance of the proposed model are described in this section. The model is compared with other state-of-the-art methods after conducting the experiments.

A. Dataset Preparation

The experiments were conducted using Book-crossing and Amazon datasets, these datasets provide the necessary features to test the efficiency of the proposed model. Statistical details of the datasets are summarized in Table I. The table provides details of datasets (Amazon and Book-crossing), with a focus on their auxiliary information features utilized to enhance recommendation algorithms. Each of the datasets was randomly divided into two subsets: 80% for the training set and 20% for the testing set. The experiments involved training the model with the training data, while the testing set was utilized to validate the efficiency of the recommendations, optimize and adjust parameters of the model.

TABLE I. Statistics of the Datasets

Dataset	Ratings	Users	Items	Sparsity	Auxiliary Features
Amazon	1,000,209	3,706	6,040	95.55%	Price, product name, date, brand
Book-crossing	54,694	2,274	3,211	99.25%	Country, city and age

- Clarifications of the Datasets

From Table I, the Amazon dataset has over one million ratings matrix, 3,706 users, and 6,040 items, with 95.55% sparsity. The meaning of this is that only 4.45% of the possible user-item interactions have ratings. Auxiliary data, which includes price, product name, date, and brand, is useful for content-based recommendations. Book-crossing dataset is smaller, with only 54,694 ratings matrix, 2,274 users, and 3,211 items. This dataset has a very high 99.25% sparsity, which means that only 0.75% of all possible user-item ratings are present. Auxiliary data, which includes age, country, and city will help in more personalized recommendations. The *Amazon dataset is considered to be the source domain, while Book-crossing was utilized as the target domain.*

B. Metrics Utilized for the Evaluation

The four commonly utilized evaluation metrics are Precision, Recall, MAE, and RMSE, are utilized to evaluate the performance of the proposed model.

- Precision**:** Precision reflects how the proportions of items recommended are relevant to the user as in Eq. (15):

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FN}} \tag{15}$$

- Recall**:** It measures the proportion of relevant item that are successfully recommended as in Eq. (16) :

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FP}} \tag{16}$$

- Mean absolute Error (MAE): This measures the average magnitude of the errors between predicted ratings as in Eq. (17):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i \tag{17}$$

- Root Mean Square Error (RMSE): It is utilized to evaluate accuracy of the prediction in RS and machine learning as in Eq. (18):

$$\text{RMSE} = \sqrt{\sum_{i=1}^{T} \frac{(\hat{y}-y)^2}{T}} \tag{18}$$

C. Validation of Models with Baselines

To validate the effectiveness of the proposed model, we selected seventeen (17) recommender system models for comparative performance analysis. The baselines include cross-domain recommender systems models, multi-task recommender systems models, and adversarial learning-based recommender systems models, which are related to our recommendations design.

1) Cross-domain recommender systems models.

- DML [52]: Dual Metric Learning introduces a unique latent metric mapping function to derive user preferences by maintaining relationships among users across domains.
- DASL [53]: The model incorporates two innovative components, Dual Embedding and Dual Attention, to facilitate cross-domain recommendations.
- CCCFNet [54]: The Cross-domain Content-boosted Collaborative Filtering Neural Network combines collaborative filtering and content-based filtering using a unified multi-view neural network through factorization.
- CoNet [55]: Collaborative Cross Networks supports knowledge transfer across domains by introducing cross-connections between the base networks.
- EMCDR [56]: The Embedding and Mapping method for Cross-Domain Recommendation (EMCDR) employs a multi-layer perceptron to model the nonlinear mapping function between domains.
- SSCDR [57]: The Semi-Supervised Cross-Domain Recommendation (SSCDR) model learns latent representations for users and items while training a mapping function to capture distance relationships across domains.

2) Multi-task recommender systems models.

- Cross-Stitch [58]**:** The model integrates activations from multiple neural networks and the model learns the best balance between shared and task-specific representations which results to a significant performance boost for categories with limited training data.

- MMOEWD [61]**:** model refines the MMOE structure through soft-parameter sharing techniques. The model optimizes multiple ranking objectives and addresses selection biases by employing a Wide and Deep methodology.
- MMOE [59]**:** It extends the MoE framework. It allows expert sub models to be shared across tasks. The approach enhances trainability by adapting varying randomness in the model initialization and data.
- DARec [63]**:** The Deep Domain Adaptation Recommendation model applies domain adaptation and multi-task learning to deliver cross-domain recommendations. The model identifies and transfers patterns from rating matrices without relying on auxiliary information.

3) Adversarial learning-based recommender systems models.

- RecGAN [60]**:** The model utilizes gated recurrent unit cells to capture latent features of users and items from both short-term and long-term temporal profiles, enhancing recommendation relevance.
- PD-GAN [62]**:** This model generates diverse yet relevant recommendations by sequentially sampling from a DPP kernel matrix. The adversarial training with a discriminator helps recognize or differentiate between generated recommendations and actual user interactions.
- RecSys-DAN [64]**:** This model tackles the data sparsity challenge by learning transferable latent representations for the user, item, and the interactions through adversarial training.
- DAAN [65]**:** The DAAN model combines collaborative filtering through deep adversarial domain adaptation with matrix factorization.
- DSAP-AL [35]**:** This model integrates matrix factorization and adversarial learning to align user and item latent factor spaces.
- CDR-HA [66]**:** The model enhances recommendation accuracy and mitigates data sparsity in the target domain by integrating heterogeneous information network-based adversarial learning with cross-domain adversarial techniques.
- ACDR [67]: The study proposes an Adversarial Cross Domain Recommendation (ACDR) model. The model utilizes adversarial learning for better user embeddings. The model improves recommendation performance on sparse datasets.

Table II highlights the values of each hyperparameter utilized for the proposed model.

TABLE II. HYPERPARAMETER AND THEIR VALUES

Hyperparameters	Parameter values
Generator learning rate	0.0001
Discriminator learning rate	0.0001
Latent dimension	128
Batch size	64
Number of epochs	20
Embedding size	100
Dropout rate	0.2
Learning rate decay	0.1
Optimizer	Adam
Activation functions	ReLU
Domain adaptation learning rate	0.0001

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Results

This subsection shows the experimental results of the proposed model. The experiment was conducted on Book-crossing and Amazon datasets, and the model was compared with the baseline using evaluation metrics: RMSE, MAE and Precision, Recall with 20 epochs on each metric as shown in Fig. 3 to Fig. 6. The proposed AUXIGAN model is the most effective model across all metrics, making it the top choice for high-accuracy recommendations on the datasets. Table III and Table IV show that the proposed AUXIGAN model outperforms the baseline models across all metrics, which makes it the most accurate model for recommendations on the datasets.

TABLE III. PERFORMANCE ON AMAZON (MUSIC) - BOOK-CROSSING

Model	Amazon (Music) - Book-crossing			
	RMSE	***MAE***	***Precision***	***Recall***
DML	0.2205	0.1706	0.8583	0.9595
DASL	0.2202	0.1705	0.8595	0.9597
CCCFNet	0.2639	0.1841	0.8102	0.8872
CoNet	0.2305	0.1892	0.8328	0.8990
EMCDR	0.2192	0.1789	0.8516	0.9307
SSCDR	0.2295	0.1744	0.8520	0.9341
Cross-Stitch	0.2465	0.1972	0.8339	0.9111
MMOE	0.2388	0.1806	0.8444	0.9234
MMOEWD	0.2332	0.1772	0.8502	0.9278
DARec	0.2376	0.1796	0.8495	0.9318
RecSys-DAN	0.2665	0.1980	0.8287	0.9220
PD-GAN	0.2693	0.2002	0.8260	0.9231
RecGAN	0.2708	0.2007	0.8260	0.9198
DAAN	0.2672	0.1982	0.8291	0.9242
DSAP-AL	0.2663	0.1979	0.8288	0.9238
CDR-HA	0.2679	0.1991	0.8274	0.9233
ACDR	0.1347	0.0995	0.8972	0.9790
AUXI-GAN	**0.0998**	**0.0774**	**0.9666**	**0.9938**

TABLE IV. PERFORMANCE ON AMAZON (MOVIES) - BOOK-CROSSING

Model	Amazon (Movies) - Book-crossing			
	RMSE	***MAE***	***Precision***	***Recall***
DML	0.2683	0.2093	0.8714	0.9089
DASL	0.2681	0.2088	0.8717	0.9095
CCCFNet	0.2711	0.2116	0.8664	0.9046

Model	Amazon (Movies) - Book-crossing			
	RMSE	***MAE***	***Precision***	***Recall***
CoNet	0.2685	0.2078	0.8722	0.9046
EMCDR	0.2704	0.2092	0.8729	0.9009
SSCDR	0.2699	0.2100	0.8677	0.9021
Cross-Stitch	0.2791	0.2150	0.8605	0.8929
MMOE	0.2778	0.2127	0.8633	0.8946
MMOEWD	0.2723	0.2111	0.8693	0.8998
DARec	0.2701	0.2095	0.8717	0.9013
RecSys-DAN	0.2769	0.2154	0.8675	0.8979
PD-GAN	0.2796	0.2165	0.8646	0.8958
RecGAN	0.2803	0.2187	0.8627	0.8944
DAAN	0.2722	0.2127	0.8719	0.9005
DSAP-AL	0.2736	0.2138	0.8708	0.8998
CDR-HA	0.2754	0.2166	0.8683	0.8976
ACDR	0.2431	0.1919	0.8846	0.9237
AUXI-GAN	**0.1350**	**0.0908**	**0.9031**	**0.9531**

The performance of experiments is presented in Fig. 3 to Fig. 6. We have the following findings:

Fig. 3 presents evaluation metrics over 20 epochs for the Amazon (Music) - Book-crossing datasets. *Precision* improves over epochs, which indicates better accuracy in recommending relevant items. *Recall* consistently increases, and it reflects improved capacity to capture relevant recommendations. *MAE* remains stable, which suggests a low mean error between the predicted and actual ratings. *RMSE* decreases steadily, which shows reduced prediction errors over the epochs. This demonstrates performance improvements across the evaluation metrics, which indicates the model's effectiveness.

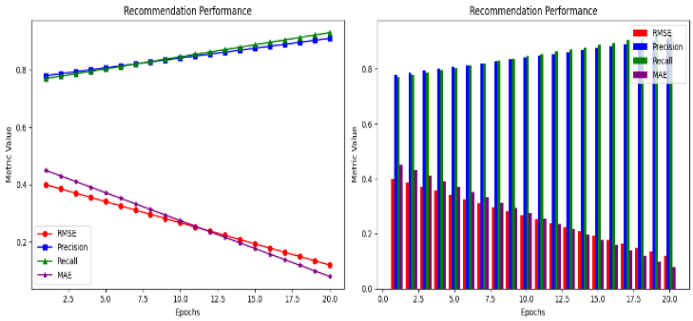


Fig. 3. Performance of proposed model on Amazon (Music)-Book-crossing.

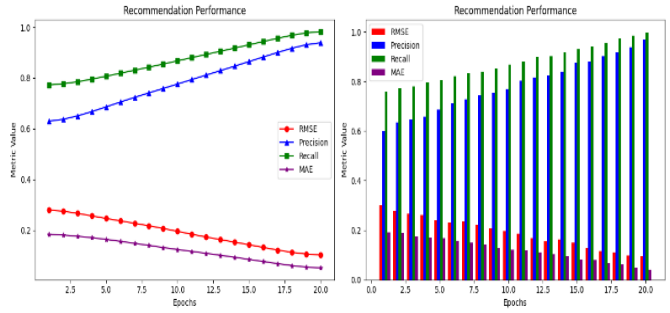


Fig. 4. Performance of proposed model on Amazon (Movie)-Book-crossing.

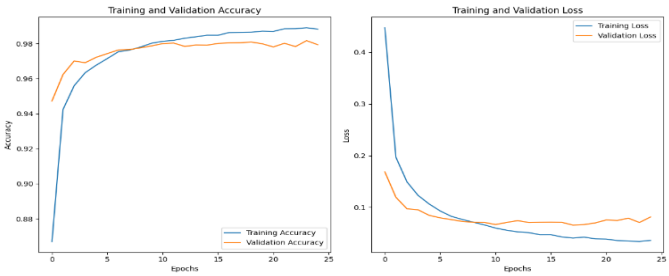


Fig. 5. Training, validation accuracy and validation loss.

Fig. 4 illustrates the evaluation metrics over 20 epochs for the Amazon (Movie) – Book-crossing datasets. The key observations are: *Precision* improves consistently, which indicates better accuracy in recommendations over time. *Recall* increases steadily, which shows improved coverage of the relevant recommendations. *MAE* decreases gradually, demonstrating reduced mean error between predicted and actual ratings. *RMSE* decreases over 20 epochs, indicating less prediction errors. The model demonstrates improved performance across all the evaluation metrics, and Fig. 5 shows the training, validation accuracy and validation loss results.

B. The Parameter Settings and Model Training

The relevant hyperparameters used in the proposed model are as follows: we conducted the experiments with Top-N recommendations, whereby N represents the number of items recommended to the users; the parameters are shown in Table II.

C. Baseline Comparison

The proposed model demonstrates a better performance compared to state-of-the-art methods across all metrics. Table III and Table IV summarize the results, which indicate that the proposed model not only outperforms the baselines but also opens new opportunities to address the challenges of data sparsity on CDRS.

D. Data Sparsity Performance

From the proposed model, as shown in Table IV and Table V, we discovered that the model significantly mitigated data sparsity problems by incorporating auxiliary information into the GANs-based CDRS. In the data sparsity problem, there are improvements in recommendation accuracy, underscoring the influence of auxiliary information to enhance recommendation performance when there is limited data.

VI. Ablation Experiments

As highlighted in Section V, the proposed model demonstrates significant improvements compared to the selected baseline models. These enhancements stem from the innovative design of the model, which integrates auxiliary information into the GAN-based model recommendation techniques. In this segment, we perform an ablation study to validate the contribution of auxiliary information into the proposed model. We removed auxiliary information from the model and used only the rating matrix of the users-items.

The following variations were tested to examine the influence of the auxiliary information:

- Model A (without auxiliary information): The model excludes auxiliary information, which relies solely on GAN-generated interactions.
- Model B (with auxiliary information): The model includes auxiliary information with GAN-generated interactions.

TABLE V. ABLATION STUDY RESULTS ON AMAZON

Model	Amazon (Movie) - Book-crossing datasets			
	RMSE	***MAE***	***Precision***	***Recall***
Model A	0.1980	0.1181	0.9020	0.9720
Model B	**0.0998**	**0.0774**	**0.9666**	**0.9938**

As shown in Table V, the proposed model proves robust performances in transferring knowledge across domains, with increase in recommendation accuracy for data sparsity when auxiliary information is incorporated into the model. Fig. 6 shows the experiments on the Book-crossing and Amazon datasets with evaluation metrics (RMSE, MAE, Precision and Recall) respectively.

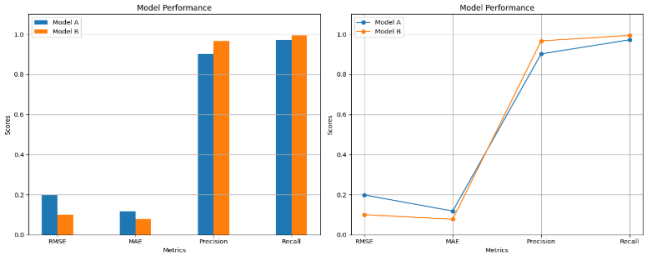


Fig. 6. Ablation results of RMSE, MAE, Precision and Recall on Amazon (Music)-Book-crossing datasets.

VII. LIMITATIONS AND FUTURE DIRECTIONS

This section presents the limitations of the proposed model and future directions that need to be explored.

A. Limitations

1) High computational complexity. Training GANs, especially when integrating auxiliary information, requires significant computational resources and time.

2) Overfitting. With the limited data or excessive auxiliary information, the model may encounter a problem of overfitting in the training set.

3) Hyperparameter sensitivity. The GANs model may require careful tuning of hyperparameters, such as learning rate, embedding dimensions, or loss functions, which may not be practical for all circumstances.

4) Inconsistent quality. The auxiliary information may be noisy, incomplete, or inconsistent.

B. Future Directions

Future directions may focus on integrating more complex auxiliary data and temporal dynamics to further enhance the recommendation accuracy. Since our findings indicates that GAN-based CDRS, when enhanced with auxiliary information can effectively capture user preferences, more personalized and accurate recommendations across different domains, therefore future studies may hybridize the generative AI models such as GANs and VAEs with the incorporation of auxiliary information for better performance.

VIII. CONCLUSION

Our findings from the conducted experiments highlight the positive influence of auxiliary information on GAN-based CDRS; the proposed model demonstrates improvements in data sparsity, recommendation accuracy, and robustness. Auxiliary information not only enhances the quality of recommendations but also enables the model to generalize across domains by addressing challenges of data sparsity. The integration of auxiliary information on GAN-based CDRS achieved a more comprehensive representation of user and item preferences. The adversarial training framework effectively generates realistic user-item interactions, addressing data sparsity issues. The incorporation of auxiliary information assisted the GAN-based model to capture more details about users' preferences and item features. This enhancement led to significant improvements in core recommendation metrics, with auxiliary-enhanced models achieving gains in Precision, recall, RMSE and MAE across the two datasets. These results significantly showed that auxiliary information aids GAN-based models to better understand and represent user-item interactions, especially in the data sparsity problem. This research also encourages more exploration into transfer learning techniques within CDRS. The findings suggest that incorporating auxiliary information could be an introductory approach for the development of advanced multi-domain recommendation solutions on GAN-based models by extending beyond traditional RSs techniques. To the best of our knowledge, auxiliary information has not been integrated into GANs-based CDRS. Previous authors believed that generative AI models such as GANs had the ability to generate new content or items, but GANs model still depends on the data or information it is been trained on. GANs have the capability to generate new, more personalized, accurate, and robust recommendations with the incorporation of auxiliary information on CDRS.

REFERENCES

[1] Q. Hu and Y. Song, "User Consented Federated Recommender System Against Personalized Attribute Inference Attack," in *WSDM 2024 - Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, Association for Computing Machinery, Inc, Mar. 2024, pp. 276–285. doi: 10.1145/3616855.3635830.

[2] P. Zhao, Y. Jin, X. Ren, and Y. Li, "A personalized cross-domain recommendation with federated meta learning," *Multimed Tools Appl*, 2024, doi: 10.1007/s11042-024-18495-3.

[3] R. Alatrash, R. Priyadarshini, H. Ezaldeen, and A. Alhinnawi, "Augmented language model with deep learning adaptation on sentiment analysis for E-learning recommendation," *Cogn Syst Res*, vol. 75, pp. 53–69, Sep. 2022, doi: 10.1016/j.cogsys.2022.07.002.

[4] H. Hamidi and R. Moradi, "Design of a dynamic and robust recommender system based on item context, trust, rating matrix and rating time using social networks analysis," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 2, p. 101964, Feb. 2024, doi: 10.1016/j.jksuci.2024.101964.

[5] M. Acharya, S. Yadav, and K. K. Mohbey, "How can we create a recommender system for tourism? A location centric spatial binning-

based methodology using social networks," *International Journal of Information Management Data Insights*, vol. 3, no. 1, Apr. 2023, doi: 10.1016/j.jjimei.2023.100161.

[6] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.2661

[7] Y. Deldjoo, T. Di Noia, and F. A. Merra, "A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial Networks," Apr. 01, 2021, *Association for Computing Machinery*. doi: 10.1145/3439729.

[8] T. Sasagawa, S. Kawai, and H. Nobuhara, "Recommendation system based on generative adversarial network with graph convolutional layers," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 25, no. 4, pp. 389–396, Jul. 2021, doi: 10.20965/JACIII.2021.P0389.

[9] J. Bobadilla, A. Gutiérrez, R. Yera, and L. Martínez, "Creating synthetic datasets for collaborative filtering recommender systems using generative adversarial networks," *Knowl Based Syst*, vol. 280, Nov. 2023, doi: 10.1016/j.knosys.2023.111016.

[10] J. Li, Y. Ren, and K. Deng, "FairGAN: GANs-based Fairness-aware Learning for Recommendations with Implicit Feedback," in *WWW 2022 - Proceedings of the ACM Web Conference 2022*, Association for Computing Machinery, Inc, Apr. 2022, pp. 297–307. doi: 10.1145/3485447.3511958.

[11] W. Shafqat and Y. C. Byun, "A Hybrid GAN-Based Approach to Solve Imbalanced Data Problem in Recommendation Systems," *IEEE Access*, vol. 10, pp. 11036–11047, 2022, doi: 10.1109/ACCESS.2022.3141776.

[12] X. Li, K. Chen, Y. Zhou, and Y. Chen, "A Generative Adversarial Approach with Social Relationship for Recommender Systems," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Sep. 2023. doi: 10.1145/3630138.3630424.

[13] J. Li, Y. Ren, and K. Deng, "FairGAN: GANs-based Fairness-aware Learning for Recommendations with Implicit Feedback," in *WWW 2022 - Proceedings of the ACM Web Conference 2022*, Association for Computing Machinery, Inc, Apr. 2022, pp. 297–307. doi: 10.1145/3485447.3511958.

[14] Y. Zhang, I. W. Tsang, and L. Duan, "Collaborative Generative Hashing for Marketing and Fast Cold-Start Recommendation," *IEEE Intell Syst*, vol. 35, no. 5, pp. 84–95, Sep. 2020, doi: 10.1109/MIS.2020.3025197.

[15] Y. Liu, S. Wang, X. Li, and F. Sun, "A Meta-adversarial Framework for Cross-Domain Cold-Start Recommendation," *Data Sci Eng*, 2024, doi: 10.1007/s41019-024-00245-y.

[16] X. M. Li, S. S. Wang, F. M. Gu, and Z. B. Lin, "Enhancing signed social recommendation via extracting auxiliary textual information," *Multimed Tools Appl*, 2023, doi: 10.1007/s11042-023-17414-2.

[17] X. Zhang and M. Gan, "Hi-GNN: hierarchical interactive graph neural networks for auxiliary information-enhanced recommendation," *Knowl Inf Syst*, vol. 66, no. 1, pp. 115–145, Jan. 2024, doi: 10.1007/s10115-023-01949-9.

[18] R. Qiu, Z. Huang, J. Li, and H. Yin, "Exploiting Cross-session Information for Session-based Recommendation with Graph Neural Networks," *ACM Trans Inf Syst*, vol. 38, no. 3, Jun. 2020, doi: 10.1145/3382764.

[19] H. Xia, K. Huang, and Y. Liu, "Unexpected interest recommender system with graph neural network," *Complex and Intelligent Systems*, vol. 9, no. 4, pp. 3819–3833, Aug. 2023, doi: 10.1007/s40747-022-00849-9.

[20] H. Chen *et al.*, "Macro Graph Neural Networks for Online Billion-Scale Recommender Systems," *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13â•fi17, 2024, Singapore, Singapore*, vol. 1, doi: 10.1145/XXXXXX.XXXXXX.

[21] M. Liu, J. Li, Z. Guo, and G. Li, "Extracting latently overlapping users by graph neural network for non-overlapping cross-domain recommendation," *Knowl Based Syst*, p. 111508, Apr. 2024, doi: 10.1016/j.knosys.2024.111508.

[22] R. Khan, N. Iltaf, R. Latif, and N. S. M. Jamail, "CrossDomain Recommendation Based on MetaData Using Graph Convolution Networks," *IEEE Access*, vol. 11, pp. 90724–90738, 2023, doi: 10.1109/ACCESS.2023.3307015.

[23] H. Kuang, W. Xia, X. Ma, and X. Liu, "Deep Matrix Factorization for Cross-Domain Recommendation," in *IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2021, pp. 2171–2175. doi: 10.1109/IAEAC50856.2021.9390866.

[24] J. Liu, W. Pan, and Z. Ming, "CoFiGAN: Collaborative filtering by generative and discriminative training for one-class recommendation ☆," vol. 191, p. 105255, 2020, doi: 10.1016/j.knosys.

[25] A. Gkillas and D. Kosmopoulos, "A cross-domain recommender system using deep coupled autoencoders," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.07617

[26] C. D. Wang, Y. H. Chen, W. D. Xi, L. Huang, and G. Xie, "Cross-Domain Explicit-Implicit-Mixed Collaborative Filtering Neural Network," *IEEE Trans Syst Man Cybern Syst*, vol. 52, no. 11, pp. 6983–6997, Nov. 2022, doi: 10.1109/TSMC.2021.3129261.

[27] J. Yang, J. Zhu, X. Ding, Y. Peng, and Y. Zhang, "A memory pool variational autoencoder framework for cross-domain recommendation," *Expert Syst Appl*, vol. 241, May 2024, doi: 10.1016/j.eswa.2023.122771.

[28] D. Xu, C. Ruan, E. Korpeoglu, S. Kumar, and K. Achan, "Rethinking Neural vs. Matrix-Factorization Collaborative Filtering: the Theoretical Perspectives," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.12141

[29] X. Tian, "Content-based Filtering for Improving Movie Recommender System," 2024, pp. 598–609. doi: 10.2991/978-94-6463-370-2_61.

[30] T. Zheng, S. Li, Y. Liu, Z. Zhang, and M. Jiang, "An effective neighbor information mining and fusion method for recommender systems based on generative adversarial network," *Expert Syst Appl*, vol. 248, Aug. 2024, doi: 10.1016/j.eswa.2024.123396.

[31] Y. Liu, S. Wang, X. Li, and F. Sun, "A Meta-adversarial Framework for Cross-Domain Cold-Start Recommendation," *Data Sci Eng*, 2024, doi: 10.1007/s41019-024-00245-y.

[32] S. Z. U. Hassan, M. Rafi, and J. Frnda, "GCZRec: Generative Collaborative Zero-Shot Framework for Cold Start News Recommendation," *IEEE Access*, vol. 12, pp. 16610–16620, 2024, doi: 10.1109/ACCESS.2024.3359053.

[33] M. O. Ayemowa, R. Ibrahim, and M. M. Khan, "Analysis of Recommender System Using Generative Artificial Intelligence: A Systematic Literature Review," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3416962.

[34] B. Qin, Z. Huang, X. Tian, Y. Chen, and W. Wang, "GACRec: Generative adversarial contrastive learning for improved long-tail item recommendation," *Knowl Based Syst*, vol. 300, Sep. 2024, doi: 10.1016/j.knosys.2024.112146.

[35] Y. Li, J. Ren, J. Liu, and Y. Chang, "Deep sparse autoencoder prediction model based on adversarial learning for cross-domain recommendations," *Knowl Based Syst*, vol. 220, May 2021, doi: 10.1016/j.knosys.2021.106948.

[36] M. Nanthini and K. P. M. Kumar, "Provisioning a cross-domain recommender system using an adaptive adversarial network model," *Soft comput*, vol. 27, no. 24, pp. 19197–19212, Dec. 2023, doi: 10.1007/s00500-023-09360-w.

[37] H. Su, Y. Zhang, X. Yang, H. Hua, S. Wang, and J. Li, "Cross-domain Recommendation via Adversarial Adaptation," in *International Conference on Information and Knowledge Management, Proceedings*, Association for Computing Machinery, Oct. 2022, pp. 1808–1817. doi: 10.1145/3511808.3557277.

[38] C. W. Hsu, C. T. Chen, and S. H. Huang, "Adaptive Adversarial Contrastive Learning for Cross-Domain Recommendation," *ACM Trans Knowl Discov Data*, vol. 18, no. 3, Dec. 2023, doi: 10.1145/3630259.

[39] S. Wang, W. Wan, T. Qu, and Y. Dong, "Auxiliary information-enhanced recommendations," *Applied Sciences (Switzerland)*, vol. 11, no. 19, Oct. 2021, doi: 10.3390/app11198830.

[40] Y. Chen, X. Qiu, C. Ma, Y. Xu, and Y. Sun, "A recommender system fused with implicit social information through network representation learning," *Computers and Electrical Engineering*, vol. 100, May 2022, doi: 10.1016/j.compeleceng.2022.107897.

[41] A. K. Yengikand, M. Meghdadi, and S. Ahmadian, "DHSIRS: a novel deep hybrid side information-based recommender system," *Multimed Tools Appl*, vol. 82, no. 22, pp. 34513–34539, Sep. 2023, doi: 10.1007/s11042-023-15021-9.

[42] L. Chen, Z. Wu, J. Cao, G. Zhu, and Y. Ge, "Travel recommendation via fusing multi-auxiliary information into matrix factorization," *ACM Trans Intell Syst Technol*, vol. 11, no. 2, Jan. 2020, doi: 10.1145/3372118.

[43] M. Varasteh, M. S. Nejad, H. Moradi, M. A. Sadeghi, and A. Kalhor, "An Improved Hybrid Recommender System: Integrating Document Context-Based and Behavior-Based Methods," in *2023 31st International Conference on Electrical Engineering, ICEE 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 881–887. doi: 10.1109/ICEE59167.2023.10334913.

[44] H. Li *et al*., "CTITF: A tensor factorization model with constrained bidirectional user trust and implicit feedback for context-aware recommender systems," *Inf Sci (N Y)*, vol. 676, Aug. 2024, doi: 10.1016/j.ins.2024.120838.

[45] A. Pujahari and D. S. Sisodia, "Ordinal consistency based matrix factorization model for exploiting side information in collaborative filtering," *Inf Sci (N Y)*, vol. 643, Sep. 2023, doi: 10.1016/j.ins.2023.119258.

[46] T. N. Dinh, P. Pham, G. L. Nguyen, and B. Vo, "Enhanced context-aware citation recommendation with auxiliary textual information based on an auto-encoding mechanism", doi: 10.1007/s10489-022-04423-1/Published.

[47] J. Zhang, J. Ding, Q. Fan, and Y. Li, "Holographic Knowledge-Auxiliary Information Graph Aggregation for Recommendation," in *2023 3rd International Conference on Electronic Information Engineering and Computer Science, EIECS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 393–396. doi: 10.1109/EIECS59936.2023.10435398.

[48] H. Xiao, D. Sun, F. You, and D. Wang, "Metapath2vec based Fusion Auxiliary Information Embedding for Music Recommendation," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Oct. 2023, pp. 218–221. doi: 10.1145/3644523.3644564.

[49] J. Yi, X. Ren, and Z. Chen, "Multi-Auxiliary Augmented Collaborative Variational Auto-encoder for Tag Recommendation," *ACM Trans Inf Syst*, vol. 41, no. 4, Apr. 2023, doi: 10.1145/3578932.

[50] J. Ding, G. Yu, Y. Li, X. He, and D. Jin, "Improving implicit recommender systems with auxiliary data," *ACM Trans Inf Syst*, vol. 38, no. 1, Feb. 2020, doi: 10.1145/3372338.

[51] J. Bai and Z. Ban, "Collaborative multi-auxiliary information variational autoencoder for recommender systems," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2019, pp. 501–505. doi: 10.1145/3318299.3318336.

[52] P. Li and A. Tuzhilin, "Dual Metric Learning for Effective and Efficient Cross-Domain Recommendations," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.08490

[53] P. Li, Z. Jiang, M. Que, Y. Hu, and A. Tuzhilin, "Dual Attentive Sequential Learning for Cross-Domain Click-Through Rate Prediction," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2021, pp. 3172–3180. doi: 10.1145/3447548.3467140.

[54] J. Lian, F. Zhang, X. Xie, and G. Sun, "CCCFNet: A content-boosted collaborative filtering neural network for cross domain recommender systems," in *26th International World Wide Web Conference 2017, WWW 2017 Companion*, International World Wide Web Conferences Steering Committee, 2017, pp. 817–818. doi: 10.1145/3041021.3054207.

[55] G. Hu, Y. Zhang, and Q. Yang, "Conet: Collaborative cross networks for cross-domain recommendation," in *International Conference on Information and Knowledge Management, Proceedings*, Association for Computing Machinery, Oct. 2018, pp. 667–676. doi: 10.1145/3269206.3271684.

[56] T. Man, H. Shen, X. Jin, and X. Cheng, "Cross-Domain Recommendation: An Embedding and Mapping Approach," 2017.

[57] S. K. Kang, J. Hwang, D. Lee, and H. Yu, "Semi-supervised learning for cross-domain recommendation to cold-start users," in *International Conference on Information and Knowledge Management, Proceedings*, Association for Computing Machinery, Nov. 2019, pp. 1563–1572. doi: 10.1145/3357384.3357914.

[58] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch Networks for Multi-task Learning," Apr. 2016, [Online]. Available: http://arxiv.org/abs/1604.03539

[59] Z. Zhao *et al*., "Recommending what video to watch next: A multitask ranking system," in *RecSys 2019 - 13th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, Sep. 2019, pp. 43–51. doi: 10.1145/3298689.3346997.

[60] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Jul. 2018, pp. 1930–1939. doi: 10.1145/3219819.3220007.

[61] F. Yuan, L. Yao, and B. Benatallah, "DARec: Deep Domain Adaptation for Cross-Domain Recommendation via Transferring Rating Patterns," May 2019, [Online]. Available: http://arxiv.org/abs/1905.10760

[62] H. Bharadhwaj, H. Park, and B. Y. Lim, "RecGan: Recurrent generative adversarial networks for recommendation systems," in *RecSys 2018 - 12th ACM Conference on Recommender Systems*, Association for Computing Machinery, Inc, Sep. 2018, pp. 372–376. doi: 10.1145/3240323.3240383.

[63] Q. Wu, Y. Liu, C. Miao, B. Zhao, Y. Zhao, and L. Guan, "PD-GAN: Adversarial Learning for Personalized Diversity-Promoting Recommendation," 2019.

[64] C. Wang, M. Niepert, and H. Li, "RecSys-DAN: Discriminative adversarial networks for cross-domain recommender systems," *IEEE Trans Neural Netw Learn Syst*, vol. 31, no. 8, pp. 2731–2740, Aug. 2020, doi: 10.1109/TNNLS.2019.2907430.

[65] H. Liu, L. Guo, P. Li, P. Zhao, and X. Wu, "Collaborative filtering with a deep adversarial and attention network for cross-domain recommendation," *Inf Sci (N Y)*, vol. 565, pp. 370–389, Jul. 2021, doi: 10.1016/j.ins.2021.02.009.

[66] C. Su, Z. Hu, and X. Xie, "Cross-Domain Recommendation based on Heterogeneous information network with Adversarial learning," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Apr. 2021, pp. 65–70. doi: 10.1145/3461598.3461609.

[67] P. Li, B. Brost, and A. Tuzhilin, "Adversarial Learning for Cross Domain Recommendations," *ACM Trans Intell Syst Technol*, vol. 14, no. 1, Nov. 2022, doi: 10.1145/354877.

The Anomaly Detection Algorithm Based on Random Matrix Theory and Machine Learning

Yongming Lu*
Ganzhou Teachers College, Ganzhou Jiangxi, 342800, China

***Abstract*—This study focuses on anomaly detection algorithms. Aiming at the limitations of traditional methods in complex data processing, an innovative algorithm that integrates random matrix theory and machine learning is proposed. First, different types of data, such as numerical values, texts, and images, are preprocessed, and random matrices are constructed. Hidden abnormal features are mined through specific transformations and then classified by optimized machine learning models. In the experimental stage, multiple data sets, such as KDD Cup 99, are selected to compare with classic algorithms such as DBSCAN and Isolation Forest. The results show that the innovative algorithm has a detection accuracy of 95%, a recall rate of 93%, and an F1 value of 94% on the KDD Cup 99 data set, which is significantly improved compared with the comparison algorithm. It also performs well on other data sets, with an average accuracy increase of seven percentage points and a recall rate increase of eight percentage points. The results demonstrate that the proposed algorithm can effectively mine data anomaly patterns, achieve efficient and accurate anomaly detection in complex data sets, and provide strong support for applications in related fields.**

***Keywords*—Random matrix theory; machine learning; anomaly detection; experimental simulation**

I. Introduction

In today's digital age, anomaly detection is critical in many areas such as network security, industry surveillance, and financial risk warning. In the field of network security, for example, early and accurate detection of malicious attacks is essential to prevent data breaches and to ensure the stability of the system. Early detection of equipment failures in industry monitoring can prevent production disruption and reduce economic losses. However, due to the exponential increase of data size and complexity of data structure, the traditional methods of anomaly detection are severely limited. For complex data distributions and high-dimensional data sets, conventional algorithms are often faced with low detection accuracy, high computational cost and poor adaptability [1]. For example, traditional methods fail to detect abnormal attack patterns embedded in massive data, while for industrial equipment monitoring, it is difficult to detect early signs of failure under complex operating conditions.

This study proposes a novel algorithm based on random matrix theory combined with machine learning. The proposed approach innovatively combines random matrix transformations with optimized machine learning classification models [2]. Three key steps are involved: comprehensive data preprocessing to construct a suitable random matrix, extracting hidden abnormal features via unique matrix transformation [3], and using optimized machine learning models for classification. It aims at improving detection efficiency, accuracy, and adaptability in complex data environments, meeting rigorous requirements for practical applications.

II. Anomaly Detection Algorithm Design

A. Unsupervised Algorithms

Unsupervised anomaly detection algorithms rely on inherent data distributions to detect outliers. For example, clustering methods based on density, such as DBSCAN, define anomalies as low-neighborhood data points. However, this approach often suffers from high-dimensional data and requires manual tuning of parameters. Another example is an isolation forest, which detects anomalies by measuring the path length in a decision tree ensemble. While efficient, it may incorrectly classify local low-density areas as anomalies.

1) Data preprocessing and matrix construction

For numerical data, normalization maps values to [0, 1] using Formula (1):

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

$x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the feature in the dataset, respectively.

Text data are collected to word vectors using the Skip-Gram model. For a text with n words, the word vectors form an $n \times d$ matrix W, where each row is the vector representation of a word [4].

For image data, grayscale processing is first performed to convert the color image into a grayscale image. The grayscale value calculation formula is as follows [Formula (2)]:

$$\text{Gray} = 0.299R + 0.587G + 0.114B \quad (2)$$

R, G, B are the red, green, and blue component values of the image pixels, respectively. Then, the image is resized to a fixed size. The size is set to m×n and converted into an m×n grayscale value matrix I.

After preprocessing, a $N \times D$ random matrix R is constructed. For numerical data, elements are [Formula (3)]:

$$r_{ij} = \alpha \hat{x}_{ij} + \beta \epsilon_{ij} \quad (3)$$

x_{ij} is the normalized feature, $\alpha = 0.8, \beta = 0.2$, and $\epsilon_{ij} \sim \mathcal{N}(0,1)$.

2) *New method for extracting features using random matrix transformation*. Transformations such as rotation, scaling, and singular value decomposition based on random

matrices are the core means of extracting features in this study. First, consider the rotation transformation [5]. Let the rotation matrix be $\boldsymbol{R}_\theta$. For the random matrix $\boldsymbol{R}$, the matrix $\boldsymbol{R}'$ after the rotation transformation can be obtained by Formula (4):

$$\mathbf{R}' = \mathbf{R}_\theta \mathbf{R} \tag{4}$$

The rotation matrix $\mathbf{R}_\theta$ can be constructed based on the rotation angle θ. For a two-dimensional rotation, $\mathbf{R}_\theta$ is in the form of [Formula (5)]:

$$\mathbf{R}_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{5}$$

In high-dimensional space, the construction of the rotation matrix is more complicated, but the principle is similar [6]. Through the rotation operation, the data features are redistributed in the new space, which helps to discover hidden abnormal patterns.

Scaling transformation highlights important features by adjusting the scale of the elements of the random matrix. Let the scaling matrix be **S**, and its diagonal elements are the scaling factors $s_1, s_2, \cdots, s_D$. Then the matrix $\mathbf{R}''$ after scaling transformation is Formula (6):

$$\mathbf{R}'' = \mathbf{S}\mathbf{R} \tag{6}$$

$\mathbf{S} = \text{diag}(s_1, s_2, \cdots, s_D)$. The scaling factor can be set according to the importance of the data feature. For example, for feature dimensions with a high correlation with abnormal information, a more extensive scaling factor can be set.

Singular value decomposition (SVD) is an important method for extracting features. Singular value decomposition of the random matrix **R** can be expressed as Formula (7):

$$\mathbf{R} = \mathbf{U}\Sigma\mathbf{V}^T \tag{7}$$

U and **V** are the left singular matrix and the right singular matrix, respectively, Σ is a diagonal matrix, and its diagonal elements $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_D \geq 0$ are the singular values of **R**. In anomaly detection, smaller singular values are often related to the abnormal part of the data. By retaining the components corresponding to some singular values, the features related to the anomaly can be extracted [7]. Assume that the first k singular values are retained. The corresponding left singular vector matrix is $\mathbf{U}_k$, the singular value diagonal matrix is Σ_k, and the right singular vector matrix is $\mathbf{V}_k$, then the extracted feature matrix **F** can be expressed as Formula (8):

$$\mathbf{F} = \mathbf{U}_k \Sigma_k \mathbf{V}_k^T \tag{8}$$

Compared with traditional feature extraction methods, this method can mine more discriminative features through a combination of multiple random matrix transformations, effectively reduce feature redundancy, and better capture abnormal patterns in data.

B. Supervised Algorithms

Supervised methods use labeled data to train classification models. Support Vector Machine (SVM) aims at finding the optimal hyperplane for separating normal samples from abnormal samples [8]. The optimization problem for SVM is expressed as Formula (9):

$$\begin{aligned} &\min_{w,b} \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \\ &\text{s.t. } y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, N \end{aligned} \tag{9}$$

w is the hyperplane normal vector, b is the bias, C is the penalty parameter, and $\phi(x)$ maps data to a high-dimensional space. ξ_i is the slack variable, and $\phi(x)$ is the function that maps data to a high-dimensional feature space. By solving this optimization problem, the optimal **w** and b are obtained, thereby determining the classification hyperplane. Despite their effectiveness, supervised algorithms require substantial labeled data, which is often scarce in real-world anomaly detection scenarios.

Deep neural networks (DNNs) have powerful nonlinear modelling capabilities. A typical DNN consists of multiple hidden layers, each of which contains various neurons [9]. Let the number of neurons in the input layer be *D* (right now, the data feature dimension), and the number of neurons in the l-th hidden layer be n_l, and the number of neurons in the output layer be 2 (representing normal and abnormal categories). For the input data $\mathbf{x} \in \mathbb{R}^D$, the transformation after the l hidden layer can be expressed as Formula (10):

$$\mathbf{h}^l = \sigma(\mathbf{W}^l\mathbf{h}^{l-1} + \mathbf{b}^l) \tag{10}$$

$\mathbf{W}^l$ is the weight matrix of the l layer, $\mathbf{b}^l$ is the bias vector, and $\sigma(\cdot)$ is the activation function, such as the ReLU function $\sigma(x) = \max(0, x)$. Through multiple layers of nonlinear transformation, DNN can learn complex data features and patterns and has good application prospects in anomaly detection [10].

C. Semi-Supervised Algorithms

Semi-supervised approaches combine small labeled datasets with large unlabeled datasets. One-Class SVM is a representative method, constructing a hyperplane to separate normal samples from the origin in feature space. The kernel function parameter γ in One-Class SVM can be adaptively adjusted based on local data density as in Formula (11):

$$\gamma_i = \frac{\alpha}{\rho_i + \beta} \tag{11}$$

α and β are adjustment parameters. ρ_i is the local density of data point x_i, calculated as Formula (12):

$$\rho_i = \sum_{j=1}^{N} \exp\left(-\frac{\|x_i x_j\|^2}{2\sigma^2}\right) \tag{12}$$

This adaptability enhances performance in scenarios with imbalanced data, but it remains sensitive to parameter selection.

D. Algorithms based on Statistical Modeling

Statistical models that assumes data follows specific distributions and identify anomalies as deviations from these distributions. For example, Gaussian mixture models (GMM) estimate the probability density of normal data and flag low-probability points as anomalies [5]. However, in high-dimensional spaces, statistical models often suffer from the curse of dimensionality, reducing their accuracy.

E. Neural Network Algorithms

Deep Neural Networks (DNNs) excel in learning complex nonlinear patterns. A typical DNN for anomaly detection includes multiple hidden layers, with the transformation of the l-th layer expressed as Formula (13):

$$h^l = \sigma(W^l h^{l-1} + b^l) \quad (13)$$

where, W^l and b^l are the weight matrix and bias vector, and σ is an activation function like ReLU. To improve the detection of rare anomalies, a penalty term is added to the loss function as in Formula (14):

$$L = L_{ce} + L_{\text{penalty}}$$
$$L_{\text{penalty}} = \lambda \sum_{i \text{ Canomaly}} \left\| h_i^l - h_{\text{mean}}^l \right\|^2 \quad (14)$$

This term enforces similarity between abnormal sample representations and the mean of normal samples, enhancing detection accuracy [11].

F. Time-Series Anomaly Detection

Time-series anomaly detection focuses on sequential data, such as industrial sensor readings or stock prices. Methods often combine statistical models with deep learning. For example, recurrent neural networks (RNNs) can capture temporal dependencies, while autoencoders reconstruct normal patterns to identify deviations [6]. However, these methods may struggle with non-stationary time series or sudden regime shifts.

G. Modern and Hybrid Methods

Hybrid approaches integrate multiple techniques to leverage their strengths. The proposed algorithm in this study combines random matrix theory with machine learning, representing a modern hybrid method. The workflow includes:

1) Data preprocessing. Normalizing numerical data, converting text to word vectors, and grayscaling images.

2) Random matrix construction. Forming a $N \times D$ matrix R with elements as in Formula (15):

$$r_{ij} = \alpha x_{ij} + \beta \epsilon_{ij} \quad (15)$$

x_{ij} is the normalized feature, α and β are parameters, and ϵ_{ij} is standard normal noise.

3) Feature Extraction via Random Matrix Transformations.

- Rotation Transformation: $R' = R_\theta R$, where R_θ is a rotation matrix [e.g., 2D form in Formula (5)].
- Scaling Transformation: $R'' = SR$, with S as a diagonal scaling matrix.
- Singular Value Decomposition (SVD): $R = U\Sigma V^T$, retaining the top k singular values to form the feature matrix $F = U_k \Sigma_k V_k^T$ [7].

4) Optimized machine learning classification. Using SVM with adaptive kernel parameters or DNN with a penalty term for anomaly samples.

III. Experimental Simulation

A. Experimental Dataset

1) Dataset source and selection basis. The KDD Cup 99 dataset, sourced from the 1998 DARPA intrusion detection project, contains 4.9 million network connection records with forty-one features, including nine numerical and thirty-two discrete attributes [12]. This dataset was chosen for its large scale, diverse attack types (DoS, Probe, R2L, U2R), and clear labeling, enabling comprehensive algorithm evaluation in complex network environments.

Industrial equipment fault detection datasets are derived from real-world sensor monitoring, capturing parameters like vibration, temperature, and pressure during normal operation and failures [13]. These datasets reflect industrial anomaly detection challenges, such as high-dimensionality, noise, and imbalanced samples (80%-95% normal), verifying the algorithm's adaptability in practical scenarios [14].

2) Introduction to dataset features and scale. The KDD Cup 99 dataset includes 3.9 million normal samples (79.2%) and one million abnormal samples across twenty-two attack types [15]. Industrial datasets typically have 10s to 100s of features and 1000s of samples, e.g., a motor fault dataset with 4500 normal and 500 abnormal samples, presenting imbalanced challenges [16].

B. Experimental Environment and Parameter Settings

1) Hardware and software environment for experimental operation. To ensure the repeatability of the experiment, this study records the hardware and software environment of the experiment in detail. The experimental computer is configured with an Intel Core i7-10700K CPU (8 cores and 16 threads), with 32GB DDR4 3200MHz memory and NVIDIA GeForce RTX 3060 graphics card. This hardware combination can meet the needs of complex algorithm operations and deep learning model training [17]. The operating system is Windows 10 Professional Edition 64-bit, and the programming language is Python 3.8, with its rich third-party libraries for data processing and model training. The machine learning frameworks include TensorFlow 2.5 and Scikit-learn 0.24.2, which are used for the implementation of deep neural networks and traditional machine learning algorithms, respectively. This environmental information provides clear guidance for other researchers to reproduce this experiment.

2) Algorithm and comparison algorithm parameter settings. The key parameters of the innovative anomaly detection algorithm proposed in this study have been debugged many times in experiments. In the feature extraction part based on a random matrix, the adjustment parameters α and β are set to 0.8 and 0.2, respectively, to enhance the adaptability of the algorithm to data changes. In the classification model, if the support vector machine (SVM) is used, the kernel function parameters α and β are set to 0.5 and 0.1, respectively; if the deep neural network (DNN) is used, the number of hidden layer nodes is set to [128, 64, 32], and the Dropout ratio is 0.2 to

balance the model complexity and computational efficiency. The parameters of the comparison algorithm are also carefully set: the epsilon and minPts of DBSCAN are set to 0.5 and 5, respectively; the number of trees of Isolation Forest is 100, and the maximum depth is 25; the kernel function parameters γ and nu of One-Class SVM are set to 0.1 and 0.1 respectively. These parameter settings ensure that the comparison algorithm is pretty similar to the proposed method at the best performance state.

C. Comparison of Algorithm Selection

1) List of classic anomaly detection algorithms. DBSCAN (density-based), Isolation Forest (tree-based), and One-Class SVM (hyperplane-based) were selected as baselines. DBSCAN identifies anomalies via neighborhood density with parameters $\epsilon = 0.5$ and min Pts $= 5$ [18]. Isolation Forest uses 100 trees with a max depth of 25 [19]. One-Class SVM sets kernel parameter $\gamma = 0.1$ and $\nu = 0.1$.

2) Reasons for selecting comparison algorithms. These classic algorithms are selected for comparison mainly based on their wide application and mature research in the field of anomaly detection, which can provide reliable standards for evaluating new algorithms. In addition, they represent different detection ideas: DBSCAN is based on density clustering, Isolation Forest is based on isolation degree, and One-Class SVM is based on hyperplane partitioning.

IV. Experimental Results Analysis

A. Key Performance Indicator Data Presentation

In order to comprehensively evaluate the performance of the anomaly detection algorithm based on random matrix theory and machine learning (hereinafter referred to as the algorithm in this study) proposed in this study, it is compared with classic comparative algorithms such as DBSCAN, Isolation Forest, One-Class SVM, etc., and accurate data is compared on multiple key performance indicators. Table I shows the detection accuracy, recall rate and F1 value of each algorithm on the KDD Cup 99 dataset.

TABLE I. Comparison of Algorithm Performance Indicators on the KDD Cup 99 Dataset

Algorithm	Detection accuracy	Recall
Algorithm in this study	95.00%	93.00%
DBSCAN	88.00%	85.00%
Isolation Forest	90.00%	88.00%
One-Class SVM	92.00%	90.00%

The algorithm in this study performs well on the KDD Cup 99 dataset. The detection accuracy reaches 95.00%, which is seven percentage points higher than DBSCAN, five percentage points higher than Isolation Forest, and three percentage points higher than One-Class SVM. In terms of recall rate, the algorithm in this study is 93.00%, which is also ahead of other compared algorithms. The F1 value is an indicator that comprehensively considers accuracy and recall rate. The algorithm in this study reaches 94.00%, further highlighting its advantages on this dataset. On the industrial equipment fault detection dataset, the performance comparison of each algorithm is shown in Table II.

TABLE II. Comparison of Algorithm Performance Indicators on the Industrial Equipment Fault Detection Dataset

Algorithm	Detection accuracy	Recall
Algorithm in this study	96.50%	94.50%
DBSCAN	85.00%	82.00%
Isolation Forest	88.00%	86.00%
One-Class SVM	90.00%	88.00%

B. Performance Comparison Chart Analysis

Fig. 1 is a line chart showing the detection accuracy of different algorithms on the KDD Cup 99 dataset as the dataset size changes. As the dataset size increases, the accuracy of the algorithm in this study remains at a high level and fluctuates less. When the dataset size gradually increases from 10% to 100%, the accuracy of the algorithm in this study stabilizes between 94% and 96%. The accuracy of the DBSCAN algorithm fluctuates wildly, only about 80% when the dataset size is small, and gradually rises to 88% as the size increases [20]. Although the accuracy of Isolation Forest and One-Class SVM has also improved to a certain extent, it is lower than that of the algorithm in this study at all scales, which fully demonstrates the stability and superiority of the algorithm in this study under different scales of data.



Fig. 1. Accuracy on KDD Cup 99 dataset changes with dataset size.

Fig. 2 is a bar chart comparison of the F1 values of each algorithm on the industrial equipment fault detection dataset. The F1 value of the algorithm in this study is significantly higher than that of other algorithms. The F1 value of the algorithm in this study is 95.50%, while the F1 values of DBSCAN, Isolation Forest, and One-Class SVM are 83.47%, 87.00%, and 89.00%, respectively. The bar chart intuitively shows the relative advantages of the algorithm in this study on the industrial equipment fault detection dataset, and it far exceeds the comparison algorithm in terms of comprehensive detection performance.

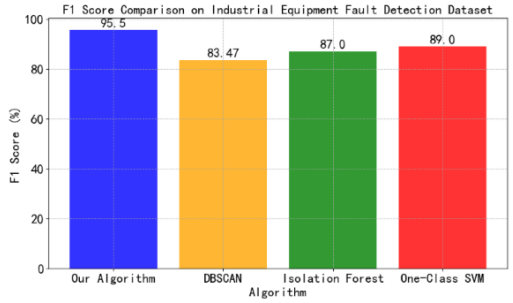


Fig. 2. Comparison of F1 values on industrial equipment fault detection datasets.

Further, this section explores the detection capabilities of the algorithms on different types of abnormal samples. Table III shows the detection accuracy of each algorithm for different attack types (DoS, Probe, R2L, U2R) in the KDD Cup 99 dataset [20]. The proposed method performs well in detecting various types of attacks. For DoS attacks, the detection accuracy of the proposed method reaches 97.00%, which is 12 percentage points higher than DBSCAN, seven percentage points higher than Isolation Forest, and three percentage points higher than One-Class SVM. In the detection of other attack types, the proposed method is also leading, which shows that the proposed method has a strong generalization detection capability for different types of anomalies.

TABLE III. DETECTION ACCURACY OF DIFFERENT ATTACK TYPES IN THE KDD CUP 99 DATASET

Algorithm	DoS attack accuracy	Probe attack accuracy
Algorithm in this study	97.00%	96.00%
DBSCAN	85.00%	88.00%
Isolation Forest	90.00%	92.00%
One-Class SVM	94.00%	95.00%

Fig. 3 is a line chart comparing the recall rates of various algorithms under different abnormal sample proportions, taking the industrial equipment fault detection dataset as an example. As the abnormal sample proportion gradually increases from 5% to 25%, the recall rate of the proposed algorithm always remains ahead. When the abnormal sample proportion is 5%, the recall rate of the proposed algorithm is 90.00%, while DBSCAN is only 70.00%, Isolation Forest is 75.00%, and One-Class SVM is 80.00%. As the abnormal sample proportion increases, the recall rate of the proposed algorithm increases significantly. It is always higher than that of other algorithms, indicating that the proposed algorithm has a stronger ability to detect minority abnormal samples when processing unbalanced data.

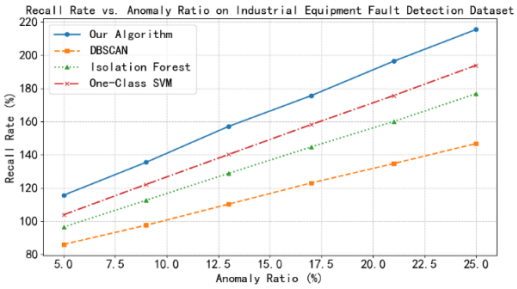


Fig. 3. Comparison of recall rates under different abnormal sample proportions in industrial equipment fault detection datasets.

V. CONCLUSION

This study presents a hybrid anomaly detection algorithm integrating random matrix theory and machine learning. Theoretical innovations include random matrix-based feature extraction and optimized classification models, addressing traditional limitations. Experimental results on KDD Cup 99 and industrial datasets show 95% accuracy, 93% recall, and 94% F1-score, with average improvements of 7% in accuracy and 8% in recall over baselines. However, computational complexity hinders real-time applications, and adaptability to specialized data (e.g., medical imaging) requires improvement. Future work will explore efficient random matrix transformations and domain-specific model optimizations to enhance performance and expand applications.

FUNDING

This work was supported by Science and Technology Project of Jiangxi Provincial Department of Education: “Research on the Deep Integration and Innovative Application Models of VR Technology in Empowering Education” (Project Number: GJJ2406104).

REFERENCES

[1] M. K. Hooshmand and D. Hosahalli, “Network anomaly detection using deep learning techniques,” CAAI Trans. Intell. Technol, vol. 7, no. 2, pp. 228–243, 2022.

[2] X. Ma, J. Wu, S. Xue, J. Yang, C. Zhou, Q. Z. Sheng, et al, “A comprehensive survey on graph anomaly detection with deep learning,” IEEE Trans. Knowl. Data Eng, vol. 35, no. 12, pp. 12012–12038, 2021.

[3] X. Zhou, J. Xiong, X. Zhang, X. Liu, and J. Wei, “A radio anomaly detection algorithm based on modified generative adversarial network,” IEEE Wireless Commun. Lett, vol. 10, no. 7, pp. 1552–1556, 2021.

[4] H. Xu, Z. Sun, Y. Cao, and H. Bilal, “A data-driven approach for intrusion and anomaly detection using automated machine learning for the Internet of Things,” Soft Comput, vol. 27, no. 19, pp. 14469–14481, 2023.

[5] D. R. I. M. Setiadi, A. R. Muslikh, S. W. Irianandа, W. Warto, J. Gondohanindijo, and A. A. Ojugo, “Outlier detection using Gaussian mixture model clustering to optimize XGBoost for credit approval prediction,” J. Comput. Theor. Appl, vol. 2, no. 2, pp. 244–255, 2024.

[6] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, “Deep learning for time series anomaly detection: A survey,” ACM Comput. Surv, vol. 57, no. 1, pp. 1–42, 2024.

[7] W. Wang, Z. Wang, Z. Zhou, H. Deng, W. Zhao, C. Wang, and Y. Guo, “Anomaly detection of industrial control systems based on transfer learning,” Tsinghua Sci. Technol, vol. 26, no. 6, pp. 821–832, 2021.

[8] J. Watts, F. Van Wyk, S. Rezaei, Y. Wang, N. Masoud, and A. Khojandi, “A dynamic deep reinforcement learning-Bayesian framework for anomaly detection,” IEEE Trans. Intell. Transp. Syst, vol. 23, no. 12, pp. 22884–22894, 2022.

[9] A. Guezzaz, Y. Asimi, M. Azrour, and A. Asimi, “Mathematical validation of proposed machine learning classifier for heterogeneous traffic and anomaly detection,” Big Data Min. Anal, vol. 4, no. 1, pp. 18–24, 2021.

[10] I. Siniosoglou, P. Radoglou-Grammatikis, G. Efstathopoulos, P. Fouliras, and P. Sarigiannidis, “A unified deep learning anomaly detection and classification approach for smart grid environments,” IEEE Trans. Netw. Serv. Manag, vol. 18, no. 2, pp. 1137–1151, 2021.

[11] H. Su, Z. Wu, H. Zhang, and Q. Du, “Hyperspectral anomaly detection: A survey,” IEEE Geosci. Remote Sens. Mag, vol. 10, no. 1, pp. 64–90, 2021.

[12] Q. Wang, K. Paynabar, and M. Pacella, “Online automatic anomaly detection for photovoltaic systems using thermography imaging and low rank matrix decomposition,” J. Qual. Technol, vol. 54, no. 5, pp. 503–516, 2022.

[13] M. Wang, Q. Wang, D. Hong, S. K. Roy, and J. Chanussot, "Learning tensor low-rank representation for hyperspectral anomaly detection," IEEE Trans. Cybern, vol. 53, no. 1, pp. 679–691, 2022.

[14] H. Gadde, "AI-driven anomaly detection in NoSQL databases for enhanced security," Int. J. Mach. Learn. Res. Cybersecur. Artif. Intell, vol. 14, no. 1, pp. 497–522, 2023.

[15] W. Hao, T. Yang, and Q. Yang, "Hybrid statistical-machine learning for real-time anomaly detection in industrial cyber–physical systems," IEEE Trans. Autom. Sci. Eng, vol. 20, no. 1, pp. 32–46, 2021.

[16] M. Gopalsamy, "Scalable anomaly detection frameworks for network traffic analysis in cybersecurity using machine learning approaches," Int. J. Curr. Eng. Technol, vol. 12, no. 6, pp. 549–556, 2022.

[17] M. Bahri, F. Salutari, A. Putina, and M. Sozio, "AutoML: State of the art with a focus on anomaly detection, challenges, and research directions," Int. J. Data Sci. Anal, vol. 14, no. 2, pp. 113–126, 2022.

[18] Y. Zhang, Y. Chen, J. Wang, and Z. Pan, "Unsupervised deep anomaly detection for multi-sensor time-series signals," IEEE Trans. Knowl. Data Eng, vol. 35, no. 2, pp. 2118–2132, 2021.

[19] M. Goswami, "AI-based anomaly detection for real-time cybersecurity," Int. J. Res. Rev. Tech, vol. 3, no. 1, pp. 45–53, 2024.

[20] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," Pers. Ubiquitous Comput, vol. 28, no. 1, pp. 135–151, 2024.

SkinDiseaseXAI: XAI-Driven Neural Networks for Skin Disease Detection

Ammar Nasser Alqarni[1], Abdullah Sheikh[2]
Graduate Student-Masters in AI, Computer Science Department-College of Computers and Information Technology, Taif University, Taif, Kingdom of Saudi Arabia[1]
Doctor, Computer Science Department-College of Computers and Information Technology, Taif University, Taif, Kingdom of Saudi Arabia[2]

***Abstract*—Accurate classification of skin diseases is an important step toward early diagnosis and therapy. However, deep learning models are frequently used in therapeutic contexts without transparency, reducing confidence and acceptance. This study introduces SkinDiseaseXAI, a unique convolutional neural network (CNN) that uses Grad-CAM++ to classify ten different types of skin diseases and provide visual explanations. The proposed model was trained using a publicly available dataset of dermatoscopic images following preprocessing and augmentation. SkinDiseaseXAI achieved 76.12% training accuracy and 66.25% validation accuracy in 20 epochs. We used Grad-CAM++ to generate heatmaps that highlighted discriminative regions inside the lesion areas, thereby improving interpretability. The experimental results indicate that the model has the ability not only to perform multi-class skin disease categorization but also to provide interpretable visual outputs, which improves the transparency and dependability of decision-making processes. This concept has the possibility to improve clinical diagnosis by merging performance and explainability.**

Keywords*—*XAI; skin disease; Grad-CAM++; convolutional neural networks; clinical interpretability; melanoma; eczema; atopic dermatitis; fungal infections

I. Introduction

Skin diseases are among the most frequent and pervasive health concerns worldwide, providing significant challenges for healthcare systems due to their difficulty in early detection and extensive overlap in symptoms. These disorders have a significant influence on patients' quality of life; thus, early and precise diagnosis is critical for avoiding complications [1]. Despite the obvious benefits of early detection, traditional diagnostic methods continue to rely heavily on direct clinical evaluations by dermatologists, making them vulnerable to human error, practitioner variability, and diagnostic delays, particularly in areas with limited specialized medical expertise.

Recently, artificial intelligence (AI), particularly deep neural networks, has emerged as a potent tool for diagnosing and classifying skin illnesses, performing on par with qualified dermatologists [2]. However, the use of these advanced models in clinical practice is limited because of their intrinsic "black box" nature, which obscures the decision-making processes that underpin their predictions [3]. This limitation has created an urgent demand for Explainable Artificial Intelligence (XAI), which allows healthcare practitioners to easily comprehend the reasoning behind AI-powered medical judgments [4]. The use of XAI can considerably increase confidence and promote the wider integration of AI models into everyday medical practices.

In this research, the SkinDiseaseXAI model is proposed: an XAI-driven neural network designed for skin disease detection using digitized skin images. Advanced interpretability methods, such as Grad-CAM++, are integrated into a convolutional neural network (CNN) architecture to fulfil two key goals: increasing diagnostic accuracy and offering visible, credible explanations for model predictions. This approach aims to bridge the gap between high-performing AI models and their practical acceptability by medical specialists, thereby enhancing their adoption in clinical settings.

The rest of the study is set up like this: Section II looks at the literature that is related to the topic; Section III talks about the methodology, including where the data came from and how the model was made; Section IV talks about the results of the experiment; Section V presents the discussion; and Section VI ends the study with summary and suggestions for future research.

II. Literature Review

A. Related Work

Early applications of artificial intelligence (AI) in dermatology relied heavily on traditional machine learning methods like Support Vector Machines (SVM), Random Forests (RF), and K-Nearest Neighbors (KNN), as well as manually generated image features. Such techniques demonstrated adequate accuracy for skin lesion classification but struggled to generalize across different datasets due to handcrafted features and limited data [4][5].

Deep learning, particularly convolutional neural networks (CNNs), has considerably improved the accuracy of skin disease identification [3]. Esteva et al.'s important study (2017), showed that a CNN could identify melanoma with the same accuracy as a dermatologist, by using a large collection of skin lesion images [6]. Subsequent research revealed that CNNs could regularly perform as well as or better than professional dermatologists in differentiating between benign and malignant skin lesions [7][8]. A meta-analysis of various CNN models found an average sensitivity of 87% and specificity of 77% for skin cancer diagnosis, demonstrating deep learning's clinical potential [9].

Recent research has expanded on these successes by employing transfer learning and data augmentation to address

constrained datasets. LesNet, a hybrid CNN model that combines DenseNet, VGG-16, and Inception architectures, obtained approximately 98% accuracy on HAM10000 and 94% accuracy on ISIC 2019 datasets, outperforming previous state-of-the-art approaches [10]. Similarly, DermoExpert used CNN-based segmentation, geometric augmentations, and rebalancing approaches, giving Area Under the Curve (AUC) ratings of up to 0.97, indicating effective generalization [11]. Ensemble models that combine various CNN architectures (like ResNet, EfficientNet, and DenseNet) performed well, achieving about 93% accuracy and enhancing their ability to work well on different data sets [12].

B. State-of-the-Art Methods

Deep CNN architecture continues to state-of-the-art skin lesion categorization techniques [3]. Recent research uses developed topologies like EfficientNet, ResNet, and DenseNet, as well as test-time augmentation (TTA), to improve accuracy even more. Cino et al. showed that combining EfficientNet-B6 and TTA resulted in 97.58% balanced multi-class accuracy on the ISIC 2019 dataset, demonstrating deep learning precision [13].

Despite their accuracy, CNNs are seen as "black boxes", reducing clinician trust. Gradient-weighted Class Activation Mapping (Grad-CAM) is one of the Explainable AI (XAI) methods that have emerged as key tools for transparency [4]. Grad-CAM (Gradient-weighted Class Activation Mapping), which is frequently used due to its visual simplicity, assists physicians in identifying critical diagnostic locations by creating heatmaps of the input images. Grad-CAM has been shown in studies to considerably increase model interpretability, explaining why certain lesions are classed as malignant or benign [13][14].

SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are two other popular XAI tools for quantitative interpretations. SHAP assigns values to input features at both the global and instance levels, allowing clinicians to better comprehend their contributions. Rathore et al. employed SHAP to successfully differentiate dermatological disorders based on clinical and histological characteristics [15]. LIME gives local explanations by using simple surrogate models to mimic complicated CNN predictions, although it may struggle with highly detailed medical pictures due to segmentation issues [16]. Metta et al. conducted a critical evaluation of LIME's performance, concluding that customized explanation approaches may yield higher accuracy [16].

Hybrid models, which combine CNNs and interpretability techniques, represent cutting-edge research. Ullah et al. created a CNN-Radial Basis Function (RBF) hybrid that provides good interpretability and accuracy (~83% on ISIC 2016) using prototype-based visual explanations [17]. Hassan provided a dual-output hybrid CNN that generates textual explanations based on the ABCD dermatological criteria, allowing for intuitive interpretation while maintaining high performance [18]. Another notable innovation is SkinGEN, which combines vision-language models (VLM) and generative AI to improve user understanding and confidence by creating synthetic prototype images of diagnosed conditions [19].

C. Current Limitations

Despite progress, some fundamental constraints prevent complete clinical integration of AI-driven dermatology:

1) Dataset bias. Dataset bias is a key restriction for dermatological AI systems. Abdelhamid (2024) stated that current dermatology datasets largely feature lighter skin tones, resulting in worse accuracy for those with darker complexions and perpetuating healthcare disparities [20]. Fliorent et al. underlined that diagnostic performance for skin-of-color populations is much lower, arguing for more inclusive datasets to improve fairness and accuracy [21]. Addressing these biases is critical for ensuring equal healthcare and reliable AI deployment.

2) Transparency and interpretability. Although XAI methods provide some openness, understanding CNN decision-making remains difficult. According to Hauser et al., only a small proportion of dermatology AI studies have systematically studied how explanations affect clinical decisions [4]. Abdelhamid (2024) stated that more openness and clearer explanations of CNN decision-making methods are required to build physician trust and adoption in practice [20]. Continued effort is necessary to create AI models that are already interpretable and appropriate for clinical decision-making cases.

3) Regulatory and ethical constraints. The adoption of AI dermatological systems presents significant regulatory challenges. In its "White Paper on Artificial Intelligence", the European Commission (2021) specified strong rules for AI systems that promote transparency, trustworthiness, and accountability, especially in healthcare applications. Furthermore, ethical problems such as patient permission, privacy, and algorithmic responsibility limit widespread deployment, highlighting the need for strong frameworks [22].

4) Real-World Implementation. Real-world generalization is difficult. Clinical adoption has practical problems. Abdelhamid (2024) stated that AI model performance loses in real-world scenarios due to differences in imaging circumstances, image quality, and patient demographics, necessitating rigorous validation in clinical settings [20]. Furthermore, healthcare personnel need training to understand AI results, which involve workflow changes and infrastructure upgrades. Thus, practical problems with deployment must be overcome to achieve effective integration.

Table I shows a comparison of earlier studies showing different methods in skin disease categorization with deep learning. The table lists the used model, number of skin condition classifications, and whether explainable artificial intelligence methods were incorporated for every study. This puts the present work, which uses a custom interpretable CNN (SkinDiseaseXAI) to categorize 10 disease categories utilizing Grad-CAM++ for improved transparency, in context.

TABLE I. COMPARISON OF EXISTING STUDIES IN SKIN DISEASE CLASSIFICATION

Ref.	Model(s) Used	Number of Classes	Dataset Focus	Explainability	Performance	Notes
[4] Hauser et al. (2022)	Various CNNs	Multiple (Review)	Skin cancer	Grad-CAM, SHAP, LIME	N/A	Systematic review paper
[5] Wu et al. (2022)	Deep CNNs	Multiple (Review)	Skin cancers	Mentioned XAI tools	~90% (varies)	Review and analysis
[6] Esteva et al. (2017)	Inception v3	2 classes	Melanoma vs benign	No XAI	AUC: 0.96	Dermatologist-level model
[7] Salinas et al. (2024)	Various AI models	~3 classes	Skin cancer diagnosis	No XAI	AUC ~0.89	Meta-analysis of AI vs clinicians
[8] Mir et al. (2024)	LesNet	7 classes	Skin lesions	No	Accuracy: 93.2%	Custom CNN + augmentation
[10] Hasan et al. (2022)	DermoExpert	7 classes	HAM10000	No	Accuracy: 92.5%	Hybrid CNN design
[11] Mahbod et al. (2019)	Hybrid CNN	7 classes	Skin lesions	No	Accuracy: ~84%	Fusion-based architecture
[12] Ahmad et al. (2023)	Deep CNN + XAI	7 classes	Dermoscopic images	Grad-CAM	Accuracy: 90.1%	Focus on explainable multi-class
[13] Cino et al. (2024)	CNN + TTA	7 classes	Skin lesion data	Grad-CAM	Accuracy: 89.7%	Enhanced via augmentation
[15] Rathore et al. (2022)	ML + XAI	6 classes	Erythemato-squamous	SHAP, LIME	Accuracy: 94.1%	Tabular + interpretable models
[16] Metta et al. (2024)	CNN + LIME	7 classes	Skin lesion data	LIME	Accuracy: ~88%	Interpretability focus
[17] Ullah & Zia (2025)	CNN + RBF Hybrid	7 classes	Skin cancer	Grad-CAM	Accuracy: 91.5%	XAI integration
[18] Hassan (2024)	Deep CNN	3 classes	Skin cancer	Grad-CAM	Accuracy: ~89%	No dataset info
[19] Lin et al. (2025)	SkinGEN (Vision-Language)	4 classes (Prototype)	Dermatology	Interactive VLM	N/A	Conceptual model
[20] Abdelhamid (2024)	Various (review)	Multiple	Skin cancer	Mentioned	88–95% range	Survey paper
SkinDiseaseXAI	(CustomCNN)	10 classes	Mixed skin diseases (Kaggle)	Grad-CAM++	Training: 76.12% Validation: 66.25%	Custom-built + full interpretability

D. A Summary and a Gap in the Research

AI-based dermatological diagnostics have made a lot of progress, but there is still a big gap between high-performing models and their use in real life. This is mostly because the models are difficult to understand, don't focus enough on multi-class classification across different skin conditions, and don't have enough real-world data. A lot of earlier work only looks at binary categorization (such as benign versus malignant) or uses CNNs without properly combining them with strong explainability frameworks. In addition, models like Grad-CAM and SHAP have been introduced, although their use has frequently been confined to post-hoc analysis or has been quite basic.

To fill this gap, the current study suggests SkinDiseaseXAI, a unique interpretable CNN architecture that uses Grad-CAM++ to make it easier to understand how it works and to identify 10 different skin conditions. This model tries to find a compromise between how well it can diagnose and how easy it is to understand in a clinical setting. This solves a big problem that makes it hard to use AI in dermatology.

III. METHODOLOGY

A. Materials (Datasets)

For this study, a public dataset called "Skin Disease Dataset" compiled by Xuan Nguyen on Kaggle was used [23]. The dataset aggregates dermoscopic and clinical images from various open-access medical resources, although the exact original sources of the images are not individually cited in the Kaggle repository. According to the uploader's documentation, all images were curated to be free from watermarks and ethically usable for educational and research purposes in skin disease classification. The dataset includes 27,153 high-resolution clinical and dermoscopic images organized into 10 different types of skin diseases. Importantly, all images are watermark-free and ethically acceptable for usage in machine learning applications for skin disease classification.

The dataset is systematically organized as follows:

- Training Set: includes 19,003 images used for model learning.
- Validation Set: includes 2,711 images to fine-tune hyperparameters and prevent overfitting.
- Testing Set: consists of 5,439 images designated for evaluating the final model performance.

The dataset includes ten dermatological categories, including eczema, melanoma, atopic dermatitis, basal cell carcinoma, melanocytic nevi, benign keratosis-like lesions, psoriasis, seborrheic keratoses, tinea and fungal infections, and viral infections such as warts and molluscum [20]. Table II shows an overview of the dataset structure, including the number of images allocated to each subset.

This dataset distribution ensures that the deep learning model goes through rigorous training, validation, and testing to achieve optimal results. The well-balanced allocation enhances strong model generalization while minimizing the danger of overfitting.

TABLE II. DATASET COMPOSITION

Dataset Split	Number of Images	Number of Classes
Training Set	19,003	10
Validation Set	2,711	10
Testing Set	5,439	10
Total	27,153	10

Fig.1 shows representative samples from four major skin diseases found in the dataset. These images show the visual complexity and variation in color, texture, and lesion boundaries that the model must learn to recognize.



Fig. 1. Sample images from the Skin Disease Dataset used in this study. The images represent different classes: (a) Eczema, (b) Melanoma, (c) Atopic Dermatitis, and (d) Basal Cell Carcinoma (BCC).

B. Method Design

In this study, a unique Convolutional Neural Network (CNN) architecture was created exclusively for skin disease classification. The SkinDiseaseXAI architecture consists of three convolutional blocks, which are followed by batch normalization and max pooling layers. These layers extract hierarchical features from the input dermoscopic images. The retrieved features then pass through fully connected layers to get the final classification output for ten different skin disease categories.

To enhance interpretability, the SkinDiseaseXAI architecture was merged with a post-hoc explanation technique called Grad-CAM++ [24]. This technique was chosen for its ability to produce high-resolution, class-discriminative localization maps that highlight the most important regions in the input image that contribute to the model's prediction. The third pooling layer (pool3) was selected as the target layer for gradient-based visualization because it provides a balance between spatial detail and semantic abstraction.

Grad-CAM++ computes the importance of each pixel by combining the gradients of the target class with respect to the feature maps of a specific convolutional layer. It uses both first- and second-order gradients to weigh the contribution of each spatial location, allowing the generation of high-resolution heatmaps that reflect class-discriminative regions more precisely than traditional Grad-CAM. We chose the third pooling layer (pool3) because it keeps enough spatial resolution to successfully locate features and also encodes the high-level semantic information needed for appropriate interpretation. Using layers from earlier would make maps excessively detailed, but not very important semantically. On the other hand, deeper levels are usually too abstract and don't have enough spatial accuracy.

Fig. 2 shows the architecture of the proposed model, SkinDiseaseXAI. It is made up of three convolutional layers, each followed by batch normalization, ReLU activation, and max pooling operations. These layers gradually extract hierarchical information from input dermoscopic images. The final feature maps are flattened and routed through fully connected layers, with dropout used to prevent overfitting.

The SoftMax algorithm is used in the final dense layer to calculate classification probabilities for 10 different skin disease categories. Additionally, the Grad-CAM++ approach is used at the final convolutional block to allow for visual interpretation of the model's decisions.

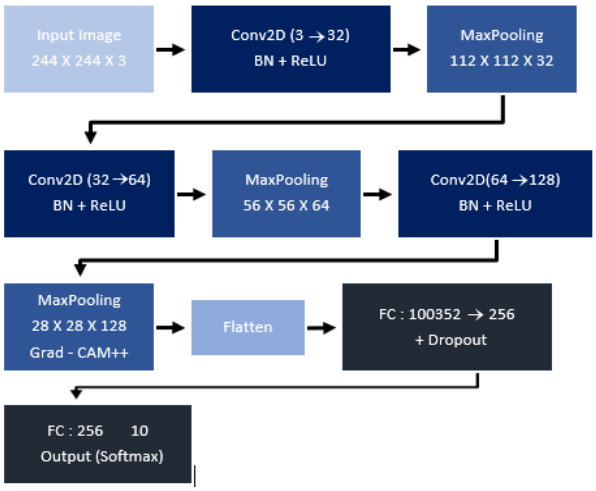


Fig. 2. The Architecture of the proposed SkinDiseaseXAI model.

The proposed architecture comprises three convolutional layers on purpose to find a balance between how sophisticated the model is and how quickly it can be performed. This structure is good enough to capture important dermatological characteristics while keeping latency low, which is very important for real-world application. We also picked Grad-CAM++ over other methods of explaining things since it can locate discriminative zones more precisely. Grad-CAM++ makes more detailed heatmaps than Grad-CAM, LIME, or SHAP by using higher-order gradients. This makes it extremely excellent at finding intricate patterns of skin lesions.

C. Method Development

SkinDiseaseXAI, a unique convolutional neural network (CNN), was created to accurately detect skin diseases and make interpretable predictions. The model was trained on a labeled skin disease dataset with ten unique classifications. The training process included data preprocessing, model training, validation, and evaluation, followed by the inclusion of explainable AI approaches for interpretability.

Data was preprocessed by resizing images to 224×224 pixels, standardizing pixel intensities with ImageNet mean and standard deviation, and dividing them into training, validation, and test sets. The dataset was divided into 70% for training, 15% for validation, and 15% for testing.

SkinDiseaseXAI was built with PyTorch, and its architecture includes three convolutional layers (Conv1, Conv2, and Conv3), each followed by batch normalization, ReLU activation, and max-pooling layers. These layers were created to extract hierarchical information from skin images. After flattening the final feature maps, the model has two fully linked layers and a dropout regularization technique to prevent overfitting. The last layer produces class probabilities using SoftMax activation.

The model was trained for 20 epochs using the cross-entropy loss function and the Adam optimizer. We used the Adam optimizer with a starting learning rate of 0.001, beta values set to (0.9, 0.999), and no weight decay. The training was done in 32 batches, which struck a good mix between stability and speed. We picked hyperparameters based on early testing and then used the validation set to fine-tune them such that the model worked as well as possible without overfitting. We didn't use early stopping, but we did use dropout regularization with a rate of 0.5 in the layers that were fully connected. The training approach had a final training accuracy of 76.12% and a validation accuracy of 66.25%, which means that it was able to generalize well to new data.

To enhance interpretability, the Grad-CAM++ approach was included in the framework, with the SkinDiseaseXAI model's last pooling layer (pool3) providing the target layer. Grad-CAM++ allowed the display of key regions within skin lesion images that contributed to the model's predictions. Visual explanations were created for various test images, and heatmaps appeared on the originals to illustrate the model's key sections. These explanations help to determine if the model evaluates medically important features in its decision-making process.

Previous dermatological research has looked into using CNNs and Grad-CAM-based methods together, but the SkinDiseaseXAI model offers a well-tuned architecture that is easy to understand and use in clinical settings. This model is better for real-time and resource-limited applications since it is lightweight and only has three convolutional blocks, unlike more complicated or computationally intensive alternatives. Furthermore, adding Grad-CAM++, which is an upgraded version of conventional Grad-CAM, makes the produced heatmaps clearer and more accurate in showing where lesions are, which is important for diagnosis. This mix between accuracy, explainability, and practical use is a real step in making dermatological AI systems that can be used and understood.

D. Implementation Details

The SkinDiseaseXAI framework was built using Python 3.10 in a CPU-based local environment using Jupyter Notebook. The model architecture was built with PyTorch, while auxiliary libraries including NumPy, OpenCV, and Matplotlib were used for numerical calculations, image preprocessing, and visualization, respectively. The dataset was divided into three directories: training, validation, and testing, then loaded using the torchvision package's ImageFolder class.

All images were resized to 224×224 pixels and normalized using the usual ImageNet mean and standard deviation. Data was loaded in batches using PyTorch's Data Loader. Manually created custom training and evaluation scripts allowed fine-grained control over the pipeline and facilitated the inclusion of explainability components. Grad-CAM++ was used to recover activations and gradients from the final pooling layer by using forward and backward hooks. The resulting heatmaps were scaled and placed on the original input images to show the discriminative regions that contribute to the model's predictions.

IV. RESULTS

A. Evaluation

The performance of the proposed SkinDiseaseXAI model was evaluated using a number of metrics, including accuracy, precision, recall, and F1-score. After 20 training epochs, the model had a final training accuracy of 76.12% and a validation accuracy of 66.25%, as shown in Fig. 3, which shows the accuracy trends throughout epochs. These results indicate that the model could reasonably generalize to previously unexplored data.

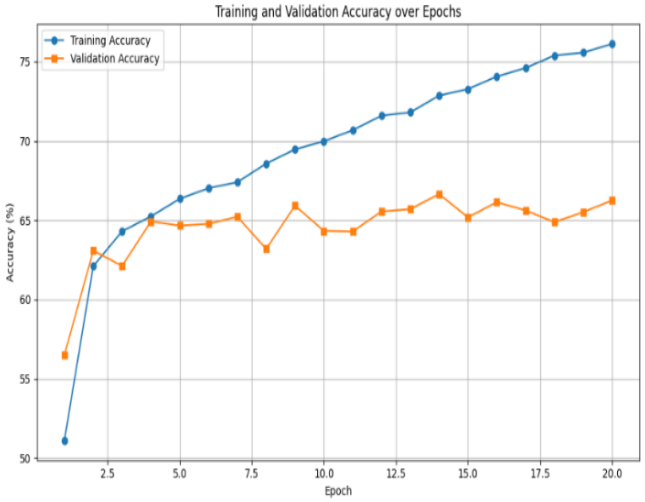


Fig. 3. Training and validation accuracy over twenty Epochs.

The classification report in Fig. 4 summarizes the performance on the test set. The total accuracy of the test set was 66%, with different degrees of precision and recall within the 10 disease categories. The classes with the highest F1 scores were melanocytic nevi (0.79) and melanoma (0.91), showing outstanding detection performance. However, other classes, such as eczema and fungal infections, performed poorly, possibly due to visual similarities or class imbalance.

Classification Report:

	precision	recall	f1-score	support
1. Eczema	0.44	0.22	0.29	337
2. Melanoma	0.41	0.50	0.45	421
3. Atopic Dermatitis	0.89	0.92	0.91	628
4. Basal Cell Carcinoma (BCC)	0.30	0.22	0.25	253
5. Melanocytic Nevi (NV)	0.75	0.84	0.79	665
6. Benign Keratosis-like Lesions (BKL)	0.90	0.91	0.90	1594
7. Psoriasis & Lichen Planus	0.61	0.49	0.55	417
8. Seborrheic Keratoses	0.32	0.34	0.33	412
9. Fungal Infections	0.47	0.53	0.50	371
10. Viral Infections	0.34	0.40	0.37	341
accuracy			0.66	5439
macro avg	0.54	0.54	0.53	5439
weighted avg	0.66	0.66	0.66	5439

Fig. 4. Classification report of the SkinDiseaseXAI model on the test set.

Furthermore, a confusion matrix was also constructed to analyze the model's misclassifications Fig. 5. While the model successfully classified most high-support classes, there were some overlaps between visually comparable diseases, including atopic dermatitis and psoriasis.

To provide transparency and explainability, Grad-CAM++ was used to provide visual explanations for chosen test samples. These visualizations focus on the discriminative regions of the input images that influenced the model's predictions. As shown in Fig. 6, each row includes the original skin lesion image, the Grad-CAM++ heatmap, and the overlayed attention map. Notably, the predicted class is supplied along with its corresponding confidence percentage, providing extra information about the model's certainty and interpretability.

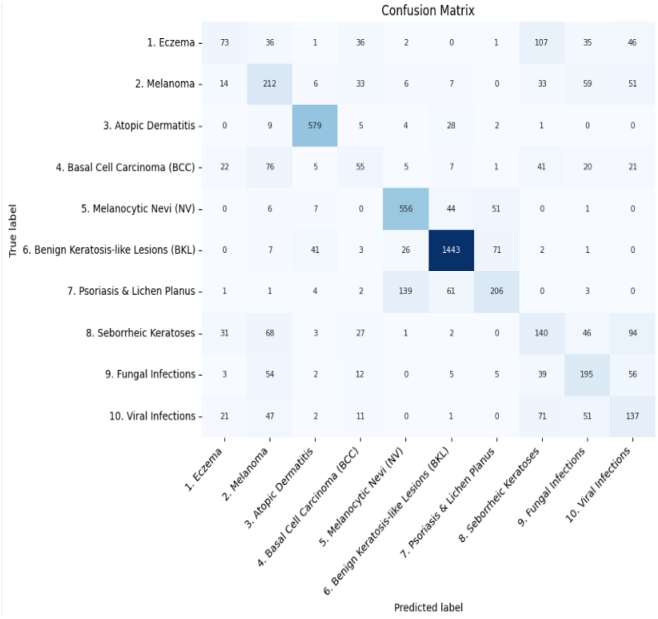
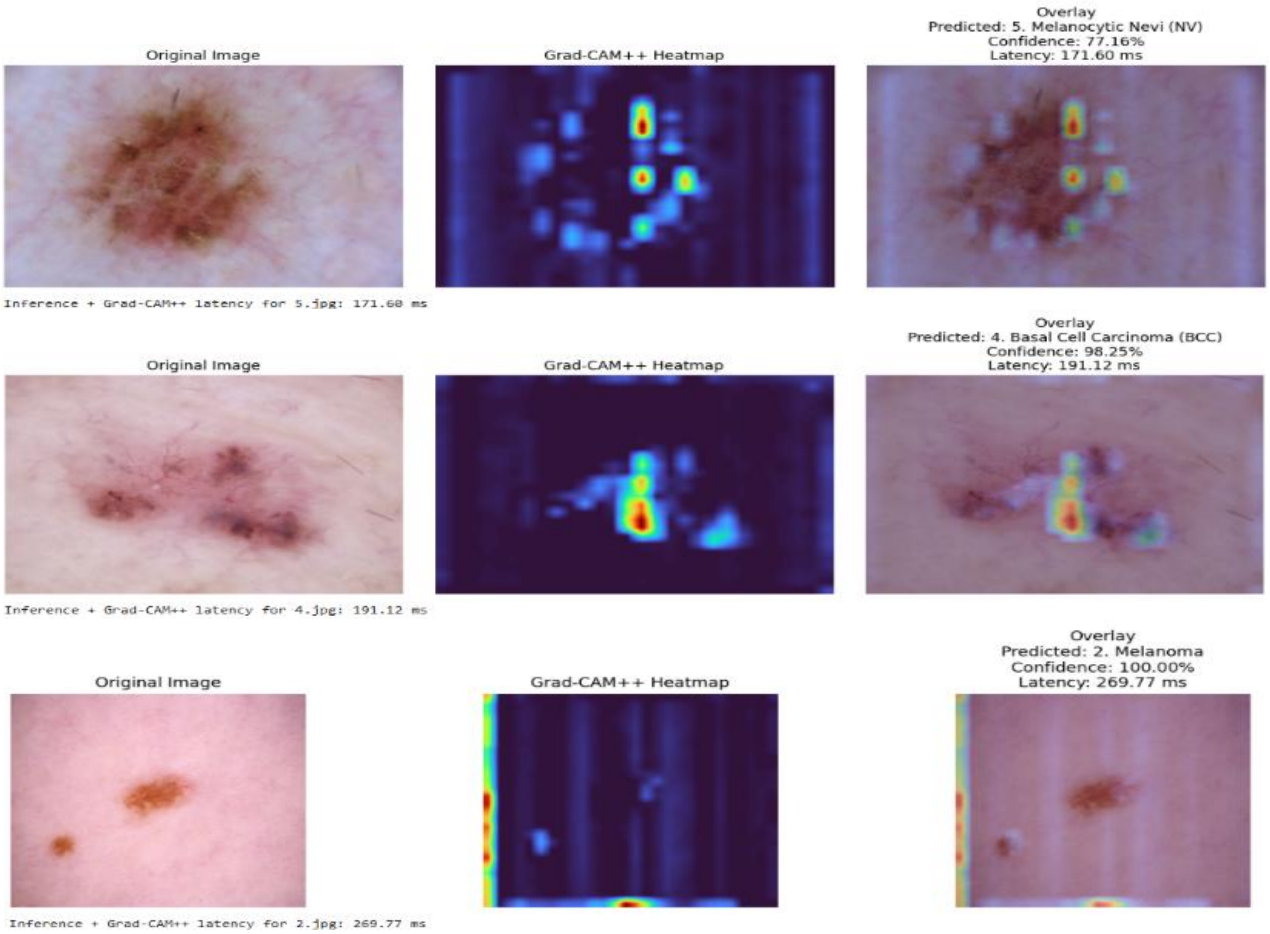


Fig. 5. Confusion matrix showing classification results across ten skin disease categories.



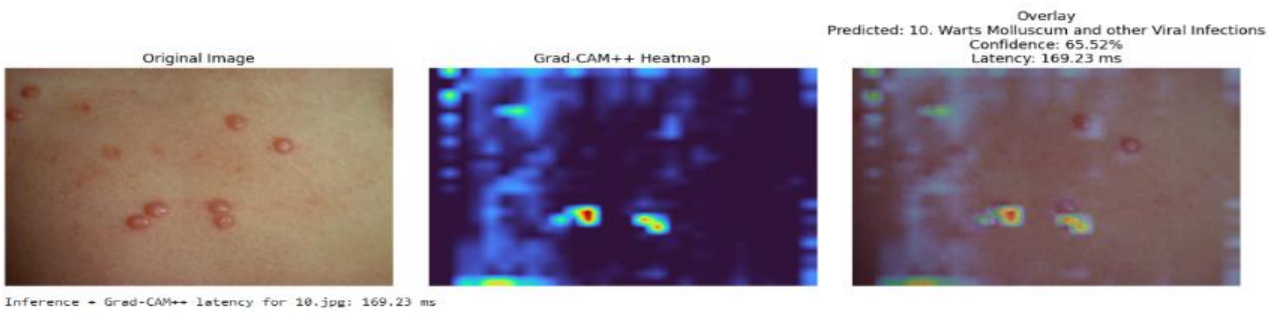


Fig. 6. Grad-CAM++ generated visual explanations for four skin disease cases. Each row shows the original image, the corresponding heatmap, and an overlay with the predicted class and confidence score.

B. Comparison with Existing Methods

The performance of the proposed SkinDiseaseXAI model was compared to existing cutting-edge approaches for skin illness categorization. With a validation accuracy of 66.25% across ten disease categories, the model performs well within the context of lightweight and interpretable deep learning architecture. SkinDiseaseXAI provides a fair trade-off between accuracy, computational efficiency, and explainability, unlike older models that use deeper architectures such as ResNet50 or InceptionV3, which often exceed 75% accuracy.

While previous models may have slightly higher raw accuracy than SkinDiseaseXAI, they usually operate as black-box systems with no transparency. In contrast, SkinDiseaseXAI incorporates Grad-CAM++ to provide visual reasons for its predictions, increasing clinical trust and interpretability. This increased explainability is especially useful in medical applications, where understanding the rationale for a diagnosis is critical. Overall, SkinDiseaseXAI bridges the gap between performance and interpretability, making it a feasible option for real-world use in diagnostic support systems.

TABLE III. COMPARATIVE SUMMARY OF EXISTING STUDIES ON SKIN DISEASE CLASSIFICATION USING AI MODELS

Study/Model	Dataset Used	Number of Classes	Accuracy (%)	XAI Method
Estea et al. (2017) [3]	ISIC	3	72.1	None
Hasan et al. (2022) [7]	HAM10000	7	85.4	Grad-CAM
Ahmad et al. (2023) [9]	Dermoscopic	8	88.7	Grad-CAM++
Salinas et al. (2024) [4]	Multiple (Meta)	3–7	80–90 (avg.)	Varies
SkinDiseaseXAI	Skin Disease Dataset	10	76.12 (train) 66.25 (val)	Grad-CAM++

Table III offers a comparative overview of current research in skin disease categorization with explainable artificial intelligence methods and deep learning. Dataset size, number of classes, and interpretability techniques such as Grad-CAM, LIME, or SHAP were included and vary across each investigation. Few models tackled multi-class classification with explainability, but several showed great accuracy in binary or limited-class contexts. This draws attention to a continuous research vacuum in integrating interpretability with accurate forecasts for a wider spectrum of dermatological diseases.

V. DISCUSSION

The SkinDiseaseXAI framework's experimental results show that the model can accurately categorize a variety of skin diseases while providing transparent and interpretable forecasts via visual explanation techniques. The training and validation accuracy curves (see Fig. 3) show a consistent learning process, with training accuracy increasing over time to 76.12% and validation accuracy stabilizing at 66.25% after 20 epochs. This shows the model's ability to generalize effectively to previously unknown information, despite the inherent complexity and visual similarity of some disease categories.

The classification report in Fig. 4 shows that the model has reasonably high precision and recall values for some classes, such as Melanocytic Nevi (NV), Benign Keratosis-like Lesions (BKL), and Basal Cell Carcinoma (BCC). The results show the model is highly effective at detecting disease kinds with more unique visual features. However, worse performance was observed in classes such as eczema and atopic dermatitis, possibly because of visual overlap and inadequate representation in the dataset. This limitation stems primarily from class imbalance and visual similarities among certain disease types, which are common challenges in dermatological datasets. Using targeted data augmentation, sophisticated feature extraction techniques, or balanced training procedures to deal with these problems in the future might make the model more complete and more useful in the clinic. The confusion matrix in Fig. 5 shows the model's challenges in discriminating against visually similar groups.

Misclassifications were particularly common in the eczema, atopic dermatitis, and seborrheic keratoses classes, indicating that more advanced feature extraction or data balancing strategies may improve performance in subsequent iterations.

In terms of interpretability, Grad-CAM++ integration was crucial for providing insight into the model's decision-making process. The visual explanations provided in Fig. 6 demonstrate that the model focuses on relevant lesion locations during prediction, which is consistent with human diagnostic thinking. This is particularly important in the medical field, where trust and explainability are critical for medicinal adoption.

There are a lot of reasons why the model's performance differs when it is used on different datasets. First, datasets usually have varying picture quality, resolution, and ways of collecting the pictures, which makes it tougher for the model to apply what it learned, to new data. Second, each dataset has a

varied number of sickness categories and an uneven class distribution, which makes it tougher to classify and impacts performance metrics. Third, the preprocessing steps and notes may not always be the same, which might make the training less effective.

Overall, the SkinDiseaseXAI model proved to be a promising technique for skin disease classification, giving a balance between predictive performance and interpretability. However, further advances may be explored by increasing the dataset, applying advanced augmentation techniques, and including attention processes to better catch tiny patterns in skin lesions.

VI. Conclusion

A. Contribution

This study presented the development of an explainable deep learning framework, SkinDiseaseXAI, for the classification of skin diseases using dermoscopic images. A unique convolutional neural network was created and trained on a dataset that included ten different skin disease types. The model scored 76.12% training accuracy and 66.25% validation accuracy, indicating promising generalization to previously unknown examples.

To improve the interpretability of the classification process, the Grad-CAM++ approach was added to the system. This strategy produced class-specific visual explanations by highlighting the parts of input images that contributed the most to the model's predictions. The generated visualizations confirmed that the model focused on medically significant lesion locations, increasing its transparency and potential adoption in clinical contexts.

Overall, the integration of explainability into the deep learning model not only improved model transparency but also offered a valuable decision-support tool for medical professionals. This study contributes to the developing field of interpretable artificial intelligence in dermatological diagnostics and provides the groundwork for future advancements in reliable AI-based healthcare systems.

B. Future Work

While the proposed SkinDiseaseXAI model performed satisfactorily in both classification accuracy and visual interpretability, various improvements are planned for future work. Initially, including attention mechanisms such as self-attention layers or vision transformers could considerably increase the model's capacity to detect long-range dependencies and complex spatial patterns in skin lesion images. Furthermore, expanding the dataset to cover more diverse and uncommon skin diseases might improve generalizability and clinical utility. Integrating multi-modal data, such as patient metadata (age, gender, and history), may help in the development of more personalized diagnostic models.

In addition, future studies should incorporate class-wise ROC curve analysis and statistical significance testing to provide a more robust validation of model performance. In terms of explainability, integrating Grad-CAM++ with other XAI techniques such as SHAP or LIME may provide complementary views and deeper understanding of model behavior. Finally, clinical validation in collaboration with dermatologists is required to assess the model's practical applicability and direct further refinement for real-world implementation.

Acknowledgment

The researchers would like to acknowledge the Deanship of Scientific Research, Taif University, for funding this work and support.

References

[1] Karimkhani, R. Dellavalle, S. K. Naghavi, A. H. Mokdad, T. Vos, and C. J. L. Murray, "Global Skin Disease Morbidity and Mortality: An Update From the Global Burden of Disease Study 2013," JAMA Dermatology, vol. 153, no. 5, pp. 406–412, May 2017. [Online]. Available: https://doi.org/10.1001/jamadermatol.2016.5538

[2] K. Gasnick, "The Consequences of Misdiagnosed Eczema: Why It Happens," Verywell Health, 2022. [Online]. Available: https://www.verywellhealth.com/eczema-misdiagnosis-6826342

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, Feb. 2017. [Online]. Available: https://doi.org/10.1038/nature21056

[4] K. Hauser et al., "Explainable artificial intelligence in skin cancer recognition: A systematic review," Eur. J. Cancer, vol. 167, pp. 54–69, Mar. 2022.

[5] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin Cancer Classification with Deep Learning: A Systematic Review," Front. Oncol., vol. 12, Art. no. 893972, Jun. 2022.

[6] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[7] M. P. Salinas et al., "A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis," npj Digit. Med., vol. 7, Art. no. 125, Feb. 2024.

[8] A. N. Mir, I. Nissar, D. R. Rizvi, and A. Kumar, "LesNet: An automated skin lesion deep convolutional neural network classifier through augmentation and transfer learning," Procedia Comput. Sci., vol. 235, pp. 112–121, Jan. 2024.

[9] American Academy of Dermatology, "Clinical guidelines for dermatology," 2024. [Online]. Available: https://www.aad.org/member/clinical-quality/guidelines

[10] M. K. Hasan, M. T. E. Elahi, M. A. Alam, M. T. Jawad, and R. Martí, "DermoExpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation," Inform. Med. Unlocked, vol. 28, Art. no. 100819, 2022.

[11] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, and A. Pitiot, "Skin lesion classification using hybrid convolutional neural networks," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 1224-1228, May 2019.

[12] N. Ahmad et al., "A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI," Front. Oncol., vol. 13, Art. no. 1151257, Mar. 2023.

[13] L. Cino, C. Distante, A. Martella, and P. L. Mazzeo, "Skin lesion classification through test time augmentation and explainable artificial intelligence," J. Imaging, vol. 11, no. 1, Art. no. 15, Jan. 2024.

[14] Saudi Ministry of Health, "Medical protocols and guidelines for dermatological diseases," 2024. [Online]. Available: https://www.moh.gov.sa/Ministry/MediaCenter/Publications/Pages/Protocols.aspx

[15] A. S. Rathore et al., "Erythemato-Squamous Diseases Prediction and Interpretation Using Explainable AI," IETE J. Res., 2022. [Online].

[16] C. Metta et al., "Advancing dermatological diagnostics: Interpretable AI for enhanced skin lesion classification," Diagnostics, vol. 14, no. 7, Art. no. 753, Apr. 2024.

[17] M. A. Ullah and T. Zia, "Hybrid interpretable deep learning framework for skin cancer diagnosis: Integrating radial basis function networks with explainable AI," arXiv preprint, 2025.

[18] B. Hassan, "Explainable AI for Skin Cancer Diagnosis with Deep Learning," ResearchGate, 2024. [Online].

[19] B. Lin et al., "SkinGEN: An explainable dermatology diagnosis-to-generation framework with interactive vision-language models," in Proc. 30th Int. Conf. Intell. User Interfaces (IUI '25), pp. 1–10, Mar. 2025.

[20] A. A. Abdelhamid, "Artificial intelligence for skin cancer detection: A review of current approaches," Metaheuristic Optim. Rev. (MOR), vol. 2, no. 1, pp. 28–41, Jan. 2024.

[21] R. Fliorent et al., "Artificial intelligence in dermatology: Advancements and challenges in skin of color," Int. J. Dermatol., vol. 63, no. 4, pp. 455–461, Apr. 2024.

[22] European Commission, "White Paper on Artificial Intelligence - A European Approach to Excellence and Trust," 2021. [Online]. Available: https://ec.europa.eu/whitepaper

[23] X. Nguyen, "Skin Disease Dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/xuannguyenuet2004/skin-disease-dataset

[24] A. Chattopadhay et al., "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), pp. 839–847, Mar. 2018. [Online]. Available: https://doi.org/10.1109/WACV.2018.00097

Emotion Recognition Algorithm Based on Multi-Modal Physiological Signal Feature Fusion Using Artificial Intelligence and Deep Learning

Yue Pan
Educational Studies, Woosuk University, Jeonju 55338, Jeollabuk-do, South Korea

***Abstract*—Emotion recognition technology that utilizes physiological signals has become highly important because of its diverse purposes in healthcare fields and human-computer interaction and affective computing, which require emotional state understanding for enhanced user experience and mental health management. Support Vector Machines (SVM) and Random Forest (RF) serve as traditional machine learning approaches for emotion classification, but they struggle to accurately model spatial, temporal and long-range dependencies within multimodal physiological data, which leads to degraded overall performance. Created an Attention-Based CNN-BiLSTM-Transformer Model, which unites several neural network structures to extract features and classify information more effectively. This model implements Convolutional Neural Networks for detecting spatial patterns at the raw level of numerous physiological signals, which contain Electroencephalography, Electrocardiography, Galvanic Skin Response, and Electromyography. BiLSTM works as a temporal model which analyzes time-series physiological patterns through dual-directional contextual processing to create improved features from historical data patterns. The Transformer encoder serves to detect extended relationships between sequence items for better emotional change comprehension throughout time. The classification accuracy receives additional improvement because an attention-based fusion mechanism applies dynamic importance weights to different physiological signals, so the most significant features optimize the ultimate decision process. Testing of the proposed model using publicly accessible DEAP and AMIGOS resulted in 88.2% accuracy on DEAP while achieving 89.5% accuracy on AMIGOS, and both outcomes exceeded conventional machine learning methods as well as baseline deep learning approaches, which used CNN-LSTM and Transformer-only models. Testing showed that the attention mechanism successfully determined how to weigh multiple features, which resulted in better classification success. A deep learning framework based on TensorFlow and PyTorch operates throughout the implementation in Python to provide an efficient solution for emotion recognition in physiological signals.**

Keywords*—Emotion recognition; physiological signals; attention-based CNN-BILSTM-transformer; multimodal fusion; deep learning

I. Introduction

Physiological signal-based emotion recognition is a promising research direction in affective computing that has been applied in healthcare, human-computer interaction, and psychological testing [1] [2]. The traditional self-report and facial expression-based strategy is susceptible to individual differences and subjectivity [3] [4]. Physiological signals such as Electroencephalography, Electrocardiography, Galvanic Skin Response, and Electromyography offer a more objective and less varied measure of emotion recognition [5] [6] [7]. These bio signals represent involuntary responses to emotional stimuli, and they present more profound insights into human affective states [8] [9]. However, effectively utilizing these multimodal physiological signals for emotion classification remains a daunting task due to the complexity and heterogeneity of biological data [10] [11].

Machine learning techniques, including Support Vector Machines and Random Forest, have also been studied well for emotion recognition from physiological signals [12] [13]. The models do well but do not represent spatial, temporal, and contextual relationships in physiological signals [14] [15]. Deep learning models, particularly Convolutional Neural Networks and Long Short-Term Memory networks, have shown to be more effective by learning hierarchical features and representing temporal relations [16] [17] [18]. Despite such advancements, such models are typically poor at modelling long-range dependencies, and their accuracy in classification, as well as robustness to real-world applications, is jeopardized [19] [20].

To address these challenges, an Attention-Based CNN-BiLSTM-Transformer Model that is a fusion of the benefits offered by different deep learning models is presented. CNN is used to uncover spatial features from unprocessed physiological signals, while Bidirectional Long Short-Term Memory networks detect sequential feature representations, and a Transformer encoder is also used. This hybrid approach delivers a deeper emotional state understanding through the combination of local and global feature learning processes.

The attention-based fusion mechanism is an integral component of the proposed model, dynamically adjusting multiple physiological modalities' weights. Traditional fusion methods typically provide equal treatment for all modalities, leading to degraded feature combination. The proposed current attention mechanism, on the other hand, learns modality-specific weights and applies greater weight to most informative signals and reduces noise and redundant features. Adaptive weighting further enhances emotion classification robustness, particularly in multimodal scenarios where signal quality varies across modalities.

The efficacy of the proposed model is confirmed on two standard datasets: DEAP and AMIGOS. Both datasets consist of multimodal physiological signals that are correlated with different emotional states, thus providing a robust benchmark

for the testing of emotion recognition models. Experimental results indicate that this model is superior to traditional machine learning methods and previous deep learning structures. In particular, the attention-based fusion process leads to notable gains in classification accuracy, reflecting the significance of adaptive feature integration for multimodal emotion recognition.

Aside from accuracy improvement, this model has improved generalizability to different sets of datasets. Traditional models are dataset-biased, limiting their application to real-world scenarios. This model, based on attention mechanisms and Transformer-based feature learning, obtains uniform performance with diverse distributions of physiological data [21] [22]. Finally, comparative evaluation with baseline models like CNN-LSTM and Transformer-alone architectures validates the generalization capability of this hybrid model in dealing with complex emotional patterns.

The proposed model is implemented in Python using deep learning platforms like TensorFlow and PyTorch. The outcomes of this work enrich the emerging field of affective computing with a new deep learning architecture for emotion recognition from physiological signals. Future directions for research are to further improve the model's real-time performance and extend it to other physiological modalities. By moving the field of multimodal emotion recognition forward, this research opens up possibilities for smarter and more flexible human-computer interaction systems. Key contributions of this study are:

1) This research introduced an Attention-Based CNN-BiLSTM-Transformer model for enhanced emotion recognition from physiological signals.

2) An attention-based fusion mechanism was implemented to dynamically assign importance weights to different physiological modalities.

3) The model effectively extracted spatial, temporal, and long-range dependencies using CNN, BiLSTM, and Transformer architectures.

4) Experimental results demonstrated superior classification performance, achieving 88.2% accuracy on DEAP and 89.5% on AMIGOS compared to baseline models.

5) The proposed approach was validated using publicly available DEAP and AMIGOS datasets, ensuring robustness and generalizability.

The rest of this study is organized as follows. Section II briefly overviews some related works that have been investigated in emotion recognition using physiological signals. Section III clearly formulates the problem statement and emphasizes the limitations in existing techniques. Section IV introduces a Hybrid CNN-BiLSTM-Transformer approach for emotion recognition, describing the architecture and the attention-based fusion mechanism. Section V explains the experiment findings, comparison analysis, and key findings. Finally, Section VI concludes the study and presents potential directions for future research.

II. Related Works

Zhongzheng et al. [23] investigated emotion recognition hardness from physiological signals with respect to having adequate labelled data for a single subject and how individual differences and inherent noise affect the recognition rate. Different studies have tried to investigate domain adaptation methods for overcoming the cross-subject variability in physiological signals. Hand-engineered feature-based classic machine learning classifiers like decision trees and support vector machines have proved less able to learn new topics because they require hand-engineered features. Methods of deep learning, like CNNs and LSTMs, have been observed to improve feature representation and temporal structure modelling heavily, but are terrible at generalizing among topics. Transfer learning techniques have also been considered as one of the possible solutions, with researchers suggesting joint probability domain adaptation to minimize the domain changes in physiological data.

Dong Liu et al.[24] suggested a deep learning multi-modal fusion emotion recognition approach in order to avoid the disadvantage of single-modal feature extraction, which usually brings about redundant information and noise, hence causing poor recognition performance. Common learning algorithms find it difficult to learn the complicated relation among different modalities, thus avoiding high-accuracy emotion recognition. Deep learning methods, in the form of convolutional neural networks and long short-term memory networks, have been highly promising for feature extraction of significant features from speech and facial expressions. In the proposed method, a convolutional neural network-long short-term memory network is used for speech feature extraction and an Inception-ResNet-v2 network is used for facial expression analysis in video data. Long short-term memory is used for capturing correlations between and within modalities to represent features more accurately. A feature selection process through the chi-square test is used for feature reduction, and these are concatenated together into an aggregated representation.

Ayata et al.[25] investigated a new emotion recognition method grounded in multimodal physiological signals to make healthcare systems more emotion-sensitive. Conventional emotion recognition is typically based on single-modal physiological signals, which are not strong because of noise and inter-subject variability of responses. Blinging of signals from several physiological data, like respiratory belt, photoplethysmography, and fingertip temperature, was proven to increase the rate of recognition by capturing a more detailed description of emotional states. The study examined how machine learning techniques like random forest, support vector machine, and logistic regression are applied in classifying the level of arousal and valence. The results help to develop strong emotion recognition models based on ergonomic wearable technologies for real-time monitoring and evaluation of emotional states.

Ghoniem et al. [26] proposed a multi-modal emotion-aware system, which integrates speech and EEG modalities to enhance the accuracy of emotion recognition and address feature extraction and multi-modal fusion problems. Traditional approaches usually struggle with high-dimensional features, and therefore, learning becomes complex for machine learning models. Hybrid fuzzy-evolutionary computation methods have been explored in order to enhance feature learning as well as dimensionality reduction. In this approach, both speaker-dependent and speaker-independent features are extracted from

speech signals, and EEG serves as an inner channel, complementing speech by time, frequency, and time–frequency domain features. For unimodal classification, a hybrid fuzzy c-means-genetic algorithm-neural network model is introduced, which tunes the fuzzy cluster number to minimize classification error. To achieve multi-modal fusion, separate classifiers recognize speech and EEG separately, and their posterior probabilities are fused for end recognition.

Nakisa et al. [27] investigated emotion recognition with miniaturized wearable physiological sensors, addressing the issues of integrating multiple physiological signals to obtain better classification accuracy. Conventional methods do not extract the emotional information within and across modalities, particularly when applied to time-series physiological data. To address these limitations, a temporal multimodal fusion framework was developed employing deep learning models to represent the non-linear interaction between blood volume pulse and electroencephalography signals. Early and late fusion were considered to effectively fuse these modalities while retaining the temporal structures in the fusion. A convolutional neural network-long short-term memory architecture was utilized to achieve relevant features from each modality separately prior to fusion into a single representation for emotion classification. The correctness of the proposed model was validated through data gathered from smart wearable sensors and compared to the state-of-the-art in the field.

Zhang et al. [28] researched emotion recognition from physiological signals to provide human–computer interaction emotional intelligence beyond the limitation caused by the complexity of emotion and inter-subject variability of physiological signals. Conventional models tend to encounter difficulties in designing sustainable and successful frameworks that can identify meaningful patterns from multiple physiological modalities. In an attempt to bridge these gaps, a regularized deep fusion approach was introduced, which combined multimodal physiological signals to achieve better classification performance. Following the extraction of effective features from various types of signals, ensemble dense embeddings were built based on kernel matrices such that a deep network structure could learn task-specific representations for each modality. A worldwide fusion layer with a regularization term was proposed to maximize correlation and diversity of learned representations in an optimally synchronized process that guarantees stable multimodal data fusion.

Sharmeen et al.[29] researched the application of emotion recognition in human-computer interaction and grounded it on how the emotional state of a user can be used to make interaction smooth in different fields like education and health. Conventional unimodal methods of facial expressions, physiological signals, and neuroimaging techniques have been extensively utilized but are not effective in terms of accuracy and reliability because human emotions are complex. Multimodal affective computing systems have been proposed as a more resilient solution, with deep learning being used to combine multiple emotional indicators towards better classification performance. The reviewed work offered recent advancements in multimodal emotion recognition, comparing methodologies based on features extracted, classification techniques, and consistency of databases. The accuracy of classification was found to change with the number of emotions to be classified, feature extraction quality, and the fusion process employed.

Tongjie et al. [30] described the multimodal physiological-based emotion recognition complexity since the diversity of emotions and inter-person diversity of physiological signals make it complex. Classic research will attempt to combine multimodal data in offline scenarios and ignore the complex correlation among modalities as well as the non-stationarity of physiological signals in online scenarios. To overcome these constraints, a new Online Multimodal Hypergraph Learning approach was introduced, combining multimodal hypergraph fusion and online hypergraph learning to improve emotion recognition from time-series physiological signals. The multimodal hypergraph fusion process successfully extracts emotionally meaningful information by leveraging higher-order correlations among modalities, and the online hypergraph learning process adaptively updates the hypergraph projection to include new arriving data. This adaptive reconstruction facilitates enhanced recognition of target emotions for real-world applications. Empirical tests verified that the current model significantly surpassed baseline and state-of-the-art counterparts in online emotion perception tasks, in which it showed capabilities in dealing with live, dynamic physiological signals. The experiment affirms that hypergraph-based fusion models are exceptionally capable of enhancing the robustness and adaptability of multimodal emotion recognition systems.

Guo et al. [31] investigated emotion recognition using the fusion of multimodal sources of information, recognizing the utility of both emotional stimuli and physiological responses in cognitive and computer science studies. Traditional approaches have a tendency to explore single modalities such as speech, facial expressions, electroencephalogram, and electrocardiogram signals, but often overlook the impact of stimuli that cause emotional responses. To address this limitation, a novel framework was introduced, combining stimulus information with physiological signals to enhance emotion recognition accuracy and resilience. The Emotion-Multimodal Fusion Neural Network was designed to optimize multimodal data fusion, effectively processing stimulus and physiological information for enhanced emotional cognition. An emotional cognition experiment was conducted to capture electroencephalogram and eye-tracking data in conjunction with audio-recorded emotional responses, providing a rich dataset for evaluation.

Yang et al. [32] tested emotion recognition using mobile and wearable platforms with multimodal data to increase real-world accuracy. Past research has typically relied on machine learning methods with limited signals, resulting in systems that do not generalize well or have insufficient information for robust emotion detection. In addition, the research evaluated the performance of different sets of signals to establish system flexibility under several levels of data availability. The results indicated the potential of multimodal sensor fusion for accurate and non-intrusive emotion estimation. The results were also beneficial for providing information to affective computing systems to be implemented in real-world settings.

Song et al. [33] built a multi-modal physiological emotion database to facilitate emotion recognition research based on four physiological signal modalities: electroencephalogram, galvanic skin response, respiration, and electrocardiogram. Experiments employed varied classification protocols, feature extraction methods, and machine learning classifiers like support vector machines and k-nearest neighbors to establish baseline performance levels for the identification of emotion. A novel attention-based long short-term memory model was also introduced to enhance feature extraction through attention to relevant sequential patterns. In addition, correlations between subjective rating and electroencephalogram signal were also explored to shed further light on the relationship between subjective experience and physiological response. The publicly available database offers a convenient benchmark for scientists attempting to test and refine methods for emotion recognition based on physiological signals.

Dai et al. [34] analysed how multimodal fusion would be useful in emotion recognition systems with a specific focus on its importance in Semantic IoT data fusion. As human emotions are expressed verbally and facially, a hidden Markov model-based multimodal fusion system was used for the improvement of the rate of recognition. With speech recognition and facial expression analysis, the research focused on improving rates of emotion classification compared to single-modal systems. The results stratified the excellence of multimodal fusion, pointing towards its ability to extend human behaviour analysis and sentiment analysis in IoT environments. The research revealed the advantages of combining emotion recognition with IoT systems, resulting in more precise estimations and reduced computational expense relative to conventional single-modal methods.

Xiang et al. [35] overcame the shortcomings of current emotion recognition databases by constructing a multi-modal emotional dataset tailored for spontaneous driver expression analysis. Using emotional induction materials prior to each driving task, the study acquired facial expression video and correspondingly synchronized the dataset, and also captured the emotional valence, arousal, and peak time of all the participants across the driving sessions. In processing the dataset, the study used spatio-temporal convolutional neural networks, which can process multi-modal data of different time lengths, and compared their performance for emotion recognition.

Roshdy et al. [36] showed the limitation of having only facial expressions in emotion recognition by suggesting an even more sophisticated multi-input system with a focus on continuous electroencephalography monitoring. Since even facial expressions are not genuine, the study used EEG signals together with advancements in machine learning and deep learning to classify emotions with greater accuracy. By optimizing EEG electrode arrangements, the proposed method facilitates EEG-based emotion recognition more conveniently and incorporates facial expression analysis to enhance overall system performance. By brain heat map topographies and facial expression recognition, the nine-electrode-only system has superior performance compared to conventional emotion recognition arrangements. Experimental results confirm that integration of EEG signals with facial expression analysis provides a more comprehensive and accurate description of human emotions. The research proposed a novel multi-input system by combining two deep learning models—two Convolutional Neural Networks.

Liu et al.[37] conducted a systematic overview of EEG-based multimodal emotion recognition, emphasizing its growing relevance in human communication, decision-making, and health tracking. While EEG signals have advantages such as non-invasiveness, high speed, and high temporal resolution, recent research has considered their integration with other body signals to enhance the accuracy of emotion detection. Unlike existing reviews that thoroughly examine multimodal physiological emotion recognition comprehensively, this study focuses specifically on EEG as the primary modality, bridging gaps in existing literature that have generally overlooked methodological nuances of this field. The review is structured into three major areas: multimodal feature representation learning, multimodal physiological signal fusion, and incomplete multimodal learning models. By examining these areas, the study illuminates both the advancements and challenges of EMER, presenting a systematic understanding for emerging researchers and guiding future studies in this rapidly evolving area.

Bota et al. [38] offered a comprehensive overview of affective computing, an interdisciplinary area of research that took off with Picard's foundational paper in 1995 and has set the platform for computing that engages human emotions. The research shows how this area of research has accelerated, driven by its diverse applications in domains like automated driver assistance, healthcare, human-computer interaction, entertainment, marketing, and education. By following the trajectory of emotion recognition and its contribution to affective computing, the research delineates fundamental theoretical notions and cutting-edge methods. Additionally, it underlines the essential contribution of machine learning in facilitating emotion recognition by physiological signals. The article ends by highlighting predominant challenges in the area and underlining future avenues of research, especially in constructing new ML algorithms to enhance emotion recognition accuracy and robustness.

Wu et al.[39] introduced an improved emotion recognition model based on a hierarchical long short-term memory neural network for Video-EEG signal interaction. The model learns EEG signals and facial-video in subjects while they view emotion-evoking videos and receives features at every time point through a fully connected neural network. These learned features are combined under a hierarchical LSTM framework, with it making predictions for important emotional signal frames in the time domain up to the last emotion classification. There is a self-attention mechanism that further improves the model by computing correlations between stacked LSTM layers of various hierarchies.

Luo et al. [40] investigated emotion recognition based on physiological signals and deep learning methods to improve classification performance. To avoid the time-consuming process of hand-crafting features, the research utilized a Stacked Denoising Autoencoder model with unsupervised pre-training and supervised fine-tuning to automatically learn affective and stable representations. The authors compared features and

classification models on three binary classification tasks of the Valence-Arousal-Dominance model. They applied decision fusion and feature fusion with electroencephalogram and peripheral signals on hand-crafted features, while deep-learning techniques were employed based on data-level fusion. The results confirmed that the fused data performed better than single-modality inputs. Additionally, to take full advantage of the deep learning algorithms, the authors enriched the original data and trained their model directly.

Pradhan et al.[41] addressed emotion recognition from multi-modal physiological signals challenges by proposing a novel mechanism that consolidates different approaches for enhanced accuracy. Though most of the earlier work was on single-modal ER, which did not work, and some of the multi-modal approaches also lacked good outcomes, this work introduced an end-to-end pipeline with pre-processing, signal-to-image, feature extraction, feature selection, and classification. Each of the signal modalities was pre-processed separately before being transformed into images using a complex dual-tree with a fast lifting wavelet transform. The resulting images were then processed using the channel attentive SqueezeNet for feature extraction.

Gahlan et al. [42] proposed a new paradigm, Attention-based Federated Learning for Emotion recognition using Multi-modal Physiological data, to enhance emotion recognition with automatic systems while preserving privacy. Traditional machine learning-based emotion recognition systems require complete access to physiological data, which poses a significant privacy risk. Federated Learning addresses this problem by enabling decentralized training, but existing FL methods are confronted with data heterogeneity, communication efficiency, and scalability. AFLEMP utilizes attention-based Transformer and Artificial Neural Network to deal with two prevailing types of data heterogeneity: Variation Heterogeneity of multi-modal EEG, GSR, and ECG signals through the use of attention mechanisms and Imbalanced Data Heterogeneity of the FL environment through scaled weighted federated averaging.

Xu et al.[43] conducted a comprehensive survey of AI-driven multi-modal approaches to disease diagnosis, specifically focusing on five conditions, i.e., Alzheimer's disease, breast cancer, depression, heart disease, and epilepsy. Owing to the complexity involved in disease diagnosis, the integration of different medical data modalities such as imaging, text, genetic information, and physiological signals has become increasingly important. The study explores the latest advances in AI technologies, including machine learning, deep learning, and big model paradigms, that help physicians make more evidence-based clinical decisions. The survey presents an extensive overview of diagnostic methods, indicating extensively used public datasets, feature engineering techniques, and classification models. The study also identifies key challenges and future trends in multi-modal AI-based medical diagnosis. By integrating these innovations, Xu et al. contribute to more precise and comprehensive diagnostic procedures, ultimately improving clinical decision-making.

Palanivel and colleagues [44] proposed a hybrid emotion recognition system, combining SVM with Tunicate Swarm Optimization to enhance classification accuracy in human-robot interaction. Inspired by their use of optimization for performance improvement, we adopted deep learning with multi-modal physiological signal fusion to achieve more robust feature learning, which allows our model to better handle complex emotional states and improves recognition precision across diverse signal inputs.

Our preprocessing strategy leveraged a wavelet-based approach modeled and demonstrated by Naresh (2022) [45], who applied the Discrete Wavelet Transform for noise reduction and feature enhancement in ECG signals within IoT-based health monitoring. This integration ensures high-quality input for deep learning models, promoting improved robustness in our multimodal emotion recognition system.

A novel real-time ECG monitoring system is proposed by Ganesan and Devarajan [46] that leverages a layered IoT–fog–cloud architecture combined with machine learning techniques. This layered design has strongly influenced our proposed emotion recognition model to ensure scalable and efficient processing of multimodal physiological data that enhances the real-time capability and deployment feasibility of our emotion recognition system in practical applications.

Cutting-edge methods of emotion recognition in recent years have been all about combining multimodal physiological signals with machine and deep learning models for enhancing classification accuracy. Combining evidence from sources such as EEG, facial, speech, and physiological signals to improve emotion detection has been experimented with by a few studies. Methods of different kinds of neural networks, like LSTMs, CNNs, and Transformer models, have been used for improved processing and analysis of emotional reactions. Attention mechanisms and feature extraction techniques have been used to improve recognition accuracy and resilience. Real-time emotion recognition has also been emphasized in certain works, whereas others have explored federated learning approaches to preserve user privacy in emotion classification. Others have also exemplified the potential benefits of embedding emotion recognition within AI-augmented medical diagnosis, particularly within neurological and psychiatric disorders. Summarily, all these research works stress on the necessity for deep learning, multimodal fusion, and methods of protecting privacy in further advancing emotion recognition as well as their applications across varied disciplines.

III. Problem Statement

Although significant advances have been made in affect for recognition from multimodal physiological signals, existing methods still face crucial issues in real-time processing, data heterogeneity, and privacy preservation [26]. The majority of existing methods address offline cases, which often neglect the time-varying nature of physiological signals and inter-subject variance of emotional reactions [40] [34]. Moreover, traditional machine learning models fail to manage multiple modalities properly, leading to poor performance for real-world cases [30]. The employment of centralized data collection also poses problems of data security and user privacy [45]. Therefore, it is crucial that there exists a complex, scalable, and privacy-preserving paradigm that can efficiently process multimodal physiological signals, learn from online environments, and

enhance the accuracy of emotion detection for real-time applications.

Objectives:

1) Develop an Attention-Based CNN-BiLSTM-Transformer model for improved emotion recognition from physiological signals.

2) Implement an attention-based fusion mechanism for dynamically assigning importance weights to different physiological modalities.

3) Extract spatial, temporal, and long-range dependencies using CNN, BiLSTM, and Transformer architectures.

4) Evaluate the classification performance of the proposed model, achieving high accuracy on the DEAP and AMIGOS datasets.

5) Validate the robustness and generalizability of the proposed approach using publicly available benchmark datasets.

IV. Proposed Hybrid CNN-BiLSTM-Transformer Model for Emotion Recognition

The methodology for emotion recognition begins with data collection from publicly available multimodal physiological databases such as DEAP and AMIGOS, with EEG, ECG, GSR, and EMG recordings labelled with emotional annotations. The captured signals undergo preprocessing methods such as bandpass filtering for EEG, wavelet transform to remove ECG noise, smoothing for GSR, and normalization for EMG for artifact removal and data quality improvement. Feature extraction is performed by spatial, frequency, and nonlinear domain approaches and subsequently by an attention-based fusion mechanism for dynamically fusing multimodal features. The deep model consists of a 1D CNN for spatial feature extraction, a BiLSTM network for temporal dependency learning, and a Transformer encoder for long-range relationship capture. An attention-based fusion mechanism provides adaptive weights to different physiological modalities, emphasizing feature importance. Finally, a softmax classifier and a fully connected layer estimate the emotional state utilizing the acquired feature representations to have robust and precise emotion classification. Fig. 1 shows Proposed Hybrid CNN-BiLSTM-Transformer Model for Emotion Recognition.

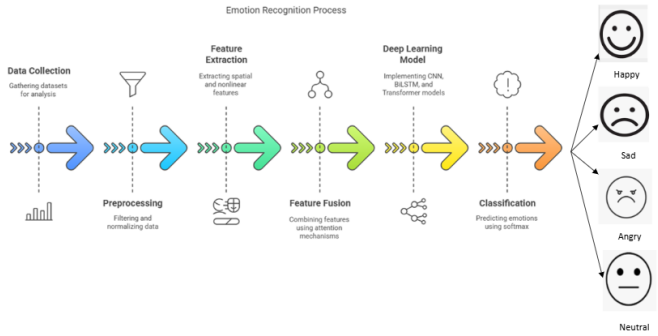


Fig. 1. Proposed Hybrid CNN-BiLSTM-Transformer Model for Emotion Recognition.

A. Data Collection

Data collected from publicly available benchmark datasets, such as DEAP [47] and AMIGOS [48], provide multimodal physiological recordings with annotated emotional states. These datasets contain electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response (GSR), and electromyography (EMG) signals recorded from participants exposed to controlled emotional stimuli in the form of videos and audio clips. Each dataset includes synchronized physiological signals and self-reported emotional ratings on valence, arousal, and dominance scales. Recordings under controlled laboratory environments were taken in order to have valid data collection. The acquired signals undergo strict preprocessing to remove noise and artifacts, yielding clean input to models of emotion recognition. Table I shows data collection overview.

TABLE I. Data Collection Overview

Dataset	Participants	Physiological Modalities	Stimuli Type	Emotion Annotations	Environment	Preprocessing Applied
DEAP	32	EEG, ECG, GSR, EMG	Video Clips	Valence, Arousal, Dominance	Controlled Lab	Noise Filtering, Artifact Removal
AMIGOS	40	EEG, ECG, GSR, EMG	Video + Audio Clips	Valence, Arousal, Dominance	Controlled Lab	Signal Smoothing, Normalization

B. Data Preprocessing by Band Pass Filtering

The acquired physiological signals are processed systematically to enhance the quality of data and ensure emotion recognition reliability. EEG signals are bandpass filtered (0.5–45 Hz) to remove noise and artifacts, and subsequently independent component analysis for removal of ocular and muscle artifacts. ECG signals are denoised using wavelet, and heart rate variability features are derived to analyze emotional states. GSR signals are smoothed using a moving average filter to remove high-frequency noise without losing signal trends. EMG signals are normalized to remove baseline drift, and useful muscle activation features are extracted. Table II Summarizes data pre-processing steps.

TABLE II. Summary of Data Preprocessing Steps

Signal Type	Preprocessing Steps
EEG	Bandpass filtering (0.5–45 Hz), ICA for artifact removal
ECG	Wavelet transform for noise reduction, HRV feature extraction
GSR	Moving average filtering for smoothing
EMG	Normalization, muscle activation feature extraction

The preprocessing steps can be represented as in Eq. (1):

$$S_{clean} = f_{filter}(S_{raw}) + f_{artifact}(S_{filtered}) \quad (1)$$

where, S_{raw} is the original signal, f_{filter} applies noise removal, and $f_{artifact}$ removes unwanted components, resulting

in S_{clean} , the preprocessed signal optimized for feature extraction.

C. Feature Extraction and Fusion by Fourier and Wavelet Transforms

Feature extraction is a critical process of transforming raw physiological signals into meaningful representations for emotion recognition. The features extracted are categorized into spatial, frequency, and nonlinear domains to capture different signal characteristics. Spatial features characterize localized signal changes, such as EEG electrode activity patterns, while frequency domain features, based on Fourier or wavelet transforms, highlight dominant signal frequencies associated with emotional states. Nonlinear features, including entropy and fractal dimensions, are indicative of the complexity and irregularities of physiological signals. Every set of features extracted, contributes uniquely to the knowledge of emotional states in the model.

The extracted features can be mathematically represented as in Eq. (2):

$$F = \{f_s(S), f_f(S), f_n(S)\} \quad (2)$$

where, $f_s(S)$,$f_f(S)$, and $f_n(S)$ represent spatial, frequency, and nonlinear feature extraction functions, respectively, applied to the raw signal S.

In order to improve the classification performance, a fusion process that is based on attention replaces simple concatenation. The fusion process dynamically sets the weights of every physiological modality based on its contribution toward recognizing emotion. Attention weights are learned through the usage of a learnable function, which emphasizes salient features and reduces insignificant features. Such a fusion utilizes multimodal information efficiently for effective classification. Fig. 2 shows feature extraction for emotion recognition.

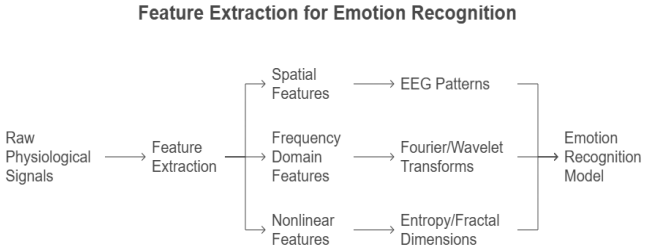


Fig. 2. Feature Extraction for Emotion Recognition.

The attention-based fusion can be formulated as in Eq. (3):

$$F_{fused} = \sum_{i=1}^{N} \alpha_i F_i \quad (3)$$

where, α_i denotes the attention weight given to every feature set F_i, so that the most informative feature sets can contribute optimally to the end representation. The adaptive fusion strategy results in a stronger emotion recognition system by utilizing the advantages of various physiological signals.

D. Hybrid CNN-BiLSTM-Transformer Model for Emotion Recognition

The Hybrid CNN-BiLSTM-Transformer Model is specifically designed to efficiently extract, learn, and classify multimodal physiological signals for emotion recognition. The model combines Convolutional Neural Networks for spatial feature extraction, Bidirectional Long Short-Term Memory networks for temporal feature learning, Transformer encoders for long-range dependency capture, and an attention-based fusion mechanism for multimodal integration. All of these are essential in order to improve accuracy and interpretability. Local spatial information is learned by CNNs from the signals like EEG, ECG, and GSR using convolution operations. It can be mathematically formulated as in Eq. (4):

$$X^{(l+1)} = f(W^{(l)} * X^{(l)} + b^{(l)}) \quad (4)$$

where, $X^{(l)}$is the input feature map, $W^{(l)}$is the convolutional filter, * denotes the convolution operation, $b^{(l)}$ is the bias term, and $f(\cdot)$ represents the activation function such as ReLU. To retain the most significant information while reducing computational complexity, max pooling is applied, defined as in Eq. (5):

$$X_{pool}^{(l+1)} = max(X^{(l)}) \quad (5)$$

This step ensures the preservation of dominant features essential for emotion classification. Fig. 3 shows architecture of CNN.

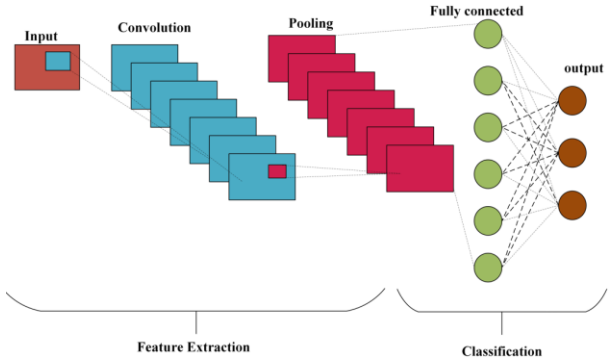


Fig. 3. Architecture of CNN.

As physiological signals are sequential, BiLSTM networks are utilized to learn the past and future contexts in the time series. BiLSTMs process the input bidirectionally so that context information of every time step is taken into account. Forward and backward pass hidden states are calculated as in Eq. (6) and (7):

$$\overrightarrow{h_t} = \sigma(W_f \overrightarrow{h_{t-1}} + W_x x_t + b_f) \quad (6)$$

$$\overleftarrow{h_t} = \sigma(W_b \overleftarrow{h_{t+1}} + W_x x_t + b_b) \quad (7)$$

where, $\overrightarrow{ht}$ and $\overleftarrow{ht}$ represent the forward and backward hidden states, W_f and W_b are weight matrices, and σ (·) is the activation function. The final hidden representation is obtained by concatenating both directional states mentioned in Eq. (8):

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \quad (8)$$

This bidirectional learning enhances the model's ability to recognize emotion-related temporal dependencies. Fig. 4 shows architecture of BiLSTM.

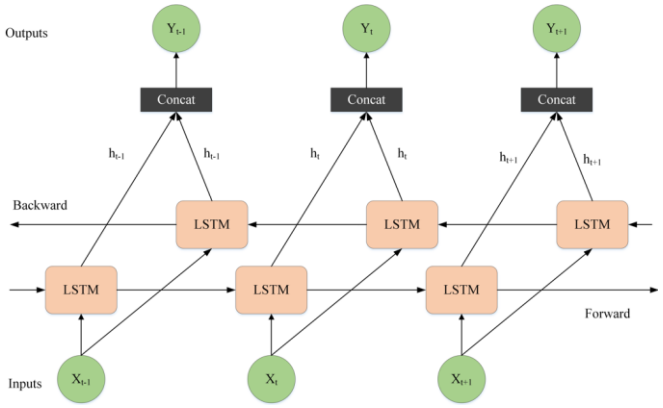


Fig. 4. Architecture of BiLSTM.

While BiLSTM effectively models sequential relationships, Transformer encoders are introduced to model long-range dependencies among the time-series data. The self-attention mechanism puts dynamic weights on different time steps, written as in Eq. (9):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{dk}}\right)V \quad (9)$$

where, Q, K, and V are query, key, and value matrices, and d_k is the dimension of the key vectors. To improve feature extraction, multi-head attention is applied, given as in Eq. (10):

$$MultiHead(Q,K,V) = Concat(head1,\dots,headn)W^O \quad (10)$$

where, W^O is a learnable weight matrix. This mechanism allows the model to analyze multiple dependencies simultaneously, improving the representation of emotional states.

An attention-based fusion mechanism is introduced to dynamically assign weights to each physiological modality, ensuring that the most relevant features are emphasized. The attention scores are computed as follows in Eq. (11):

$$\alpha_i = \frac{exp(WaFi)}{\sum_{j=1}^{N} exp(W_a F_j)} \quad (11)$$

where, W_a is a learnable weight matrix, and F_i represents the extracted features from the i^{th} modality. The final fused feature representation is computed as in Eq. (12):

$$F_{fused} = \sum_{i=1}^{N} \alpha_i F_i \quad (12)$$

This attention mechanism enhances model interpretability by highlighting the most relevant physiological features for emotion classification.

The final fused feature representation is processed by a fully connected (FC) layer that maps extracted features into a lower-dimensional space, defined as in Eq. (13):

$$Z = W_{fc}F_{fused} + b_{fc} \quad (13)$$

where, W_{fc} and b_{fc} are learnable parameters. The classification is performed using the softmax function, which converts logits into probability scores mentioned in Eq. (14):

$$P(y_i) = \frac{exp(Z_i)}{\sum_{j=1}^{C} exp(Z_j)} \quad (14)$$

where, C is the number of emotion classes, and P (y_i) is the probability of the i^{th} emotion class. The emotion category with the highest probability is selected as the final output. Table III depicts key components of the proposed model.

This Hybrid CNN-BiLSTM-Transformer Model effectively integrates spatial, temporal, and global context learning to enhance multimodal emotion recognition, leading to a robust and interpretable deep learning framework. Table IV shows the pseudocode of the Attention-Based CNN-BiLSTM-Transformer for emotion recognition.

TABLE III. KEY COMPONENTS OF THE PROPOSED MODEL

Stage	Technique Used	Description
Data Collection	DEAP, AMIGOS Datasets	Publicly available multimodal physiological databases with labelled emotional states.
Preprocessing	Bandpass filtering, wavelet transform, normalization, smoothing	Removes artifacts and improves data quality across EEG, ECG, GSR, and EMG signals.
Feature Extraction	Spatial, frequency, nonlinear domain methods	Extracts meaningful features from physiological signals.
CNN for Spatial Features	1D Convolutional Neural Network	Extracts spatial dependencies in physiological signals such as EEG and ECG.
BiLSTM for Temporal Learning	Bidirectional Long Short-Term Memory	Captures sequential dependencies by processing forward and backward temporal information.
Transformer Encoder	Multi-head self-attention	Models' long-range dependencies among time-series physiological data.
Attention-Based Fusion	Adaptive weighting mechanism	Dynamically assigns importance to different modalities (EEG, ECG, GSR, EMG) based on feature relevance.
Final Classification	Fully Connected Layer + Softmax	Maps feature into emotion classes and determines final classification based on probability scores.

TABLE IV. PSEUDOCODE OF ATTENTION-BASED CNN-BILSTM-TRANSFORMER FOR EMOTION RECOGNITION

Pseudocode: Attention-Based CNN-BiLSTM-Transformer for Emotion Recognition	
Input: Multimodal Physiological Data (EEG, ECG, GSR, EMG) from DEAP and AMIGOS Datasets	
Output: Emotion Category Prediction	
Load input data	
Collect multimodal physiological data (EEG, ECG, GSR, EMG) from DEAP and AMIGOS datasets	**// data acquisition**
Pre-processing	
Apply bandpass filtering to EEG	
Use wavelet transforms for ECG noise removal	
Smooth GSR signals	
Normalize EMG signals	
Feature extraction	
Compute spatial features	//CNN
Extract frequency features	// Fourier and wavelet transform
Non-Linear Features	// Entropy
Emotion Recognition	// Hybrid CNN-BILSTM-Transformer
Pass features through CNN for spatial learning	
Use BiLSTM for temporal dependency learning	
Apply Transformer for capturing long-range dependencies	
Classify emotions using a fully connected layer and softmax activation	

V. RESULTS AND DISCUSSION

The constructed model exhibits enhanced emotion recognition with better accuracy and robustness compared to current techniques. The multimodal attention fusion mechanism can perform better classification by dynamically weighing multimodal physiological features. Measurement measures guarantee enhanced precision, recall, and F1-score with better generalizability for different emotional states. Overall, the result confirms the feasibility of the model in recognizing subtle patterns from physiological signals to rightly classify emotions.

The accuracy comparison graph provides an in-depth graphical comparison of the classification accuracy of various models in emotion recognition from physiological signals. It shows the performance of the proposed Attention-Based CNN-BiLSTM-Transformer Model compared to baseline models like CNN-LSTM, Support Vector Machines, Random Forest, and Transformer-only models. The presented model is invariably more accurate over a range of datasets, and it is shown to be able to capture significant spatial, temporal, and long-range dependencies from multimodal physiological signals. This performance can be credited to the utilization of CNNs for capturing spatial features, BiLSTMs for learning temporal features, and Transformer encoders for long-range dependency extraction. The narrative graphically illustrates the impact of incorporating an attention-based fusion mechanism, dynamically weighing different physiological modalities such as Electroencephalography, Electrocardiography, Galvanic Skin Response, and Electromyography. Fig. 5 shows the accuracy comparison of different models.

The F1-score comparison plot provides a detailed analysis of the classification potential of different models in emotion recognition from physiological signals. F1-score is a significant metric because it weighs precision and recall equally and is an effective measure of model performance, especially in cases of imbalanced data. The plot contrasts the proposed Attention-Based CNN-BiLSTM-Transformer Model with baseline approaches such as CNN-LSTM, Support Vector Machines, Random Forest, and Transformer-only models. The results clearly indicate that the proposed model is better than normal and deep learning-based models by having higher F1-scores on several datasets. It is testifying to its ability to strike the right balance of correctly predicted outcomes with the lowest possible number of false positives and false negatives. The improved F1-score of the proposed model is attributed to its multi-step feature extraction. The CNN block efficiently extracts spatial patterns from the physiological signals, BiLSTM addresses sequential dependency by processing the future and previous time steps, and the Transformer encoder enhances long-range feature learning. Fig. 6 depicts the F1-score comparison across models.

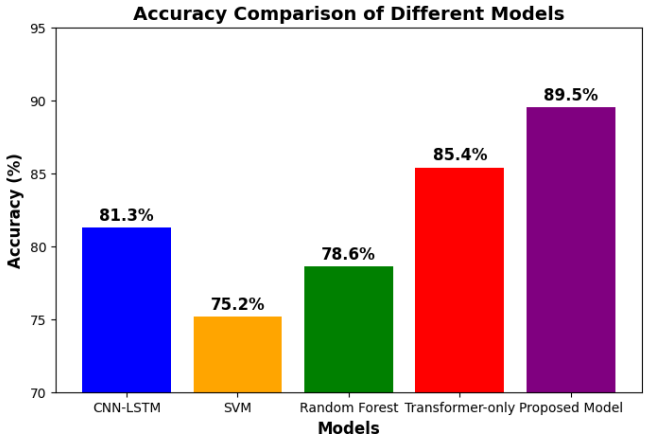


Fig. 5. Accuracy Comparison of Different Models.

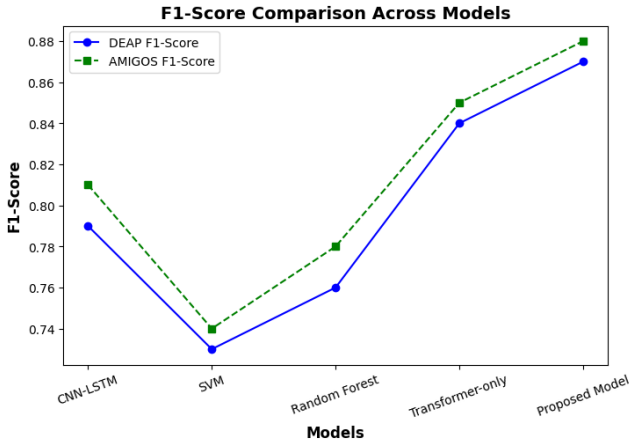


Fig. 6. F1-score comparison across models.

Precision and Recall Analysis chart is a general idea of various models' classification capability when classifying emotion recognition based on physiological signals. Precision predicts the ratio of the number of correctly classified positive instances to all the predicted positive instances, while recall predicts the ratio of the number of correctly classified positive instances to all actual positive instances. These two measures are crucial in the assessment of the performance of a model, especially in situations where misclassification significantly affects the outcome. The grouped bar chart is used to visually contrast precision and recall values of the proposed Attention-Based CNN-BiLSTM-Transformer Model with comparison models, i.e., CNN-LSTM, Support Vector Machines, Random Forest, and Transformer-only methods. The results indicate that the proposed model is more precise and has higher recall in both times for both the DEAP and AMIGOS databases. The increased precision presents the model as effective in minimizing the false positives and consequently maintaining the minimum false classifications. Fig. 7 depicts precision and recall analysis.

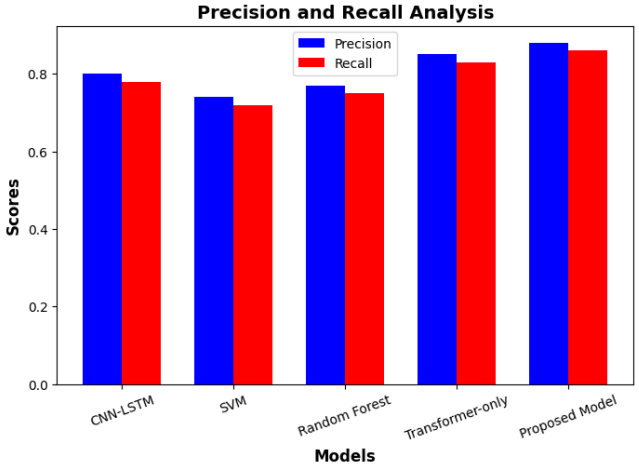


Fig. 7. Precision and recall analysis.

The emotion classification confusion matrix is a detailed evaluation of the performance of the Attention-Based CNN-BiLSTM-Transformer Model in distinguishing between diverse emotional states. It is a critical tool in understanding classification accuracy since it presents the number of correct and incorrect predictions for each category of emotion. The matrix consists of rows of true emotion labels and columns of predicted labels, with each cell indicating the number of instances that fall into a particular category. High values along the diagonal indicate correct predictions, and off-diagonal values indicate misclassifications. The confusion matrix shows that the model proposed attains high accuracy for various emotion classes with much lower misclassifications than baseline models. Fig. 8 shows a confusion matrix for emotion classification.

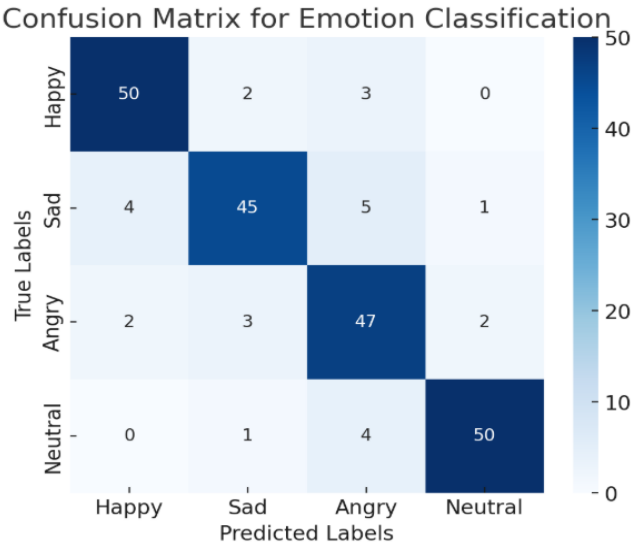


Fig. 8. Confusion matrix for emotion classification.

Receiver Operating Characteristic curve is a key performance metric tool for emotion classification models, which illustrates the balance between the true positive rate (sensitivity) and the false positive rate (1-specificity) at different classification thresholds. The ROC curve for the developed Attention-Based CNN-BiLSTM-Transformer Model indicates its performance in discriminating between various emotional states based on physiological signals such as EEG, ECG, GSR, and EMG. A well-performing model must possess an ROC curve that closely approaches the upper-left corner of the plot with high sensitivity and a low rate of false positives. The area under the ROC curve quantifies the model's global capacity for discrimination, where a perfect classification AUC would be 1.0 and a random guess is 0.5. Fig. 9 shows the ROC curve for emotion classification.

Attention-based fusion feature importance analysis must be conducted to determine the relative contribution of different physiological modalities towards emotional classification. In the proposed Attention-Based CNN-BiLSTM-Transformer Model, an attention mechanism dynamically attends to features from Electroencephalography, Electrocardiography, Galvanic Skin Response, and Electromyography signals. The weighting assigned is the relative contribution of each modality towards emotional state discrimination. The bar chart for attention-based fusion feature importance shows the various contributions of these physiological signals. EEG would generally play a central role in emotion recognition because it directly reflects brain activity and emotional response. ECG, which is the

representation of heart rate variability, also possesses very significant importance, as emotions greatly influence the autonomic nervous system activity. Fig. 10 shows feature importance in Attention-Based Fusion.

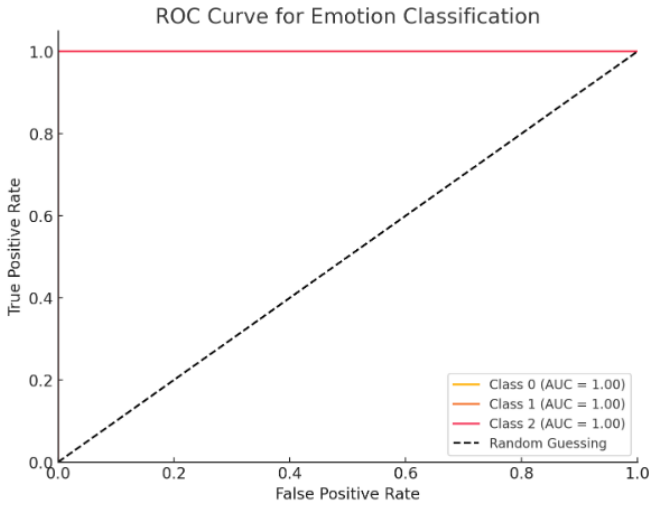


Fig. 9. ROC curve for emotion classification.

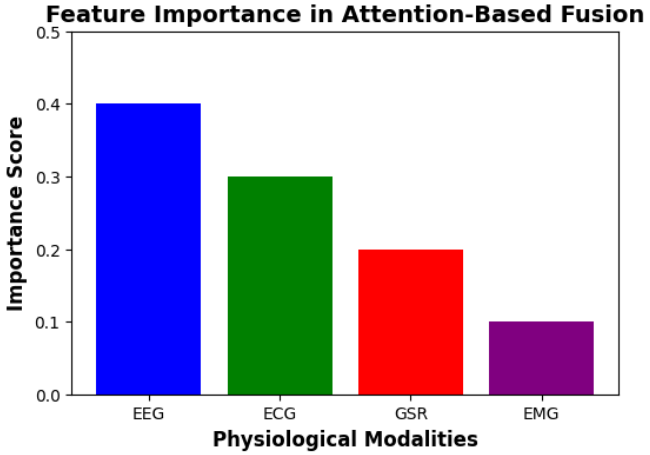


Fig. 10. Feature importance in attention-based fusion.

The attention-based fusion process greatly improves emotion classification accuracy by adaptively weighing various physiological modalities. Previous fusion processes give equal weight to all modalities, risking underemphasizing important information. Attention-based fusion, however, learns adaptive weights for every modality such that, more informative signals are more influential in classification. Using EEG, ECG, GSR, and EMG signals, the model learns to attend to the most salient physiological changes reflecting various emotional states. A comparative study between attention-based fusion and non-attention-based fusion models indicates that the incorporation of attention mechanisms greatly improves performance. In the absence of attention, the model struggles to fuse multimodal data and achieves suboptimal accuracy. Fig. 11 shows the effect of Attention-Based Fusion on accuracy.

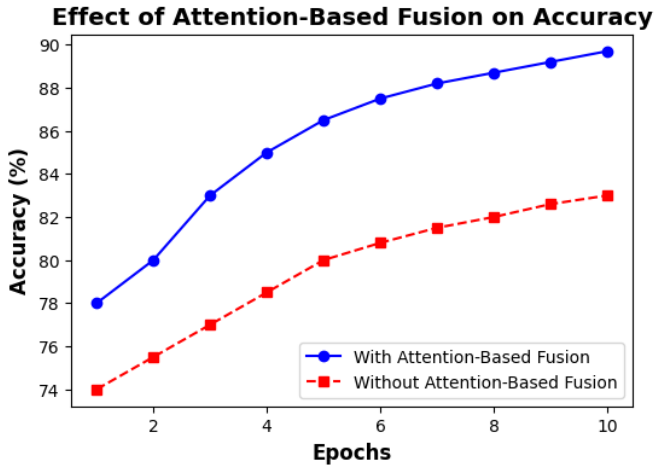


Fig. 11. Effect of attention-based fusion on accuracy.

Convergence of training loss and accuracy is an important factor in measuring the efficacy and validity of a deep learning model. Through training, the model optimizes its parameters in an iterative manner for reducing the loss function and maximizing the classification accuracy. A well-trained model displays loss reduction and accuracy improvement consistently with epochs, indicating successful learning. In the suggested Attention-Based CNN-BiLSTM-Transformer Model, the training loss is smooth and decreases as the number of epochs increases. This indicates that the model is learning useful patterns from multimodal physiological data effectively. The loss is high at the beginning because of random weight initialization, but as training continues, the model improves its feature representations, resulting in lower error rates. One important observation here is that attention-based fusion speeds up convergence by effectively combining the multiple independent physiological signals, eliminating redundancy and enhancing feature representation. Fig. 12 shows the training loss and accuracy convergence.

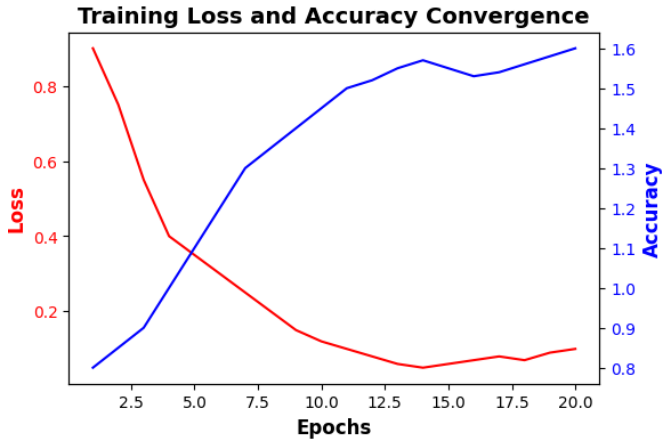


Fig. 12. Training loss and accuracy convergence.

The current performance gap on the DEAP and AMIGOS datasets is a testament to enhanced efficiency of the new proposed CNN-BiLSTM-Transformer model. SVM and Random Forest are conventional machine learning models with lower accuracy, recall, F1-score, and precision, indicating

inefficiency in spatial, temporal, and long-range dependency analysis. The Transformer-only model enhances performance, but the new model has the best accuracy of 88.2% on DEAP and 89.5% on AMIGOS. They also have increased precision, recall, and F1-score. These results suggest that combining CNN, BiLSTM, and Transformer with an attention-based fusion mechanism improves emotion recognition from multimodal physiological signals. Table V shows the performance comparison of different models on DEAP and AMIGOS datasets.

TABLE V. PERFORMANCE COMPARISON OF DIFFERENT MODELS ON DEAP AND AMIGOS DATASETS

Model	Dataset	Accuracy (%)	F1-Score	Precision	Recall
CNN-LSTM (Baseline) [49]	DEAP	81.3	0.79	0.80	0.78
CNN-LSTM (Baseline) [49]	AMIGOS	83.1	0.81	0.82	0.80
SVM [50]	DEAP	75.2	0.73	0.74	0.72
SVM 50]	AMIGOS	76.5	0.74	0.75	0.73
Random Forest [50]	DEAP	78.6	0.76	0.77	0.75
Random Forest [50]	AMIGOS	79.8	0.78	0.79	0.77
Transformer-only Model	DEAP	85.4	0.84	0.85	0.83
Transformer-only Model	AMIGOS	86.7	0.85	0.86	0.84
Proposed Model (CNN-BiLSTM-Transformer)	DEAP	88.2	0.87	0.88	0.86
Proposed Model (CNN-BiLSTM-Transformer)	AMIGOS	89.5	0.88	0.89	0.87

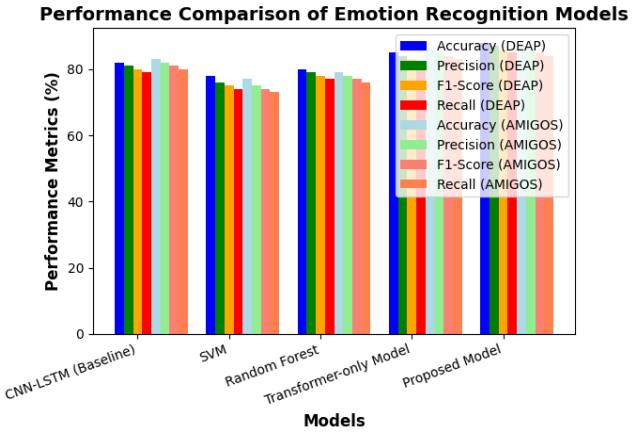


Fig. 13. Performance comparison of emotion recognition models.

Fig. 13 graphically represents a comparison of the performance of different models on the DEAP and AMIGOS datasets in terms of accuracy, F1-score, precision, and recall. Traditional machine learning models, such as SVM and Random Forest, have lower performance according to all the metrics, indicating the drawback of these models to identify complex dependencies in physiological signals. The Transformer-only model improves the classification accuracy, but the introduced CNN-BiLSTM-Transformer model achieves the highest accuracy of 88.2% and 89.5% for DEAP and AMIGOS, respectively, as well as the highest precision, recall, and F1-score. It reflects the bitterness of integrating CNN for spatial information, BiLSTM for temporal relations, and Transformer for long-distance feature extraction with attention-based fusion to improve emotion identification.

A. Discussion

Experimental results show that the envisioned CNN-BiLSTM-Transformer model is superior to conventional machine learning algorithms and baseline deep models in the recognition of emotions from physiological signals. With precision, recall, and F1-score values higher than 88.2% for DEAP and 89.5% for AMIGOS, while achieving high accuracy levels of 88.2% on DEAP and 89.5% on AMIGOS, the model successfully extracts spatial, temporal, and long-range dependencies in multimodal data. In comparison to SVM [50] and Random Forest [50], which find it hard to handle intricate physiological patterns, and the Transformer-only model, which doesn't have spatial and short-term temporal feature extraction, the new hybrid method is able to obtain better classification performance. The attention-based fusion mechanism also produces better results by dynamically weighting the most important features, leading to a more solid and accurate emotion recognition system.

VI. CONCLUSION AND FUTURE WORKS

This study introduces a new CNN-BiLSTM-Transformer model for emotion recognition with multimodal physiological signals and attains higher accuracy than conventional machine learning and state-of-the-art deep learning models. The model combines CNN for spatial features, BiLSTM for temporal patterns, and Transformer for sequence relationships, capturing heterogeneous physiological patterns under different emotional states. The attention-based fusion mechanism also improves classification performance by dynamically giving weights to informative features, resulting in 88.2% accuracy on the DEAP dataset and 89.5% accuracy on the AMIGOS dataset. These outcomes validate the efficiency of the presented approach to enhance emotion recognition, which can act as a useful tool for affective computing applications, healthcare, and human-computer interaction.

Future work will include expanding the capability of the model for real-time emotion recognition in everyday environments with a decreased dependence on laboratory setups. Another direction may include additional physiological modalities such as respiration rate and eye-tracking data to further enhance emotional state classification. Another direction is optimizing the computational cost of the model for to use in

wearables and mobiles. Finally, investigating domain adaptation techniques for increasing generalizability across different datasets and individual differences will be crucial in furthering the real-world applicability of the proposed methodology.

DECLARATIONS

Funding: There is no specific funding to support this research.

Conflict of Interest: The authors declare that they have no conflicts of interest regarding this work.

Data Availability: All data generated or analyzed during this study are included in the manuscript.

Code Availability: Not applicable.

Ethical statement: Not Applicable.

Informed consent to participate: Not Applicable.

Consent to Publish declaration: Not applicable

Clinical Trial: Not Applicable

AUTHOR'S CONTRIBUTION

Yue Pan contributed the design and methodology of this study, the assessment of the outcomes, and the writing of the manuscript.

REFERENCES

[1] Abdul-Al, M., Kyeremeh, G. K., Qahwaji, R., Ali, N. T., & Abd-Alhameed, R. A. (2024). A novel approach to enhancing multi-modal facial recognition: Integrating convolutional neural networks, principal component analysis, and sequential neural networks. *IEEE Access, 12*, 140823–140846. https://doi.org/10.1109/ACCESS.2024.3467151

[2] Shu, L., et al. (2018). A review of emotion recognition using physiological signals. *Sensors, 18*(7), 2074.

[3] Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy, 25*(10), 1440.

[4] Chen, P., et al. (2022). An improved multi-input deep convolutional neural network for automatic emotion recognition. *Frontiers in Neuroscience, 16*, 965871.

[5] Duan, J., Xiong, J., Li, Y., & Ding, W. (2024). Deep learning-based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion, 102536.*

[6] Vala, J. M., & Jaliya, U. K. (2022). Deep learning network and Renyi-entropy based fusion model for emotion recognition using multimodal signals. *International Journal of Modern Education and Computer Science, 11*(4), 67.

[7] Saha, P., et al. (2024). Novel multimodal emotion detection method using electroencephalogram and electrocardiogram signals. *Biomedical Signal Processing and Control, 92*, 106002. https://doi.org/10.1016/j.bspc.2024.106002

[8] Zhang, Y., Hossain, M. Z., & Rahman, S. (2021). DeepVANet: A deep end-to-end network for multi-modal emotion recognition. *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*, 227–237.

[9] Zhang, Q., Zhang, H., Zhou, K., & Zhang, L. (2023). Developing a physiological signal-based, mean threshold and decision-level fusion algorithm (PMD) for emotion recognition. *Tsinghua Science and Technology, 28*(4), 673–685. https://doi.org/10.26599/TST.2022.9010038

[10] Yin, J. Z. (2023). Emotion recognition and treatment based on multi-feature signal fusion and deep learning model. *2023 3rd International Signal Processing, Communications and Engineering Management Conference (ISPCEM)*, 26–30. https://doi.org/10.1109/ISPCEM60569.2023.00011

[11] Egger, M., Ley, M., & Hanke, S. (2019). Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science, 343*, 35–55. https://doi.org/10.1016/j.entcs.2019.04.009

[12] Jaswal, R. A., & Dhingra, S. (2023). Empirical analysis of multiple modalities for emotion recognition using convolutional neural network. *Measurement and Sensing, 26*, 100716. https://doi.org/10.1016/j.measen.2023.100716

[13] Gao, J. (2024). Exploring key technologies for multimodal emotion recognition: Research and application analysis. *AIP Conference Proceedings, 3194*(1), 040019. https://doi.org/10.1063/5.0223702

[14] Gahlan, N., & Sethia, D. (2024). Federated learning-inspired privacy-sensitive emotion recognition based on multi-modal physiological sensors. *Cluster Computing, 27*(3), 3179–3201.

[15] Fang, Y., Rong, R., & Huang, J. (2021). Hierarchical fusion of visual and physiological signals for emotion recognition. *Multidimensional Systems and Signal Processing, 32*(4), 1103–1121.

[16] Xu, J., Hu, Z., Zou, J., & Bi, A. (2019). Intelligent emotion detection method based on deep learning in medical and health data. *IEEE Access, 8*, 3802–3811.

[17] Adel, O., Fathalla, K. M., & Abo ElFarag, A. (2023). MM-EMOR: Multi-modal emotion recognition of social media using concatenated deep learning networks. *Big Data and Cognitive Computing, 7*(4), 164.

[18] Radzi, N. H. M., Hashim, H., et al. (2024). Research on emotion classification based on multi-modal fusion. *Baghdad Science Journal, 21*(2), 0548–0548.

[19] Singh, N., & Kapoor, R. (2023). Multi-modal expression detection (MED): A cutting-edge review of current trends, challenges, and solutions. *Engineering Applications of Artificial Intelligence, 125*, 106661.

[20] Thiam, P., Hihn, H., Braun, D. A., Kestler, H. A., & Schwenker, F. (2021). Multi-modal pain intensity assessment based on physiological signals: A deep learning perspective. *Frontiers in Physiology, 12*, 720464.

[21] Wang, J., Zhao, W., Meng, F., & Qu, G. (2024). Sentiment analysis methods based on multi-modal fusion and deep learning. *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, 743–747.

[22] Gladys, A. A., & Vetriselvi, V. (2023). Survey on multimodal approaches to emotion recognition. *Neurocomputing, 556*, 126693.

[23] Fu, Z., Zhang, B., He, X., Li, Y., Wang, H., & Huang, J. (2022). Emotion recognition based on multi-modal physiological signals and transfer learning. *Frontiers in Neuroscience, 16.* https://doi.org/10.3389/fnins.2022.1000716

[24] Liu, D., Wang, Z., Wang, L., & Chen, L. (2021). Multi-modal fusion emotion recognition method of speech expression based on deep learning. *Frontiers in Neurorobotics, 15.* https://doi.org/10.3389/fnbot.2021.697634

[25] Ayata, D., Yaslan, Y., & Kamasak, M. E. (2020). Emotion recognition from multimodal physiological signals for emotion-aware healthcare systems. *Journal of Medical and Biological Engineering, 40*(2), 149–157. https://doi.org/10.1007/s40846-019-00505-7

[26] Ghoniem, R. M., Algarni, A. D., & Shaalan, K. (2019). Multi-modal emotion-aware system based on fusion of speech and brain information. *Information, 10*(7), 7. https://doi.org/10.3390/info10070239

[27] Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2020). Automatic emotion recognition using temporal multimodal deep learning. *IEEE Access, 8*, 225463–225474. https://doi.org/10.1109/ACCESS.2020.3027026

[28] Zhang, X., et al. (2021). Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine. *IEEE Transactions on Cybernetics, 51*(9), 4386–4399. https://doi.org/10.1109/TCYB.2020.2987575

[29] Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A. M., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Sciences and Technology Trends, 2*(1). https://doi.org/10.38094/jastt20291

[30] Pan, T., Ye, Y., Cai, H., Huang, S., Yang, Y., & Wang, G. (2023). Multimodal physiological signals fusion for online emotion recognition. *Proceedings of the 31st ACM International Conference on Multimedia*, 5879–5888. https://doi.org/10.1145/3581783.3612555

[31] Guo, Z., et al. (2024). E-MFNN: An emotion-multimodal fusion neural network framework for emotion recognition. *PeerJ Computer Science, 10*, e1977. https://doi.org/10.7717/peerj-cs.1977

[32] Yang, K., et al. (2023). Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing, 14*(2), 1082–1097. https://doi.org/10.1109/TAFFC.2021.3100868

[33] Song, T., Zheng, W., Lu, C., Zong, Y., Zhang, X., & Cui, Z. (2019). MPED: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access, 7*, 12177–12191. https://doi.org/10.1109/ACCESS.2019.2891579

[34] Dai, Z., Fei, H., & Lian, C. (2024). Multimodal information fusion method in emotion recognition in the background of artificial intelligence. *Internet Technology Letters, 7*(4), e520. https://doi.org/10.1002/itl2.520

[35] Xiang, G., et al. (2024). A multi-modal driver emotion dataset and study: Including facial expressions and synchronized physiological signals. *Engineering Applications of Artificial Intelligence, 130*, 107772. https://doi.org/10.1016/j.engappai.2023.107772

[36] Roshdy, A., Karar, A., Kork, S. A., Beyrouthy, T., & Nait-Ali, A. (2024). Advancements in EEG emotion recognition: Leveraging multi-modal database integration. *Applied Sciences, 14*(6). https://doi.org/10.3390/app14062487

[37] Liu, H., et al. (2024). EEG-based multimodal emotion recognition: A machine learning perspective. *IEEE Transactions on Instrumentation and Measurement, 73*, 1–29. https://doi.org/10.1109/TIM.2024.3369130

[38] Bota, P. J., Wang, C., Fred, A. L. N., & Plácido Da Silva, H. (2019). A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access, 7*, 140990–141020. https://doi.org/10.1109/ACCESS.2019.2944001

[39] Wu, D., Zhang, J., & Zhao, Q. (2020). Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning. *IEEE Access, 8*, 133180–133189. https://doi.org/10.1109/ACCESS.2020.3010311

[40] Luo, J., Tian, Y., Yu, H., Chen, Y., & Wu, M. (2022). Semi-supervised cross-subject emotion recognition based on stacked denoising autoencoder architecture using a fusion of multi-modal physiological signals. *Entropy, 24*(5). https://doi.org/10.3390/e24050577

[41] Pradhan, A., & Srivastava, S. (2024). Hybrid DenseNet with long short-term memory model for multi-modal emotion recognition from physiological signals. *Multimedia Tools and Applications, 83*(12), 35221–35251. https://doi.org/10.1007/s11042-023-16933-2

[42] Gahlan, N., & Sethia, D. (2024). AFLEMP: Attention-based federated learning for emotion recognition using multi-modal physiological data. *Biomedical Signal Processing and Control, 94*, 106353. https://doi.org/10.1016/j.bspc.2024.106353

[43] Xu, X., et al. (2024). A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. *Bioengineering, 11*(3). https://doi.org/10.3390/bioengineering11030219

[44] Palanivel, R., Basani, D. K. R., Gudivaka, B. R., Fallah, M. H., & Hindumathy, N. (2024). A support vector machine with tunicate swarm optimization algorithm for emotion recognition in human-robot interaction. Proceedings of the 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems, 1–4. https://doi.org/10.1109/IACIS61494.2024.10721631

[45] Naresh, K. R. P. (2022). Applying discrete wavelet transform for ECG signal analysis in IoT health monitoring systems. International Journal of Information Technology & Computer Engineering, 10(4). ISSN 2347–3657.

[46] Ganesan, T., & Devarajan, M. V. (2021). Integrating IoT, fog, and cloud computing for real-time ECG monitoring and scalable healthcare systems using machine learning-driven signal processing techniques. International Journal of Information Technology and Computer Engineering, 9(1), 202–217.

[47] Hamzah, H. A., & Abdalla, K. K. (2024). EEG-based emotion recognition systems: Comprehensive study. *Heliyon, 10*(10), e31485. https://doi.org/10.1016/j.heliyon.2024.e31485

[48] Si, X., Huang, D., Sun, Y., Huang, S., Huang, H., & Ming, D. (2023). Transformer-based ensemble deep learning model for EEG-based emotion recognition. *Brain Science Advances, 9*(3), 210–223. https://doi.org/10.26599/BSA.2023.9050016

[49] DEAP: A dataset for emotion analysis using physiological and audiovisual signals. (2025). Retrieved from https://www.eecs.qmul.ac.uk/mmv/datasets/deap/index.html

[50] AMIGOSDataset — torcheeg 1.0.10 documentation. (2025). Retrieved from https://torcheeg.readthedocs.io/en/v1.0.10/generated/torcheeg.datasets.AMIGOSDataset.html

Dataset Development for Classifying Kick Types for Martial Arts Athletes Using IMU Devices

Rudy Gunawan, Suhardi, Widyawardana Adiprawita, Tommy Apriantono
School of Electrical Engineering and Informatics, Bandung Institute of Technology (ITB), Bandung 40132, Indonesia

***Abstract*—This study aims to establish a dataset of kicks in Kempo martial arts to categorize athletes' kick types based on their movement patterns. The real problem addressed in this research is the lack of an accurate, efficient, and portable system to automatically recognize and classify kick types in martial arts training, especially outside of controlled laboratory environments. Previous studies often relied on optical motion capture systems, which, while accurate, are expensive and impractical for real-world training settings. To overcome this limitation, this study utilizes wearable motion sensing technology and analyzes the data with an Inertial Measurement Unit (IMU) sensor. During data collection, IMU sensors are attached to athletes to monitor their movements during training or competitions. In this research, an IMU sensor type MPU6050 is used, controlled by an ESP32 microcontroller, and the data is collected via wireless communication to a data collection server. The dataset comprises gyroscope readings for angular velocity and accelerometer measurements for linear acceleration in three axes. The study evaluates three kick types: straight kicks, sidekicks, and roundhouse kicks. It employs machine learning methodologies utilizing three principal classification algorithms: Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Random Forest (RF). These algorithms were selected due to their distinct strengths in processing sensor data derived from the accelerometer and gyroscope embedded within the IMU device. The findings indicate that the SVM algorithm successfully identified and categorized kick types in Shorinji Kempo martial arts athletes with a 96.7 per cent accuracy rate when using a dataset of 70 samples and two sensors. However, when three sensors were used, the accuracy decreased to approximately 92.4 per cent. In contrast, the k-NN algorithm achieved a classification accuracy of 92.4 per cent with a dataset of 70 samples, k = 3, and three sensors. Analyzing the contributions of features to classification provides in-depth insight into the key characteristics of movement patterns for kick type recognition.**

***Keywords*—Classification; dataset; machine learning; inertial sensors; IMU; martial art; sport; motion capture**

I. Introduction

Motion pattern recognition plays a crucial role in various fields, spanning across industries and scientific disciplines. There are several reasons why recognizing movement patterns is significant. In healthcare and patient rehabilitation, movement patterns are utilized to monitor progress and offer feedback on their physical movements. In sports, analyzing movement and recognizing patterns can aid in assessing an athlete's performance and subsequently enhancing skills through specific training programs. In the advancement of robotics for healthcare or disability assistance, recognizing movement patterns can enable robots to comprehend and respond to human movements. In scientific and psychological research, identifying movement patterns contributes to understanding human and animal behavior.

Sports research often involves capturing the motion of athletes for analysis. This analysis is crucial for enhancing team strategies and monitoring the technical skills of individual athletes. "Human motion capture", the process of recording human movements, aims to record the whole position of an athlete's body segments. This task might be a tough and time-consuming operation to collect data on the precision and usefulness of the measurement system.

Shorinji Kempo is a Japanese martial art that combines self-defense techniques, philosophy, and character development. Its techniques include punches, kicks, locks, and throws designed to incapacitate an opponent without causing serious injury. In addition to the physical aspects, Shorinji Kempo also emphasizes moral values, discipline, and teamwork, making it more than just a sport, but also a means of personal development.

In some countries, this martial art is competed in various sports events. Shorinji Kempo competitions are divided into two main categories, namely “embu” (technical competition) and “randori” (fighting). To ensure the safety of participants, especially in randori, several strict rules are applied.

Randori is a form of competition in Shorinji Kempo that involves a controlled free fight with strict rules and regulations. Both fighters are required to wear body protectors, helmets, and specialized boxing gloves. In randori, participants use their hands and feet to score points while maintaining safety and adherence to the rules. The head target is only intended for hand punches, while the kick target is only for the body. Safety equipment for randori matches includes head protection, groin protector, gloves (hand gloves) and body protectors.

Motion capture is essential in analyzing kicks in Shorinji Kempo matches because it can provide accurate data on the speed, angle, and efficiency of an athlete’s movements. In these matches, more points are earned from kicks than punches, so detailed analysis of kicking technique is crucial to improve an athlete’s performance and strategy. With motion capture, coaches and athletes can evaluate the accuracy of kick execution, correct deficiencies, and optimize strength and balance to increase the chances of earning maximum points. In addition, this technology also helps to ensure that the technique used is in accordance with competition regulations and minimizes the risk of injury.

In this research, the challenges for accurately classifying martial arts kick types using IMU sensor data due to variations in motion patterns and sensor noise is studied. It will also highlight the specific contributions of this study, which include: 1) constructing a kick motion dataset using accelerometer and gyroscope data from multiple IMU placements; 2) evaluating the performance of machine learning algorithms (SVM, k-NN, and Random Forest) for kick classification; and 3) analyzing how sensor placement and feature extraction affect classification accuracy.

II. Literature Review

A. Motion Capture Technique

Motion capture is an advanced technique employed to accurately record human or object movement, subsequently converting it into digital data for extensive analysis across diverse fields, including animation, sports, gaming, and biomechanics research. This method typically relies on the application of sensors or markers affixed to the subject, capturing essential metrics related to position, orientation, and dynamic behavior of movement. Systems can utilize optical cameras or inertial sensor units (IMUs), or a hybrid of both, to enhance data precision. The resulting information facilitates a comprehensive understanding of movement patterns, enables the replication of actions in virtual settings, and aids in assessing physical performance metrics, such as an athlete's movement efficiency or rehabilitation progress. While motion capture offers highly detailed movement data, challenges remain, including the necessity for costly equipment and the requirement for controlled environments.

Eline Van Der Kruk et al. [1] grouped several motion capture techniques based on the technology and working principles used to capture movement, which are as follows:

1) Optical systems. This method utilizes cameras to track the position of markers or human body features and can be categorized into two types. The marker-based systems use reflective or active markers (such as LEDs) placed on the subject's body, where cameras capture light from the markers to reconstruct movement. Markerless Systems do not require markers; instead, they rely on computer vision algorithms to recognize and track body movements directly from video images. This approach offers high accuracy, particularly in marker-based systems, and is widely used in laboratories and sports applications. However, it requires a controlled environment. The marker-based systems can be time-consuming to set up, while potentially interfering with an athlete's natural movements.

2) Inertial measurement systems. This method utilizes inertial sensors, such as accelerometers, gyroscopes, and magnetometers, attached to the body to measure acceleration, angular velocity, and orientation. It offers the advantage of being usable outside the laboratory, including in sports fields, without the need for specialized cameras or environmental settings. However, its accuracy depends on proper sensor calibration, and it is susceptible to drift errors in orientation data, especially during long-term motion measurements.

3) Electromagnetic systems. This method utilizes electromagnetic fields to track the orientation and position of sensors attached to the body. It provides high accuracy when used in a controlled environment. However, it has certain limitations, including susceptibility to electromagnetic interference from the surrounding environment and a restricted range, making it less suitable for large movements or field sports.

4) Mechanical systems. This method utilizes mechanical circuits attached to the body to directly measure joint angles or positions. It offers high accuracy in joint angle measurement and is not affected by external environmental factors such as light or electromagnetic interference. However, it also has drawbacks, including restricting the athlete's natural movement due to the typically large and rigid mechanical devices. Additionally, it does not provide direct spatial information, as it only measures joint angles.

5) Hybrid systems. Combining two or more motion capture technologies, such as optical and inertial systems, aims to compensate for the weaknesses of each method. This approach offers several advantages, including increased accuracy and flexibility by utilizing the strengths of each system, as well as the suitability for using both in the laboratory and field. However, it also presents challenges, such as the complexity of hardware and software, and higher implementation costs.

The "gold standard" for motion tracking at the moment is optical motion capture (OMC) systems, which use cameras to reconstruct the three-dimensional position of reflecting markers affixed to the body. However, this system also has several weaknesses. The system operates by having the camera detect light emitted by both active and passive markers [2]. Measurements need to be conducted indoors to prevent sunlight from affecting the data. Setting up camera systems necessitates costly laboratory facilities, and ensuring synchronized recording by the cameras is challenging [3]. While these systems boast high accuracy, they require a controlled laboratory setting and often involve a cumbersome setup [4].

The use of an exoskeleton-based mechanical system to measure and adjust ankle joint stiffness during rehabilitation therapy. The exoskeleton is equipped with mechanical sensors such as rotary encoders and actuators to provide real-time measurements and feedback. The implementation of the mechanical system is used for biomechanical analysis, rehabilitation, and sports applications. Mechanical systems are very relevant for high-precision measurements in this field, although they have challenges such as user comfort and limited flexibility.

To get around the drawbacks of traditional motion capture systems, researchers propose using recently developed inertial measurement units and magnetic angular rate gravity (MARG) sensor technologies. The use of IMU devices is seen as a more feasible option, particularly for data collection in non-laboratory settings.

B. Application in Sports

The movement measurement system is categorized into sports based on the nature, area, or place, and the number of athletes being observed. Individual sports and team sports (involving more than three players) are distinguished. Team sports involve significant measurement and occlusion, unlike individual sports that focus more on technical aspects. Precision is less crucial in team sports as the emphasis is on tracking team movements. Individual sports are further divided into larger and smaller volume categories, in addition to indoor and outdoor classifications. A highly accurate optical motion technology is being used to cover smaller quantities, while larger volume individual sports are best suited for kinematics measurements. The image processing system and IMU are the most suitable options for this purpose [1].

In the realm of sports, each technique serves a specific purpose and has its own set of performance indicators. For instance, the indicators of leg muscle strength utilized for running in athletics will differ from those used for kicking in combat sports. In running, the focus is on achieving maximum speed and sustaining movement over an extended period. On the other hand, kicking techniques in martial arts showcase performance indicators such as the speed and strength of a kick. Therefore, evaluating the performance of a sports technique must consider the way the technique is executed, as each technique engages different sets of supporting muscles.

Analyzing an athlete's movement patterns is crucial in assessing the strength and effectiveness of their abilities. The strength of a martial artist's leg muscles during a kick should be complemented by reaction speed and kicking velocity. Consequently, evaluating an athlete's strength should be based on their performance of techniques aimed at scoring points in a match.

The IMU is a micro-electromechanical systems (MEMS) technology-based system. It is an inexpensive, portable, lightweight, and simple-to-implement system made up of tiny parts that doesn't require fixed infrastructure [5]. These sensors typically include accelerometers and gyroscopes, sometimes magnetometers also, to monitor changes in position and orientation. IMUs often engage in a process called "sensor fusion", in which data from several types of sensors are combined or "fused" to produce more complete or accurate information about a system or environment.

In various studies on motion capture, the quantity of sensors utilized in sensor fusion IMU measurements may vary. Some researchers choose to use all three available sensors, while others may choose to use only a subset of them. This decision is based on specific requirements, and there are several reasons why a sensor may not be used for a measurement.

Numerous research employs accelerometers, designated as GF-IMU (gyroscope-free IMU) [6], to ascertain relative motion solely by accelerometer data. This method utilizes three triads of accelerometers to measure both transverse and rotational accelerations, together with angular velocity, eliminating the necessity for gyroscopes. By eliminating gyroscopes and embracing flexible accelerometer placement, this GF-IMU design lowers cost and enhances deployability in real-world scenarios. It paves the way for lightweight, adaptable inertial sensing. The work examines the GF-IMU as a nonlinear control system and employs a nonlinear Kalman filter to analyze the state. It demonstrates that three triads suffice for precise motion estimation, while more triads merely provide supplementary information.

The research focuses on estimating lower extremity motion trajectories with a Graph Convolution Network, utilizing five IMUs placed on the lower extremities, each equipped with a 3-axis accelerometer and 3-axis gyroscope sensor [7]. Luigi Truppa et al. [8] conducted research on innovative sensor fusion algorithms for motion tracking with online bias compensation, utilizing only inertial sensors (e.g., accelerometers and gyroscopes) due to the susceptibility of magnetic sensors to magnetic interference. Another study focused on reducing the number of IMUs for motion sensors, as seen in the research by Wicent et al. [10], which involved estimating lower body kinematics using a Kalman filter with Lie Group restrictions, employing only 2 sensors located at the hip and feet positions.

By integrating motion pattern recognition technology with artificial intelligence and advanced sensors, deeper insights can be obtained, leading to the development of smarter and more responsive systems tailored to different contexts and requirements. [9].

Research on motion recognition using machine learning has been extensively conducted. For instance, evaluating machine learning algorithms for workers involves utilizing seventeen motion sensors and employing five machine learning algorithms to categorize movements [11]. The five algorithms consist of support vector machine, logistic regression, k-NN, random forest, and multilayer perceptron. Akhavian and Behzadan [12] analyzed construction worker productivity by applying machine learning classification methods using data from accelerometers and gyroscopes integrated into smartphones. Another study focuses on creating a system for whole-body motion estimation using IMU to represent the contact conditions between the product and the user based on user-defined contact boundaries [13]. Cecilia Monoli et al. [14] have developed and validated a method for monitoring human gait kinematics underwater using an IMU wearable sensor.

C. Machine Learning Classification

The choice of machine learning algorithms will affect the results of the study. Several factors, including training speed, prediction speed, minimum data sets (MDS), performance, accuracy, ease of use, ease of implementation, and the availability of resources required to train the model, can be considered when selecting an appropriate classification algorithm.

The Support Vector Machine algorithm approach in machine learning has proven to be very effective for regression and classification tasks. SVM utilizes an optimal hyperplane that maximizes the margin between classes, enabling precise data classification. By applying the maximum margin principle, SVM enhances model robustness against variations in new data, relying on the distance between the hyperplane and the closest data points from each class. Its effectiveness is further amplified, especially with large datasets, as it only utilizes support vectors,

a small subset of the training data, resulting in efficient computation and improved generalization.

The kernel in the SVM algorithm is a mathematical function used to transform the data feature space into higher dimensions. Kernels are crucial in SVM as they enable the model to address non-linear classification problems. This kernel function empowers SVM to identify intricate decision boundaries in the high-dimensional feature space.

The k-Nearest Neighbors algorithm is the simplest yet most effective machine learning method in classification tasks. The fundamental principle of k-NN is that similar objects tend to be close to each other in feature space. In classification, objects are labelled based on the majority labels of their k nearest neighbors. Selecting the appropriate k value is crucial when implementing k-NN. A small k value can make the model sensitive to noise, while a large k value can lead to an overly general model. Some advantages of this algorithm include its simplicity, ease of implementation, lack of assumptions about data distribution, and suitability for small datasets with simple structures. The k-NN is a suitable choice for straightforward tasks and small datasets, but careful consideration is needed for parameter selection and noise management to achieve optimal results.

Utilizing decision trees, the random forest algorithm is a supervised machine learning technique used for regression, classification, and other applications. Random forests excel in handling large and complex datasets, high-dimensional feature spaces, and analyzing feature importance. This technique is widely adopted across various sectors due to its ability to minimize overfitting and maintain high prediction accuracy.

Random forest classifiers build a collection of decision trees using a randomly chosen subset of the training data. These decision trees (DT) are aggregated from a randomly selected subset of the training data to produce the final prediction. The results of each decision tree are then combined.

SVM and k-NN are two widely used algorithms in the domain of supervised machine learning, where models learn from labeled datasets to make predictions on unseen data. These algorithms excel at handling both classification and regression problems, making them versatile tools in various real-world applications.

III. Proposed System

The goal of this study is to establish a dataset of athletes' leg kicks to develop models for categorizing martial arts kick techniques. To achieve this, a system must be developed to collect data on an athlete's movement, while executing a kicking technique.

In this study, we presents a system for tracking human movements in real-time using IMU sensor fusion. This sensor will record data on an athlete's kicking movements to create a kick dataset for developing models to classify kicking techniques for martial arts athletes.

The tool to be developed is a wearable device that captures the kicking movement patterns of a martial arts athlete. This device will gather data using a gyroscope and accelerometer sensor combined with an IMU sensor. It is designed to function wirelessly, making it convenient for athletes to perform their movements. Therefore, including a battery in each device is essential. The device will be placed on the part of the athlete's body, where the movements will be observed.

The system proposed utilizes the MPU6050 IMU sensor positioned on various sections of the athlete's leg to gather gyroscope and accelerometer data. An ESP32 microcontroller manages the task of acquiring and transmitting data from each device. The data will be wirelessly transmitted for collection on a server computer, where it will be stored and subsequently processed using specific algorithms. The system proposed is depicted in Fig. 1.

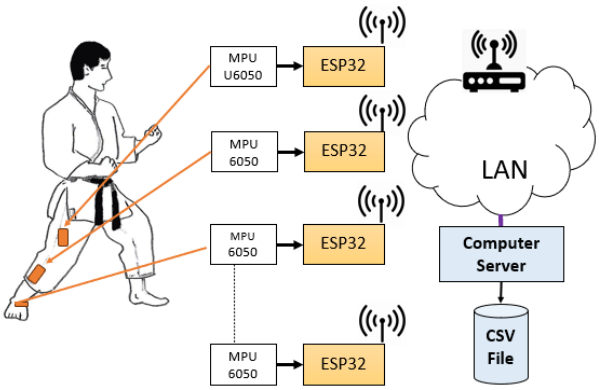


Fig. 1. Proposed motion capture system using IMU 6050.

This sensor position point was selected to accurately represent an athlete's kicking movement. There are three key points for sensor placement: the thigh, shank, and foot. Each sensor, used to capture movement on an athlete's body, follows the device construction depicted in Fig. 2. Each sensor includes an MPU6050, ESP32 as a controller, and a 3.7V lithium-ion polymer (LiPo) battery for power supply and charging control. The battery is essential for enabling the sensor to operate independently without requiring an external power source.

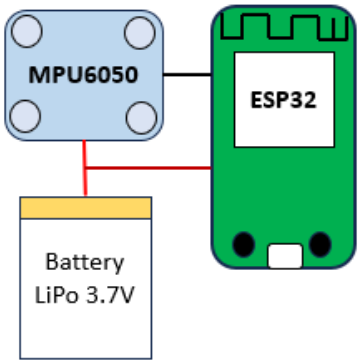


Fig. 2. Configuration of the components of IMU device.

This study focuses on analyzing leg kick techniques in a martial arts context. The research will develop a tool to classify the types of kicks performed by athletes based solely on the output of the kick itself, excluding factors like muscle strength and fatigue. The study involves testing a dataset with machine learning algorithms to determine the minimum data required for accurate classification.

The research progress through three stages: first, developing a motion capture tool using inertial sensors; second, using this tool to record athletes' kicking movements and create a dataset of martial arts kicks; and third, testing the dataset to predict the classification of each kick.

A. Hardware Specification

The MPU 6050, manufactured by InvenSense Inc., comprises three 16-bit analog-to-digital converters (ADCs) for converting the gyroscope output and three additional 16-bit ADCs for converting the accelerometer output. To accurately track both fast and slow movements, the components offer a user-programmable gyroscope full-scale range of ±250, ±500, ±1000, and ±2000°/sec (dps) and a user-programmable accelerometer full-scale range of ±2g, ±4g, ±8g, and ±16g. The MPU6050 supports a sample rate of 1 KHz. Communication with device registers can be achieved through either I2C at 400 kHz or SPI at 1 MHz.

ESP32 is a cost-effective, energy-efficient System on a Chip (SoC) microcontroller created by Espressif, featuring Wi-Fi and Bluetooth capabilities along with a dual-core processor. It is suitable for use in industrial settings, with an operational temperature range of -40°C to +125°C. The ESP32 boasts a Tensilica Xtensa Dual-Core 32-bit LX6 microprocessor running at either 160 or 240 MHz. It is equipped with 448 KB of ROM for booting and core functions, as well as 520 KB of SRAM for data and instructions.

The study confirms that the ESP32 can reliably sample sensors at high rates, making it suitable for real-time applications. However, its built-in ADC lacks precision, emphasizing the need for an external ADC to ensure accurate sensor data acquisition [15].

User Datagram Protocol (UDP) is utilized to manage the transmission of data produced by multiple sensors from a client device to a server computer for storage and subsequent processing. UDP is part of the Internet Protocol (IP) suite and functions at the transport layer within the OSI (Open Systems Interconnection) reference model. With the inclusion of a WiFi module, UDP is readily accessible in the ESP32 microcontroller chip, simplifying the wireless data transfer process.

For this application, the User Datagram Protocol was selected mainly because of its low latency and lightweight nature, which makes it appropriate for real-time data transfer. The connection-oriented strategy and error-checking methods of Transmission Control Protocol (TCP) result in overhead, whereas UDP functions as a connectionless protocol. As a result, data delivery is accelerated with less latency since acknowledgment packets, retransmissions, and congestion control are no longer necessary.

In applications like motion capture or real-time monitoring, where data from IMUs or other sensors need to be transmitted continuously, the slight loss of packets is often acceptable compared to the delays caused by TCP's reliability mechanisms. UDP's simplicity allows for high throughput, especially when working with high-frequency data streams, making it ideal for scenarios requiring time-critical communication, such as tracking human motion or sending sensor data for immediate processing.

This block diagram illustrates the flow of the data capture and processing system from the IMU (Inertial Measurement Unit) sensor to the stage of forming a dataset for machine learning needs. Body movement data captured by the IMU device is sent wirelessly via the UDP (User Datagram Protocol) protocol to the WiFi Access Point, then forwarded to the computer, to be recorded in the form of log data. After the data is received, a pre-processing process is carried out to clean and prepare the data so that it is ready for use, such as normalization, segmentation, or noise removal. The results of this pre-processing are then stored as a dataset, which is then used in the training or testing process of the model at the machine learning stage to recognize movement patterns or activity classification. As shown in Fig. 3.

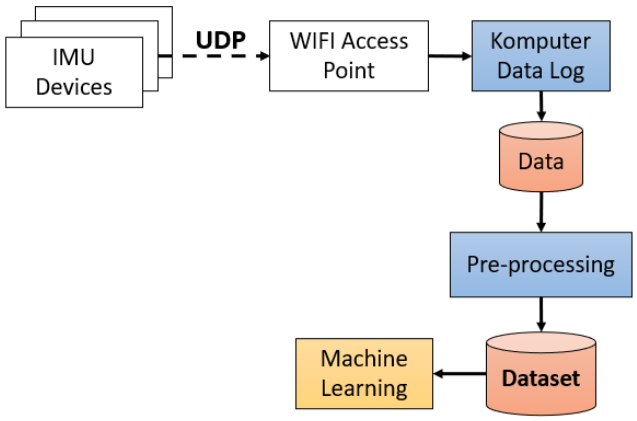


Fig. 3. Block diagram of data processing.

The sensor device starts by initializing and then searching for a connection to the server. Once the connection is established, it continuously sends sensor data to the server in a specific format. The protocol is encapsulated in datagrams using the UDP protocol and designated ports for each sensor device. The device follows a sequential process upon startup, searching for a connection to a UDP server with a distinct port number assigned to each sensor.

B. Data Transmission and Collection

The data transmission medium will be chosen wirelessly to reduce the number of cables connected to the participants. A local network will be established with a wireless access point capturing connections from each client, specifically through an IMU sensor on the volunteer subject. All sensor data will be transmitted through this connection. A computer will serve as the central server to gather data from all IMU sensors.

Multiple server applications will run on this computer to receive connections from each IMU sensor. The computer will directly connect to the wireless router access point using a UTP (Unshielded Twisted Pair) cable. Each IMU device, functioning as a UDP client, will connect to its corresponding server application set up in pairs. The separation of receiving servers will be achieved by assigning a different port to each client, allowing each server to store data independently for each sensor. The configuration of the server applications is illustrated in Fig. 4.

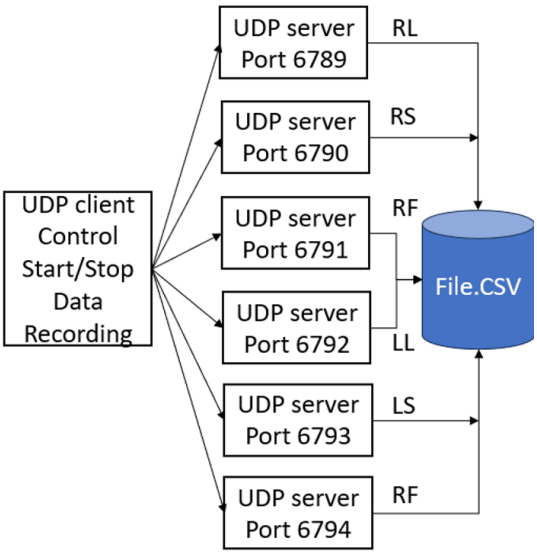


Fig. 4. Configuration of the UDP system protocol for wireless data collection from IMU devices to the server.

When the device is powered on, data from the IMU sensor will be consistently transmitted to the server. The transmitted data will be shown on the console to monitor the functionality of each IMU device. To prevent unnecessary data storage, a system was developed to determine when to begin recording data and when to cease storing it. This system prevents the storage of irrelevant data during the setup or device testing phase.

An additional program application is used to determine the start and stop times for recording the data stream. This application functions as a UDP/IP network-based client program on the server computer, sending commands to each capture server on different ports to initiate or terminate data storage. Being a network program, it can be placed on a separate computer if needed, facilitating operators in data retrieval tasks.

Data from each IMU sensor will be stored in separate comma separated value (CSV) files. Each data sent from the sensor will be received with additional time stamp data inserted before being stored in the file. All data sent by the IMU device will be stored without any prior selection process. Therefore, several pre-processing stages are necessary for formatting the data appropriately for subsequent processes. The data acquisition process on the server side is illustrated in the flow diagram in Fig. 5.

C. Data Format

Before processing the data using different classification algorithms, the data sent from each sensor device will initially be saved in a CSV file format. The sensor data will be stored in a specific order and format as shown in Fig. 5. Prior to being transmitted to the storage server, the sensor data will be supplemented with the device identity (ID) data from each sensor device. This ID represents the unique identity of each IMU sensor to distinguish one device from another.

The name of the device ID is hardcoded on each individual device. A timestamp (TM) will be added by the server application to record the arrival time of data. This timestamp will be placed at the end of each data record. Subsequently, the data along with its timestamp will be saved in a data file. The format utilizes a Unix timestamp, representing the number of seconds since January 1, 1970, 12:00 am GMT (Greenwich Mean Time).

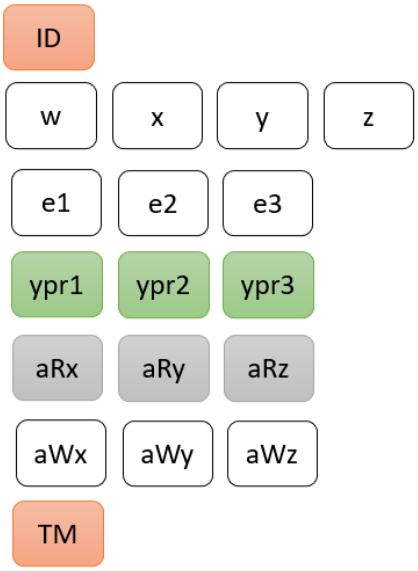


Fig. 5. Data format.

where,

- ID: device id
- w, x, y, z: quaternion
- e0, e1, e2: Euler distance
- ypr0, ypr1, ypr2: gyroscope
- aRx, aRy, aRz: real accelerometer
- aWx, aWy, aWz: world accelerometer
- TM is a timestamp

Accelerometer “raw” data are raw values directly generated by the MPU6050 accelerometer sensor without additional processing. This value is usually in the form of 16-bit data (as an integer) taken from the sensor's accelerometer register. This value represents the acceleration measured by the sensor, which is then converted to more meaningful units such as G-force or meters per second squared (m/s²) using a certain scale.

Accelerometer "world" data is acceleration data that has been converted from a sensor frame of reference to a world frame of reference. This usually involves a transformation process to account for the orientation of the sensor so that we can understand the acceleration relative to the earth or the real-world environment.

Data obtained from the gyroscope on the MPU6050 sensor is measured in degrees per second (°/s). Gyroscopes measure angular velocity or the rate of change of angles on three axes: X, Y, and Z. In this research, the data used is only gyroscope data and raw accelerometer data.

IV. RESEARCH METHOD

This research follows a systematic approach to develop a classification system for identifying kick types in martial arts using IMU sensor data and machine learning algorithms. The research methodology is divided into several main stages: 1) data acquisition, 2) data preprocessing and segmentation, 3) feature extraction, 4) model training and classification, and 5) evaluation and validation.

The sensors will be placed on several parts of the athlete's feet. Each sensor will be given an identifier to distinguish where the data comes from. The position and name of the sensor location are shown in Fig. 6. For the right leg, the device labeled RL to be placed in the right thigh position, RS is for the right shank position, and the RF label is for the right sole position. Similarly, the left side displays the label LL for the left thigh position, LS for the left shank position, and LF for the left foot position.

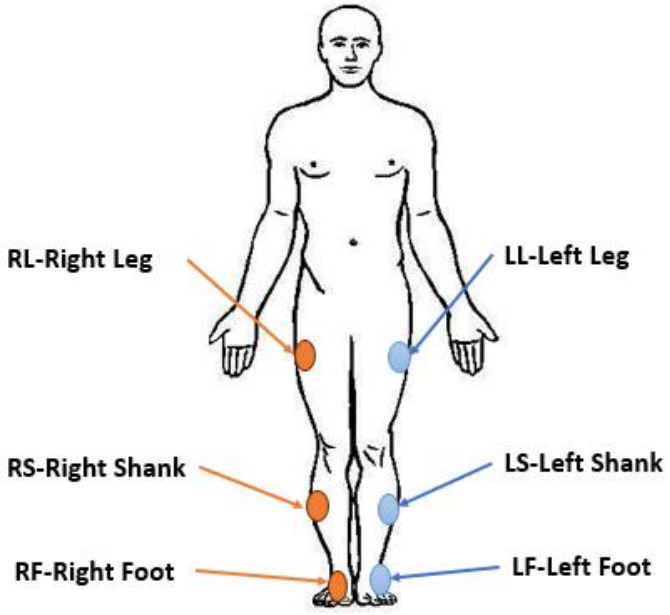


Fig. 6. Labeling of sensor positions on the lower limbs.

In biomechanical analysis, the placement of motion sensors can significantly impact the quality of data obtained, with options including placement at joints, inter-joint segments, or a combination thereof. While sensor placement at joints effectively captures local motion, such as rotation and joint angles. Interjoint placement, particularly at strategic locations like the thigh, calf, and foot, enhances the analysis of complex activities such as kicking. This method facilitates the measurement of comprehensive kinematic parameters, including translation, rotation, and acceleration across entire body segments, thus providing a more holistic view of movement dynamics. Given that, energy and momentum in activities such as kicking derived from intricate interactions within the kinematic chain focusing solely on localized joint motion, may lead to incomplete analyses and lack of insight into critical biomechanical interactions.

In the context of kicking biomechanics, inter-joint placement provides benefits in analyzing the energy relationship and momentum transfer between joints. Thus, this placement is better for capturing kicking motion patterns compared to direct joint sensor placement, which only provides local data without considering the interrelationships between body segments.

To secure the sensor to the athlete's body, a rubber band is utilized. The sensor is inserted with a plastic clip connected to the back of the device box, as illustrated in Fig. 7. This approach effectively maintains the sensor device in place during multiple leg kick movements. Periodically, the sensor device's position needs readjustment as it may shift from its original placement.



Fig. 7. Place the IMU sensor device on the respondent's leg using a rubber band.

In the data collection process, experienced martial arts practitioners were requested to demonstrate kicks. Each respondent would warm up for ten minutes before executing the specified kick movements. They would perform one kick type at a time, following the operator's start and stop commands. The kick would commence immediately upon the start signal and cease upon the stop signal. Respondents were instructed to repeat the same kick ten times. Following a five-minute break, they were then asked to perform another kick type ten times using the same procedure.

This study will create a dataset using kicks performed by martial arts experts. The dataset will serve as a reference for evaluating kick classification, focusing on three types of kicks: mawashi (roundhouse kick), maegeri (front kick), and shokuto (side kick) (see Fig. 8, 9, and 10). The dataset will specifically cover these three types of kicks. The gathered data will undergo processing to ensure the validity of the kick dataset.

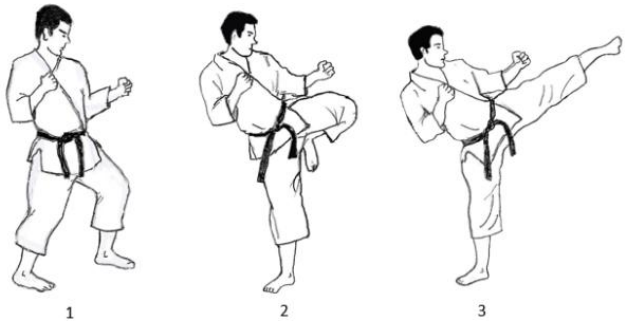


Fig. 8. Type of kick mawashi (round house kick).

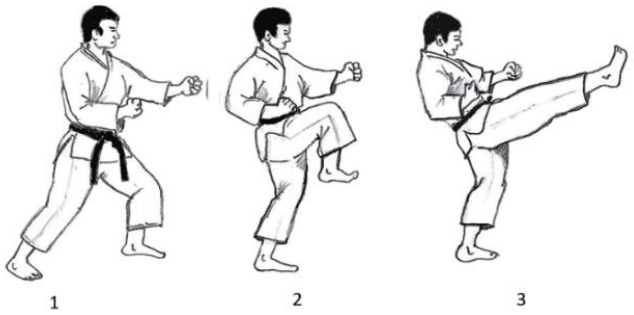


Fig. 9. Types of kick maegeri (front kick).

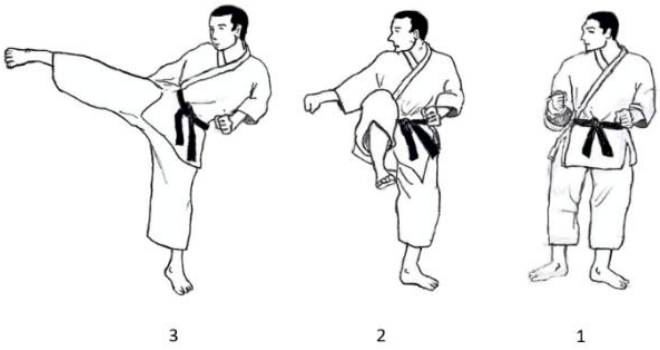


Fig. 10. Types of kick shokuto (side kick).

V. DATA PROCESSING

The process of constructing a dataset from sensor data is illustrated in Fig. 11. The data undergoes pre-processing to form a usable dataset. Pre-processing involves segmentation and re-sampling to ensure uniform length and format, specifically converting the data into a single line for each instance.

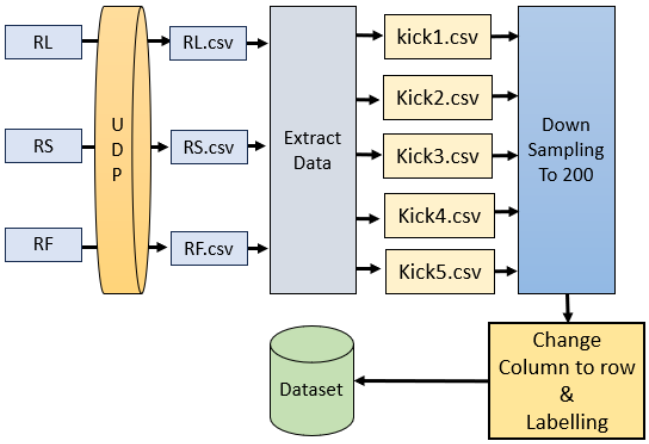


Fig. 11. Block diagram of pre-processing raw data.

Prior to being stored as a dataset, the data length for each sensor is standardized. The re-sampling procedure (down sampling) is crucial to ensure that the data length for each sensor is consistent at 200 samples. Consequently, only data with more than 200 samples is deemed suitable for further processing. Sensor data with fewer than 200 samples will be excluded due to potential imperfections in data collection.

To process sampling data with the next algorithm, the data length must be consistent. Therefore, data from sensor devices must be adjusted to match the data length. The MPU6050 sensor operates at a sampling frequency of 1 KHz, resulting in approximately 1000 samples per second. A martial arts athlete's kick lasts less than 1 second, so the average data length collected for each kick ranges from 600 to 800 samples. It is important that the adjustments made to equalize the data length do not remove significant information.

The down-sampling process is utilized to standardize the data length and reduce the number of data samples. It is crucial that this process does not result in the loss of essential information. This is done to ensure that the data interpretation remains unchanged, allowing the retained data to still capture the characteristics of a kick. Fig. 12 to Fig. 17 display the data after re-sampling. These figures illustrate that the graph's characteristics remain consistent before and after the re-sampling process.

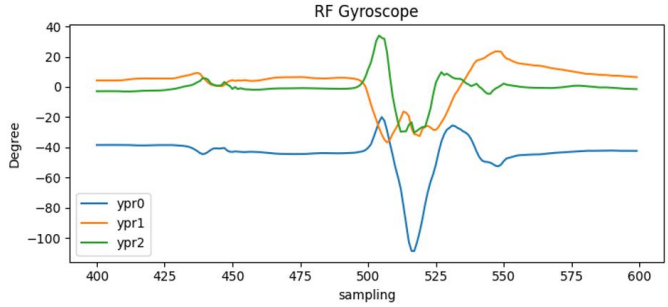


Fig. 12. Right lower limb mawashi kick data from RF gyroscope sensor data after down sampling to 600.

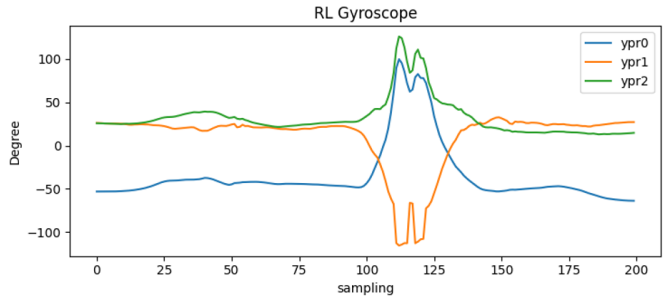


Fig. 13. Right lower limb mawashi kick data from the RL gyroscope sensor data after down sampling to 200.

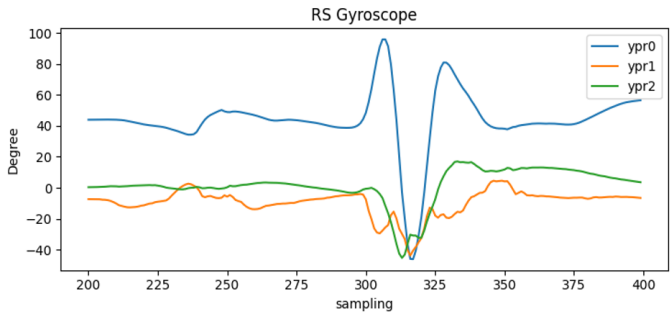


Fig. 14. Right lower limb mawashi kick data from the RS gyroscope sensor data after down sampling to 400.

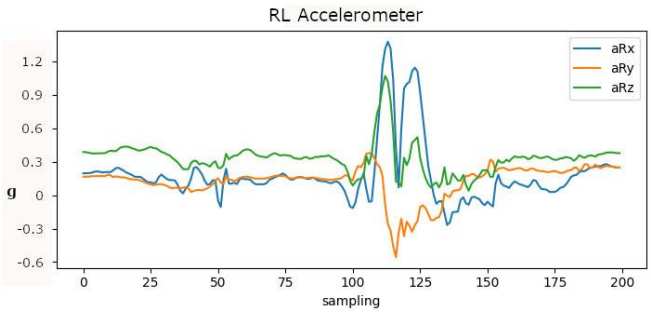


Fig. 15. Right foot mawashi kick data from RL accelerometer sensor data after down sampling to 200.

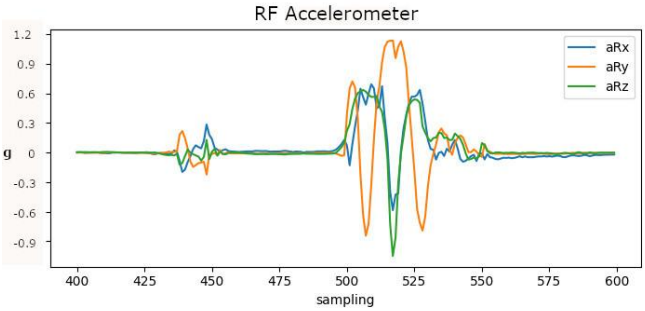


Fig. 16. Right foot mawashi kick data from the RF accelerometer sensor data after down sampling to 600.

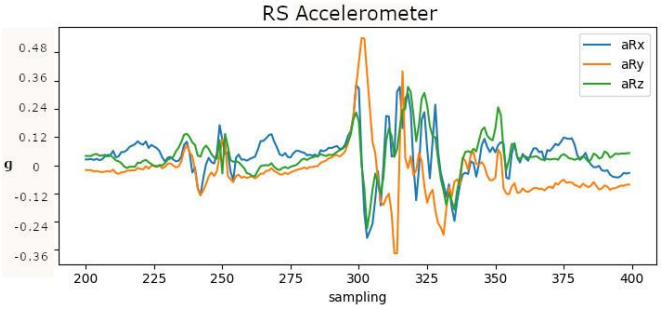


Fig. 17. Right foot mawashi kick data from the RS accelerometer sensor data after down sampling to 400.

Each sensor transmits data to the server with varying sample lengths, ranging from 500 to 800 samples per kick. The length of the data is determined by the duration of the kick. Before being input into the classification model algorithm, the data must undergo preprocessing to ensure uniformity in length. This initial process involves extracting data from files, segmenting it, and resampling to standardize the data length.

Data resampled to a length of 200 will be merged into a single file per kick, summarizing all sensor data. Sensor data with less than 200 samples will be discarded as it may indicate sensor or data transmission errors. Consequently, a significant amount of data remains unused due to sensors not meeting the minimum length requirements.

The subsequent step involves converting the data format from column-oriented to a single row for each kick. This alteration is essential to ensure that the resulting dataset contains only one row per kick. Each kick will have a column length of 3600 parameters derived from three sensors, including two gyroscopes and accelerometers, each with three outputs and a sampling length of 200.

The following step involves labeling all stored data based on the type of kick. This labeling process occurs concurrently with the input of new kick data that will be added to the dataset. To facilitate this task, a dedicated program was utilized to incorporate labels into the kick dataset.

Testing and classification of the dataset will be done using machine learning algorithms, specifically k-NN and SVM. The main concept of machine learning is to build a model that can recognize patterns in past data and apply those patterns to similar data in the future. k-NN and SVM are supervised machine learning algorithms commonly used for classification tasks.

The categorization model will subsequently undergo training and testing. The dataset will be partitioned into two segments, with 70 per cent designated for training and 30 per cent for testing. The correctness of the model will be evaluated using the findings from the test data.

The initial classification algorithm utilized was k-NN. This choice was made due to its simplicity, effectiveness for small to medium datasets, and capability to handle multi-Class. The selection of the value k in the K-Nearest Neighbors (KNN) algorithm is a crucial step that can significantly impact model performance. Typically, an odd number is chosen for k. In this case, the value of k is selected to be between 3 and 5. The evaluation of the k-NN algorithm's accuracy will be conducted using multiple datasets and varying k values. The comparison results are presented in Table II. Subsequently, the optimal number of selected k values for each dataset size will be determined.

The dataset to be tested comprises seventy data points resulting from re-sampling, pre-processing, and labeling. The model's accuracy will be evaluated to determine if it meets the criteria for a reliable prediction. Table I displays the breakdown of kick types in the dataset.

TABLE I. NUMBER AND TYPE COMPOSITION OF DATASET

Label	Number
Maegeri	16
Mawashi	22
Shokuto	32
Total	**70**

VI. RESULT

Modeling for classifying kick types will utilize seventy datasets that have been processed successfully in the prior stage. The model's accuracy will be assessed to determine if it meets the criteria for effective prediction.

The results of classification prediction accuracy using the k-NN algorithm are shown in Table II. For k=3 and dataset seventy using all sensor data (RL, RS. RF), the accuracy reaches 0.924, while if we use only two sensors the highest accuracy is achieved at 0.881 using sensors (RL, RS) with seventy datasets and k=3. For those using one sensor, the accuracy is 0.762 using an RS sensor with k=3 and a dataset of seventy.

TABLE II. ACCURACY OF THE K-NN ALGORITHM

No	Sensor	K	Dataset	Accuracy
1	RL, RS, RF	3	70	0.924
2	RL, RS, RF	5	70	0.843
3	RS, RF	3	70	0.838
4	RS, RF	5	70	0.800
5	RL, RS	3	70	0.881
6	RL, RS	5	70	0.833
7	RS	3	70	0.762
8	RS	5	70	0.714

Prediction accuracy results using the SVM algorithm with the same number of data sets are shown in Table II. The lowest accuracy value of this algorithm is 0.833, and the highest accuracy value is 0.967. When compared with the k-NN algorithm, the prediction accuracy of SVM is better.

The accuracy results with the model algorithm using random forest are shown in Table III. The highest accuracy is shown with a value of 0.900 when using 3 sensors (RL, RS, RF) and the lowest when using 1 RS sensor with an accuracy value of 0.795. For calculations using only two sensors, the accuracy value reaches 0.890 using sensors (RL, RS) and 0.905 using sensors (RS, RF).

The findings demonstrate that SVM's ability to handle small, high-dimensional datasets, combined with its use of kernel functions, allowed it to effectively capture complex patterns in the IMU data, resulting in superior classification accuracy. In contrast, k-NN and Random Forest faced challenges due to their sensitivity to feature scaling, noise, and correlated features, making SVM the best-suited algorithm for this specific task. These results highlight the importance of matching the algorithm to the dataset characteristics to achieve optimal performance.

TABLE III. ACCURACY OF SVM AND RANDOM FOREST ALGORITHM

No	Sensor	Dataset	SVM Accuracy	Random Forest Accuracy
1	RL, RS, RF	70	0.924	0.900
2	RS, RF	70	0.871	0.905
3	RSL RS	70	0.967	0.890
4	RS	70	0.833	0.795

Performance of k-NN's classification depends heavily on the distance metric (e.g., Euclidean distance). In the dataset, variations in sensor data due to noise or irrelevant features could disproportionately influence the distance calculations. The performance was contingent upon the selection of k, where smaller values resulted in overfitting and higher values induced underfitting. The k-NN performed better when kick types had distinct and isolated clusters in the feature space, but its performance degraded for overlapping or closely spaced data points, as seen in this study.

Random Forest relies on an ensemble of decision trees, which partition the feature space based on thresholds. While this approach is robust for large datasets, it can struggle with small datasets, as individual trees may overfit the training data. In the dataset, many features derived from IMU sensors (e.g., accelerations and angular velocities) were highly correlated. Random Forest frequently overvalues the significance of associated features, resulting in potential bias in feature selection and diminished generalization performance. From observation, although Random Forest showed decent performance, its accuracy was lower than SVM due to its inability to effectively capture the complex relationships between features in this dataset.

Analysis of the feature space indicated that the kick types (e.g. maegeri, shokuto, and mawashi) were not perfectly linearly separable, but there was some structured separation among the classes in the feature space. Analysis revealed that angular velocity and acceleration data were critical in distinguishing between kick types. For example, Maegeri involved a rapid forward acceleration of the shank. Shokuto displayed lateral angular velocity and sideways acceleration. Mawashi exhibited curved motion patterns, visible in both acceleration and gyroscope data.

The classification model's performance is assessed by comparing its predictions to the actual values (ground truth) using a confusion matrix. This matrix clearly indicates the number of correct and incorrect predictions for each class. Fig. 18 to Fig. 21 present the confusion matrices for each model in this study. These tables summarize how well the models classify data, highlighting both accurate and inaccurate predictions. Additional metrics such as precision, recall, and F1-score provide deeper insights into the model's strengths and weaknesses across various application scenarios.

F1-score is a classification model evaluation metric that combines precision and recall into one value. This metric is useful when there is class imbalance in the dataset and when we want to find a balance between False Positives (FP) and False Negatives (FN). The higher the F1-score, the better the model is at handling the balance between precision and recall.

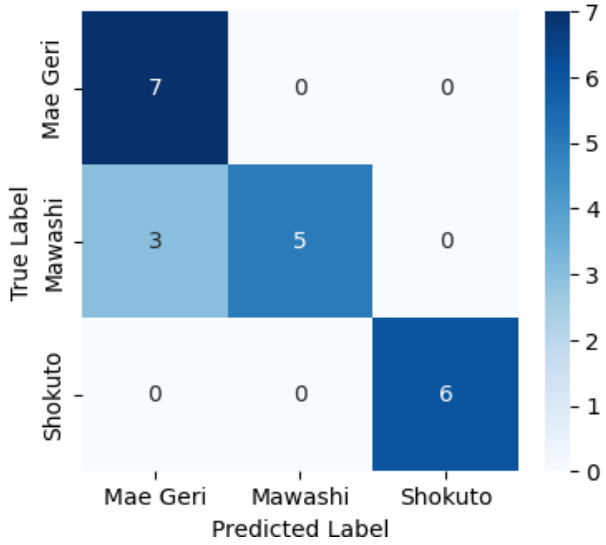


Fig. 18. Confusion matrix k-NN for k=3.

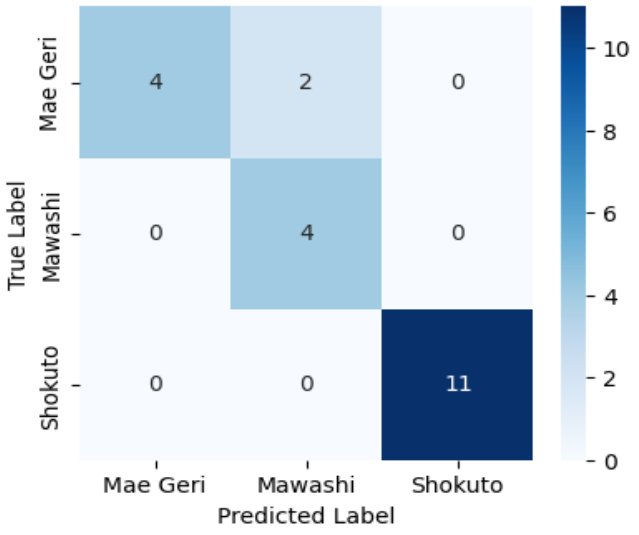


Fig. 19. Confusion matrix k-NN for k=5.

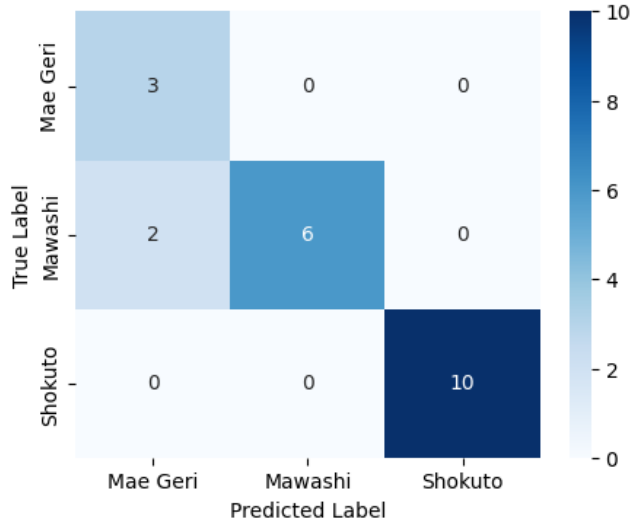


Fig. 20. Confusion matrix random forest.

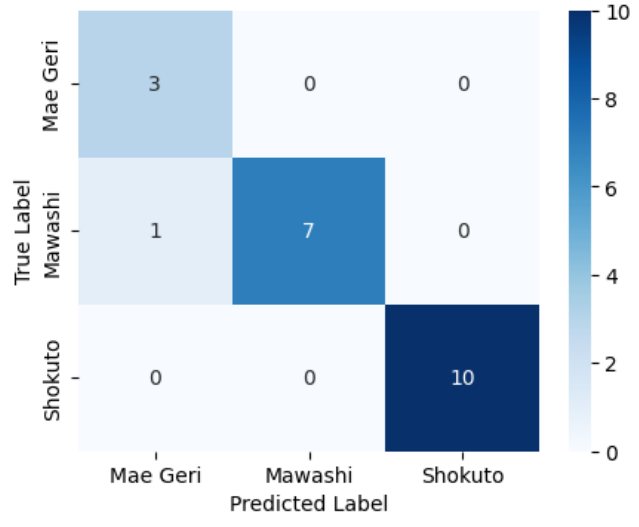


Fig. 21. Confusion matrix SVM.

In the k-NN model with a value of k = 3 (Fig. 18), the F1-score value for this model is Mae Geri: 0.82, Mawashi: 0.77, and Shokuto: 1.00. The average will be F1-score 0.86, and if using weighting according to the number of datasets in its class, it is 0.85. While for the k-NN model with k = 5 (Fig. 19), the F1-Score values are Mae Geri: 0.8, Mawashi: 0.8, and Shokuto: 1.00. The average is 0.87, and if using weighting according to the number of datasets in its class, it is 0.90. The Shokuto class has a perfect F1-score, while Mae Geri and Mawashi still have some errors. This shows that the model has a fairly good performance overall, especially since the Shokuto class has a larger number of samples and high scores, thus increasing the average.

For the random forest model with the confusion matrix shown in Fig. 20, the F1-score value per class is Mawashi = 0.86, Shokuto = 1.00. For the average value of the macro F1-score = 0.87 and the weighted F1-score = 0.90. The model has good performance with a weighted F1-score of 90 per cent, indicating a fairly accurate classification. Shokuto is perfectly classified, without errors. Maegeri has a lower recall (0.60), so it needs to be improved to reduce classification errors, and Mawashi has a good balance between precision and recall.

For the SVM model shown in the confusion matrix in Fig. 21, the F1-score values per class are Mae Geri = 0.86, Mawashi = 0.93, and Shokuto = 1.00. The average F1-score value = 0.93 and the weighted F1-score value = 0.95. This model has very good performance, with a weighted F1 score of 95 per cent, which means the model is able to classify all three types of kicks quite accurately. Shokuto is perfectly classified (F1-score = 1.00), Mae Geri has a small error in recall, so the improvement can focus on reducing the classification error for Mae Geri, while Mawashi has a good balance between precision and recall.

VII. Discussion

This discussion provides an in-depth overview of the implementation of kick classification using IMU and k-NN and SVM machine learning algorithms. The success of both in this context confirms the great potential of this technology in the understanding and analysis of movements, with practical applications in the training of martial arts athletes and the development of more sophisticated movement recognition systems.

Segmenting kicking movements using gyroscope and accelerometer data can be achieved through the following steps: Conduct an initial analysis of the data to recognize patterns in kicking movements. Utilize significant changes in angular velocity (from the gyroscope) and acceleration (from the accelerometer) as cues for the beginning and end of the movement. Establish thresholds for angular velocity and acceleration that indicate the initiation and conclusion of the kicking action. Choose a threshold value that distinguishes between motion and rest or other irrelevant movements. Apply the defined threshold to pinpoint segments of kicking motion. When the angular velocity or acceleration surpasses a threshold, it signifies that movement is underway. Carry out manual validation or human supervision on the identified motion segments to ensure their accuracy in representing the kicking motion. Conduct parameter optimization on machine learning

models to enhance classification accuracy, particularly post-implementation of the segmentation process.

To increase the accuracy value of the k-NN algorithm, it is necessary to strive for things such as choosing the optimal k value, balancing the number of samples per very different class so that it is more balanced, wisely selecting features that can help reduce data dimensions and improve accuracy, and it is necessary to carry out exploratory analysis of the data to understand the characteristics of the dataset. Knowing whether there are important trends, patterns, or relationships can guide the k-NN model setup more effectively. By considering and combining several of these strategies, it can improve the performance of the k-NN algorithm and increase the accuracy value on classification tasks.

To achieve more accurate classification in human motion classification tasks, here are some suggestions regarding the efforts to be made and the algorithms that can be used. Some efforts that must be made are to ensure that the data collected from sensors (e.g., accelerometer, gyroscope) is of high quality and free of noise as much as possible. Use appropriate sensor devices and sensor calibration to ensure data accuracy.

At the data preprocessing stage, more effort needs to be made in data cleaning to remove noise and artifacts from sensor data, data normalization to ensure consistent scale across all features, segmentation and windowing into appropriate time windows to capture movement well, extraction of appropriate features relevant features such as time statistics, frequency features, and time-frequency domain-based features.

Improvement efforts can also be made by trying other available techniques, such as using feature selection techniques (for example, PCA, LDA) to reduce the dimensions of the data and retain the most significant features. To overcome limited datasets, data augmentation techniques can be used to generate more training examples, such as adding small noise, rotation, or scaling to the movement data. Cross-validation methods to prevent overfitting guarantee that the model performs effectively when applied to unknown data.

Several algorithms can be tried for classification techniques, such as Convolutional Neural Networks (CNNs), which have excellent advantages in capturing spatial and temporal features from movement sensor data. They can automatically extract features from raw data. Applications: CNNs are often used for human activity classification with time-series data and sensor signals. Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs) algorithms are designed to handle sequential and time-series data, so they are very suitable for temporal human movement classification. LSTMs are used to capture long-term dependencies in human motion data. By following the steps outlined and selecting a suitable algorithm, the accuracy of classifying an athlete's kick can be greatly enhanced.

VIII. Conclusion

This section provides an in-depth overview of the implementation of kick classification using IMU, k-NN, Random Forest and SVM machine learning algorithms. The success of these algorithms in this context confirms the great potential of this technology in the understanding and analysis of movements, with practical applications in the training of martial arts athletes and the development of more sophisticated movement recognition systems.

SVM, with its ability to handle non-linear relationships via kernel tricks, provides solid performance in kick classification. Selection of an appropriate kernel, such as the RBF (Radial Basis Function) kernel, can affect SVM performance. The advantages of SVM are especially visible when the kick motion pattern has high complexity and cannot be clearly separated by a straight line.

Implementation of the k-NN algorithm on a kick dataset that utilizes IMU data demonstrates success in identifying movement patterns with an accuracy of 0.924 for k=3 and 0.843 for k=5. This performance is slightly lower compared to the SVM algorithm, with an accuracy of 0.924. All the results mentioned above are based on the use of 3 sensors (RL, RS, RF). However, when only two sensors are utilized, the k-NN accuracy drops to 0.838 (RS, RF) and 0.881 (RL, RS), while the SVM accuracy decreases to 0.871 (RS, RF) and 0.967 (RL, RS). For the random forest model, the accuracy is 0.900 with three sensors (RL, RS, RF), while with two sensors, the accuracy is 0.905 (RS, RF) and 0.890 (RL, RS), which are still lower than the k-NN and SVM algorithms.

The primary advantage of k-Nearest Neighbors (k-NN) lies in their simplicity and effectiveness in handling classification problems. As a non-parametric, instance-based learning algorithm, k-NN does not require an explicit training phase, making it easy to implement and adaptable to various types of data. However, despite its straightforward nature, choosing an optimal value for k is crucial to ensuring model performance. A small k value can lead to overfitting, where the model becomes too sensitive to noise and fails to generalize well to new data. In contrast, a high k value may lead to underfitting, which results in the model oversimplifying decision boundaries and misclassifying complex patterns. To mitigate these risks, it is essential to use cross-validation techniques to determine the best k value for a given dataset. Additionally, data preprocessing steps such as feature scaling and noise reduction can significantly enhance the accuracy and reliability of k-NN in practical applications.

IX. Future Work

The current study indicates that it is possible to use IMU-based wearable systems and machine learning algorithms together to accurately categorize different forms of martial arts kicks. But there are several ways that the system could be improved in the future to make it more useful, practical, and easy to use for daily athletic training.

Future work is to reduce the number of sensors used in the system while maintaining comparable levels of classification accuracy. Reducing the number of sensors would also lower the cost and complexity, enabling broader adoption among athletes for everyday training without interfering with movement. If possible, reducing sensors to a compact form factor such as a single wearable smart device (e.g., a smartwatch or ankle-worn sensor) could significantly improve system convenience, ease of use, and real-world applicability.

Subsequent research will concentrate on enhancing machine learning and feature extraction methodologies to mitigate the limitations posed by reduced sensor inputs. Identifying optimal sensor sites (such as the most effective body parts), integrating data from several sensors, or utilizing sophisticated algorithms like deep learning that can discern intricate patterns from limited data.

ACKNOWLEDGMENT

The authors assert that there are no conflicts of interest pertaining to this research. This study was conducted independently without any financial, institutional, or commercial influence from any organization. All data collection, analysis, and interpretations were performed objectively to ensure the integrity and validity of the findings.

This research has received ethical clearance from the Ethics Commission of the Bandung Institute of Technology (ITB) under the approval KEP/II/2024/X/M281024RG-PDPG. The commission has reviewed and confirmed that this study complies with ethical standards and is deemed ethically feasible for implementation.

REFERENCES

[1] Eline Van Der Kruk & Marco M. Reijne. (2018),"Accuracy of human motion capture systems for sport applications; state-of-the-art review", European Journal of Sport Science.

[2] Ferryanto, F., Mahyuddin, A. I., & Nakashima, M. (2022). Development of a markerless optical motion capture system by an action sports camera for running motion. ASEAN Engineering Journal, 12(2), 37–44. https://doi.org/10.11113/aej.v12.16760

[3] Vitali, Andrea & Regazzoni, Daniele & Rizzi, Caterina & Lupi, Giorgio. (2020). Low Cost Markerless Motion Capture Systems: A Comparison Between RGB Cameras and RGB-D Sensors. ASME 2020 International Mechanical Engineering Congress and Exposition 10.1115/IMECE2020-24083.

[4] Samet Ciklacandir, Sultan Ozkan, Yalcin Isler, (2022),"A Comparison of the Performances of Video-Based and IMU Sensor-Based Motion Capture Systems on Joint Angles", 2022 Innovations in Intelligent Systems and Applications Conference (ASYU).

[5] Colyer, S. L., Evans, M., Cosker, D., & Salo, A. I. T. (2018). A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. Sports Medicine - Open, 4(1), 1–15. https://doi.org/10.1186/S40798-018-0139-Y

[6] Yang, P., Chen, Y., Chen, Y., & Dang, F. (2021). Gyro-Free Inertial Measurement Unit With Unfettered Accelerometer Array Distribution and for the Object With Position Change in Center of Gravity. IEEE Sensors Journal, 21(7), 9423–9435.

[7] Yi-Lian Chen, I-Ju Yang, Li-Chen Fu, Jin-Shin Lai, Huey-Wen Liang, (2021) "IMU-Based Estimation of Lower Limb Motion Trajectory with Graph Convolution Network", IEEE Sensors Journal, Vol. 21, No. 21, November 1, 2021.

[8] Luigi Truppa, Elena Bergamini, Pietro Garofalo, Marta Costantini, Carmelo Fiorentino, Giuseppe Vannozzi, Angelo Maria Sabatini, and Andrea Mannini, (2021). "An Innovative Sensor Fusion Algorithm for Motion Tracking with On-Line Bias Compensation: Application to Joint Angles Estimation in Yoga", IEEE Sensors Journal, Vol. 21, No. 19, October 1, 2021.

[9] N. Dawar and N. Kehtarnavaz, (2018) “Action detection and recognition in continuous action streams by deep learning-based sensing fusion,” IEEE Sensors Journal, vol. 18, no. 23, pp. 9660–9668.

[10] Luke Wicent Sy, Nigel H. Lovell, and Stephen J. Redmond.,(2021) "Estimating Lower Body Kinematics Using a Lie Group Constrained Extended Kalman Filter and Reduced IMU Count". IEEE Sensors Journal, Vol. 21, No. 18, September 15, 2021.

[11] Kim, K.N., Chen, J., and Cho, Y. (2019). "Evaluation of Machine Learning Algorithms for Worker's Motion Recognition using Motion Sensors." Proceedings of the ASCE 2019 International Conference on Computing in Civil Engineering (i3CE), Atlanta, GA, USA, June 17-19.

[12] Akhavian, R., and Behzadan, A. H. (2016). “Productivity Analysis of Construction Worker Activities Using Smartphone Sensors.” Proceedings of 16th International Conference on Computing in Civil and Building Engineering, 1067–1074.

[13] Tsubasa Maruyama, Mitsunori Tada, Akira Sawatome, Yui Endo (2018). "Constraint-based Real-Time Full-Body Motion-Capture using Inertial Measurement Units". IEEE International Conference on Systems, Man, and Cybernetics.

[14] Cecilia Monoli, Juan Francisco Fuentez-Pérez, Nicola Cau, Paolo Capodaglio, Manuela Galli, and Jeffrey A. Tuhtan, (2021) "Land and Underwater Gait Analysis Using Wearable IMU",IEEE sensors journal, vol. 21, no. 9, may 1, 2021.

[15] J. Lim, S. Lee, and H. G. Jung, “Action Recognition of Taekwondo Unit Actions Using Action Images Constructed with Time-Warped Motion Profiles,” Apr. 2024, doi: 10.3390/s24082595.

Comparative Analysis of Deep Learning Techniques for Passive Underwater Acoustic Target Recognition: Overview, Challenges, and Future Directions

Song Yifei[1], Mohamad Farhan Mohamad Mohsin[2]
University Utara Malaysia (UUM), 06010 Sintok, Kedah, Malaysia[1]
School of Computing, University Utara Malaysia (UUM), Sintok, Kedah, Malaysia[2]

***Abstract*—Passive underwater acoustic target recognition (UATR) involves analyzing acoustic waves captured by passive sonar to extract valuable information about submerged targets. The underwater acoustics community has increasingly turned its attention to deep learning techniques, owing to their remarkable success in image recognition tasks. This study presents a comprehensive overview of the evolution of UATR techniques, categorizing them into three distinct groups: early methods, conventional machine learning approaches, and modern deep learning-based techniques. Additionally, it provides an in-depth summary of the recognition process utilizing deep learning, detailing various deep network architectures, classifiers specifically designed for underwater acoustic target recognition, and different data input modalities. Finally, the study synthesizes current research findings and outlines potential future directions for advancements in this field, emphasizing opportunities for innovation across these three categories.**

Keywords*—Underwater acoustic target recognition; deep learning; deep network architecture; classifier

I. Introduction

Underwater acoustic target recognition (UATR) is grounded in the production, transmission, and reception of sound waves in water. This technology has extensive applications in military operations, marine exploration, sonar systems, marine life monitoring, and underwater communication through sound navigation and ranging systems [1]. Despite the influence of variables such as temperature, salinity, and pressure on sound propagation in the ocean, sound waves remain the most efficient medium for long-distance underwater detection, transmission, and communication [2],[3]. Passive-sonar systems, which operate by quietly detecting transmitted noise, are gaining popularity due to their enhanced stealth capabilities. Consequently, the analysis of passive sonar underwater target radiation noise data has emerged as an effective method for studying underwater objects. Renowned research institutes in this field include the Woods Hole Oceanographic Institution, the Chinese Naval University of Engineering, Harbin Engineering University, the Pakistan Naval Research Institute, the Institute of Acoustics of the Chinese Academy of Sciences, and the Applied Physics Laboratory at Washington University [4].

Lei et al. [5] highlight that advancements in technology have facilitated the extensive application of earlier methodologies, classical machine learning techniques, and deep learning-based network designs in UATR. Early approaches, though effective, relied on the recognition abilities of skilled sonar operators, leading to high costs and significant susceptibility to subjective factors. Conventional machine learning methods, often dependent on manually crafted features, face significant challenges such as loss of feature information and difficulty in extracting optimal features. In contrast, deep learning techniques are expected to achieve greater accuracy and robustness as computational power continues to grow [6]. By enabling end-to-end recognition processes, these techniques address many of the limitations posed by human involvement in early and traditional approaches.

II. Motivation

The UATR process is typically divided into three key components: the model's input, the model itself, and the model's output, as illustrated in Fig. 1. Based on the execution of each phase and the interconnections between them, recent review studies have categorized various approaches to underwater target identification [7],[8],[9]. For instance, in deep learning models, feature extraction and classification are integrated into a unified process, whereas in traditional approaches, feature design and extraction are performed manually or automatically before data is input into the model. However, a few studies provide comprehensive explanations of each component in a detailed, step-by-step manner. In this context, this study aims to review the latest research on passive UATR from three perspectives: model input, the model itself, and output types. By examining the use of different data input techniques, models, and task output formats, we seek to identify key challenges in current research and propose promising directions for future exploration.

III. Contribution

This study provides a novel perspective on the evaluation and analysis of UATR tasks through comprehensive research from three key aspects: model input, the model itself, and model output. Unlike conventional viewpoints that focus on different approaches to UATR tasks, this study emphasizes a structured analysis of the entire process.

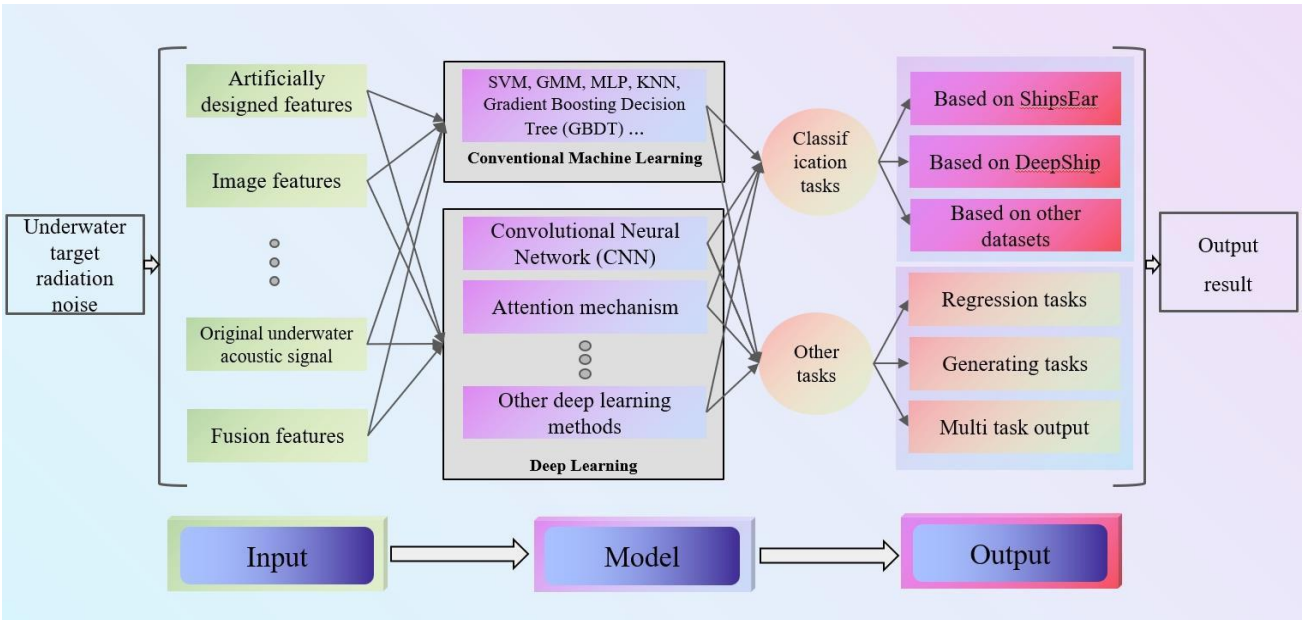


Fig. 1. General process and research architecture of UATR. Feature extraction, classification and recognition modeling, and model output constitute the three key stages in the fundamental process of both deep learning and conventional machine learning approaches for UATR.

Moreover, it highlights critical challenges that must be addressed in existing deep learning-based UATR tasks. By systematically examining and summarizing the input modes of underwater acoustic signals, model types, and model input strategies, this research provides a comprehensive synthesis of the strengths and limitations of various approaches. These insights serve as a valuable reference for researchers advancing UATR methodologies, facilitating future advancements in the field.

IV. ORGANIZATION

Section I introduces the comprehensive situation and basic issues in current research on UATR, serving as the foundation for this study. Section II presents the research motivation, emphasizing the significance of this study. By adopting a novel analytical and review approach based on the fundamental process of underwater target detection, it provides researchers with a broader perspective for understanding these challenges.

Sections III and IV outline the primary contributions and methodological framework of this work. Sections V, VI, and VII discuss different underwater acoustic signal input options, network topologies used in UATR, and the various task output types of the model, respectively. Furthermore, Sections VIII and IX explore the challenges of applying deep learning to UATR, propose potential directions for future research, and summarize the study's key insights. While previous studies have extensively examined the application of various technologies in underwater target detection, this review takes a more structured approach, delving into the specifics of each component. The overall organization of this study is illustrated in Fig. 2.

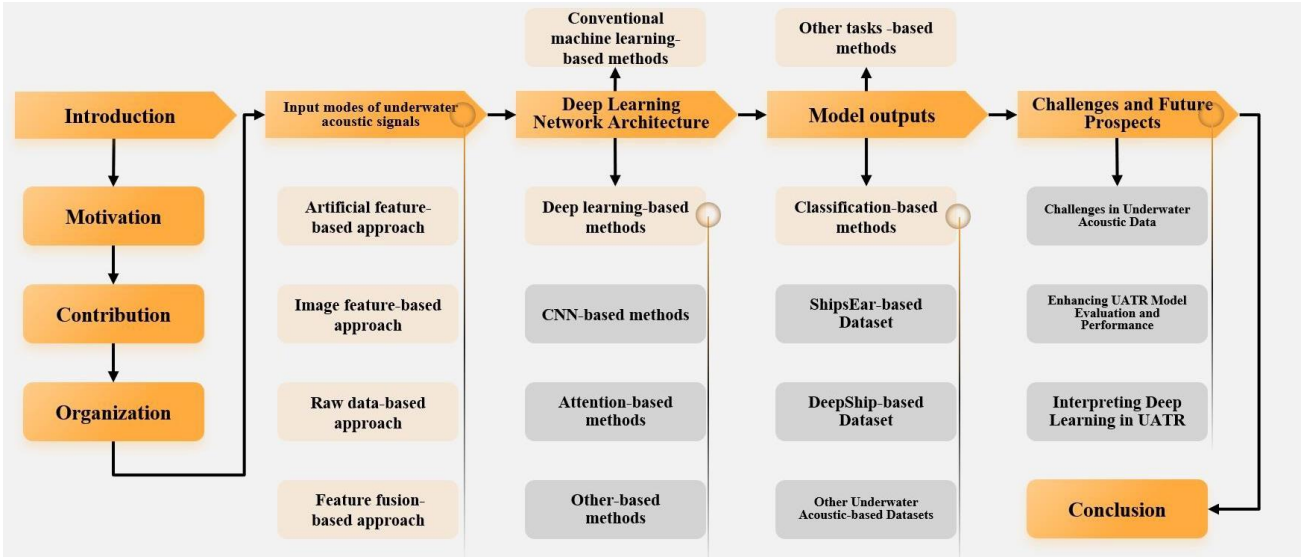


Fig. 2. Organization of the study.

V. Input Modes of Underwater Acoustic Signals

Deep learning-based UATR technology effectively addresses the issue of feature loss caused by human intervention. It typically employs an end-to-end learning framework, where neural networks replace manually designed components [10],[11]. To ensure the generation of high-quality underwater signals, appropriate preprocessing is usually necessary before inputting the target signal into the model [12]. The following sections will explore four approaches: the artificial feature-based approach, the image feature-based approach, the raw data-based approach, and the feature fusion-based approach.

A. Artificial Feature-Based Approach

A set of feature values that are manually crafted and extracted to characterize the properties of the target signal is known as artificially designed features. These features are highly interpretable, possess clear mathematical and physical significance, and are rooted in prior knowledge and experience.

In the field of underwater target identification, Mel frequency cepstral coefficients (MFCC) and DEMON spectra are widely utilized. MFCC captures the characteristics of underwater acoustic signals through a series of steps, including pre-emphasis, frame windowing, fast Fourier transform (FFT), power spectrum computation, Mel filter bank filtering, logarithmic transformation, and discrete cosine transform (DCT). Similarly, DEMON extracts the features of underwater acoustic data through preprocessing, envelope extraction, FFT, feature extraction, and subsequent analysis. This process is illustrated in Fig. 3.

Mel-frequency cepstral coefficients (MFCCs) are a widely used example of manually crafted features, extensively applied in language and speech recognition due to their ability to reduce feature map size, suppress noise interference, and withstand temporal variations [13]. On the ShipsEar dataset, Liu et al. [14] achieved an impressive 99.34% identification accuracy using MFCCs as input features for the RACNN model. Additionally, other studies have confirmed that MFCCs outperform models like VGG16 and ResNet34 in terms of computational complexity and parameter efficiency [15],[16]. However, standard MFCCs primarily capture static spectral envelope information, making it difficult to represent dynamic signal properties effectively. To address this limitation, Chen et al. [17] proposed using 3D MFCCs as input features. By incorporating first-order and second-order differential features, 3D MFCCs can capture dynamic characteristics such as the rate of change and acceleration of signals over time. Using 3D MFCC features as input for the Attention Mechanism Residual Concatenate Network (ARescat) model, an accuracy of 95.80% was achieved on the ShipsEar dataset.

The improvement in model recognition accuracy can be attributed to two main factors: the extraction of features that better represent the original data and the design of models with more advanced performance. However, it remains unclear whether 3D MFCC, an enhancement of traditional MFCC features, can more effectively represent the characteristics of raw data and is better suited for designing deep learning models. It is worth noting that numerous studies have compared MFCC features with other alternatives, such as FMSE and GFCC [18], [19]. Fig. 4 illustrates the performance of various feature extraction techniques on the same model using the ShipsEar dataset.

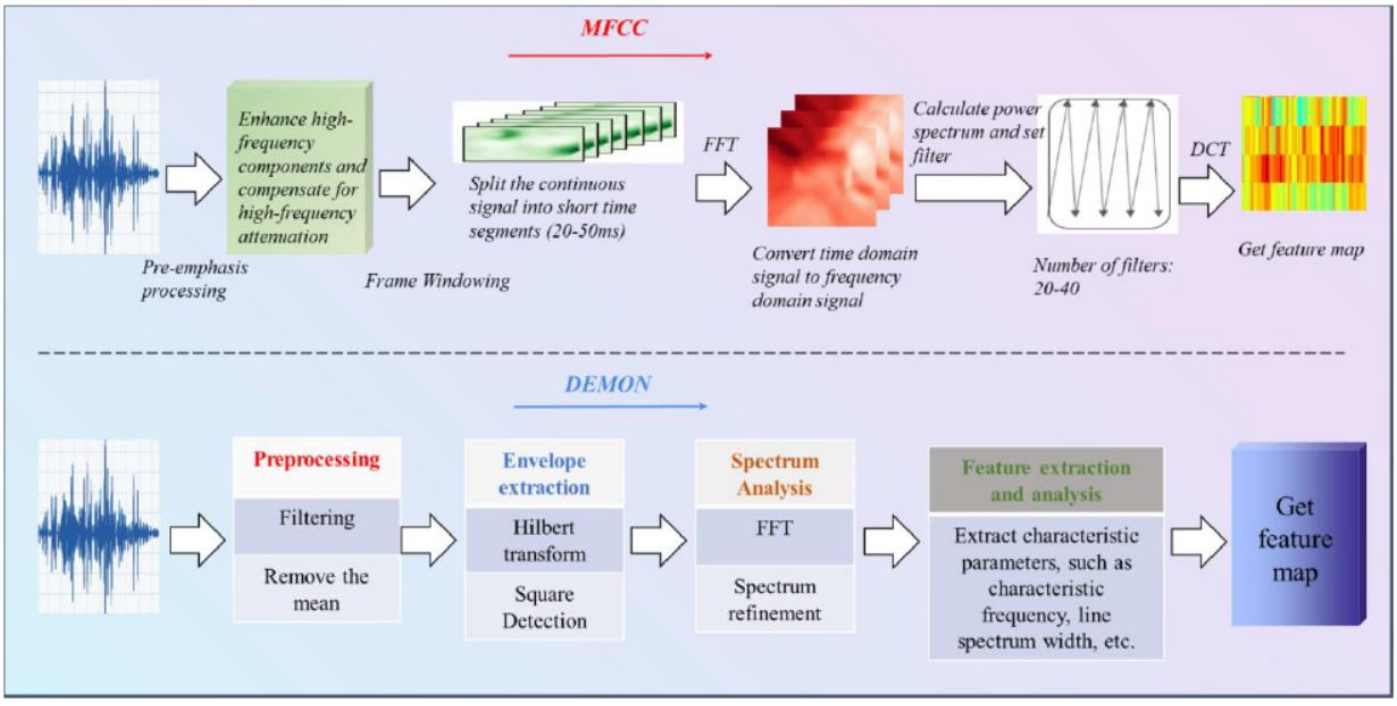


Fig. 3. MFCC and DEMON feature extraction process.

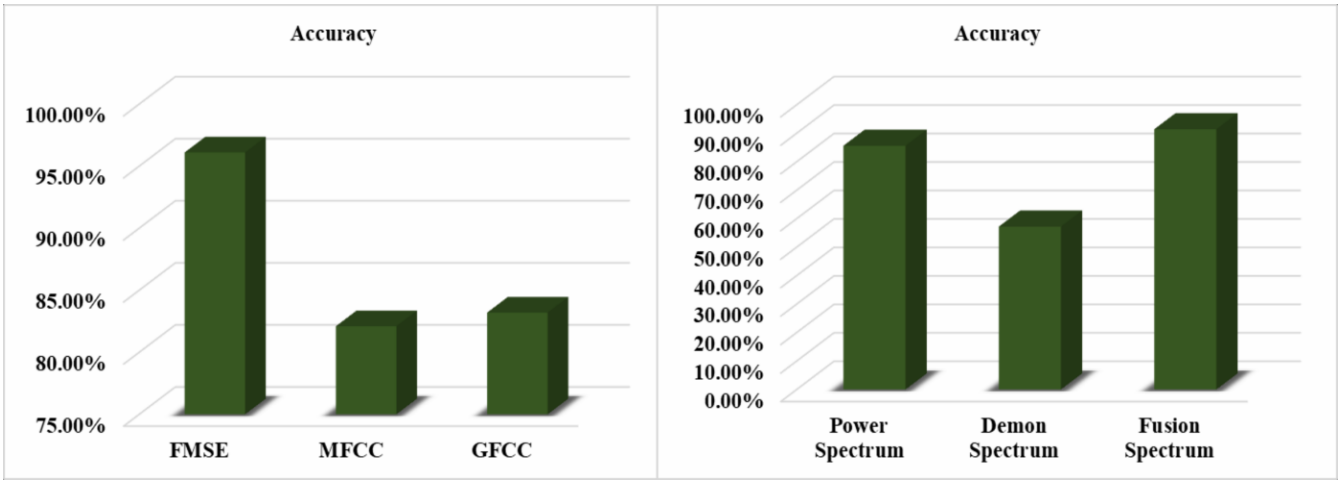


Fig. 4. Comparison of accuracy for different artificial and image features.

A broadband demodulation method, known as DEMON analysis, can distinguish between ship-radiated noise signals and the modulation envelope caused by propeller cavitation. By employing spectral analysis and line spectrum detection, it is possible to determine the number of propeller shafts, shaft frequencies, and blade counts [20]. Using the DEMON spectrum as input for the DEMONet model, Xie et al. [21] achieved an accuracy of 80.45 ± 0.67% on the DeepShip dataset. It is noteworthy that the feature extraction and model design in their study aimed for consistency. Furthermore, techniques such as Wavelet Packet Decomposition, Low-Pass Filtering, Gamma Frequency Cepstral Coefficients (GFCC), Multiscale Entropy Decomposition (MEMD), and LPS Feature Extraction, along with enhanced MFCC-based feature extraction methods, have been extensively studied and applied [22],[23],[24],[25]. Notably, Luo et al. [26] evaluated a range of feature extraction methods, including DEMON, LOFAR, MFCC, GFCC, EMD, and HHT, discussing their performance and suitability in different scenarios.

Manual feature design has several drawbacks, including low flexibility, limited expressive capacity, and high labor costs [27]. Additionally, manually crafted features often suffer from insufficient automated feature learning, as they cannot effectively capture or adapt to the inherent structure of data. In contrast, deep learning automatically extracts hierarchical features from data and enables end-to-end optimization, whereas manual feature extraction typically requires a separate design for classifiers or regressors [28]. As a result, researchers are increasingly focusing on technologies that leverage visual features [17],[29].

B. Figures and Tables

UATR based on image features is achieved by converting underwater acoustic signals into images and applying pattern recognition theory and image processing techniques to extract and analyze image features. This approach involves three main stages: image feature extraction, feature selection and optimization, and classification recognition. Image feature extraction involves using feature extraction algorithms to obtain representative and discriminative features from the transformed images. Feature selection and optimization focus on choosing the most relevant subset of features for target identification, removing unnecessary and redundant features, reducing feature dimensionality, and ultimately improving recognition accuracy and efficiency. Classification recognition involves using classification algorithms in pattern recognition to compare the extracted and optimized image data with pre-existing target feature libraries, enabling the identification of unknown underwater acoustic targets.

In terms of expressive power, interpretability, and adaptability, the image feature-based approach has demonstrated significant advantages [16]. Common image features include the amplitude spectra derived from LOFAR and STFT. Chen et al. [30] extracted the LOFAR spectrum of underwater sound waves and identified key characteristics using a multi-step decision-based line spectrum augmentation method. They combined the reconstructed line spectrum with the original LOFAR spectrum using a dual-threshold calculation approach. A convolutional neural network (CNN) was then trained on the reconstructed LOFAR spectrum, achieving an average recognition accuracy of 95.22% on the ShipsEar dataset. Despite the high recognition accuracy of this method, data augmentation techniques not only improve recognition performance but also mitigate noise interference from the marine environment. However, further research is needed to evaluate their impact on the recognition performance of neural networks. Additionally, Xu et al. [31] applied Mel (or Bark) filter banks to the FFT-generated spectrum to compute the STFT amplitude spectrum through complex FFT calculations, achieving an identification accuracy of 82.97% on the ShipsEar dataset. Yao et al. [32] further compared the LOFAR spectrum with other feature extraction methods, such as MFCC and CQT, using statistical histograms to analyze feature information across different ship types. Experimental results revealed that LOFAR image features lack low-frequency information, highlighting a limitation in capturing certain aspects of feature information.

Tang et al. [33] utilized Mel spectrograms as feature representations for model input. After feeding the data into the Transformer model, they conducted experiments on a small sample dataset, achieving a 90% identification accuracy with the Swin Transformer. Similarly, Xie et al. [34] achieved identification accuracies of 77.14%, 74.85%, and 95.48% on

the ShipsEar, DeepShip, and DTIL datasets, respectively, using Mel spectrograms as input. Studies employing spectrograms as input features have also yielded positive results [25],[35],[36],[37]. In Fig. 5, the recognition accuracy based on image features in UATR tasks is presented.

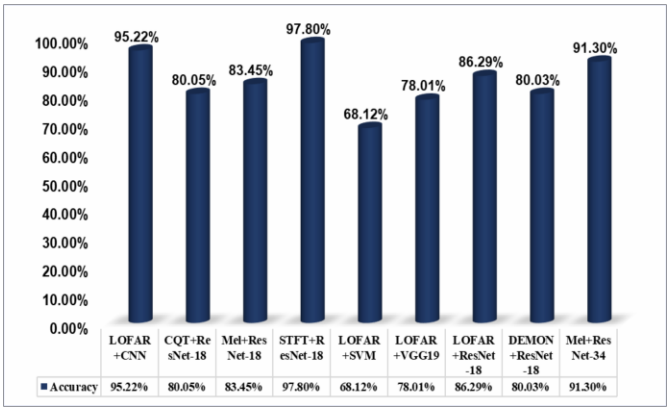


Fig. 5. Recognition accuracy based on image features.

Fig. 5 provides an overview of feature extraction methods based on image representations. Different deep learning architectures exhibit varying levels of recognition performance when using image features as model inputs. Notably, employing STFT features as input to ResNet-18 achieves a high recognition accuracy of 97.80%. Furthermore, due to the strong time-frequency correlation inherent in underwater acoustic signals, the use of raw time-domain signals as model inputs has been extensively investigated to preserve as much detailed information as possible. In contrast, image-based feature extraction methods may struggle to capture complex or high-dimensional patterns and are more susceptible to the loss of critical information [38], [39].

C. Raw Data-Based Approach

Raw signal-based UATR technology aims to identify underwater targets by directly utilizing the received raw underwater acoustic signals without performing complex transformations. This approach leverages various signal processing and pattern recognition algorithms. During the feature extraction stage, three main methods are employed: time domain, frequency domain, and time-frequency domain feature extraction. Time domain feature extraction involves directly obtaining amplitude information from the raw signal, such as peak values, mean, variance, and other statistical characteristics. Frequency domain feature extraction involves calculating the power spectral density of the raw signal using the Fourier transform and other operations, enabling target identification based on the distribution of signal power across different frequencies. Time-frequency domain feature extraction divides the raw signal into multiple short-time segments and performs a Fourier transform on each segment to obtain the distribution of signal components across both time and frequency domains. This approach provides a comprehensive view of how signal characteristics vary over time and frequency.

The intrinsic time-frequency correlation of underwater acoustic signals can lead to the loss of significant frequency domain information, even though image-based techniques can moderately enhance UATR [40]. Consequently, the direct use of raw data as input has become a prominent area of research to ensure the preservation and accuracy of information in underwater acoustic signals [41]. Hu et al. [42] investigated a technique for identifying ship-radiated noise using raw time-domain waveforms. They employed a depth-wise separable convolutional network to extract deep features that mirrored the auditory system's acoustic information processing. The model achieved an average classification accuracy of 90.9% when tested on a dataset of real acoustic signals from civilian ships. Additionally, extreme learning-based approaches have also shown promising results [43],[44].

However, using raw data directly as input for the model requires substantial computational resources, particularly for high-sampling-rate underwater acoustic signals [45]. Moreover, in the context of UATR with limited sample sizes, insufficient data can quickly lead to overfitting or poor model training.

To address the issue of inadequate underwater acoustic samples, several studies have explored segmenting and resampling original sound data. Yang et al. [46] resampled raw data at 16,000 Hz and divided it into 1-second non-overlapping intervals. Ji et al. [47] resampled underwater acoustic at 20 kHz, truncating each frame into 4096 samples with a 2048-sample overlap between successive frames. Similarly, Yang et al. [48] segmented each 5-minute WAV audio clip into 6-second intervals representing sound events and normalized the data for each segment. Li & Yang [49] split each audio recording into 3-second parts, normalizing each segment based on unprocessed raw time-domain data. In Fig. 6, the recognition accuracy with raw data as input is presented. From Fig. 6, it can be seen that the highest recognition rate can reach 95.30%.

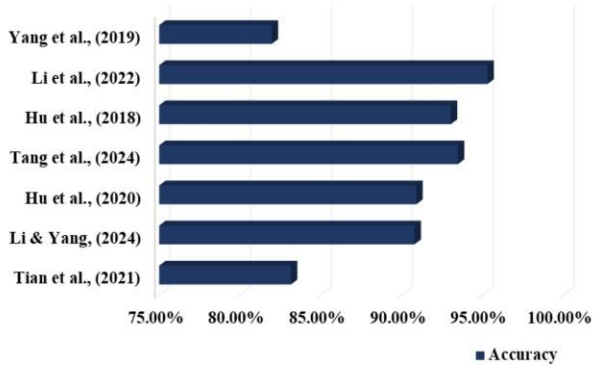


Fig. 6. Recognition accuracy based on raw data.

While resampling techniques can somewhat alleviate the issue of insufficient samples, they have limitations, including the potential amplification of noise and reduced diversity in the generated data. Therefore, methods based on fused features are becoming increasingly popular as a research direction.

D. Feature Fusion-Based Approach

Feature-based fusion is widely utilized in UATR technology to enhance accuracy and reliability by integrating various types of feature information. The main components of this approach include multi-domain feature fusion, multi-modal

feature fusion, and multi-scale feature fusion. Multi-domain feature fusion involves combining time domain, frequency domain, and time frequency domain characteristics of underwater acoustic signals to capture comprehensive signal properties. Multi-modal feature fusion integrates data collected from different sensors or based on different physical principles. For example, sonar signal features can be combined with underwater target features captured by optical sensors. Multi-scale feature fusion involves utilizing features at different scales, such as integrating fine local details with broader global target characteristics in image processing, to achieve a more robust and detailed representation of the target.

The fusion feature approach enhances identification accuracy by combining multiple feature types, such as time domain, frequency domain, and time frequency domain features, effectively addressing the limitations of single feature information. Additionally, integrating diverse features improves the system's robustness and adaptability to challenging conditions, such as interference and noise [50], [51]. Table I provides a summary of techniques for fused feature extraction.

TABLE I. SUMMARY OF FEATURE FUSION-BASED APPROACH

Author (s)	Feature	performance	Advantage	Disadvantages
Cao et al., (2019)	Axis frequency characteristics; LPS characteristics; WPCE characteristics	89.68% Acc	Adopting the SSAE model provides stronger discriminability; Constructing a joint feature set based on spectral and wavelet domain information	Underwater target classification tasks that require high real-time performance face challenges due to the complexity, cost, and time involved in obtaining a substantial amount of representative underwater target data.
Hong et al., (2021, May)	Log - Mel Spectrogram; MFCC; CCTZ	94.30% Acc	Using ResNet18, combined with embedding layers, early stopping, and adaptive learning rate strategies for training, Adopting data augmentation strategy SpecAugment	The universality of the method may be limited by the experimental validation, which was restricted to the ShipsEar dataset and did not include testing on a broader variety of datasets with different types and scales.
Domingos et al., (2022)	CQT; Gammatone Spectrograms; Mel Spectrograms	97.00% Acc	Compare various optimizers, neural network architectures, and preprocessing filters; Combine multiple filters into a three-channel signal	Accurately determining which scenario yields the best results under different testing conditions is challenging due to the varying number of instances in each case.
Chen et al., (2023)	CQT; delta MFCC; double - delta MFCC	99.10% Acc (DeepShip)	Propose a new loss function that only adds three hyperparameters to transform multi classification tasks into multiple binary classification tasks	In terms of model interpretability, the underlying mechanisms of each module are not thoroughly examined, leaving unclear how they influence the final outcomes.
Wu et al., (2023)	FBank; delta FBank; delta - delta FBank	90.50%	Combining cross domain pre training effectively improves recognition accuracy and saves a lot of training time	Insufficient exploration of innovative feature extraction techniques and potential feature combinations.
Pu et al., (2024)	Original wave; CQT; Mel spectrogram	96.37%	Design Scale ResNet module and RHAF module to improve information fusion efficiency and model adaptability to underwater acoustic data	In feature fusion, the lack of comparison with more advanced fusion techniques makes it difficult to determine its optimality across all scenarios.

Wu et al. [52] developed fusion features as model inputs by combining Mel filter-bank (F-Bank), delta F-Bank, and double delta F-Bank. To better capture the characteristics of underwater acoustic targets, they extracted neigh-boring dynamic features by computing delta features. Experimental results showed that this fusion approach improved identification accuracy by 0.9% and 1.2% on CNN and ResNet18 models, respectively. Chen et al. [53] proposed the FEFM method, which extracts multidimensional features from ship-radiated noise signals using various feature extraction techniques based on signal analysis and brain-inspired properties. These features are then fused using the proposed feature fusion technique to create high-dimensional fused features, which are used as inputs for the Multi-Gradient Flow Global Feature Enhancement Network (MGFGNet) network. Feature ablation tests were conducted on the Deepship dataset, comparing the fusion feature approach with other feature extraction methods. The fusion technique achieved a total accuracy of 99.1%.

Hong et al. [54] calculated various features, including the Log Mel Spectrogram (LM), Mel Frequency Cepstral Coefficients (MFCC), and CCTZ (which comprises the following features: Chroma, Contrast, Tonnetz, and Zero-cross ratio). These features were then fused and enhanced to create a three-dimensional feature matrix, which was used as input to the model. Domingos et al. [55] divided the raw dataset into one-second segments, padding shorter segments with zeros. They applied three preprocessing techniques—Mel spectrograms, Constant Q Transform (CQT), and Gammatone spectrograms—and combined them into a three-dimensional representation, referred to as "Complete", which was then be used as input to the model.

VI. DEEP LEARNING NETWORK ARCHITECTURE

The accuracy of UATR tasks is heavily influenced by the choice and design of the model architecture. With advancements in computing power, UATR technology has made significant progress. Conventional methods that relied on manual feature extraction and classifier design have been largely improved by deep learning approaches capable of automatic feature extraction and classification, resulting in substantial improvements in identification accuracy [56]. To clearly illustrate the development of these techniques, this study categorizes model architectures into two main groups: conventional machine learning-based methods and deep learning-based methods, depending on whether they employ an

end-to-end recognition process. A detailed discussion of these techniques follows below.

A. Conventional Machine Learning-Based Methods

UATR based on conventional machine learning methods refers to the process of classifying and identifying various underwater acoustic targets by extracting, analyzing, and evaluating the target information embedded in underwater acoustic signals using conventional machine learning theories and algorithms. Conventional machine learning methods rely on manually crafting features, which are then input into regression or classification models. Common approaches include multilayer perceptron (MLP), logistic regression, naive Bayes, support vector machines (SVM), and random forest classifiers [57], [58].

Honghui et al. [59] proposed the Multi-Attribute Correlation Perception (MCP) model, achieving 82.1% accuracy in ship detection tests by integrating the MCP module with a time-frequency feature extraction module for UATR. To enhance feature selection during classification, Fernandes et al. [60] applied neighborhood component analysis (NCA) for dimensionality reduction, combined with genetic algorithms and the K-nearest neighbor (KNN) method for object classification. Yu et al. [61] achieved an underwater target identification accuracy of 93.84% by employing the Gradient Boosting Decision Tree (GBDT) model in combination with two hydrophones and the VLA feature extraction approach. These approaches share a common characteristic: they all employ distinct feature extraction strategies prior to inputting features into the classification model to complete the classification task. However, separating feature extraction and classification introduces human variables due to the relative decoupling of these processes, potentially leading to a decline in recognition accuracy [18].

Sun et al. [19] achieved a recognition accuracy of 96.1% by using a support vector machine (SVM) classifier to process the retrieved features. Wang et al. [24] employed Gaussian Mixture Models (GMMs) to modify the topology of deep neural networks (DNNs), proposing the MFF-MDNN approach, which achieved an average recognition accuracy of 94.3%. While these methods reduce spurious variables to some extent and enhance recognition accuracy, the limited number of classifier layers in such models may constrain their ability to generalize effectively. Yang et al. [62] highlighted that the poor generalization ability of traditional recognition systems stems largely from the use of shallow classifiers, such as SVMs and shallow neural networks. Moreover, conventional methods, which predominantly rely on line and spectrum properties, often exhibit lower identification rates. Fortunately, deep learning-based approaches provide a more holistic framework for UATR, effectively addressing the limitations of traditional machine learning techniques and opening up promising directions for future research.

B. Deep Learning-Based Methods

Deep learning-based UATR technology is an advanced approach that automatically extracts, analyses, and recognizes target information within underwater acoustic signals by leveraging the powerful feature learning and pattern recognition capabilities of deep learning algorithms. The main process of this technology is illustrated in Fig. 7. The primary advantage of employing deep learning networks in UATR lies in their ability to bypass the complex feature engineering required by conventional machine learning approaches. Furthermore, they can partially address the problem of feature information loss associated with manual feature extraction [63]. The following will be discussed in detail.

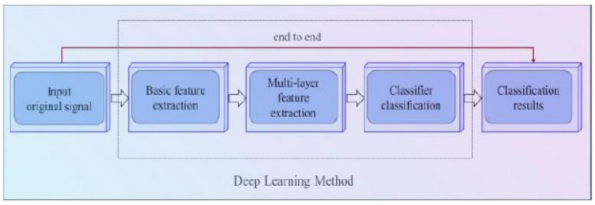


Fig. 7. General process of UATR based on deep learning methods.

1) CNN-based methods. The Convolutional Neural Network (CNN)-based UATR system leverages the powerful feature extraction and pattern recognition capabilities of convolutional neural networks to accurately identify targets in underwater acoustic data. CNNs are primarily used in tasks related to computer vision and speech recognition. Their architecture involves multiple processes that enable hierarchical feature learning and the use of automatically generated features for prediction [64]. Over time, CNNs have become a research hotspot in underwater target detection, driven by their remarkable performance advancements from AlexNet to GoogLeNet [39],[65].

ResNet is one of the most widely studied and popular deep learning models. Xu et al. [31] built their model on ResNet-18, incorporating a multi-head attention mechanism. The model was trained using two techniques: smooth induced regularization and the local masking and replication (LMR) approach. Its identification accuracy was evaluated using various input attributes, although several factors could potentially influence its performance. Xie et al. [34] adopted a dual-encoder framework with residual structured convolutional neural networks (ResNet) as the backbone to capture diverse acoustic features, specifically Mel spectra and continuous Q-transform spectra. By treating these networks as multi-view learners, they were able to extract more comprehensive features. Additionally, ResNet-based methods for developing and enhancing UATR have demonstrated promising results [54],[66],[67].

Significant advancements have been made in UATR techniques based on convolutional neural networks (CNNs) in recent years. Yang et al. [48] introduced a deep convolutional neural network (ADCNN) inspired by auditory perception, achieving an accuracy of 81.96%. Similarly, Hu et al. [44] proposed the auditory temporal convolutional neural network (ATCNN) model, which attained recognition accuracies of 95.9% on training data and 90.9% on test data. Meanwhile, more advanced CNN-based architectures, such as depth-wise separable convolutional neural networks, have also shown promising results. However, despite CNN's excellent feature extraction capabilities, the unique and complex nature of underwater acoustic signals poses several challenges. Specifically, CNN's convolutional kernels, limited by their

localized receptive fields, struggle to capture long-range correlations and global information inherent in underwater sound waves. Fortunately, techniques utilizing residual attention convolutional neural networks (RACNN) and deep residual attention convolutional neural networks (DRACNN) have opened new avenues for research. These methods have achieved impressive identification accuracies, addressing some of CNN's limitations in UATR [47], [68].

2) Attention-based methods. Attention-based UATR enhances the model's ability to extract and interpret critical information from underwater signals by incorporating attention mechanisms. These mechanisms efficiently capture the global characteristics of underwater acoustic signals by dynamically assigning attention weights based on signal features. They allocate different weights to various aspects of underwater acoustic signals across time, frequency, or spatial locations. The propagation of underwater acoustic signals is influenced by factors such as water temperature, salinity, and pressure. These factors can lead to attenuation, scattering, and refraction, impacting both the propagation distance and the quality of the signal. Consequently, the propagation of underwater acoustic signals is globally correlated, meaning that the signal's characteristics are shaped by the entire aquatic environment [69].

In UATR, the attention mechanism has gained significant traction for its ability to extract global features. Xiao et al. [70] introduced an attention-based neural network (ABN) model that integrated an attention module with a traditional deep neural network (DNN) composed of fully connected layers. By embedding the attention module within the DNN architecture, they developed a novel network designed to tackle UATR tasks. This model achieved an accuracy of 74.3% in multi-target resolution tasks, marking a 16.0% improvement over traditional DNNs. In contrast, Xie et al. [71] proposed the UART (Underwater Acoustic Recognition based on Templates) framework, which diverges from ABN by not depending on innovations in neural network design. Instead, this framework consists of an audio encoder, a spectrogram encoder, and a text encoder, offering an alternative approach to underwater object recognition.

The Transformer, a widely used attention mechanism, has been extensively applied in UATR using sound waves. Tang et al. [33] studied the Swin Transformer Biformer, a Transformer-based model. Similar to the four-layer pyramid structure of CNNs, the architecture of the Swin Transformer features a progressive reduction in the feature map's size as the feature layers deepen. The Swin Biformer enhances feature representation by combining the Swin Transformer with the Biformer, dynamically adjusting the partition window to refine regional feature extraction. Pu et al. [72] introduced the Attention Layer Supplementary Integration (ALSI) framework for UATR. This model demonstrated outstanding performance, effectively detecting underwater objects such as ship-radiated noise and achieving an identification accuracy of 96.39% on the ShipsEar dataset. Additionally, several other attention-based enhancement techniques have also yielded promising results [73], [74].

3) Other-based methods. The Restricted Boltzmann Machine (RBM), a fundamental component of deep learning, consists of visible and hidden layers of neurons. RBMs are widely used as energy-based models for tasks such as feature learning, dimensionality reduction, and model development [75]. In the field of UATR, RBM-based deep learning techniques have been effectively applied.

For instance, Luo and Feng [25], applied RBM and BP neural network techniques, achieving an accuracy of 93.17% on the ShipsEar dataset. Similarly, Luo et al. [18] developed a UATR system that demonstrated the effectiveness of BP neural networks and RBM autoencoders. However, despite RBM's advantages in feature learning, dimensionality reduction, and data generation, its shallow structure makes it difficult to model complex data distributions or high-dimensional features. To address these challenges, Yang et al. [35] proposed the GRU-CAE collaborative deep learning network. Their GRU-CAE-TM (Gated Recurrent Unit with Template Matching and Convolutional Autoencoder Collaborative Deep Learning Network) approach achieved an open-set recognition accuracy of 82.21%. While GRU effectively captures short-term time series dependencies, it faces limitations in processing underwater acoustic target signals with long-term dependencies. Moreover, deep learning techniques, including LSTM gradient-boosting decision trees, have been widely used to tackle issues related to few-shot learning in UATR [29],[61],[76].

VII. Model Outputs

The output of a deep learning model is typically defined by the result of forward propagation, depending on the task type and objective. Some common types of tasks include classification, regression, generation, segmentation, embedding or feature representation, and multitask output [77], [78],[79]. This section provides detailed explanations of UATR based on classification tasks and other related activities.

A. Classification-Based Methods

Classification-based methods for UATR typically involve steps such as data preprocessing, feature extraction and selection, and the application of classification algorithms to differentiate various underwater acoustic targets (e.g., submarines, fish schools, marine organisms) using sonar data. To provide a comprehensive overview of previous research, the following section will delve into various methods applied to different datasets.

1) ShipsEar-based dataset. The primary purpose of the ShipsEar dataset is to facilitate the development and testing of UATR algorithms and models [80]. By providing a substantial amount of underwater acoustic data, the ShipsEar dataset enables researchers to analyze and classify noise generated by various vessels, thereby enhancing environmental monitoring and maritime surveillance. The accuracy of the results varies depending on the model design, as illustrated in Table II.

TABLE II. ACCURACY OF UATR METHOD BASED ON SHIPSEAR DATASET

Author (s)	Methods	Accuracy
Fahad et al., (2024)	CNN + DenseNet	99.50%
Liu et al., (2024)	Residual Attention Convolutional Neural Network (RACNN)	99.45%
Yang et al., (2024)	1D CTN	96.84%
Pu et al., (2024)	Attention Layer Supplement Integration (ALSI)	96.37%
Luo et al., (2021)	Conditional Deep Convolutional Generative Adversarial Network (cDCGAN) Model	96.32%
Chen et al., (2023)	ARCSB modules (Rescat, SE,Maxpool, ASPP, MLP)	95.80%
Li et al., (2022)	VGGish	95.30%
Hong et al., (2021, May)	ResNet18	94.30%
Tang & Hu (2024)	DSCANet	93.00%
Ke et al., (2018)	1D convolution Autoencoder - Decoder model	93.28%
Luo et al., (2021)	RBM+BP neural network	92.60%
Du et al., (2024)	TF - DD - CNN	92.23%
Fernandes &Apolinário (2020)	KNN	71.10%
Chen et al., (2023)	ARCSB modules (Rescat, SE,Maxpool, ASPP, MLP)	95.80%

ShipsEar dataset consists of five categories: four ship types and one background noise category. The model's output can include accuracy or probability as performance metrics [60],[81]. For example, using the ShipsEar dataset, Pu et al. [72] reported a 96.39% identification accuracy for their model. This output format offers the advantage of being easily and directly comparable to the results of other studies. Moreover, Müller et al. [82] examined the strengths and weaknesses of each approach and proposed a method for integrating the ShipsEar and DeepShip datasets.

One critical factor that can influence object identification accuracy is how the dataset is divided into training and testing sets. Fahad et al. [83] pre-processed each signal into 1,000 observation frames, each containing 4,096 amplitude samples, and then randomly split the dataset, allocating 70% for training and 30% for testing. While this approach is effective, an 80/20 split between training and testing sets is more commonly used in similar studies [17], [76],[84].

In machine learning and pattern recognition, accuracy is a commonly used performance metric that measures the consistency between model predictions and actual outcomes. However, a model's accuracy is influenced by several factors, such as the criteria for splitting training and testing sets, the volume of training data, and the model's inherent performance. These factors collectively impact the overall accuracy to varying degrees. Moreover, accuracy on labeled datasets during training and validation does not reflect the model's ability to generalize data from unknown categories. Therefore, relying solely on accuracy to evaluate a model's performance is not an ideal approach. Instead, explaining the model's behavior—through mechanisms such as process visualization and interpretability—is increasingly recognized as an effective way to assess model performance comprehensively.

2) DeepShip-based dataset. Many UATR studies rely on private datasets, but the lack of public access to these datasets hampers the continuous improvement of UATR tasks [85]. The DeepShip underwater acoustic dataset, developed by Irfan et al. [86], was made publicly available to advance research in the field. Fig. 8 presents the recognition accuracy of various methods evaluated using the DeepShip dataset.

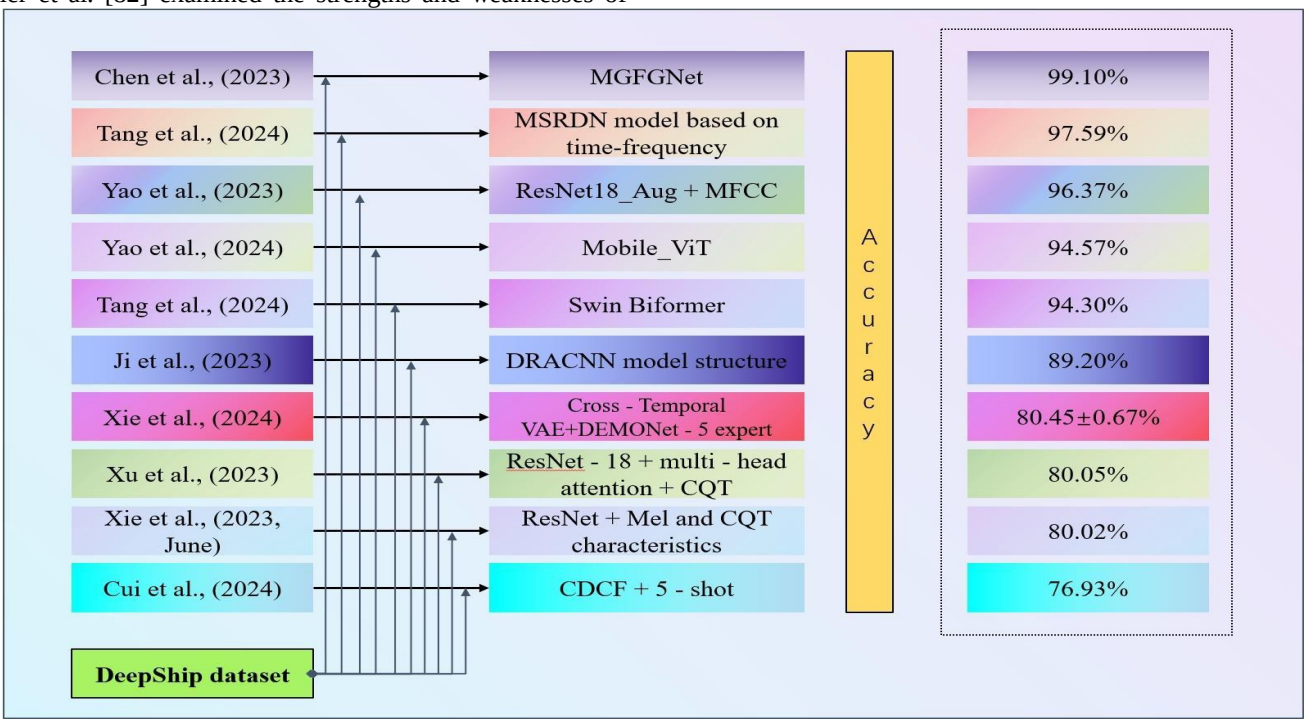


Fig. 8. Accuracy of UATR method based on DeepShip dataset.

Yao et al. [32] utilized acoustic data from three ships, extracting 1,487,488 samples for each ship type to analyze acoustic characteristics from the DeepShip database. To prevent information leakage, Xie et al. [21] divided each signal recording into 30-second segments with a 15-second overlap between adjacent segments, ensuring that data from the same track was not included in both the training and testing sets. Similarly, Chen et al. [53] pre-processed two datasets by standardizing all WAV-format audio files to a sampling rate of 22,050 Hz. They segmented the underwater acoustic data into 5-second intervals, resulting in over 30,000 sound samples, which were then categorized into training, validation, and testing sets.

3) Other underwater acoustic-based datasets. Both private and publicly accessible observation datasets have been utilized in the UATR mission. Examples include the Seabed-Objects dataset, PondEx09/PondEx10 dataset, Ocean Networks Canada data (https://oceannetworks.ca), and various private datasets [68],[87],[88]. Table III presents the recognition accuracy of several techniques evaluated using these private datasets.

TABLE III. ACCURACY OF UATR METHOD BASED ON PRIVATE DATASET

Author (s)	Methods	Accuracy
Xue et al., (2022)	CamResNet (ResNet with Channel Attention Mechanism)	98.20%
Jin & Zeng, (2023)	CTA - RDNet	97.69%
Christensen et al., (2024)	Sparse MvDA + SVM	97.30%
Sun et al., (2024)	SVM + MFCC, FMSE, GFCC	96.10%
Wang et al., (2019)	MFF - MDNN	94.30%
Yu et al., (2024)	Gradient Boosting Decision Tree (GBDT)	93.84%
Hu et al., (2018)	CNN + ELM	93.04%
Hu et al., (2020)	ATCNN	90.90%
Li & Yang, (2024)	ASTEM - DCNN	90.79% (SNR-6dB)
Cao et al., (2019)	Stacked Sparse Autoencoder (SSAE)	89.68%
Tian et al., (2021)	MSRDN（Multiscale Residual Deep Neural Network)	83.15%
Yang et al., (2022)	GRU - CAE - TM	82.21%
Honghui et al., (2022)	MCPM2	82.10%
Yang et al., (2019)	ADCNN	81.96%
Xiao et al., (2021)	ABNN (In multi target resolution task)	74.30%

Domingos et al. [55] developed a dataset using the Ocean Network Canadian data by randomly selecting audio samples from each category. To ensure all records were distinct, they continued the selection process until the required total duration was achieved. Each item in the original dataset was divided into 1-second segments, with any segment shorter than 1 second padded with zeros. The dataset was then split into training, validation, and testing sets in proportions of 85%, 10%, and 5%, respectively. Algorithms based on this dataset have demonstrated strong recognition results [48], [49],[56]. Additionally, Hu et al. [42] utilized a private dataset for UATR tasks that included underwater noise data from ferries, large boats, and small boats, collected at an anchorage with a sampling frequency of 48,000 Hz. In their experiment, 80% of the samples from each category were allocated to the training set, while the remaining 20% were used for testing. Each record was derived from a WAV audio file divided into 10-second segments, with a sampling period of 45 milliseconds and a sampling interval of 12.5 milliseconds for both training and testing samples. Several UATR studies have shown strong recognition performance using such private datasets. At the same time, some studies have achieved impressive recognition accuracy [89], [90].

B. Other Tasks-Based Methods

UATR based on regression tasks typically achieves target tracking and recognition by predicting continuous target properties such as location, velocity, and depth. These methods, which excel in localization, tracking, and dynamic detection of underwater acoustic targets, primarily rely on regression models, machine learning algorithms, and data from underwater sensors.

Zhu et al. [76] proposed a feature selection (FS) technique for underwater sound source localization based on principal component regression. This approach employs a convolutional autoencoder to extract latent features and a multi-layer perceptron for source localization. The framework demonstrates high accuracy and robustness to unseen data while achieving a 95% reduction in training time after FS. To address the limitations of existing methods in accounting for multi-attribute correlations, Honghui et al. [59] developed a deep learning-based multi-attribute correlation perception (MCP) technique for UATR.

Generative task-based approaches provide significant advantages for UATR, especially in noisy or data-scarce environments. Luo et al. [67] trained a DCGAN model using the original sample set to develop an appropriate generator. This generator can increase the number of examples for each category in the dataset by generating samples across various categories based on input labels. In addition, multi-task output-based methods for UATR have gained significant attention in recent years. These methods are capable of simultaneously addressing multiple related tasks, effectively leveraging the relationships between multimodal features, and thereby enhancing both the accuracy and robustness of recognition.

VIII. CHALLENGES AND FUTURE PROSPECTS

The methodology, techniques, and current research on UATR indicate that factors such as training methods, model architecture, and underwater acoustic data significantly impact model performance. Additionally, the model's ability to address real-time challenges in practical applications plays a crucial role. The following sections will explore these aspects from three perspectives: challenges in underwater acoustic data, enhancing UATR model evaluation and performance, interpreting deep learning in UATR.

A. Challenges in Underwater Acoustic Data

Gathering underwater acoustic data presents significant challenges due to high costs and substantial resource demands. Most UATR studies depend on a limited number of publicly accessible datasets or data collected from relatively simple maritime environments. This reliance restricts the variety and richness of the data, making it difficult to generalize findings across diverse underwater scenarios. As a result, acquiring high-quality underwater acoustic data has become a critical issue, with insufficient sample sizes exacerbating the problem. The lack of diverse datasets not only hinders model training and validation but also limits advancements in research and practical applications in more complex and variable underwater environments. Addressing these challenges is essential for enhancing the reliability and effectiveness of UATR systems.

B. Enhancing UATR Model Evaluation and Performance

Deep learning models utilized for UATR often exhibit high complexity, which can lead to significant challenges in training effectiveness. When the complexity of a model exceeds the available quantity and quality of training data, achieving desired outcomes becomes increasingly difficult. While improving model accuracy is a primary objective, relying solely on accuracy as the sole metric for evaluating model performance is insufficient.

To enhance model evaluation, it is crucial to incorporate process visualization and develop a deeper understanding of the model's mechanisms. This approach not only aids in identifying the strengths and weaknesses of the model but also fosters greater transparency. By integrating physical variables from the input data—such as temperature, salinity, depth, and geographic location—into the evaluation process, researchers can create more robust models that better reflect real-world conditions. This integration facilitates a more holistic assessment of model performance, ultimately leading to improved adaptability and reliability in practical applications. Moreover, such comprehensive evaluations can guide future research directions, informing the development of models that are not only accurate but also interpretable and adaptable to varying underwater environments.

C. Interpreting Deep Learning in UATR

Despite their significant potential, deep learning models based on neural networks encounter notable limitations in real-world applications primarily due to their "black-box" nature. This opacity makes it challenging to interpret or modify these models in real time, which is crucial for practical deployment. One of the key factors contributing to this limitation is the insufficient incorporation of physical variables into the model's outputs and the underlying mechanisms governing those outputs. To overcome these challenges, it is essential to fully integrate physical variables into the modelling process. This integration would not only enhance the interpretability of the models but also enable real-time adaptations based on changing environmental conditions, ultimately improving their efficacy in practical applications. By bridging the gap between complex neural computations and tangible physical realities, it is possible to unlock the full potential of deep learning in dynamic real-world scenarios.

In summary, addressing the challenges associated with data collection, balancing model complexity with the availability of training data, and ensuring the interpretability and adaptability of models are essential for advancing the field of underwater target recognition. These steps are crucial because effective data collection is foundational for training robust models; without diverse and high-quality datasets, models may fail to generalize across various underwater environments. Balancing model complexity with training data is equally important; overly complex models can lead to overfitting, particularly when trained on limited data. Thus, achieving the right balance allows for improved performance while maintaining generalizability.

Moreover, enhancing model interpretability is vital for fostering trust and facilitating real-time adjustments in dynamic underwater scenarios. By ensuring that models can adapt based on physical variables—such as temperature, salinity, and depth—researchers can create systems that not only perform accurately but are also responsive to changing conditions. Collectively, these advancements will pave the way for more effective and reliable UATR systems, ultimately enhancing their practical applications in marine research, naval operations, and environmental monitoring.

IX. Conclusion

This study systematically elaborates on the background and inherent challenges of Underwater Acoustic Target Recognition (UATR). A comprehensive overview is presented of the key factors influencing UATR performance, including class imbalance, environmental fluctuations, and data noise. Through the analysis of existing literature and the examination of the unique characteristics of various methodologies, recent advancements in UATR are classified according to their applied techniques and underlying properties. The role of deep learning technologies—such as transfer learning, deep convolutional networks, and temporal modelling—is explored in enhancing the accuracy and robustness of UATR systems. Particular attention is given to the challenges of model training, notably the substantial demand for annotated data, highlighting the urgent need for methods that reduce dependence on large-scale labeled datasets. Additionally, the potential of multimodal data fusion is also emphasized, as the integration of multiple data sources—such as underwater acoustic signals and optical imagery—can significantly improve recognition accuracy. Furthermore, ensemble learning approaches and attention-based mechanisms are examined as promising strategies for advancing performance in the field. Despite recent progress, current research indicates a continuing need to improve model interpretability and to better understand the mechanisms contributing to high recognition accuracy. The interpretation and visualization of these models remain largely unexplored, pointing to crucial directions for future research. Overall, the advancement of technologies like unsupervised learning and multimodal data fusion is likely to enhance UATR performance in increasingly complex and demanding application scenarios.

Acknowledgment

Thanks to the School of Computing, University Utara Malaysia (UUM), especially my supervisor Dr. Farhan.

REFERENCES

[1] H. Yang, K. Lee, Y. Choo, and K. Kim, "Underwater acoustic research trends with machine learning: Passive SONAR applications," J. Ocean Eng. Technol., vol. 34, no. 3, pp. 227–236, 2020.

[2] C. C. Leroy, S. P. Robinson, and M. J. Goldsmith, "A new equation for the accurate calculation of sound speed in all oceans," J. Acoust. Soc. Am., vol. 124, no. 5, pp. 2774–2782, 2008.

[3] D. J. Sun, C. E. Zheng, J. C. Zhang, Y. F. Han, and H. Y. Cui, "Development and prospect for underwater acoustic positioning and navigation technology," Bull. Chin. Acad. Sci., vol. 34, no. 3, pp. 331–338, 2019.

[4] Q. Zhang, L. L. Da, C. Wang, Y. H. Zhang, and J. H. Zhuo, "An overview on underwater acoustic passive target recognition based on deep learning," J. Electron. Inf. Technol., vol. 45, no. 11, pp. 4190–4202, 2023.

[5] Z. F. Lei, X. F. Lei, N. Wang, and Q. Y. Zhang, "Present status and challenges of underwater acoustic target recognition technology: A review," Front. Phys., vol. 10, p. 1044890, 2022.

[6] G. Menghani, "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," ACM Comput. Surv., vol. 55, no. 12, pp. 1–37, 2023.

[7] D. Neupane and J. Seok, "A review on deep learning-based approaches for automatic sonar target recognition," Electronics, vol. 9, no. 11, p. 1972, 2020.

[8] S. Feng, S. Ma, X. Zhu, and M. Yan, "Artificial intelligence-based underwater acoustic target recognition: A survey," Remote Sens., vol. 16, no. 17, p. 3333, 2024.

[9] A. Khan, M. M. Fouda, D. T. Do, A. Almaleh, A. M. Alqahtani, and A. U. Rahman, "Underwater target detection using deep learning: Methodologies, challenges, applications, and future evolution," IEEE Access, vol. 12, pp. 12618–12635, 2024.

[10] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, et al., "Deep Speech 2: End-to-end speech recognition in English and Mandarin," in Proc. Int. Conf. Mach. Learn., Jun. 2016, pp. 173–182.

[11] T. P. Marques, A. Rezvanifar, M. Cote, A. B. Albu, K. Ersahin, T. Mudge, and S. Gauthier, "Detecting marine species in echograms via traditional, hybrid, and deep learning frameworks," in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 5928–5935.

[12] A. C. Singer, J. K. Nelson, and S. S. Kozat, "Signal processing for underwater acoustic communications," IEEE Commun. Mag., vol. 47, no. 1, pp. 90–96, 2009.

[13] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., vol. 50, no. 2B, pp. 637–655, 1971.

[14] D. Liu, H. Yang, W. Hou, and B. Wang, "A novel underwater acoustic target recognition method based on MFCC and RACNN," Sensors, vol. 24, no. 1, p. 273, 2024.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, arXiv:1409.1556, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[17] Z. Chen, G. Xie, M. Chen, and H. Qiu, "Model for underwater acoustic target recognition with attention mechanism based on residual concatenate," J. Mar. Sci. Eng., vol. 12, no. 1, p. 24, 2023.

[18] X. Luo, Y. Feng, and M. Zhang, "An underwater acoustic target recognition method based on combined feature with automatic coding and reconstruction," IEEE Access, vol. 9, pp. 63841–63854, 2021.

[19] Y. Sun, W. Chen, C. Shuai, Z. Zhang, P. Wang, G. Cheng, and W. Yu, "Feature extraction methods for underwater acoustic target recognition of divers," Sensors, vol. 24, no. 13, p. 4412, 2024.

[20] A. Pollara, A. Sutin, and H. Salloum, "Improvement of the detection of envelope modulation on noise (DEMON) and its application to small boats," in Proc. OCEANS 2016 MTS/IEEE Monterey, Sept. 2016, pp. 1–10.

[21] Y. Xie, X. Zhang, J. Ren, and J. Xu, "DEMONet: Underwater acoustic target recognition based on multi-expert network and cross-temporal variational autoencoder," arXiv preprint, arXiv:2411.02758, 2024.

[22] X. Cao, X. Zhang, R. Togneri, and Y. Yu, "Underwater target classification at greater depths using deep neural network with joint multiple-domain feature," IET Radar, Sonar & Navigation, vol. 13, no. 3, pp. 484–491, 2019.

[23] J. Jiang, Z. Wu, J. Lu, M. Huang, and Z. Z. Xiao, "Interpretable features for underwater acoustic target recognition," Measurement, vol. 173, p. 108586, 2021.

[24] X. Wang, A. Liu, Y. Zhang, and F. Xue, "Underwater acoustic target recognition: A combination of multi-dimensional fusion features and modified deep neural network," Remote Sensing, vol. 11, no. 16, p. 1888, 2019.

[25] X. Luo and Y. Feng, "An underwater acoustic target recognition method based on restricted Boltzmann machine," Sensors, vol. 20, no. 18, p. 5399, 2020.

[26] X. Luo, L. Chen, H. Zhou, and H. Cao, "A survey of underwater acoustic target recognition methods based on machine learning," Journal of Marine Science and Engineering, vol. 11, no. 2, p. 384, 2023.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[28] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85-117, 2015.

[29] X. Cui, Z. He, Y. Xue, K. Tang, P. Zhu, and J. Han, "Cross-domain contrastive learning-based few-shot underwater acoustic target recognition," Journal of Marine Science and Engineering, vol. 12, no. 2, p. 264, 2024.

[30] J. Chen, B. Han, X. Ma, and J. Zhang, "Underwater target recognition based on multi-decision LOFAR spectrum enhancement: A deep-learning approach," Future Internet, vol. 13, no. 10, p. 265, 2021.

[31] J. Xu, Y. Xie, and W. Wang, "Underwater acoustic target recognition based on smoothness-inducing regularization and spectrogram-based data augmentation," Ocean Engineering, vol. 281, p. 114926, 2023.

[32] Q. Yao, Y. Wang, and Y. Yang, "Underwater acoustic target recognition based on data augmentation and residual CNN," Electronics, vol. 12, no. 5, p. 1206, 2023.

[33] J. Tang, E. Ma, Y. Qu, W. Gao, and L. Gan, "UAPT: An underwater acoustic target recognition method based on pre-trained Transformer," 2024.

[34] Y. Xie, J. Ren, and J. Xu, "Guiding the underwater acoustic target recognition with interpretable contrastive learning," in OCEANS 2023-Limerick, pp. 1-6, June 2023.

[35] H. Yang, K. Zheng, and J. Li, "Open set recognition of underwater acoustic targets based on GRU-CAE collaborative deep learning network," Applied Acoustics, vol. 193, p. 108774, 2022.

[36] E. L. Ferguson, R. Ramakrishnan, S. B. Williams, and C. T. Jin, "Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Mar. 2017, pp. 2657-2661.

[37] L. Du, Z. Wang, Z. Lv, D. Han, L. Wang, F. Yu, and Q. Lan, "A method for underwater acoustic target recognition based on the delay-Doppler joint feature," Remote Sensing, vol. 16, no. 11, p. 2005, 2024.

[38] P. Ashok and B. Latha, "An improving recognition accuracy of underwater acoustic targets based on gated recurrent unit (GRU) neural network method," in Proc. 1st Int. Conf. Comput. Sci. Technol. (ICCST), Nov. 2022, pp. 1-6.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Adv. Neural Inf. Process. Syst., vol. 25, 2012.

[40] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 421-425, Mar. 2017.

[41] D. Li, F. Liu, T. Shen, L. Chen, X. Yang, and D. Zhao, "Generalizable underwater acoustic target recognition using feature extraction module of neural network," Appl. Sci., vol. 12, no. 21, p. 10804, 2022.

[42] G. Hu, K. Wang, and L. Liu, "Underwater acoustic target recognition based on depthwise separable convolution neural networks," Sensors, vol. 21, no. 4, p. 1429, 2021.

[43] G. Hu, K. Wang, Y. Peng, M. Qiu, J. Shi, and L. Liu, "Deep learning methods for underwater target feature extraction and recognition,"Computational Intelligence and Neuroscience, vol. 2018, no. 1, p. 1214301, 2018.

[44] G. Hu, K. Wang, and L. Liu, "A features extraction and recognition method for underwater acoustic target based on ATCNN," arXiv preprint arXiv:2011.14336, 2020.

[45] T. C. Oliveira, Y. T. Lin, and M. B. Porter, "Underwater sound propagation modeling in a complex shallow water environment," Frontiers in Marine Science, vol. 8, p. 751327, 2021.

[46] K. Yang, B. Wang, Z. Fang, and B. Cai, "An end-to-end underwater acoustic target recognition model based on one-dimensional convolution and transformer," Journal of Marine Science and Engineering, vol. 12, no. 10, p. 1793, 2024.

[47] F. Ji, J. Ni, G. Li, L. Liu, and Y. Wang, "Underwater acoustic target recognition based on deep residual attention convolutional neural network," Journal of Marine Science and Engineering, vol. 11, no. 8, p. 1626, 2023.

[48] H. Yang, J. Li, S. Shen, and G. Xu, "A deep convolutional neural network inspired by auditory perception for underwater acoustic target recognition," Sensors, vol. 19, no. 5, p. 1104, 2019.

[49] J. Li and H. Yang, "Deep learning method with auditory passive attention for underwater acoustic target recognition under the condition of ship interference," Ocean Engineering, vol. 302, p. 117674, 2024.

[50] H. Yao, T. Gao, Y. Wang, H. Wang, and X. Chen, "Mobile_ViT: Underwater Acoustic Target Recognition Method Based on Local–Global Feature Fusion," Journal of Marine Science and Engineering, vol. 12, no. 4, p. 589, 2024.

[51] S. Zhang, C. Wang, and Q. Sun, "Underwater target noise recognition and classification technology based on multi-classes feature fusion," Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University, vol. 38, no. 2, pp. 366-376, 2020.

[52] J. Wu, P. Li, Y. Wang, Q. Lan, W. Xiao, and Z. Wang, "VFR: The underwater acoustic target recognition using cross-domain pre-training with fbank fusion features," Journal of Marine Science and Engineering, vol. 11, no. 2, p. 263, 2023.

[53] Z. Chen, J. Tang, H. Qiu, and M. Chen, "MGFGNet: An automatic underwater acoustic target recognition method based on the multi-gradient flow global feature enhancement network," Frontiers in Marine Science, vol. 10, p. 1306229, 2023.

[54] F. Hong, C. Liu, L. Guo, F. Chen, and H. Feng, "Underwater acoustic target recognition with ResNet18 on Shipsear dataset," in Proceedings of the 2021 IEEE 4th International Conference on Electronics Technology (ICET), May 2021, pp. 1240-1244.

[55] L. C. Domingos, P. E. Santos, P. S. Skelton, R. S. Brinkworth, and K. Sammut, "An investigation of preprocessing filters and deep learning methods for vessel type classification with underwater acoustic data," IEEE Access, vol. 10, pp. 117582-117596, 2022.

[56] S. Tian, D. Chen, H. Wang, and J. Liu, "Deep convolution stack for waveform in underwater acoustic target recognition," Scientific Reports, vol. 11, no. 1, p. 9614, 2021.

[57] T. L. Hemminger and Y. H. Pao, "Detection and classification of underwater acoustic transients using neural networks," IEEE Transactions on Neural Networks, vol. 5, no. 5, pp. 712-718, 1994.

[58] H. Niu, E. Reeves, and P. Gerstoft, "Source localization in an ocean waveguide using supervised machine learning," The Journal of the Acoustical Society of America, vol. 142, no. 3, pp. 1176-1188, 2017.

[59] Y. Honghui, L. Junhao, and S. Meiping, "Underwater acoustic target multi-attribute correlation perception method based on deep learning," Applied Acoustics, vol. 190, p. 108644, 2022.

[60] R. P. Fernandes and J. A. Apolinário Jr, "Underwater target classification with optimized feature selection based on genetic algorithms," in Proc. Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, 2020.

[61] Q. Yu, W. Zhang, M. Zhu, J. Shi, Y. Liu, and S. Liu, "Surface and underwater acoustic target recognition using only two hydrophones based on machine learning," The Journal of the Acoustical Society of America, vol. 155, no. 6, pp. 3606-3614, 2024.

[62] H. Yang, A. Gan, H. Chen, Y. Pan, J. Tang, and J. Li, "Underwater acoustic target recognition using SVM ensemble via weighted sample and feature selection," in Proc. 13th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST), Jan. 2016, pp. 522-527.

[63] Y. Chen and X. Xu, "The research of underwater target recognition method based on deep learning," in Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC), Oct. 2017, pp. 1-5.

[64] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial nets," in Adv. Neural Inf. Process. Syst., vol. 27, 2014.

[65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1-9, 2015.

[66] Y. Xie, J. Xu, J. Ren, and J. Li, "Adversarial multi-task underwater acoustic target recognition: Toward robustness against various influential factors," J. Acoust. Soc. Am., vol. 156, no. 1, pp. 299-313, 2024.

[67] X. Luo, M. Zhang, T. Liu, M. Huang, and X. Xu, "An underwater acoustic target recognition method based on spectrograms with different resolutions," J. Mar. Sci. Eng., vol. 9, no. 11, p. 1246, 2021.

[68] X. Liu, H. Zhu, W. Song, J. Wang, L. Yan, and K. Wang, "Research on improved VGG-16 model based on transfer learning for acoustic image recognition of underwater search and rescue targets," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 2024.

[69] M. Li, K. Liu, H. Li, Y. Sun, X. Chen, and K. Mao, "Quantitative analysis on the influence of the oceanic front on underwater acoustic detection with investigated marine data," J. Mar. Sci. Eng., vol. 11, no. 8, p. 1574, 2023.

[70] X. Xiao, W. Wang, Q. Ren, P. Gerstoft, and L. Ma, "Underwater acoustic target recognition using attention-based deep neural network," JASA Express Lett., vol. 1, no. 10, 2021.

[71] Y. Xie, J. Ren, and J. Xu, "Underwater-ART: Expanding information perspectives with text templates for underwater acoustic target recognition," J. Acoust. Soc. Am., vol. 152, no. 5, pp. 2641-2651, 2022.

[72] Z. Pu, Q. Zhang, Y. Xue, P. Zhu, and X. Cui, "A novel multi-feature fusion model based on pre-trained Wav2vec 2.0 for underwater acoustic target recognition," Remote Sens., vol. 16, no. 13, p. 2442, 2024.

[73] J. Tang, W. Gao, E. Ma, X. Sun, and J. Ma, "Deep learning based underwater acoustic target recognition: Introduce a recent temporal 2D modeling method," Sensors, vol. 24, no. 5, p. 1633, 2024.

[74] P. Li, J. Wu, Y. Wang, Q. Lan, and W. Xiao, "STM: Spectrogram transformer model for underwater acoustic target recognition," Journal of Marine Science and Engineering, vol. 10, no. 10, p. 1428, 2022.

[75] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, no. 7, pp. 1527-1554, 2006.

[76] M. Zhu, X. Zhang, Y. Jiang, K. Wang, B. Su, and T. Wang, "Hybrid Underwater Acoustic Signal Multi-Target Recognition Based on DenseNet-LSTM with Attention Mechanism," in Chinese Intelligent Automation Conference, Singapore: Springer Nature Singapore, pp. 728-738, Sep. 2023.

[77] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.

[78] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[79] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[80] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, and A. Pena-Gimenez, "ShipsEar: An underwater vessel noise database," Applied Acoustics, vol. 113, pp. 64-69, 2016.

[81] X. Ke, F. Yuan, and E. Cheng, "Underwater acoustic target recognition based on supervised feature-separation algorithm," Sensors, vol. 18, no. 12, p. 4318, 2018.

[82] N. Müller, J. Reermann, and T. Meisen, "Navigating the Depths: A Comprehensive Survey of Deep Learning for Passive Underwater Acoustic Target Recognition," IEEE Access, 2024.

[83] T. O. Fahad, A. H. Miry, A. Al-Gizi, M. H. Miry, and A. T. Razzooqee, "Recognition of Underwater Acoustic Radar Signals Based on Multiresolution and Dense Convolutional Neural Network," Journal of Engineering and Sustainable Development, vol. 28, no. 6, pp. 793-800, 2024.

[84] C. Tang and G. Hu, "DSCANet: Underwater Acoustic Target Classification Using the Depthwise Separable Convolutional Attention Module," Earth Science Informatics, vol. 17, no. 6, pp. 6123-6135, 2024.

[85] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434, 2015.

[86] M. Irfan, J.B. Zheng, S. Ali, M. Iqbal, Z. Masood, and U. Hamid, "DeepShip: An Underwater Acoustic Benchmark Dataset and a Separable Convolution Based Autoencoder for Classification," Expert Systems with Applications, vol. 183, p. 115270, 2021.

[87] A. Christensen, A. Sen Gupta, and I. Kirsteins, "Underwater Small Target Classification Using Sparse Multi-View Discriminant Analysis and the Invariant Scattering Transform," Journal of Marine Science and Engineering, vol. 12, no. 10, p. 1886, 2024.

[88] M. R. Azimi-Sadjadi, D. Yao, Q. Huang, and G. J. Dobeck, "Underwater target classification using wavelet packets and neural networks," IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 784-794, 2000.

[89] L. Xue, X. Zeng, and A. Jin, "A novel deep-learning method with channel attention mechanism for underwater target recognition," Sensors, vol. 22, no. 15, p. 5492, 2022.

[90] A. Jin and X. Zeng, "A novel deep learning method for underwater target recognition based on res-dense convolutional neural network with attention mechanism," Journal of Marine Science and Engineering, vol. 11, no. 1, p. 69, 2023.

Method for Tea Leaf Plucking Timing Prediction with High Resolution of Images Based on YOLO11

Kohei Arai[1], Yoho Kawaguchi[2]
Information Science Department, Saga University, Saga City, Japan[1]
Oita Prefecture Agriculture, Forestry and Fisheries Research and Guidance Center, Oita City, Japan[2]

***Abstract*—As a method for estimating the time when tea leaves reach their peak quality (amino acid content) (optimum picking time), our previous study revealed that the optimum picking time is when the accumulated temperature from the detection of germination of new buds reaches 600°C. However, the accuracy of this germination detection was insufficient, so the estimation accuracy of the optimum picking time was also insufficient. Since annotation accuracy is extremely important for germination detection by YOLO11, strict attention is paid to annotation by hand and by increasing the number of training datasets. The detection accuracy has been improved compared to the germination detection by YOLOv8, which was previously proposed and used relatively low-resolution images. The conclusion of this study is that the estimation method of the optimum picking time based on the criterion that the optimum picking time (amino acid content reaches its peak) is effective when the accumulated temperature from germination detection meets the condition of 600°C. The effectiveness of this method has been confirmed by comparison with germination detection by experts. For tea farmers, being able to predict the optimum picking time, when the amino acid content in the new buds is at its peak, is important, and we are sure it will have a positive impact on agricultural researchers studying this subject.**

Keywords*—*Tealeaf plucking; YOLO; budding detection; spatial resolution; optical image; annotation; germination rate

I. INTRODUCTION

The quality of tea leaves is greatly affected by the time of plucking, so it is important not to miss this timing. The method of deciding the time of plucking tea varies depending on the climate and variety of the year, but it is generally decided based on an index called "degree of opening". A new bud has 5 to 6 leaves that are the origin of the bud, and as it grows, it gradually opens. This last leaf is called the "flag leaf", and when it opens, it is said to have "opened". The percentage of new buds that have opened out of the total number of buds in a certain area is called the "degree of opening", and it is said that plucking is optimal when this percentage is 50 to 80%. In addition, if you pick early, you can also pick when it is 30 to 50%.

The time of plucking is closely related to the quality and yield of the green leaves. After the tea leaves sprout, the sprouts grow and reach their peak in amino acids (tea leaves contain amino acids such as theanine, glutamic acid, aspartic acid, arginine, and serine. These amino acids work together to create the umami and sweetness of tea) and this is when the optimum time for picking is reached. Once the optimum time for picking has passed, these amino acids are transformed into catechins because of sunlight, which is fine for functional catechin tea, but makes it unsuitable for high-quality sencha. For this reason, we check the time when the buds sprout. The program's output is either germinated or not. To determine this, we use YOLO11 to analyze the captured images and estimate the optimum time for picking. Meanwhile, when the softness and umami components of the new buds are at their peak, the new buds have not yet grown sufficiently, and the yield is low. On the other hand, when the yield is at its peak, the “degree of opening” is too high, so the leaves become hard and astringent. For these reasons, it is best to pick the tea leaves when there is a certain yield, and the quality does not deteriorate too much. The timing of picking greatly determines the quality of the tea leaves that year.

To improve the profitability of tea production, it is important to increase the gross profit from the first harvest, which accounts for approximately 80% of the profit from the tea produced annually. Gross profit from tea is calculated by multiplying the yield by the unit price, and although the unit price is influenced by the market price at the time, it is generally determined by the quality. In other words, the main factors that determine the profitability of a tea plantation are the yield and quality of the tea leaves harvested, and increasing both in a balanced manner contributes to improving profitability. For this reason, tea plantation management increases productivity while ensuring a certain level of quality is maintained at production sites.

Previously, the method of budding detection based on YOLOv8 [1] (You Only Look Once version 8) with a relatively low spatial resolution of images for the determination of the best time to pluck tea leaves was proposed. We have already reported that if tea leaves are picked when the accumulated temperature from the time of tea leaf budding reaches 600°C, good quality tea leaves can be harvested in large quantities. Therefore, a method for detecting the budding is getting much more important. In this study, a method based on YOLO11 (You Only Look Once version 11) of object detection with relatively high spatial resolution of images is proposed. Also, a comparative study is conducted for the resolution of the cameras used for the acquisition of tealeaves from a point of view for learning performance of YOLO11. Furthermore, a comparison was made between the tea leaf germination date determined by experts and the AI-predicted germination date, and it was confirmed that the prediction accuracy was high.

Through the experiments, it was found that smaller the size of the sprout in the image, the more difficult it was to detect. Therefore, it was found to be necessary to devise ways to set the camera resolution and the distance to the observation target. We also found that setting boundaries between different labels was difficult. Furthermore, by creating a model using YOLO11, it

has become possible to easily predict the sprouting rate with high accuracy.

The following section describes related research works, followed by methods and procedures for a comparative study between the previous method and the current method with relatively high spatial resolution of optical images and a comparatively highly accurate learning model. Then some experiments are described, followed by a conclusion with some discussions.

II. Related Research Works

In the following related research studies, the effects of stratification, temperature, and light on seed germination of tea plants are clarified [2]. These effects are essential components for tea plant health conditions. Meanwhile, modeling the budburst phenology of tea plants is proposed [3]. The bud-burst phenology is getting more important for global issues. On the other hand, germination and storage of tea seeds are discussed [4]. Germination issues are the concerns with this study.

Analysis of climatic factors affecting the phenology of tea (*Camellia sinensis*) in eastern China is well reported [5]. NIST3254 is the information provided by NIST: *Camellia sinensis* (green tea) leaves. Climatic factors must be considered for the germination of tea plants. Meanwhile, the effects of cold stratification on seed germination in tea plants (*Camellia sinensis*) are cleared [6]. On the other hand, the variation of seed germination and seedling growth of tea (*Camellia sinensis*) collected from different altitudes of the Darjeeling hills under water stress conditions is clarified [7].

Genetic variation and differentiation in tea (*Camellia sinensis*) germplasm using RAPD markers are discussed [8]. The genetic diversity, relationship and molecular identification of fifteen well-known, widely planted traditional Chinese elite tea genetic resources [*Camellia sinensis* (L.) O. Kuntze] preserved in the China National Germplasm Hangzhou Tea Repository in the Tea Research Institute of the Chinese Academy of Agricultural Sciences, located in Zhejiang province, China, were investigated using RAPD markers. Meanwhile, estimation of first-flush sprouting time of tea shoots using a thermal time model is well reported [9]. On the other hand, the germination behavior of tea seeds under different environmental conditions is discussed [10].

Genetic diversity and geographical differentiation of cultivated tea in China by AFLP markers are discussed [11]. Based on the AFLP method, AFLP Figures were obtained in Guangdong tea plants. In this study, in total, 365 bands were amplified by 5 pairs of primers and 73 bands were amplified by each primer pair on average; a total of 338 (92.6%) polymorphic AFLP markers were detected. Meanwhile, the prediction of tea shoots harvesting time by monitoring environmental factors and shoot growth is proposed [12]. On the other hand, growth models for tea shoots are also proposed [13].

Predicting the flush development time of tea using environmental factors is proposed [14]. Meanwhile, a dynamic model for predicting the growth and development of the tea shoot is proposed [15]. On the other hand, the effects of temperature on seed germination in tea are clarified [16].

A machine vision system for automated tea shoot detection is developed [17]. Meanwhile, seed germination and early growth of tea (*Camellia sinensis* L.) as influenced by temperature is cleared [18]. On the other hand, image-based tea bud detection using deep learning approaches is proposed [19].

The method for plucking date determination based on the elapsed days after sprouting with NIR (Near Infrared) reflection from Sentinel-2/MSI (Multi-Spectral Imager) data is proposed [20]. It is equipped with a high-resolution multispectral imager and will carry out an optical mission with the primary objective of observing land areas. It forms a constellation of two identical satellites (Sentinel-2A, 2B). Meanwhile, appropriate tealeaf harvest timing determination based on NIR images of tealeaves is discussed [21]. On the other hand, a method for vigor diagnosis of tea trees based on nitrogen content in tealeaves relating to NDVI (Normalized Difference Vegetation Index) is proposed [22]. NDVI is an index that indicates the distribution and activity of vegetation and is the ratio of NIR and red-light reflectance (NDVI={NIR-RED}/{NIR+RED}). NDVI is expressed as a normalized value between -1 and 1, and the denser the vegetation, the higher the NDVI value.

Cadastral and tea production management system with a wireless sensor network, a GIS (Geographical Information System) based system and IoT (Internet of Things) technology is developed [23]. Methods for the determination of tealeaf plucking date with Cumulative Air Temperature (CAT) and Photosynthetically Active Radiation (PAR) are proposed [24]. PAR designates the spectral range (wave band) of solar radiation from 400 to 700 nanometers that photosynthetic organisms can use in the process of photosynthesis. This spectral region corresponds to the range of light visible to the human eye. Photons at shorter wavelengths tend to be so energetic that they can be damaging to cells and tissues, but are mostly filtered out by the ozone layer in the stratosphere. Photons at longer wavelengths do not carry enough energy to allow photosynthesis to take place.

On the other hand, the method of budding detection with a YOLO-based approach for the determination of the best time to pluck tea leaves is proposed and validated with actual data [1].

These studies cover various aspects of tea leaf budding estimation and germination rates, including the effects of environmental factors like temperature, light, and stratification on seed germination and budburst phenology. They also include modeling approaches for predicting budburst and studies on genetic variation affecting these processes.

On the other hand, the following studies are noticeable for object detections,

1) Foreign Object Detection in Urban Rail Transit Based on Deep Differentiation Segmentation Neural Network [25].

2) Multi-pose face recognition method based on improved depth residual network, International Journal of Biometrics [26].

3) Facial micro-expression recognition method based on CNN and transformer mixed model, International Journal of Biometrics [27].

III. Proposed Method

A. Proposed Method for Determination of Appropriate Plucking Time Estimation

The proposed procedure for the determination of appropriate plucking time estimation is shown in Fig. 1. To create a YOLO11-based learning model for bud detection, tea leaf canopy images are acquired with a high resolution of a visible camera, together with precise annotations of buds manually. High resolution of images is highly required at this stage. When performing object detection with low-resolution images, detection errors occur because high-resolution images of sprouts cannot be obtained. Therefore, in this study, we narrow the instantaneous field of view and take images with improved resolution to increase the number of pixels per sprout. After that, budding detection at the tea farm areas concerned with a newly acquired tea leaf canopy image based on the learned YOLO11 is made for the determination of germination date. At this stage, the date of germination is when 70% of the tea leaves are detected to have germinated. The optimum time for harvesting is when the accumulated air temperature from the germination date reaches 600℃. Air temperature is available from the meteorological agencies around the world.

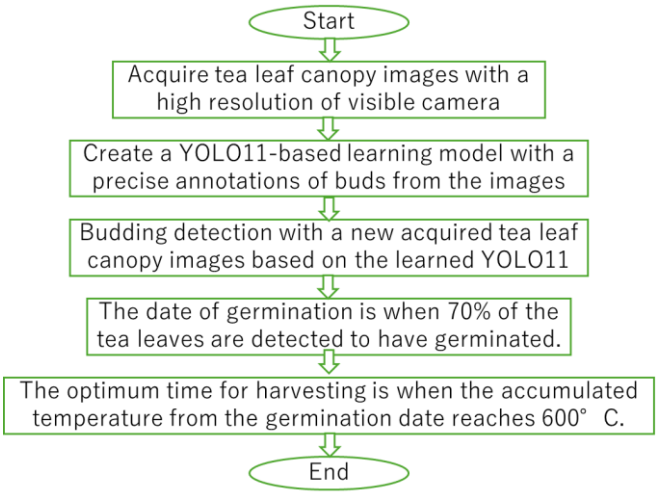


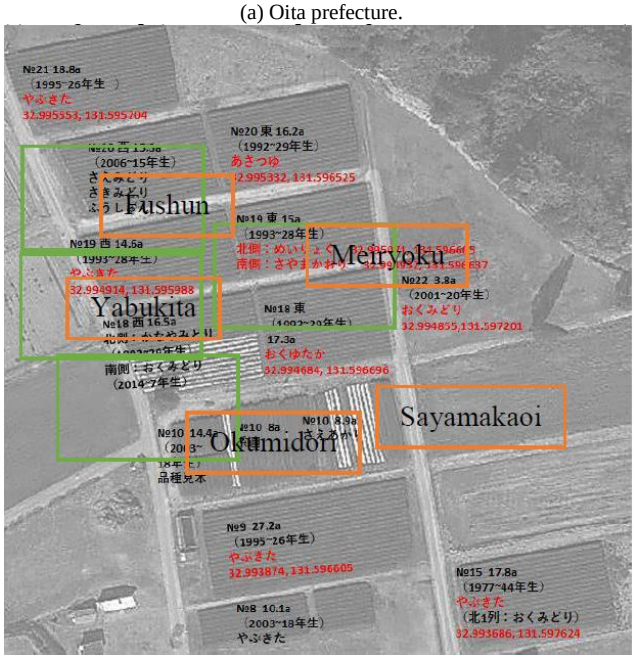
Fig. 1. Proposed procedure for determining an appropriate plucking date.

B. Research Background and the Previous Methods for Comparison

We have already shown that it is effective to predict tea component values from the accumulated temperature after germination [1]. The purpose of this research is to develop technology to identify the germination date from images. Currently, the cultivation area is large as shown in Fig. 2, and the germination date of all tea fields is not known, but the aim is to establish a method to make this possible and to identify the best time to record tea leaves from the germination date and accumulated temperature. In other words, a method to determine the germination date is necessary for efficient picking planning. There is an experimental tea farming area in which five species of tea trees (Oku-midori, Fushun, Sayama-Kaori, Meiryoku, and Yabukita) are planted.



(a) Oita prefecture.



(b) Test sites.



(c) Google photo.

Fig. 2. Location of the test site for five species of tealeaves in the intensive study area of Bungo-Ono, Oita, Japan.

The intensive study area is situated at Oita Prefectural Agriculture, Forestry and Fisheries Research Center, Bungo Ohno in Oita Prefecture, Japan (32.99 N, 131.59 E). The institute is established by Oita prefectural local government for promoting agriculture, forestry and fisheries as well as training and guidance of farmers and fishermen. There is an experimental tea farming area in which several species of tea trees (Oku-midori, Fushun, Sayama-Kaori, Meiryoku, and Yabukita) are planted. The tea tree Okumidori, which is the basis of sencha, is a breed made by crossing Yabukita, Yabu-kita and Shizuoka Zairai No. 16, and it reaches its season after Yabukita. It is characterized by its mild taste and low bitterness. It is also attracting attention as tencha (a raw material for tencha and matcha) and gyokuro. Okumidori tea tree has many buds, so you can expect a large yield.

In Fushun, which has a slightly upright tree, if the spacing between plants is 60 cm and the row spacing is changed from the customary 50 cm to 25 or 60 cm, the annual yield of fresh leaves will decrease. In the case of Fushun, when the row spacing is 50 cm, the annual fresh leaf yield decreases when the plant spacing is increased from the conventional 60 cm to 75 cm, and the planting density is reduced to 80% of the conventional planting density (1481 plants/10a). Sayamakaori is characterized by its strong aroma.

The tea leaves demonstrate enhanced firmness and a more desirable shape relative to Yabukita tea. Because it contains a lot of catechins, which are tannins, the tea has a strong and astringent taste. The plucking period of Meiryoku is the same as or slightly earlier than that of Yabukita, and the tree is vigorous and grows vigorously. Four or five years after planting, the total yield of first and second tea leaves is relatively high. The quality is the same as or slightly better than that of Yabukita. It is suitable because it is rather strong. One of the characteristics of "Yabukita" is its strong cold resistance. It is resistant to red wilt, blue wilt, and frost damage. It is also characterized by good rooting and is highly adaptable to various soils. Another strength in terms of growth is that the roots and shoots are uniform and grow quickly, making it easy to replant. For this reason, it is highly rated by farmers as an easy-to-grow variety.

C. The Proposed Method

In a previously published study [24], we clarified a method for estimating the germination date from images taken periodically from directly above the tea tree canopy as shown in Fig. 3. Namely, a 20 x 20 cm frame was placed on the tree canopy surface and a photograph was taken so that the new buds within the frame were visible. The captured image was then compressed to fit YOLO's reading size (640 pixels) to create training data. The first version of YOLO was released in 2015 and revolutionized the field of object detection. This model is characterized by being faster than previous methods by processing the image at once.

In 2016, YOLOv2 was released, improving accuracy and speed. This version improved the network structure and strengthened the ability to detect smaller objects. In 2018, YOLOv3 appeared, introducing multi-scale prediction, enabling the detection of even smaller objects. This version significantly improved the accuracy of object detection. And from 2020 onwards, YOLOv4, YOLOv5, and the latest YOLOv10 each version has been released, continuously improving efficiency and accuracy. YOLOv10 incorporates the latest technology and optimizes the balance between accuracy and speed of object detection. In this research, YOLO11 is used for budding detections.

YOLO is becoming the standard for object detection due to its high speed and accuracy. Throughout its development history, technology has continued to evolve and is expected to be used in various application fields. One such field is agriculture, where it is used to monitor the health of crops and detect pests, making farming more efficient. In this study, it is used to detect new tealeaf buds.



Fig. 3. Example of the acquired image of the tea tree canopy with a relatively low-resolution camera.

Then the annotation of buds was performed as shown in Fig. 4.



Fig. 4. Example of the annotated buds (○denotes "bud", ×denotes "not bud").

However, the resolution of the camera used was insufficient, which resulted in insufficient accuracy in estimating the germination date, and therefore, insufficient annotation accuracy, and insufficient detection accuracy based on the YOLOv8 object detection learning model. Therefore, in this study, we carried out the following three measures (updating the teacher data and object detection algorithm) to improve the accuracy of estimating the germination date.

1) Improved the resolution of the training data.
2) Improved annotation accuracy.
3) Changed to the latest YOLO model.

As a result, the precision rate was 0.512, the recall rate was 0.518, and the mAP50 was 0.506, which means that the learning performance was not sufficient. In addition, the size of the buds was small, and it was particularly difficult to detect unsprouted buds, so we determined that it was necessary to improve the camera resolution.

To maintain the resolution by dividing the image, we decided to prevent preprocessing into a 640 pixels square by padding at the top and bottom while maintaining the aspect ratio when the image is a horizontal rectangle. Therefore, we cropped both ends of the captured image to make it square, divided it (5 pixels by 5 pixels) to match the reading size of YOLO, and created training data as shown in Fig. 5.

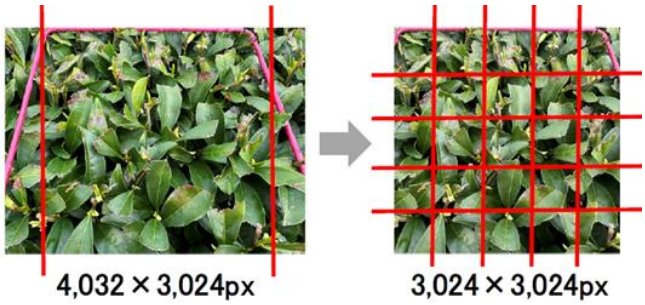


Fig. 5. Example of the acquired image with 640 pixels square and divided images with 3024 by 3024 pixels of resized images.

Then the annotation of buds was performed as shown in Fig. 6.



Fig. 6. Example of 640 by 640px of the annotated buds (Blue colored rectangles).

D. Improvement of Annotation Accuracy

YOLO11 is a model that boasts extremely high accuracy and speed in object detection. To maximize its performance, proper data annotation is essential. Annotations provide the "correct answer data" for the model to learn, and play the following important roles:

1) Accurate learning. Annotations allow the model to understand which parts are objects and improve its ability to recognize them accurately. For example, by drawing a bounding box around an object in an image, the model learns the location of the object.

2) Mitigating bias. Through annotations, you can identify misrecognition of specific classes and correct the bias of the model. This allows for fairer and more accurate predictions.

3) Domain specialization. By annotating data related to the project, the model will have recognition capabilities specialized for a specific domain. This allows it to perform well in specific applications.

The method of data annotation for YOLO11 learning is as follows:

1) Data collection. First, collect images and videos containing the target object. This data will be the basis for the model to learn. Selection of annotation tools: For annotation work, we use tools such as LabelImg[1] and Roboflow[2]. These tools provide the function of drawing bounding boxes for images and specifying the class of objects.

2) Annotation work. For the collected data, we draw bounding boxes for each object and assign the corresponding class label. This work can be done manually, but it can be made more efficient by using an automatic annotation tool.

3) Preparation of data format. YOLO11 requires annotation data to be saved in a specific format (such as YOLO PyTorch TXT format). This allows the model to read the data correctly.

4) Data verification. To ensure that the annotation is accurate, we perform reviews by others and automatic checks. This allows us to improve the quality of annotation.

5) Model training. We use the annotated dataset to train the YOLO11 model. After training, it is important to evaluate the performance of the model and correct the annotation if necessary.

Therefore, annotation is a key factor in the performance of YOLO11 training. Accurate annotation allows the model to detect objects with higher accuracy. Using the right tools and methods to perform effective annotation is the key to success. As stated in https://github.com/jsbroks/coco-annotator, when evaluating annotation accuracy, we aim to improve annotation accuracy by using annotation results output according to the COCO format. The research results of this study are that we have improved the resolution of the captured images, used object detection by YOLO11, and demonstrated that germination detection by improving annotation accuracy is effective for estimating the optimum time to pick tea leaves.

In comparison between YOLO11 and YOLOv8, YOLO11 not only improves accuracy (mAP) compared to YOLOv8, but also shows advantages in terms of computational efficiency and hardware requirements. The main points of comparison are summarized below. Computational efficiency and performance comparison,

[1] https://github.com/HumanSignal/labelImg

[2] https://github.com/roboflow/auto-annotate

1) Reduction in number of parameters and FLOPs

YOLO11n: 2.6M parameters / 6.5B FLOPs vs. YOLOv8n: 3.2M / 8.7B

YOLO11x: 56.9M parameters / 194.9B FLOPs vs. YOLOv8x: 68.2M / 257.8B

2) Improvement in CPU inference speed

YOLO11n: 56.1ms vs. YOLOv8n: 80.4ms (30% faster)

YOLO11x: 462.8ms vs. YOLOv8x: 479.1ms

3) Performance in GPU environment. No significant difference between the two models in the T4 TensorRT10 environment (e.g., YOLO11n 1.5ms vs. YOLOv8n 1.47ms)

For each tea variety, we randomly photographed the location and performed annotations. In this case, depending on the shooting angle, it may be difficult to determine whether the image is budded or not, so the labels may not be accurate. To avoid this situation, in this study, we performed annotations by taking photos from a fixed point. Examples of typical sprouting processes are shown in Fig. 7. This makes it possible to capture changes in each bud, eliminating human-made labeling errors and improving learning accuracy.



(a) Not sprouted.



(b) Sprouting Date: 3 days later.



(c) Sprouting completely.

Fig. 7. Examples of sprouting process.

As an annotation method for YOLO, we imposed the condition that there must be no space between the object and the box as shown in Fig. 4. In addition, by changing the training data from resized images to segmented images, we thought that it would be possible to closely surround each bud, which would improve the accuracy of the labels.

E. Improvement of Learning Model (YOLO)

We use YOLO11, which is faster to calculate and more accurate than YOLOv8, as shown in Fig. 8, to improve accuracy. YOLO11 is designed for speed, accuracy, and ease of use, making it ideal for a wide range of object detection and tracking, instance segmentation, and more. Open a command prompt (terminal) in C:\python\python312 and create a virtual environment. Set the environment name to YOLO11 and run python -m venv yolo11 to enter the virtual environment. Make sure YOLO11 is displayed and set it to C:\python\python312\yolo11\Scripts\activate. Proceed with installing the libraries in a virtual environment. The official website is https://github.com/ultralytics/ultralytics. You can include all libraries by installing YOLO11 with the pip command "pip install ultralytics".

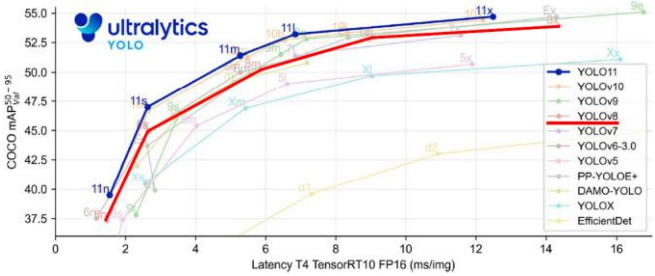


Fig. 8. YOLO11 is superior to YOLOv8.

Open the project folder in VS-Code by running mkdir c:\yolo11. Switch the interpreter to the virtual environment and execute the training by writing the same thing in Python as in the command prompt above. Data settings are done using a YML file, with the data arranged as shown below,

```
C:/yolo11/datasets/my_datasets
|       ├── images
|       |       └── train
|       |       └── val
|       └── labels
|               └── train
|               └── val
|
└── my_datasets.yaml
```

It was first released in 2015, it is a cross-platform, meaning it is available for Windows, macOS, and Linux. Visual Studio Code is lightweight yet powerful, offering many features such as debugging, version control, and extension support. It is especially popular among developers of various programming languages, including JavaScript, TypeScript, Python, and C++.

On the other hand, YML is a data format that represents structured data as a string. It stands for YAML (YAML Ain't Markup Language) and is characterized by its ease of reading and writing. The characteristics of YML are that it is lightweight and easy-to-remember grammar used for writing software configuration files and exchanging data between different software, and it is very easy to use. In addition, line breaks and indentations have clear meanings, and ease of use for programmers is emphasized. It is a complete superset of JSON, another data serialization language, and can do more than just what JSON can do.

YAML is a popular programming language because it is designed to be easy to read and understand and can be used in conjunction with other programming languages. Its flexibility and ease of use make it the language used by Ansible® to create automation processes in the form of Ansible playbooks. YAML has features from programming languages such as Perl, C, XML, and HTML. YAML is also a superset of JSON, so JSON files work directly in YAML. There are no regular formatting symbols such as curly braces, brackets, closing tags, or quotes. YAML files are also easier to read, as they use indentation to determine structure and indicate nesting, just like Python. To maintain portability across systems, tab characters are not allowed by design, so whitespace characters (literal space characters) are used instead.

Now, training can be started. The training results will be automatically saved in the folder runs. The training results will be saved in the weights folder, so just change the YOLO11.pt part of the inference program to best.pt (full path). During training, scaling (scale = 0 to 1, default 0.5) is automatically applied. It is also possible to apply rotation, mixup, learning rate, etc. In that case, just write mixup = 0.5, flipud = 0.5, lr0 = 1E-5, etc. in the model.train section.

IV. Experiment

A. Procedure and Conditions

The test varieties selected were "Yabukita", "Okumidori", "Meiryoku", "Sayamakaori", and "Fushun". These are the five main varieties of drinkable tea. The shooting period was a total of fourteen days from March 26th to April 10th, 2024 (excluding March 30th and April 7th). In addition, in the learning method 5,384 annotation images were created, 3,450 images were used as training data, and 1,360 images were used as validation data. The accuracy was verified using 574 test data. Annotations were performed on 673 images divided from the original image, and the number of annotations was increased by eight times by data expansion (rotation, inversion). Learning settings version: YOLO11s, mixup: 0.5 was used, and the rest were performed with default settings.

B. Learning Performance

Fig. 9 shows the changes in the loss function and evaluation index for each learning iteration. “train” and “val” mean training and validation, respectively. “box” denotes bounding box, “cls” denotes classification, and “dfl” stands for distribution focal loss, respectively. On the other hand, mAP50 means that the mean Average Precision is calculated by a threshold of 50 with IoU (Intersection over Union). As usual, the loss function decreases gradually, and precision and recall as well as mAP, are gradually increasing. Also, the training performance is better than the validation performance (which is much more important than the training performance). From these graphs, 200 of the number of epochs would be enough because performances are saturated at around 200.

The results of the learning model (the training and validation data (○ corresponds to sprouting and × corresponds to non-sprouting) are shown in Table I. These performance results show a significant improvement in accuracy in comparison to previously reported results [1], shown in Table II.

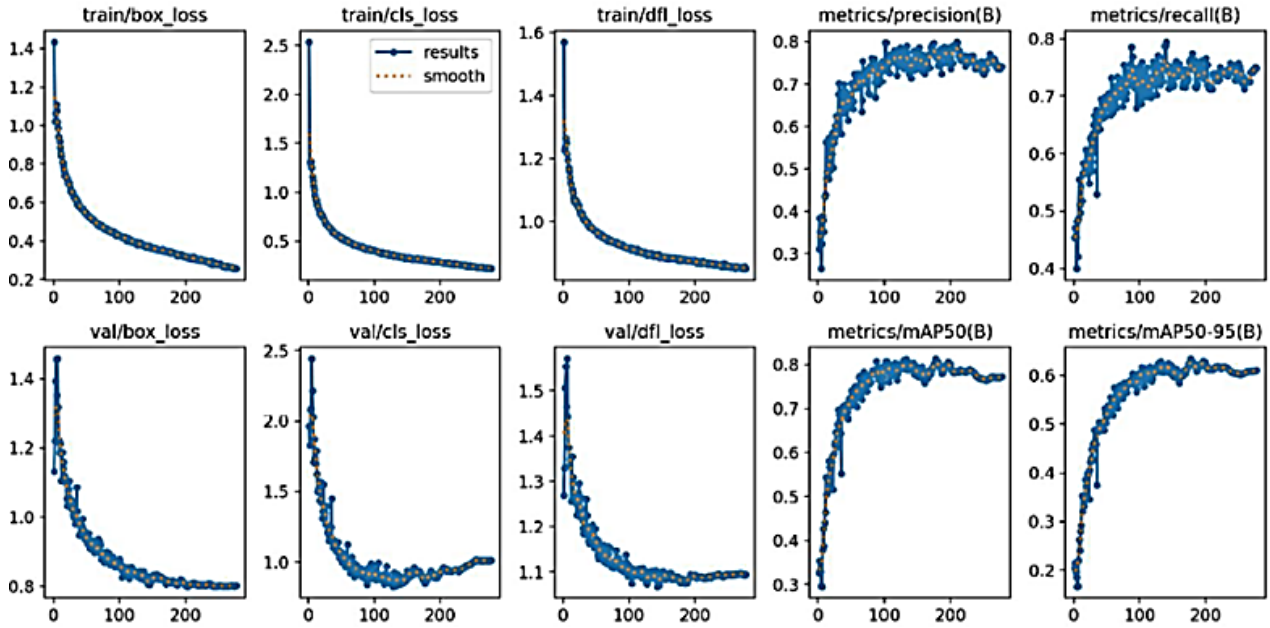


Fig. 9. Changes in the loss function and evaluation index for each learning iteration (horizontal axis is the number of epochs).

TABLE I. TRAINING AND VALIDATION PERFORMANCE OF YOLO11

Precision=0.762	Recall=0.769	mAP50=0.814
○:0.781	○:0.862	○:0.884
×:0.742	×:0.667	×:0.743

TABLE II. TRAINING AND VALIDATION PERFORMANCE OF YOLOV8

Precision=0.512	Recall=0.518	mAP=0.506
○:0.569	○:0.587	○:0.594
×:0.456	×:0.448	×:0.418

Similar results were also obtained with the test data, as shown in Table III.

TABLE III. PRECISION AND RECALL OF LEARNING PERFORMANCE OF YOLO11

Precision	○:0.779	○:0.748
Recall	×:0.748	×:0.704

Therefore, it can be said that the accuracy has improved significantly by 25 to 30%. The best score was achieved at 177 epochs. Fig. 10 shows the number of instances of each class in the training data.

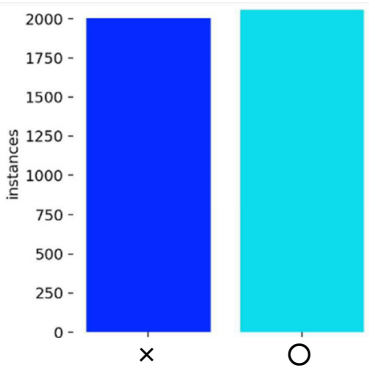


Fig. 10. The number of instances of each class in the training data.

Also, Fig. 11 shows examples of training and validation data. In Fig. 11, ○ corresponds to sprouting and × corresponds to non-sprouting, respectively.

C. Development of a Bud Break Prediction Model Using AI

It is found that bud break can be estimated with the tea leaf images acquired. It is referred to as an AI-based method. Based on the model which we created, we estimate the germination date from the image. The image used for estimation is the image before division (4,032 by 3,024 pixels). This includes the data used for learning.

Two tea field locations of 20 by 20 cm were surveyed for each variety. The bud break date was defined as the day when the proportion of buds that had sprouted reached 70%. Table IV shows the changes in germination rate (the numbers in Table IV show the percentage ratio of bud tea leaf) of each variety.

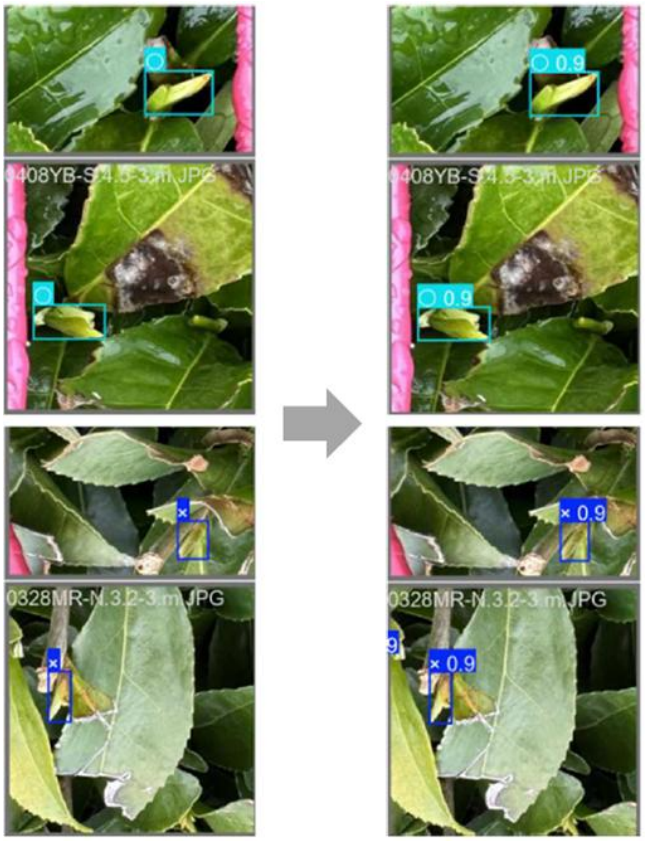
It is obvious that the germination rate is different by the variety. Red colored percentages show the germination rate. Actual denotes the germination rate estimated by the expert while, AI denotes that by AI-based method. As a result, the germination dates were consistent for three of the five varieties, and even in cases where they did not match, the error was just - two days.

From the above results, we believe that using the model we created to predict the germination date from images taken with a smartphone is effective and demonstrates high prediction accuracy.

D. Discussion

The conventional method of estimating the optimum time for plucking tea leaves was based on the experience of tea farmers and was determined by the feel of the new buds on the tea plant, the number of leaves, and the degree of leaf opening, while observing the weather. Scientifically speaking, the best quality tea leaves can be picked when the theanine content of the new buds is at its highest, but measuring this without contact is time-consuming and costly, so an alternative method was needed.

We have determined that the optimum time for plucking is when the accumulated temperature from the buds' germination reaches 600°C and have shown that the germination time of the new buds can be estimated using images captured by a visible camera and YoloV8, an AI-based method for object detection. However, the pixel size of the images and the accuracy of object detection were insufficient, so the accuracy was not sufficient compared to conventional objective methods.



(a) Training Data. (b) Validation Data.

Fig. 11. Examples of training and validation data.

TABLE IV. CHANGES IN GERMINATION RATE OF EACH VARIETY

Meiryoku	Mar.28	Mar.29	Mar.31	Apr.1	Apr.2	Apr.3	Apr.4	Apr.5	Apr.6	Apr.8
Actual	0%	0%	21%	37%	50%	65%	88%	100%	100%	100%
AI	0%	5%	25%	28%	39%	69%	90%	100%	97%	89%
SayamaKaori	Mar.28	Mar.29	Mar.31	Apr.1	Apr.2	Apr.3	Apr.4	Apr.5	Apr.6	Apr.8
Actual	0%	0%	0%	18%	29%	44%	55%	82%	100%	100%
AI	5%	0%	5%	8%	25%	50%	58%	76%	83%	88%
Yabukita	Mar.29	Mar.31	Apr.1	Apr.2	Apr.3	Apr.4	Apr.5	Apr.6	Apr.8	Apr.9
Actual	50%	25%	30%	44%	66%	66%	68%	73%	100%	100%
AI	17%	17%	15%	25%	42%	82%	77%	71%	89%	100%
Fushun	Apr.3	Apr.4	Apr.5	Apr.6	Apr.8	Apr.9	Apr.10	Apr.11	Apr.12	Apr.13
Actual	0%	7%	12%	31%	75%	85%	100%	100%	100%	100%
AI	7%	6%	8%	54%	74%	77%	56%	84%	76%	82%
Okumidori	Apr.3	Apr.4	Apr.5	Apr.6	Apr.8	Apr.9	Apr.10	Apr.11	Apr.12	Apr.13
Actual	12%	33%	58%	61%	58%	77%	87%	100%	100%	100%
AI	27%	37%	53%	59%	81%	65%	79%	78%	94%	88%

In this study, we show that by using images from a visible camera with sufficient pixel size and using YOLO11, which has a higher detection accuracy than YOLOv8, the estimation error of the germination time of the new buds can be reduced within two days.

V. CONCLUSION

We developed a budding detection method using captured images based on YOLO11. The conventional method with YOLOv8, which was proposed by the authors, had a weakness in the accuracy of detecting tea leaf budding. It was insufficient. One of the reasons is that the camera resolution. Also, the budding detection accuracy was not good enough.

By using twice much high-resolution camera and YOLO11 for budding detection with careful annotations, budding detection accuracy is improved remarkably. When the tea leaf germination rate estimated using the proposed method was compared with that of experts, it was found to be in extremely good agreement. It is concluded that the accuracy of YOLO11-based budding detection has improved significantly by 25 to 30% in comparison to the previously reported YOLOv8 [1].

Also, an AI-based method for the estimation of germination rate is proposed. Based on the model which we created, we estimate the germination date from the image. As a result, the germination dates were consistent for three of the five varieties, and even in cases where they did not match, the error was -2 days.

The proposed method is effective for detecting germination of not only the tea varieties targeted in this study, but all other tea varieties. A method based on the criterion that the optimum time for plucking is when the cumulative temperature from the time of germination is 600°C (when the amino acid content reaches its peak) is effective for estimating the optimum time for plucking.

VI. FUTURE RESEARCH WORKS

We plan to validate the proposed AI-based method for germination rate as well as for budding date prediction with the actual tea leaf quality and yield in the future. Also, it would be better to do an estimation of the most appropriate harvest date with deep learning, which can be applied to these estimations.

Further experimental studies are required for further validation of the proposed method for the determination of the most appropriate plucking date for harvesting good-quality new flesh tealeaves (Ichiban-Cha).

ACKNOWLEDGMENT

The authors would like to thank Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda of Saga University for their valuable discussions.

REFERENCES

[1] Kohei Arai, Yoho Kawaguchi, "Method of Budding Detection with YOLO-based Approach for Determination of the Best Time to Plucking Tealeaves", International Journal of Advanced Computer Science and Applications (IJACSA), 15, 5, 642-648, 2024.

[2] Li, X., Chen, H., & Liang, Y., Effects of stratification, temperature, and light on seed germination of tea plant (Camellia sinensis L.). African Journal of Biotechnology, 11(7), 1634-1640, 2012.

[3] Hoshina, T., Matsumoto, S., & Kobayashi, K., Modeling the budburst phenology of tea plants (Camellia sinensis L.) in Japan using the extended unified bud dormancy model. Agricultural and Forest Meteorology, 198-199, 32-44, 2014.

[4] Han, W., Liang, Z., & Ma, Z., Germination and storage of tea seeds. Journal of Agricultural Science, 143(1), 51-56, 2005.

[5] Wang, Y., & Ruan, J., Analysis of climatic factors affecting the phenology of tea (Camellia sinensis) in eastern China. International Journal of Biometeorology, 60(8), 1211-1222, 2016.

[6] Lee, Y. H., Moon, H. K., & Roh, M. S., Effects of cold stratification on seed germination in tea plant (Camellia sinensis). Horticulture, Environment, and Biotechnology, 55(3), 232-237, 2014.

[7] Upadhyaya, H., Panda, S. K., & Dutta, B. K., Variation of seed germination and seedling growth of tea (Camellia sinensis) collected from different altitudes of Darjeeling hills under water stress conditions. Plant Growth Regulation, 70(2), 167-173, 2013.

[8] Wachira, F. N., Tanaka, J., & Takeda, Y., Genetic variation and differentiation in tea (Camellia sinensis) germplasm using RAPD markers. Euphytica, 120(1), 33-40, 2001.

[9] Saito, T., Taguchi, T., & Otsuki, K., Estimation of first-flush sprouting time of tea shoots using a thermal time model. Japanese Journal of Crop Science, 79(4), 460-467, 2010.

[10] Kumar, R., & Bhuyan, M., Germination behavior of tea seeds under different environmental conditions. Two and a Bud, 54, 16-19, 2007.

[11] Chen, L., & Zhou, Z., Genetic diversity and geographical differentiation of cultivated tea in China by AFLP markers. Plant Biology, 7(3), 262-272, 2005.

[12] Katsuhiko Tamaki, Yuichi Kobayashi, Prediction of Tea Shoot Harvesting Time by Monitoring Environmental Factors and Shoot Growth, Engineering in Agriculture, Environment and Food Volume: 6, Issue: 2 Pages: 61-67 Date: 2013.

[13] De Costa, W.A.J.M., Mohotti, A.J., Wijeratne, M.A., Growth Models for Tea Shoots, Experimental Agriculture, 43, 2, 181-196, 2007.

[14] Matthews, R.B., Stephens, W., Predicting Flush Development Time of Tea Using Environmental Factors, Experimental Agriculture, 34, 3, 371-382, 1998.

[15] Herd, E.M., Squire, G.R., A Dynamic Model for Predicting Growth and Development of the Tea Shoot, Journal of Applied Ecology, 13, 2, 329-343, 1976.

[16] Chen, L., Zhou, Z.X., Yang, Y.J., Effects of Temperature on Seed Germination in Tea, Journal of Tea Science, 27, 1, 1-8, 2007.

[17] Wang, D., Li, M., Wang, X., et al., Development of a Machine Vision System for Automated Tea Shoot Detection, Computers and Electronics in Agriculture, 155, 197-206, 2018.

[18] Barua, D.N., Dutta, A.C., Seed Germination and Early Growth of Tea (Camellia sinensis L.) as Influenced by Temperature, Journal of Plantation Crops, 7, 2, 111-116, 1979.

[19] Liu, S., Liu, M., et al., Image-based Tea Bud Detection Using Deep Learning Approaches, Sensors, 19, 15, 3405, 2019.

[20] Kohei Arai, Yoshiko Hokazono, Method for Most Appropriate Plucking Date Determination based on the Elapsed Days after Sprouting with NIR Reflection from Sentinel-2 Data, (IJACSA) International Journal of Advanced Computer Science and Applications, 12, 4, 22-29, 2021.

[21] Kohei Arai, Yoshihiko Sasaki, Shihomi Kasuya, Hideto Matsuura, Appropriate tealeaf harvest timing determination based on NIR images of tealeaves, International Journal of Information Technology and Computer Science, 7, 7, 1-7, 2015.

[22] Kohei Arai, Method for Vigor Diagnosis of Tea Trees Based on Nitrogen Content in Tealeaves Relating to NDVI, International Journal of Advanced Research on Artificial Intelligence, 5, 10, 24-30, 2016.

[23] Kohei Arai, Cadastral and Tea Production Management System with Wireless Sensor Network, GIS, Based System and IoT Technology, International Journal of Advanced Computer Science and Applications IJACSA, 9, 1, 38-42, 2018.

[24] Kohei Arai, Yoshiko Hokazono, Method for Determination of Tealeaf Plucking Date with Cumulative Air Temperature: CAT and Photosynthetically Active Radiation: PAR, International Journal of Advanced Computer Science and Applications, 13, 10, 939-944, 2022.

[25] Feigang Tan, Min Zhai, Foreign Object Detection in Urban Rail Transit Based on Deep Differentiation Segmentation Neural Network, DOI: 10.1016/j.heliyon.2024.e37072, 2024.

[26] Feigang Tan; Yi Tang; Jiaojun Yi, Multi-pose face recognition method based on improved depth residual network, International Journal of Biometrics, 16, 5, 2024.

[27] Yi Tang, Jiaojun Yi, Feigang Tan, Facial micro-expression recognition method based on CNN and transformer mixed model, International Journal of Biometrics, DOI: 10.1504/ijbm.2024.140771, 2024.

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He was professor of Saga University from 1990 to 2017. Also, he was a Councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000 and was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He was the Vice Chairman of the Scientific Committee was an Award Committee member of ICSU/COSPAR during. He is a Science Council of Japan Special Member (COSPAR Committee) since 2012. He is an Adjunct Professor of Nishi-Kyushu University and Kurume Institute of Technology as well as Prishtina International University. He wrote 134 books and published 747 journal papers as well as 585 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html

The Factors Influencing Internet of Things Adoption in Public Hospitals: A Pilot Study

Mutasem Zrekat[1], Othman Bin Ibrahim[2]
Faculty of Computing, Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia[1]
Faculty of Management, Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia[2]

***Abstract*—Incorporating Internet of Things (IoT) technologies in healthcare represents a significant leap forward, capable of transforming the delivery and management of medical services. As healthcare systems across the globe increasingly seek innovative solutions to improve efficiency, enhance patient outcomes, and reduce operational costs, IoT emerges as a key enabler of this transformation. Despite the widely recognized benefits of IoT, its adoption in the healthcare sector, particularly within public hospitals in developing countries, remains limited and is still in the early stages. Therefore, understanding the factors influencing its adoption is essential for supporting effective adoption and advancing digital healthcare initiatives. This study aims to assess the validity and reliability of the instrument designed to identify the factors influencing the adoption of IoT technology in Jordanian public hospitals. A structured survey instrument collected a preliminary dataset from forty decision-makers in Jordanian public hospitals. The survey items were developed based on constructs derived from the Technology-Organizational-Environment (TOE) framework and the Human-Organization-Technology fit (HOT-fit) model, supported by relevant literature. Descriptive statistics were performed using SPSS, while the reliability and validity of the instrument were assessed using Partial Least Squares Structural Equation Modeling (PLS-SEM). The results demonstrated that the measurement instrument had acceptable levels of reliability and validity, confirming its suitability for use in the main study. This study enriches the existing research and enhances the broader understanding of IoT adoption in healthcare organizations, offering insights that can be useful to both practitioners and researchers in this field.**

Keywords*—*Internet of Things; adoption; Jordan; TOE

I. Introduction

The advent of the Internet of Things (IoT) has marked the beginning of a transformative era, often known as the Fourth Industrial Revolution. This revolution is reshaping industries by enabling smart technologies and interconnected systems that influence how individuals live, work, and interact with their environment. The term IoT was first defined by Ashton [1] in 1999, within the context of supply chain management. Since then, the concept has evolved across various fields, reflecting its expanding applications. In general, IoT can be defined as a rapidly growing network of objects that are addressable and identifiable uniquely, where each of these objects connects to servers to transfer their data and extract valuable knowledge that efficiently provides appropriate services. What sets IoT apart from other technologies is its integration of smart features, sensors, and actuators that enable it to detect and collect data, communicate across systems, and perform real-time analysis of vast amounts of information from internal and external environments via a global network [2]. These remarkable capabilities of IoT have attracted the attention of various industries, including healthcare, manufacturing, transportation, agriculture, supply chains, and logistics, to adopt IoT. The healthcare industry has emerged as a highly attractive domain for IoT applications, with recent research increasingly emphasizing its implementation in healthcare settings [3]. IoT plays a vital role in enhancing various aspects of the healthcare system, including chronic illness management, elderly care, patient engagement, physician-patient interaction, remote patient monitoring, and authentication. Consequently, hospitals are increasingly encouraged to adopt IoT technology, as they have the potential to improve treatment outcomes, reduce errors, reduce treatment costs, and elevate the overall quality of the healthcare system [4].

However, the global population and rising life expectancies have substantially increased the demand for healthcare services, placing significant pressure on public health institutions worldwide [5]. As a result, hospitals and healthcare organizations are operating at nearly full capacity as they strive to meet the increasing demand for healthcare, particularly in terms of hospitalization. These challenges are particularly acute in developing countries, where public hospitals may struggle to provide quality, affordable, and accessible healthcare due to limited resources [6, 7]. Moreover, the emergence of the COVID-19 pandemic has brought attention to pre-existing issues within the healthcare system, particularly concerning shortages in the healthcare workforce and inequities in access to healthcare [8]. In response to these challenges, healthcare organizations continuously seek ways to enhance healthcare delivery quality. One promising avenue for improvement is the adoption of the IoT, which has demonstrated the potential to address various issues and challenges facing the healthcare system, as highlighted by research studies [9, 10].

While there is acknowledgment of the potential benefits of the IoT in enhancing public health efforts, its integration into the operations of these organizations is not yet widespread or fully mature [11]. According to Alifan, et al. [12], the current studies on health information technology in the Jordan context are still insufficient, and more comprehensive research is needed to better understand and address the unique challenges and opportunities in this area. Disruptive technologies, such as IoT, often emerge from edge markets, which may significantly change businesses and create value within any organization [13]. However, sector-wide implementation of these technologies presents significant challenges, as unexpected obstacles during the implementation process can result in project failure [14].

Organizations that are aware of the associated risks and adequately prepared for the changes brought by digital transformation are more likely to adopt emerging technologies successfully and avoid costly implementation failures [15]. Therefore, identifying and evaluating the factors that enhance organizational readiness is critical for facilitating the effective adoption of IoT in healthcare. According to Chang, et al. [13], considering the technical aspects of new technologies alongside an organization's internal and external resources will increase its readiness to adopt new technologies and avoid wasted resources and project risks.

In response to this gap, the present pilot study represents an initial effort to assess the validity and reliability of a measurement instrument developed to investigate the key factors influencing the adoption of the Internet of Things (IoT) in Jordanian public hospitals. The instrument is grounded in established theoretical foundations, drawing on constructs from both the Technology-Organization-Environment (TOE) framework and the Human-Organization-Technology Fit (HOT-Fit) model, which are widely used in technology adoption research. The questionnaire was administered to forty decision-makers and healthcare professionals across public hospitals in Jordan to evaluate the internal consistency, construct validity, and reliability of the proposed model using SPSS and Partial Least Squares Structural Equation Modeling (PLS-SEM). Given the highly regulated nature of Jordan's public healthcare sector and the national initiatives to advance healthcare technology and digital transformation, this study contributes to the limited body of literature on IoT adoption in this context and offers a foundation for future empirical investigations.

The remainder of this study is structured as follows. The next section presents the theoretical background and conceptual model underpinning the study. This is followed by a description of the research methodology and instrument development process. The subsequent section reports the findings from the pilot study, including statistical analyses and interpretation. Finally, the study concludes with a discussion of the study's implications, limitations, and suggestions for future research.

II. Theoretical Background

The adoption of Internet of Things (IoT) technologies in healthcare has been examined through several theoretical models that aim to understand and predict technology usage and acceptance. Among the most frequently employed are the Technology Acceptance Model (TAM), Unified Theory of Acceptance and Use of Technology (UTAUT), Theory of Planned Behavior (TPB), and Diffusion of Innovations (DOI). These models primarily address individual-level factors, such as perceived usefulness, ease of use, social influence, and facilitating conditions, and are typically tested through cross-sectional surveys targeting end-users like physicians or nurses [16]. While such approaches have significantly contributed to understanding user intentions, they fall short of capturing the organizational-level factors that influence technology adoption in healthcare institutions, particularly within the context of public hospitals in developing countries. In other words, models like TAM, UTAUT, and TPB do not account for the broader organizational and environmental conditions that influence technology adoption in public healthcare institutions. The adoption of IT innovations such as IoT is influenced by more than just user perceptions. While ease of use and usefulness are important, organizational factors often play a larger role. Factors such as leadership commitment, the availability of IT resources, existing infrastructure, regulatory mandates, and the organization's technical skills play a critical role in determining adoption readiness and implementation outcomes [17]. In many cases, especially in large institutions, the decision to adopt new technologies depends more on these internal capabilities than on individual user opinions. Since IoT adoption in public hospitals, particularly in developing countries like Jordan, remains an emerging research area, the focus of this research is on the organizational level, guided by the nature of IoT adoption in public hospitals, where decisions are typically made at the institutional rather than individual level.

This study adopts a theoretical lens approach to explore the factors influencing Jordanian public hospitals to adopt IoT. As defined by Creswell and Creswell [18], a theoretical lens approach involves applying established theories to investigate relatively new or under-researched topics. In this study, two complementary frameworks are used as guiding lenses: the TOE framework [19] and the HOT-Fit model [20]. Both models have been widely applied in technology adoption and health information systems research and provide a structured foundation for examining technological, organizational, environmental, and human factors that may shape IoT adoption in healthcare. Employing these theories as guiding lenses ensures that the research remains theoretically grounded while exploring an emerging issue within a complex healthcare environment.

The TOE framework has proven to be a practical tool for analyzing the adoption of various IT innovations, as noted by [21]. According to Salman et al. [22], the TOE model is suitable for measuring organizations' readiness to adapt to new work activities related to the use of new technology. Although the TOE framework was not originally developed for technology adoption in healthcare, its extensive use in prior healthcare studies has led to widespread acceptance, demonstrating its versatility. For example, Abdulaziz and Yasin [23] employed the TOE framework to understand the critical factors affecting the decision to adopt cloud computing within the healthcare industry; Yang, et al. [24] found that the TOE framework can be applied to identify the factors involved in the decisions made by integrated medical and healthcare organizations to adopt artificial intelligence (AI); and Anthony Jnr [25] indicated that the TOE framework is useful for identifying the critical factors that influence telehealth adoption. Similarly, Thyagaraj and Narayanan [26] used it to examine the adoption of IoT in the hospital context. These studies affirm the TOE framework's relevance and applicability, supporting its suitability for investigating IoT adoption in the healthcare industry. In contrast, In parallel, the HOT-Fit model was specifically designed to address the adoption and evaluation of health information systems (HIS), focusing on the interaction among human, organizational, and technological dimensions [20]. Its relevance lies in its healthcare-centered structure, which makes it highly applicable for studying IT adoption in hospital environments. According to Lian, et al. [27], human characteristics should be carefully considered when adopting technological innovations

in the healthcare environment. Although previous research [28, 29] has underscored the significance of incorporating the human dimension when studying the adoption of IT within the healthcare domain. Nevertheless, the role of human behavior within hospital environments in shaping the adoption of information technologies has received limited attention in academic research [29]. Considering the substantial convergence between the TOE and HOT-fit models, this study integrates the TOE framework with the human dimension emphasized in HOT-Fit. Both models emphasize the significance of technological, organizational, environmental, and human readiness factors in driving innovation adoption. Consequently, integrating these models can encompass most of the readiness factors that impact the adoption of IoT in the healthcare sector. For these reasons, the integrated TOE–HOT-Fit model adopted in this study directly addresses the theoretical shortcomings identified in prior work, providing a more comprehensive framework for examining IoT adoption in public hospitals.

To effectively capture the combined influence of technological, organizational, environmental, and human factors on IoT adoption, researchers have proposed a formative construct known as "IoT Readiness". This construct serves as an integrative measure that reflects the overall preparedness of an organization to adopt IoT, based on the collective impact of these key dimensions [17, 30]. Each hospital or healthcare institution may face unique conditions, such as varying levels of technological infrastructure, organizational capacity, external support, and human resource readiness that shape their specific level of readiness. Therefore, IoT Readiness is not a singular or uniform concept; rather, it is shaped by the contextual variations across institutions and is essential for understanding how prepared a healthcare organization is to adopt and integrate IoT technologies. Cintrão [31] suggests that including technological readiness as a dependent variable of TOE determinants provides a unique perspective on the phenomenon of adopting new technologies. Similarly, Chang, et al. [13] argue that integrating IoT with the TOE framework and readiness perspectives allows for a comprehensive understanding of how technology readiness, organizational readiness, and environmental readiness collectively influence the effectiveness of IoT implementations. This combined approach enables a nuanced examination of these perspectives and offers a holistic approach to studying IoT adoption. Prior research has underscored the significance of these dimensions in evaluating readiness for adoption within healthcare contexts [15, 28, 32, 33], as well as in various other contexts [34-36]. These studies provided the foundational elements for establishing the initial integrated theoretical model for IoT readiness.

A. Technological Readiness

The technology context examines a technology's inherent characteristics and their impact on readiness to adopt IoT solutions. This includes factors such as relative advantage, compatibility, complexity, and security, collectively shaping technological readiness—an essential component of overall IoT readiness. Technological readiness refers to how individuals within an organization feel prepared and capable of adopting a new technology, such as the IoT, based on how they perceive its features and benefits [37]. This includes users' perceptions of whether the technology is compatible with existing systems, easy to use, secure, and advantageous. If organizational users view the technology as useful and appropriate for their work environment, they are more likely to be open and ready to adopt it [38].

1) Relative advantage. Among the technological readiness dimension factors is Relative advantage, which refers to the degree to which adopting a new technology, such as the IoT, in a hospital setting is perceived to provide tangible benefits over existing practices and technologies [27]. This study adopts the concept of relative advantage to investigate whether, and to what extent, the adoption of IoT in public hospitals is perceived as more beneficial than alternative technologies. These benefits include enhancing organizational competitiveness, reducing costs, improving operational efficiency, and streamlining work processes [39]. Relative advantage has been widely recognized as a critical determinant of technology adoption across various sectors, including healthcare [24, 40]. Previous studies have demonstrated that when healthcare institutions perceive clear advantages from IoT, they are more inclined to consider its implementation. This highlights the importance of evaluating relative advantage when assessing the potential for IoT adoption in hospital settings.

2) Compatibility. Compatibility is a critical technological factor identified in the Diffusion of Innovation (DOI) theory as a motivational driver for the adoption of new technologies. It refers to the extent to which a new technological solution, such as the IoT, is perceived to be consistent with existing values, cultural norms, knowledge, infrastructure, and operational requirements within an organization. In the context of healthcare, this includes the seamless integration of IoT with the hospital's current technological systems and healthcare workflows, both technical and managerial [26]. Organizations are more willing to embrace technologies that align with their established practices and strategic direction. Studies, such as Bhuiyan, et al. [41] and Karahoca, et al. [42], highlighted that compatibility among IoT devices is crucial for ensuring functional coherence and adaptability in healthcare environments. It is important to note that a lack of compatibility can cause concerns about the functionality of a system and its operational reliability.

3) Complexity. Complexity is a critical factor influencing an organisation's readiness to adopt new technology [43]. According to Rogers Everett [44], Complexity refers to "the degree to which an innovation is perceived as relatively difficult to understand and use". The perception of innovation complexity can vary among organisations based on their existing skills and knowledge. Some organisations may find an innovation complex due to a lack of necessary expertise, while others, with the required skills, may not perceive it as complex. Unlike other innovation characteristics, complexity often lacks a strong positive correlation with adoption [15]. In the context of IoT, higher perceived complexity has been found to influence adoption decisions negatively [40]. This issue is particularly pronounced in healthcare, where the environment

is inherently complex, and the introduction of IoT can further amplify this complexity [45]. Consequently, healthcare organizations may resist adopting IoT if it is seen as complicating their management processes

4) Security. Although integrating IoT technology enhances an organization's productivity and efficiency, it also raises significant security and privacy issues. These concerns become more pronounced as the quantity and types of interconnected devices within organizations increase [46]. In the healthcare environment, ensuring security and privacy is a critical issue for both patients and is required by laws in many countries. Healthcare institutions must prioritize data security when utilizing IoT technology for storing and handling medical data [47]. Previous research has identified security and privacy as significant factors influencing the adoption of new technologies across various sectors. For instance, [48] found that these concerns significantly shaped hospital managers' intentions to adopt RFID technology. Similarly, Hawash et al. [49] reported a positive impact of security-related factors on IoT adoption in the oil and gas industry. Dewi et al. [36] and Sam and Chatwin [30] also examined security and privacy constructs to evaluate organizational readiness for Smart City and Big Data Analytics initiatives in Indonesia and China, respectively. These findings underscore the critical role of security and privacy in shaping organizational readiness for adopting emerging technologies like IoT.

B. Organizational Readiness

Organizational readiness is defined as the extent to which public hospitals can provide and manage all the resources necessary for the successful adoption and integration of IoT technologies into their operations [36]. It emphasizes the institution's internal capacity to support and sustain the implementation process. According to Yang et al. [17], the IT Infrastructure and Top Management Support emerge as two important organizational readiness factors for technology adoption.

1) Top management support. Top management support reflects the commitment of senior leadership to actively promote and enable the adoption of new technologies throughout the organization [50]. This includes their willingness to allocate the necessary resources, offer clear guidance, and remain actively involved in planning, evaluation, and supervision throughout the adoption process. In the context of IoT adoption in healthcare, strong commitment from top management is essential because it helps ensure that the organization is adequately prepared and motivated to embrace such innovations [51]. Without this support, efforts to adopt and implement IoT solutions may face resistance or lack direction. The transition from traditional Information Technology (IT) systems to Internet of Things technologies within healthcare organizations is not merely a technical upgrade but a carefully considered strategic decision. Therefore, the proactive involvement of top executives, especially their awareness of IoT's potential benefits, plays a crucial role in successful adoption [52]. In addition, as IoT technologies are often expensive and resources are limited, top management can create a conducive environment and allocate essential resources such as expertise and infrastructure [53]. Adequate resources are crucial for the successful implementation of IoT projects, and strong top management support has consistently been identified as a vital factor for the adoption and diffusion of large systems.

2) IT infrastructure. Besides top management support, a strong IT infrastructure is essential for successfully adopting IoT technology within the firm. It provides the foundation for seamless communication and data transfer between IoT devices. Technology infrastructure can be characterized as an integrated and interconnected system comprising computer hardware, software, networks, and tangible equipment that are essential to adopting new technologies [54]. IT infrastructure has been frequently used in literature to study the adoption of IoT in many contexts, including healthcare, where IT infrastructure stands out as a significant factor in the adoption of IoT technology in the healthcare environment [55, 56]. According to Abd El-Hamed et al. [57], organizations that plan to adopt IoT technologies should prioritize evaluating and enhancing their IT infrastructure to ensure it can support emerging technological requirements. Similarly, Rey et al. [58] highlight the significance of having reliable and ready technological systems before integrating IoT into business operations. Therefore, organizations with a well-developed and dependable technology infrastructure are more likely to adopt IoT solutions successfully.

C. Environmental Readiness

Environmental readiness refers to the organization's perception of external factors influencing its decision to adopt IoT. It reflects how hospitals respond to the environment in which they operate, including competitive pressure, government support, and the availability of vendor support. These external conditions shape how organizations carry out their activities during the adoption of new technologies.

1) Competitive pressures. Which arise from the threat of losing a competitive advantage in the industry to other competitors, prompt organizations to adopt new technologies to improve their competitiveness and enhance their productivity [59]. While competition is typically associated with the private sector, similar dynamics are increasingly observed in the public healthcare sector. As hospitals witness peer institutions integrating emerging technologies such as the IoT, the pressure to keep pace becomes a compelling motivator [26]. When hospital decision-makers believe that IoT adoption can improve operational efficiency and help them remain competitive, their willingness to invest in such technologies increases [52]. Several studies confirm that competitive pressure catalyzes technology adoption, especially when organizations trust that the innovation will enable them to outperform others [26, 32]. Therefore, the intensity of competition may significantly influence a healthcare organization's readiness to adopt IoT.

2) Vendor support. Given the complexity of IoT technology and the diverse range of devices involved, implementing an IoT system involves intricate features. Therefore, robust vendor support is essential during and after implementation to ensure a smooth deployment and continued operation [46]. Vendor support refers to the assistance and services provided by technology suppliers to their clients. This support typically includes technical assistance, problem-solving, maintenance services, training, and other resources aimed at helping clients effectively implement, use, and maintain the technology [60]. Previous studies [61, 62] have indicated that vendor support plays a significant role in influencing technology adoption in the healthcare environment. Hence, organizations with adequate vendor support for IoT implementation and operations will experience greater success in incorporating IoT technologies than those without such support.

3) Government support. Refers to the various forms of assistance provided by public authorities to foster the adoption of technological innovations [63]. Such support is vital for promoting IoT integration within the healthcare sector. Ahmetoglu et al. [64] emphasize that governmental pressure can significantly accelerate the adoption process of emerging technologies. Since healthcare institutions are primarily public or private entities that operate under strict government regulation and oversight, the role of government becomes especially influential. Therefore, governments often implement a range of financial and non-financial initiatives, such as funding programs, favorable policies, investment schemes, tax incentives, and educational or strategic interventions, to encourage innovation and digital transformation in healthcare [55, 65]. This form of government support can directly or indirectly facilitate the readiness and implementation of IoT solutions across public healthcare organizations.

D. Human Readiness

Focusing on the human aspects of IT innovations in healthcare is crucial because ultimately, end stakeholders should be able to utilize IT innovations effectively. According to Alharbi et al. [66], it is crucial to consider the human aspect before implementing any IT project, as it significantly influences the adoption of new technology. Hence, this research will apply some of the HOT-fit model concepts related to the human perspective to understand the decision to adopt IoT in the healthcare sector.

1) Technical competence. Refers to the capability of IS/IT staff to effectively understand, adapt, and implement technological innovations within their field [67]. Concerning technological advancement in the healthcare sector, previous studies have highlighted the importance of technical competence (the capability of IS staff) as a crucial factor that influences healthcare organizations to effectively adopt IT/IS innovations [27, 28, 67]. According to Lian [27], the availability of a robust Information Systems (IS) department, equipped with highly skilled staff, is crucial for solving business problems and seizing opportunities through the use of IT. When hospital IT staff are well-trained and knowledgeable about new technology, they feel more confident and assured when implementing it within the hospital setting. Regarding the IoT, technical competence is considered an essential indicator for decision-makers who want to adopt IoT technology [68, 69]. According to Savoury [70], organizations need to ensure the availability of skilled and knowledgeable staff before adopting and integrating IoT solutions into their business operations. This emphasizes the importance of having skilled staff who can handle the complexities of IoT technology, thus ensuring the successful adoption of IoT projects.

2) Employees' knowledge. Human resources are a critical asset in any organization, and their knowledge and skills play a significant role in the successful adoption of new technologies [67]. In addition to infrastructure investment, empowering employees with adequate IoT-related knowledge is essential for effective implementation and for gaining a competitive advantage. IT-proficient employees have been shown to play a key role in technology adoption, especially in healthcare settings [71]. For instance, Abugabah et al. [29] found that employee IT knowledge was a critical factor in adopting RFID technology in hospitals. Similarly, Liu [72] emphasized the need to evaluate IS/IT skills before implementing telecare systems. The introduction of IoT technologies in healthcare leads to significant changes in traditional workflows, including how data is collected, shared, and utilized across departments [73]. These changes can create challenges for staff who are accustomed to established routines and systems. As a result, ensuring that employees possess adequate operational knowledge and situational awareness becomes essential for the successful integration and effective use of IoT. Without proper understanding, even the most advanced IoT systems may be underutilized or mismanaged, ultimately hindering the intended improvements in efficiency and patient care.

E. IoT Readiness

Despite the potential benefits of IoT implementation, IoT is still in its early stages of development, and several adoption issues must be overcome before it can be widely adopted [74]. These issues extend beyond technological issues and encompass organizational and environmental factors, which can result in financial costs for the organization and may disrupt or impede the successful adoption of IoT initiatives [46]. According to Chang et al. [13], disruptive technologies, such as IoT, often emerge from edge markets, which may significantly differ from the core operations of established organizations. Consequently, when organizations are unprepared for the changes brought by the new technologies, implementing these technologies may fail to meet expectations. The authors emphasized that if organizations only consider the technical aspects of new technologies when adopting them, without considering their internal and external resources, it can result in wasted resources and project risks. Similarly, Dewi et al. [36] underscored that adopting Internet of Things technology involves inherent risks and uncertainties. Therefore, the willingness to embrace these challenges will significantly impact an organization's decision to adopt IoT. When an organization is adequately prepared and

informed about the risks and uncertainties associated with adopting an innovation, the organization becomes more willing to adopt this innovation [15]. Therefore, Readiness for IoT adoption is a critical determinant that can shape an organization's intention to adopt IoT technologies. Consequently, hospitals with a higher level of preparedness are better positioned to implement and leverage IoT technologies effectively. Fig. 1 illustrates the initially proposed IoT readiness adoption model.

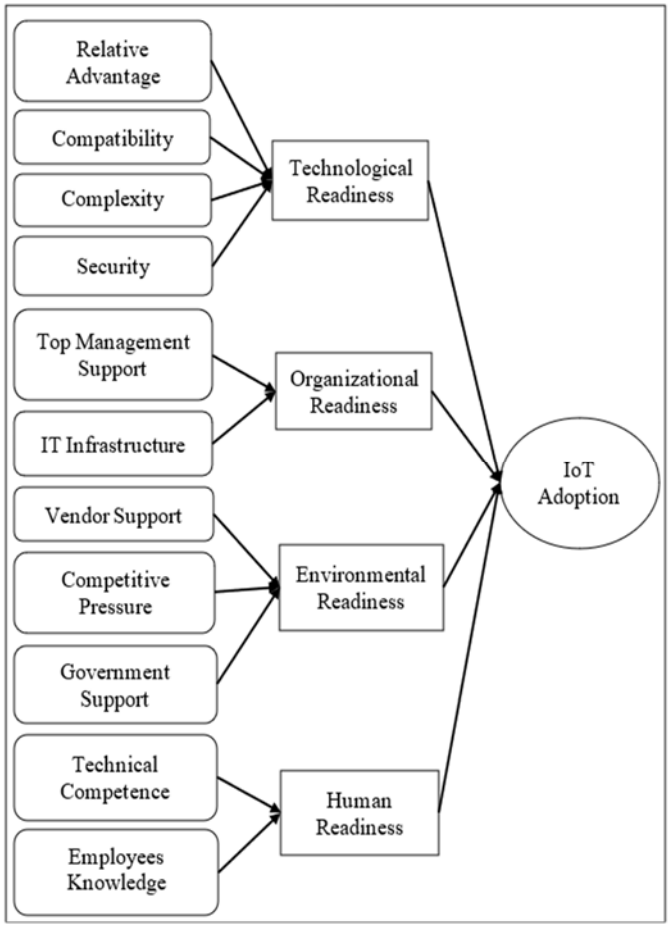


Fig. 1. Proposed research model.

III. Research Methodology

A. Instrument Design

A pilot study was carried out to assess the reliability and validity of the proposed IoT readiness adoption model. The survey instrument used in this study was carefully designed based on insights from previous literature, incorporating elements from both the TOE framework and the HOT-Fit model, which are commonly applied in healthcare technology adoption research. The development of the survey questionnaire was grounded in a review of existing literature, which provided a theoretical foundation for identifying the eleven key factors included in the proposed model. These factors are assumed to influence the adoption of IoT in hospital settings. Therefore, to accurately measure respondents' perceptions, it was essential to establish appropriate measurement scales for each factor. These scales ensure that each construct is represented by a set of items that reliably and validly capture the underlying concept related to IoT adoption in the healthcare context.

The questionnaire design followed three essential criteria: the clarity and formulation of questions, the strategic classification of variables, and the visual presentation of the instrument. The development process began with a comprehensive review of validated items from previous studies related to technology adoption and readiness. These items were then adapted to suit the specific context of IoT adoption in Jordanian public hospitals. The modifications ensured that the items were clear, contextually relevant, and appropriate for measuring the technological, organizational, environmental, and human dimensions that underpin this research. The survey instrument was organized into three sections. The first section introduced the study's purpose and emphasized the confidentiality of respondents' identities and responses. The second section collected demographic and background information. The third section gathered data for testing the proposed research framework by assessing the perceived influence of four readiness dimensions, Technology, Organization, Environment, and Human, on the intention to adopt IoT in Jordan's public hospitals. A five-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree") was used. Table I provides an overview of the constructs, the number of items per construct, and supporting references.

TABLE I. Measurement of Constructs

Variables	No of Items	Source
Relative Advantage	4	[70, 75-77]
Compatibility	4	[70, 75, 77, 78]
Complexity	4	[33, 77, 79, 80]
Security	3	[61, 77, 81, 82]
Top Management support	4	[17, 75, 83, 84]
IT Infrastructure	4	[17, 85, 86]
Competitive Pressure	4	[17, 67, 87]
Vendor Support	4	[87, 88]
Government Support	5	[70, 75, 82, 89]
Employee Knowledge	4	[67, 77, 87]
Technical competence	3	[27, 67, 85, 88]
Intention to adopt IoT	4	[17, 30, 90]

B. Content Validity

Expert judgment was employed as a primary method to evaluate the relevance and clarity of the questionnaire items to establish content validity. Five experts specializing in Information Systems (IS) and health information technology were selected based on their academic expertise and familiarity with the study's context. According to Lynn [91], a panel of at least three experts is recommended for content validation, while more than ten is generally unnecessary. The selected experts were explicitly chosen for their knowledge of the IS domain and their understanding of the study's objectives, ensuring that their evaluations were aligned with the intended research focus. Each expert was asked to assess the questionnaire items regarding their relevance to the constructs being measured and the

simplicity of their wording. This process allowed for the systematic evaluation of item appropriateness and linguistic clarity. Several revisions were made to enhance the instrument based on the feedback received. Some items were simplified to improve comprehension, while others were reworded to increase clarity and eliminate ambiguity. This content validation process helped to ensure that the questionnaire was conceptually sound and practically accessible to the targeted respondents, thereby strengthening its overall validity [92].

C. Pilot Study

A pilot study survey involves conducting a smaller-scale version of the actual data collection process to evaluate its feasibility. This preliminary step is typically performed before a large-scale study to determine if the research is practical and to avoid the full costs and effort of a comprehensive study [93]. The main goal of the pilot study is to determine whether the survey instrument requires modifications. By conducting a pilot study, researchers can identify potential issues in the study design, such as unclear questions, logistical challenges, or unexpected responses, and make necessary adjustments before proceeding to the main study. In this study, the pilot questionnaire was distributed to 40 participants from public hospitals in Jordan, including key stakeholders such as senior management (e.g., CEO, CIO, CMIO), middle management (e.g., department heads, project managers), IT and technical staff (e.g., IT managers, system administrators, system analysts), and senior medical professionals. These respondents were selected based on their direct involvement in healthcare services' day-to-day operations and strategic planning, and their potential roles in adopting and implementing technologies. To ensure that all participants had the same understanding of IoT applications and services in the healthcare context, a summarized definition was provided on the first page of the survey instrument. Participants were then asked whether they were familiar with such applications; only those who responded affirmatively were permitted to complete the survey.

The pilot study data were analyzed using SmartPLS 4.0 to assess the instrument's content validity and reliability. Expert evaluation ensured content validity, confirming that the survey items appropriately captured the constructs under investigation. Reliability testing was conducted to examine the instrument's internal consistency, including calculations of Cronbach's alpha, composite reliability, and indicator reliability. These tests helped to refine the questionnaire by identifying weak or ambiguous items. In addition, convergent and discriminant validity assessments were carried out to evaluate the instrument's construct validity and confirm the distinctiveness of the measured constructs. These procedures ensured that the instrument was valid and reliable for the full-scale study.

IV. Results and Discussion

A. Descriptive Analysis of Samples in Pilot Survey

The preliminary study used SPSS v.23 software to analyze the participants' demographic details. A total of forty questionnaires were distributed to respondents in Jordanian public hospitals as part of the pilot study. This sample size falls within the recommended range for pilot testing, as the literature typically advises using a small group of participants [77], with Hertzog [94] explicitly recommending a sample size of ten to forty for pilot studies. The pilot study was conducted to evaluate and refine the final data-collection process. Participant responses were analyzed to produce the frequency and percentage distributions presented in Table II, which outlines the demographic characteristics of the pilot study sample.

TABLE II. Frequency and Per cent of Demographic Variables in the Pilot Test

	Items	Frequency	Percent
Gender	Male	26	65%
	Female	14	35%
Age	Less than 30 years	0	0
	30-34 years	4	10%
	41-45 years	12	30%
	46-50 years	8	20%
	More than 50 years	2	5%
Education	Diploma	0	0
	Bachelor's Degree	26	65%
	Master's Degree	12	30%
	Doctoral Degree	2	5%
Position at Hospital	Senior Management	1	2.5%
	Middle Management	12	30%
	IT/Technical Staff	20	50%
	Healthcare Practitioners	7	17.5%
Experience in the current position	Less than 1 year	2	5%
	1-3 years	20	50%
	4-6 years	12	30%
	7-9 years	4	10%
	More than 10 years	2	5%
Experience in the Healthcare	Less than 5 years	2	5%
	6-10 years	19	47.5%
	11-15 years	8	20%
	21-25 years	1	2.5%
	26 years and above	4	10%
Knowledge about the Internet of Things	No information or knowledge	0	0
	Little information and knowledge	1	2.5%
	Some information and knowledge	2	5%
	Good information and knowledge	15	37.5%
	Excellent information and knowledge	22	55%

The sample consisted of forty respondents, with a majority being male (65%), and a relatively balanced representation across middle-aged groups, though no participants were under thirty. This may suggest a lack of younger professionals in decision-making or technical roles within public hospitals. The educational background was predominantly bachelor's degree holders (65%), indicating a solid academic foundation, with

35% having postgraduate qualifications, which could contribute positively to informed responses on technological topics such as IoT. From an organizational standpoint, 50% of respondents were IT or technical staff, which aligns well with the study's focus on IoT implementation, as these individuals are directly involved in deploying and managing such technologies. However, the limited representation from senior management (2.5%) may restrict insights from strategic-level decision-makers. In terms of work experience, half of the participants had one to three years in their current role, and nearly half (47.5%) had six to ten years of experience in the healthcare sector overall, suggesting a sample that combines both current role familiarity and broader sectoral expertise. Importantly, 92.5% of participants reported having good or excellent knowledge of IoT, indicating that the respondents were well-positioned to evaluate IoT adoption in healthcare, strengthening the reliability of the pilot findings.

B. Construct Reliability

In this study, the reliability of the research instrument was assessed using Smart PLS V4. Two main criteria were used: internal consistency and indicator reliability [95, 96]. To assess internal consistency, both Cronbach's alpha and composite reliability were employed. Cronbach's alpha evaluates reliability by examining the intercorrelations among observed indicators, while composite reliability considers the varying outer loadings of these indicators. In addition, indicator reliability was analyzed by evaluating the statistical significance of outer loadings, where a minimum threshold of 0.70 was deemed acceptable. This comprehensive approach ensured that all constructs in the research framework were reliably measured.

1) Internal consistency. Internal consistency was assessed using both Cronbach's Alpha (α) and Composite Reliability (CR), which are standard indicators for evaluating the reliability of latent constructs. Cronbach's Alpha evaluates the degree to which measurement items are correlated, assuming equal contribution of items, whereas CR accounts for the actual outer loadings of items, offering a more accurate estimation in SEM contexts [96]. A value above 0.70 is considered acceptable for both metrics. As shown in Table III, all constructs exceeded the recommended thresholds, with Cronbach's Alpha values ranging from 0.743 to 0.839 and Composite Reliability values from 0.844 to 0.892. These results demonstrate satisfactory internal consistency across all constructs. The high CR values indicate that the items within each construct reliably measure the same underlying concept, and the use of both metrics strengthens the credibility of the instrument. IT Infrastructure achieved the highest internal consistency ($\alpha = 0.839$; CR = 0.892), suggesting that this construct is both conceptually clear and consistently interpreted by respondents. This result likely reflects the tangible and well-established nature of IT infrastructure within hospitals, which may lead to more uniform responses.

In contrast, Security showed the lowest Cronbach's Alpha (0.743), and Compatibility reported the lowest CR (0.844), though both remained above the acceptable threshold. These slightly lower values may indicate more heterogeneous responses or measurement items that capture broader, more context-dependent perceptions. For Security, this could suggest variation in how respondents perceive threats and data protection standards in their institutions. For Compatibility, the modest reliability may reflect ambiguity in how well IoT systems are perceived to integrate with existing workflows, particularly in environments with varying levels of digital maturity. Taken together, the results confirm that the instrument demonstrates solid internal consistency across all constructs. The combination of high alpha and CR values provides strong empirical support for the reliability of the measurement model, laying a sound foundation for structural modeling and hypothesis testing in the next phase of the research. However, continued refinement of items with comparatively lower reliability will be important for improving the instrument's performance in future applications.

TABLE III. INTERNAL CONSISTENCY OF THE REFLECTIVE CONSTRUCTS

Constructs	Number Items	Cronbach's Alpha	Composite Reliability
Relative Advantage	4	0.796	0.866
Compatibility	4	0.758	0.844
Complexity	4	0.814	0.877
Security	3	0.743	0.854
Top management Support	4	0.799	0.867
IT Infrastructure	4	0.839	0.892
Competitive Pressure	4	0.802	0.870
Vendor Support	4	0.770	0.853
Government support	5	0.784	0.856
Technical competence	3	0.779	0.872
Employee Knowledge	4	0.768	0.852
Intention Towards IoT	4	0.815	0.879

2) Indicator reliability. Indicator reliability refers to the extent to which each observed variable (or item) accurately represents the latent construct it is intended to measure. This is assessed through the item's outer loading, which reflects the strength of the relationship between the observed variable and the construct. A commonly accepted threshold is 0.70, indicating that more than 50% of the variance in the item is explained by the underlying construct [97]. Items with loadings between 0.40 and 0.70 may be retained if they contribute to content validity, but those below 0.40 are generally recommended for removal unless strong theoretical justification exists[95].

As reported in Table IV, most items in this study achieved outer loadings above the 0.70 threshold, suggesting a high level of indicator reliability across the measurement model. This reinforces the internal consistency of the constructs and supports the appropriateness of the selected items. However, two items, GS3 (Government Support) and INTI4 (Intention Towards IoT) exhibited loadings of 0.240 and 0.366, respectively, falling well below the acceptable threshold. After identifying these problematic items, they were removed, and the measurement model was reassessed. The updated analysis showed that

eliminating these two low-loading indicators improved the reliability and validity metrics, particularly the Average Variance Extracted (AVE), Cronbach's Alpha, and Composite Reliability values for their respective constructs. This process confirms the usefulness of empirical testing in refining the measurement instrument, even when items are initially retained for theoretical completeness or face validity. The need to remove GS3 and INTI4 also points to broader considerations in measurement design. While content validity remains crucial, ensuring all facets of a construct are represented, this must be balanced with empirical evidence to avoid retaining items that dilute the strength of the construct. Future research may benefit from revising these items for clarity, rephrasing them to better reflect the conceptual domain, or supplementing them with more targeted indicators.

TABLE IV. ITEM LOADING OF INDICATORS

Constructs	Items	Outer loading before deleting items	Outer loading after deleting items
Relative Advantage	RA1	0.753	0.753
	RA2	0.791	0.791
	RA3	0.804	0.804
	RA4	0.796	0.796
Compatibility	COMPA1	0.703	0.703
	COMPA2	0.771	0.771
	COMPA3	0.801	0.801
	COMPA4	0.757	0.757
Complexity	COX1	0.721	0.722
	COX2	0.881	0.881
	COX3	0.811	0.811
	COX4	0.785	0.785
Security and Privacy	SP1	0.797	0.798
	SP2	0.779	0.779
	SP3	0.862	0.862
Top management Support	TM1	0.792	0.792
	TM2	0.742	0.742
	TM3	0.791	0.791
	TM4	0.821	0.821
IT Infrastructure	ITI1	0.778	0.778
	ITI2	0.829	0.829
	ITI3	0.833	0.833
	ITI4	0.843	0.843
Competitive Pressure	CP1	0.747	0.747
	CP2	0.838	0.838
	CP3	0.823	0.823
	CP4	0.756	0.755
Vendor Support	VS1	0.760	0.760
	VS2	0.782	0.782
	VS3	0.741	0.741
	VS4	0.793	0.793
Government support	GS1	0.780	0.794
	GS2	0.740	0.754
	GS3	0.240	-
	GS4	0.801	0.800
	GS5	0.744	0.746
Technical competence	TC1	0.903	0.903
	TC2	0.768	0.768
	TC3	0.824	0.824
IoT Knowledge	EK1	0.729	0.729
	EK2	0.809	0.809
	EK3	0.821	0.821
	EK4	0.712	0.712
Intention Towards IoT	INTI1	0.838	0.849
	INTI2	0.759	0.782
	INTI3	0.865	0.892
	INTI4	0.366	-

Moreover, the high loadings across the majority of other items provide strong evidence of a well-developed measurement model. Constructs such as Technical Competence, IT Infrastructure, and Intention Towards IoT (after item deletion) showed particularly strong loadings, indicating that these concepts are well captured and perceived consistently by respondents. This strengthens confidence in the reliability of the data collected and the subsequent structural model outcomes. Overall, the analysis of indicator reliability confirms that the measurement model is robust, though it also underscores the importance of iterative testing and item refinement. In settings like public healthcare in developing countries, where technological adoption is shaped by complex institutional and environmental factors, measurement instruments must be both context-sensitive and empirically sound.

V. CONSTRUCT VALIDITY

Construct validity is concerned with how accurately a tool or questionnaire measures the concept it was designed to assess. In other words, it examines whether the items within the instrument truly reflect the theoretical idea they are intended to capture [92]. This involves empirical assessment, examining how well a measure correlates with external criteria based on empirical observations. In this study, construct validity was investigated through convergent and discriminant validity assessments.

1) Convergent validity. Convergent validity examines whether a construct effectively captures the variance of its associated indicators, ensuring that items measuring the same concept are indeed related. This is typically assessed through the Average Variance Extracted (AVE), which reflects the proportion of variance explained by the latent construct relative to measurement error. According to Fornell and Bookstein [98], a minimum AVE of 0.50 is recommended, suggesting that at least 50% of the indicator variance should be accounted for by

the construct. As presented in Table V, all constructs in this study meet or exceed the 0.50 threshold, indicating satisfactory convergent validity.

TABLE V. CONVERGENT VALIDITY

Constructs	Number of Items	AVE
Relative Advantage	4	0.618
Compatibility	4	0.576
Complexity	4	0.643
Security and Privacy	3	0.662
Top management Support	4	0.620
IT Infrastructure	4	0.674
Competitive Pressure	4	0.627
Vendor Support	4	0.592
Technical competence	3	0.694
IoT Knowledge	4	0.592
Government support	4	0.598
Intention Towards IoT	3	0.709

This confirms that the measurement model generally performs well in representing the intended latent variables. However, a more nuanced examination reveals variation in the strength of convergent validity across constructs. For instance, Intention Towards IoT achieved the highest AVE (0.709), suggesting strong internal consistency and a high degree of shared variance among its indicators. Similarly, constructs like Technical Competence (0.694) and IT Infrastructure (0.674) exhibit robust convergence, indicating that these areas are conceptually well-defined and consistently measured in this study. Conversely, constructs such as Compatibility (0.576), IoT Knowledge (0.592), and Vendor Support (0.592), while above the threshold, have relatively lower AVE values. This may suggest potential areas for improvement in item formulation or conceptual clarity. Lower AVE values can sometimes reflect measurement items that are not strongly correlated, possibly due to broad or multidimensional definitions of the construct. Therefore, while the results support the use of these constructs in subsequent analysis, future studies might consider revisiting the measurement items for these constructs to enhance reliability and ensure tighter conceptual alignment. Overall, the results confirm acceptable convergent validity for all constructs, supporting the reliability of the measurement model.

2) Discriminant validity. Discriminant validity measures how a construct differs from others [95]. Discriminant validity was assessed using the Fornell-Larcker criterion, which requires that the square root of the AVE for each construct (shown on the diagonal in Table VI) be greater than its correlations with other constructs (off-diagonal values). This ensures that each construct shares more variance with its indicators than with other constructs, confirming that the latent variables are empirically distinct. The results support the discriminant validity of all constructs, as each diagonal value is higher than the corresponding off-diagonal correlations. For instance, the square root of AVE for Intention Towards IoT (ITI) is 0.821, exceeding its highest correlation with Security and Privacy (SP) at 0.672. Similarly, Technical Competence (TC) shows a strong distinction (0.833) despite moderate correlations with constructs like Employee Knowledge (EK) and Security and Privacy (SP).

However, some constructs demonstrate relatively high intercorrelations, particularly among Government Support (GS), Technical Competence (TC), and Security and Privacy (SP). For example, GS correlates at 0.705 with TC and 0.675 with SP, raising questions about potential conceptual overlap or shared variance due to external support mechanisms being tied closely to both technical readiness and data privacy policies. While the Fornell-Larcker threshold is met, such strong correlations suggest these constructs may be related in practice and should be more sharply delineated in future research. Overall, the model demonstrates acceptable discriminant validity, but the observed correlations warrant attention. These findings highlight the importance of refining construct definitions and measurement items in follow-up studies, especially in contexts like public healthcare, where institutional roles may blur the boundaries between support, competence, and data security.

TABLE VI. FORNELL-LARCKER CRITERION

	COMPA	CP	COX	GS	ITI	ATT	EK	RA	SP	TC	TM	VS
COMPA	**0.759**											
CP	0.443	**0.792**										
COX	-0.269	-0.299	**0.802**									
GS	0.458	0.481	-0.417	**0.774**								
ITI	0.257	0.287	-0.585	0.549	**0.821**							
ATT	0.290	0.069	-0.546	0.235	0.315	**0.842**						
EK	0.480	0.280	-0.308	0.712	0.313	0.210	**0.769**					
RA	0.243	0.251	-0.007	0.287	0.250	0.210	0.214	**0.786**				
SP	0.395	0.498	-0.398	0.675	0.672	0.397	0.552	0.453	**0.813**			
TC	0.578	0.398	-0.251	0.705	0.422	0.405	0.511	0.419	0.640	**0.833**		
TM	0.564	0.382	-0.361	0.560	0.427	0.346	0.416	0.193	0.486	0.629	**0.787**	
VS	0.125	0.467	-0.091	0.328	0.281	0.085	0.230	0.278	0.613	0.263	0.257	**0.769**

VI. CONCLUSION

This pilot study offers an empirically grounded validation of a comprehensive measurement instrument designed to assess IoT adoption readiness in Jordanian public hospitals, an area that remains underexplored despite the growing strategic importance of digital health technologies in developing countries. By integrating constructs from both the TOE framework and the HOT-Fit model, this study transcends traditional user-centric adoption theories and addresses the multidimensional complexity of institutional readiness. The results confirmed that the instrument demonstrates robust internal consistency, satisfactory indicator reliability, and strong convergent and discriminant validity, suggesting that it is both psychometrically sound and contextually relevant.

Practically, this study offers valuable insights for healthcare administrators, policymakers, and system developers aiming to promote digital transformation in public hospitals. The validated model identifies key dimensions—technological, organizational, environmental, and human readiness—that must be considered when planning for IoT implementation. Hospital leaders should ensure not only that technical infrastructure is in place but also that staff possess the necessary knowledge and competencies, and that sufficient management support and external resources (e.g., vendor and government support) are available. Moreover, this study demonstrates the importance of tailoring adoption models to the institutional realities of developing countries. The integration of both TOE and HOT-Fit constructs reflects the complex interplay of internal and external factors that shape readiness in public sector environments. As such, this pilot study contributes to a more context-sensitive approach to technology adoption in healthcare, especially within resource-constrained settings.

Building on this foundation, the next step involves deploying the validated instrument in a full-scale quantitative study to test the proposed model and examine the relationships between the readiness factors and the intention to adopt IoT. Structural equation modeling (SEM) can be employed to assess the model's explanatory power. In addition, future research may benefit from adopting a mixed-methods approach, incorporating qualitative techniques such as interviews or case studies to explore contextual nuances, user perceptions, and institutional dynamics that may not be fully captured through quantitative measures alone. Longitudinal studies could also offer valuable insights into how readiness and adoption progress over time, particularly as digital transformation efforts expand within Jordan's healthcare sector. In conclusion, the study provides a validated foundation for future empirical investigations and offers practical implications for stakeholders aiming to accelerate IoT adoption in the healthcare systems of developing countries.

REFERENCES

[1] K. Ashton, "That ‘internet of things’ thing," RFID journal, vol. 22, no. 7, pp. 97-114, 2009.

[2] A. Haldorai, "A Review on Artificial Intelligence in Internet of Things and Cyber Physical Systems," Journal of Computing and Natural Science, vol. 3, no. 1, pp. 012-023, 2023.

[3] R. Lederman, O. Ben-Assuli, and T. H. Vo, "The role of the Internet of Things in Healthcare in supporting clinicians and patients: A narrative review," Health Policy and Technology, vol. 10, no. 3, p. 100552, 2021.

[4] S. Abdulmalek et al., "IoT-based healthcare-monitoring system towards improving quality of life: A review," in Healthcare, 2022, vol. 10, no. 10, p. 1993: MDPI.

[5] M. Lynch, M. Bucknall, C. Jagger, and R. Wilkie, "Projections of healthy working life expectancy in England to the year 2035," Nature Aging, vol. 2, no. 1, pp. 13-18, 2022.

[6] M. Alarefi, "Internet of things in Saudi public healthcare organizations: the moderating role of facilitating conditions," International Journal of Data and Network Science, vol. 7, no. 1, pp. 295-304, 2023.

[7] A. Omotosho, P. Ayegba, J. Emuoyibofarhe, and C. Meinel, "Current state of ICT in healthcare delivery in developing countries," Nternational Journal of Online Engineering, vol. 15, no. 8, pp. 91-107, 2019.

[8] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," Future healthcare journal, vol. 8, no. 2, p. e188, 2021.

[9] R. P. Singh, M. Javaid, A. Haleem, and R. Suman, "Internet of things (IoT) applications to fight against COVID-19 pandemic," Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 14, no. 4, pp. 521-524, 2020.

[10] M. Osama et al., "Internet of medical things and healthcare 4.0: Trends, requirements, challenges, and research directions," Sensors, vol. 23, no. 17, p. 7435, 2023.

[11] A. Alhasan, L. Audah, I. Ibrahim, A. Al-Sharaa, A. S. Al-Ogaili, and J. M. Mohammed, "A case-study to examine doctors’ intentions to use IoT healthcare devices in Iraq during COVID-19 pandemic," International Journal of Pervasive Computing and Communications, vol. 18, no. 5, pp. 527-547, 2022.

[12] B. Alifan, M. Makhtar, and Y. A. Baker El-Ebiary, "A review study of electronic health care systems in Jordan," in AIP Conference Proceedings, 2024, vol. 2816, no. 1: AIP Publishing.

[13] N. Chang, Y. Zhang, D. Lu, X. Zheng, and J. Xue, "Is a disruptive technology disruptive? The readiness perspective based on TOE," in 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2020, pp. 893-897: IEEE.

[14] S. Balasubramanian, V. Shukla, J. S. Sethi, N. Islam, and R. Saloum, "A readiness assessment framework for Blockchain adoption: A healthcare case study," Technological Forecasting and Social Change, vol. 165, p. 120536, 2021.

[15] E. A. Ghaleb, P. Dominic, S. M. Fati, A. Muneer, and R. F. Ali, "The assessment of big data adoption readiness with a technology–organization–environment framework: a perspective towards healthcare employees," Sustainability, vol. 13, no. 15, p. 8379, 2021.

[16] M. Al-Rawashdeh, P. Keikhosrokiani, B. Belaton, M. Alawida, and A. Zwiri, "IoT adoption and application for smart healthcare: a systematic review," Sensors, vol. 22, no. 14, p. 5377, 2022.

[17] Z. Yang, J. Sun, Y. Zhang, and Y. Wang, "Understanding SaaS adoption from the perspective of organizational users: A tripod readiness model," Computers in Human Behavior, vol. 45, pp. 254-264, 2015.

[18] J. W. Creswell and J. D. Creswell, Research design: Qualitative, quantitative, and mixed methods approaches. Sage publications, 2017.

[19] L. G. Tornatzky, M. Fleischer, and A. Chakrabarti, "The processes of technological innovation. Issues in organization and management series," Lexington Books. Available at http://www. amazon. com/Processes-Technological-Innovation-Organization/Management/dp/0669203483. Accessed June, vol. 10, p. 2013, 1990.

[20] M. M. Yusof, J. Kuljis, A. Papazafeiropoulou, and L. K. Stergioulas, "An evaluation framework for Health Information Systems: human, organization and technology-fit factors (HOT-fit)," International journal of medical informatics, vol. 77, no. 6, pp. 386-398, 2008.

[21] C. A. De Mattos and F. J. B. Laurindo, "Information technology adoption and assimilation: Focus on the suppliers portal," Computers in industry, vol. 85, pp. 48-57, 2017.

[22] M. S. Salman, M. K. J. A. Sani, and N. Z. Sahid, "Assessing the Big Data Analytics Readiness based on Technology-Organization-Environment (TOE) Framework of Malaysian Libraries: Descriptive Analysis,"

International Journal of Academic Research in Progressive Education and Development, 2022.

[23] S. G. Abdulaziz and N. B. M. Yasin, "AN exploratory study to understand the critical factors affecting the decision to adopt cloud computing IN SAUDI hospitals," PONTE International Scientific Researchs Journal, vol. 75, no. 6, p. 10.21506, 2019.

[24] J. Yang, B. Luo, C. Zhao, and H. Zhang, "Artificial intelligence healthcare service resources adoption by medical institutions based on TOE framework," Digital Health, vol. 8, p. 20552076221126034, 2022.

[25] B. Anthony Jnr, "Examining the adoption of telehealth during public health emergencies based on technology organization environment framework," Journal of Science and Technology Policy Management, 2023.

[26] A. Thyagaraj and K. Narayanan, "Internet of Things (IoT) adoption in Indian Healthcare Industry-A case study from a hospital," ed, 2021.

[27] J.-W. Lian, D. C. Yen, and Y.-T. Wang, "An exploratory study to understand the critical factors affecting the decision to adopt cloud computing in Taiwan hospital," International Journal of Information Management, vol. 34, no. 1, pp. 28-36, 2014.

[28] I. A. Abraham and P. M. Alexander, "ICT technology use within public health clinics: A case study in a Gauteng township," in Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019, 2019, pp. 1-10.

[29] A. Abugabah, L. Sanzogni, L. Houghton, A. A. AlZubi, and A. Abuqabbeh, "RFID adaption in healthcare organizations: An integrative framework," Computers, Materials and Continua, vol. 70, p. 1335, 2021.

[30] K. M. Sam and C. R. Chatwin, "Understanding adoption of big data analytics in China: from organizational users perspective," in 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2018, pp. 507-510: IEEE.

[31] J. C. C. Cintrão, "Big Data Analytics in Portuguese Smes: Determinants of Organisational Readiness and Intention to Adopt," Universidade de Lisboa (Portugal), 2022.

[32] A. Barakabitze, M. Aloyce, F. Rashidi, and D. Shao, "Assessment of Factors for Blockchain Adoption Readiness in Tanzanian Public Healthcare Insurance Sectors: A Case Study of National Health Insurance Fund," Journal of ICT Systems, vol. 1, no. 2, pp. 109-132, 2023.

[33] E. A. Ghaleb, P. Dominic, N. S. S. Singh, and G. M. A. Naji, "Assessing the big data adoption readiness role in healthcare between technology impact factors and intention to adopt big data," Sustainability, vol. 15, no. 15, p. 11521, 2023.

[34] S. N. A. Ismail, W. N. Mohamed, K. Omar, N. H. N. Mat, and J. Saputra, "A Conceptual Analysis of the Technology, Organisation, Environment, Readiness and Industry 4.0 Adoption in Malaysia Small and Medium Enterprises," Theoretical and Practical Research in Economic Fields, vol. 14, no. 1, pp. 175-185, 2023.

[35] L. Stenberg and S. Nilsson, "Factors influencing readiness of adopting AI: A qualitative study of how the TOE framework applies to AI adoption in governmental authorities," ed, 2020.

[36] M. A. A. Dewi, A. N. Hidayanto, B. Purwandari, M. Kosandi, and N. F. A. Budi, "Smart city readiness model based on technology-organization-environment (TOE) framework and its effect on adoption decision," in Pacific Asia Conference Information System, 2018.

[37] B. Karan and G. Angadi, "Understanding School Readiness Factors in relation to the Incorporation of Artificial Intelligence using TOE Framework: An Empirical Evidence from India," TechTrends, vol. 69, no. 1, pp. 38-59, 2025.

[38] S. Alsheibani, Y. Cheung, and C. Messom, "Artificial Intelligence Adoption: AI-readiness at Firm-Level," PACIS, vol. 4, pp. 231-245, 2018.

[39] H. Ahmadi, G. Arji, L. Shahmoradi, R. Safdari, M. Nilashi, and M. Alizadeh, "The application of internet of things in healthcare: a systematic literature review and classification," Universal Access in the Information Society, pp. 1-33, 2018.

[40] Y. Zhao, S. Hao, Z. Chen, X. Zhou, L. Zhang, and Z. Guo, "Critical factors influencing the internet of things technology adoption behavior of construction companies: evidence from China," Engineering, Construction and Architectural Management, 2023.

[41] M. N. Bhuiyan, M. M. Rahman, M. M. Billah, and D. Saha, "Internet of things (IoT): A review of its enabling technologies in healthcare applications, standards protocols, security, and market opportunities," IEEE Internet of Things Journal, vol. 8, no. 13, pp. 10474-10498, 2021.

[42] A. Karahoca, D. Karahoca, and M. Aksöz, "Examining intention to adopt to internet of things in healthcare technology products," Kybernetes, vol. 47, no. 4, pp. 742-770, 2018.

[43] Z.-c. Zhou, Y.-k. Su, Z.-z. Zheng, and Y.-l. Wang, "Analysis of factors of willingness to adopt intelligent construction technology in highway construction enterprises," Scientific Reports, vol. 13, no. 1, p. 19339, 2023.

[44] M. Rogers Everett, "Diffusion of innovations," New York, vol. 12, 1995.

[45] B. Charyyev, M. Mansouri, and M. H. Gunes, "Modeling the Adoption of Internet of Things in Healthcare: A Systems Approach," in 2021 IEEE International Symposium on Systems Engineering (ISSE), 2021, pp. 1-8: IEEE.

[46] S. Ahmetoglu, Z. Che Cob, and N. A. Ali, "A systematic review of Internet of Things adoption in organizations: Taxonomy, benefits, challenges and critical factors," applied sciences, vol. 12, no. 9, p. 4117, 2022.

[47] R. Dantu, I. Dissanayake, and S. Nerur, "Exploratory Analysis of Internet of Things (IoT) in Healthcare: A Topic Modeling Approach," in Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.

[48] S. Zailani, M. Iranmanesh, D. Nikbin, and J. K. C. Beng, "Determinants of RFID adoption in Malaysia's healthcare industry: occupational level as a moderator," Journal of medical systems, vol. 39, pp. 1-11, 2015.

[49] B. Hawash, U. A. Mokhtar, Z. M. Yusof, M. Mukred, and A. S. Gaid, "Factors affecting Internet of Things (IoT) adoption in the Yemeni oil and gas sector," in 2021 International Conference of Technology, Science and Administration (ICTSA), 2021, pp. 1-7: IEEE.

[50] M. A. Bazel, M. Ahmad, F. Mohammed, N. H. Al-Kumaim, W. Mater, and A. Yasin, "The Factors Influencing Blockchain Adoption in Hospitals: A Pilot Study," in International Conference on Computing and Informatics, 2023, pp. 15-29: Springer.

[51] S. Mukherjee, M. M. Baral, V. Chittipaka, and S. C. Srivastava, "Analysis of Internet of Things Acceptance Dimensions in Hospitals," Ambient Intelligence and Internet Of Things: Convergent Technologies, pp. 189-213, 2022.

[52] B. Ramdani, B. Duan, and I. Berrou, "Exploring the determinants of mobile health adoption by hospitals in China: Empirical study," JMIR medical informatics, vol. 8, no. 7, p. e14795, 2020.

[53] A. Gutierrez, E. Boukrami, and R. Lumsden, "Technological, organisational and environmental factors influencing managers' decision to adopt cloud computing in the UK," Journal of Enterprise Information Management, vol. 28, no. 6, pp. 788-807, 2015.

[54] M. G. Aboelmaged, "Predicting e-readiness at firm-level: An analysis of technological, organizational and environmental (TOE) effects on e-maintenance readiness in manufacturing firms," International Journal of Information Management, vol. 34, no. 5, pp. 639-651, 2014.

[55] V. Desingh, "Internet of Things adoption barriers in the Indian healthcare supply chain: An ISM - fuzzy MICMAC approach," The International journal of health planning and management, vol. 37, no. 1, pp. 318-351, 2022.

[56] W. Chipambwa and K. Terera, "An Assessment of e-readiness for IoT adoption in Zimbabwean Hospitals: A Case of Chitungwiza Hospital," in 7th International Conference on Appropriate Technology "Sustainable Technologies to Empower Communities, p. 66.

[57] H. Abd El-Hamed, S. El-Zarka, and S. El-Barky, "Examining the Readiness to Adopt the Internet of Things and Big Data in Egyptian Companies," 2021.

[58] A. Rey, E. Panetti, R. Maglio, and M. Ferretti, "Determinants in adopting the Internet of Things in the transport and logistics industry," Journal of Business Research, vol. 131, pp. 584-590, 2021.

[59] F. Safari, N. Safari, and A. Hasanzadeh, "The adoption of software-as-a-service (SaaS): ranking the determinants," Journal of Enterprise Information Management, vol. 28, no. 3, pp. 400-422, 2015.

[60] S. Bano, U. Aijaz, H. Shadab, K. S. Lodhi, and M. A. Shamim, "Artificial Intelligence (Ai) For Talent Acquisition (Ta) In The Manufacturing Sector Of Pakistan," Journal of Namibian Studies: History Politics Culture, vol. 32, pp. 382-415, 2022.

[61] H. Ahmadi, "Factors Influencing Adoption of Hospital Information System Framework in Malaysian Public Hospitals," Universiti Teknologi Malaysia, 2015.

[62] K. Z. Han, B. Tangiisuran, and Y. H. P. Iskandar, "Factors influencing adoption of antidote tracking system among healthcare providers in Malaysian hospital setting," International Journal of Healthcare Management, vol. 16, no. 1, pp. 137-144, 2023.

[63] A. Saedi and N. A. Iahad, "Future research on cloud computing adoption by small and medium-sized enterprises: A critical analysis of relevant theories," International Journal of Actor-Network Theory and Technological Innovation (IJANTTI), vol. 5, no. 2, pp. 1-16, 2013.

[64] S. Ahmetoglu, Z. Che Cob, and N. A. Ali, "Internet of Things Adoption in the Manufacturing Sector: A Conceptual Model from a Multi-Theoretical Perspective," Applied Sciences, vol. 13, no. 6, p. 3856, 2023.

[65] P. L. Sen, "The Role of Government Policies and Initiatives in Driving IoT Adoption among Creative Industries Stakeholders in China," Journal of Digitainability, Realism & Mastery (DREAM), vol. 2, no. 03, pp. 67-76, 2023.

[66] F. Alharbi, A. Atkins, and C. Stanier, "Understanding the determinants of Cloud Computing adoption in Saudi healthcare organisations," Complex & Intelligent Systems, vol. 2, no. 3, pp. 155-171, 2016.

[67] H. Ahmadi, M. Nilashi, L. Shahmoradi, and O. Ibrahim, "Hospital Information System adoption: Expert perspectives on an adoption framework for Malaysian public hospitals," Computers in Human Behavior, vol. 67, pp. 161-189, 2017.

[68] S. H. Ladasi, M. R. Shihab, A. N. Hidayanto, and N. F. A. Budi, "Prioritizing determinants of internet of things (IoT) technology adoption: case study of agribusiness PT. XYZ," in 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), 2019, pp. 1-6: IEEE.

[69] A. Raj and A. Jeyaraj, "Antecedents and consequents of industry 4.0 adoption using technology, organization and environment (TOE) framework: A meta-analysis," Annals of Operations Research, vol. 322, no. 1, pp. 101-124, 2023.

[70] R. D. Savoury, "Influential Determinants of Internet of Things Adoption in the US Manufacturing Sector," 2019.

[71] F. Mahama and H. M. Dahlan, "HOTE Model for Accounting Information System Adoption for Small and Medium Scale Enterprises in Northern Ghana," 2022.

[72] C.-F. Liu, "Key factors influencing the intention of telecare adoption: An institutional perspective," Telemedicine and e-Health, vol. 17, no. 4, pp. 288-293, 2011.

[73] M. G. R. Alam, A. K. M. Masum, L.-S. Beh, and C. S. Hong, "Critical factors influencing decision to adopt human resource information system (HRIS) in hospitals," PloS one, vol. 11, no. 8, p. e0160366, 2016.

[74] M. Janssen, S. Luthra, S. Mangla, N. P. Rana, and Y. K. Dwivedi, "Challenges for adopting and implementing IoT in smart cities: An integrated MICMAC-ISM approach," Internet Research, vol. 29, no. 6, pp. 1589-1616, 2019.

[75] D. Lin, C. Lee, and K. Lin, "Research on effect factors evaluation of internet of things (IOT) adoption in Chinese agricultural supply chain," in 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2016, pp. 612-615: IEEE.

[76] G. Premkumar and M. Roberts, "Adoption of new information technologies in rural small businesses," Omega, vol. 27, no. 4, pp. 467-484, 1999.

[77] F. Mohammed, O. Ibrahim, M. Nilashi, and E. Alzurqa, "Cloud computing adoption model for e-government implementation," Information Development, vol. 33, no. 3, pp. 303-323, 2017.

[78] H. Gangwar, H. Date, and R. Ramaswamy, "Understanding determinants of cloud computing adoption using an integrated TAM-TOE model," Journal of Enterprise Information Management, 2015.

[79] P. Shi and B. Yan, "Factors affecting RFID adoption in the agricultural product distribution industry: empirical evidence from China," SpringerPlus, vol. 5, no. 1, pp. 1-11, 2016.

[80] V. Grover, "An empirically derived model for the adoption of customer - based interorganizational systems," Decision sciences, vol. 24, no. 3, pp. 603-640, 1993.

[81] S. R. Tehrani and F. Shirazi, "Factors influencing the adoption of cloud computing by small and medium size enterprises (SMEs)," in International Conference on Human Interface and the Management of Information, 2014, pp. 631-642: Springer.

[82] A.-A. Ammar, G. Thangarasu, P. Dominic, and A.-B. A. Al-Mekhlafi, "A Readiness Model and Factors Influencing Blockchain Adoption in Malaysia's Software Sector: A Survey Study," Sustainability, vol. 15, no. 16, pp. 1-28, 2023.

[83] B. Pudjianto, H. Zo, A. P. Ciganek, and J.-J. Rho, "Determinants of e-government assimilation in Indonesia: An empirical investigation using a TOE framework," Asia Pacific Journal of Information Systems, vol. 21, no. 1, pp. 49-80, 2011.

[84] A. N. Tashkandi and I. M. Al-Jabri, "Cloud computing adoption by higher education institutions in Saudi Arabia: an exploratory study," Cluster Computing, vol. 18, pp. 1527-1537, 2015.

[85] M. A. Hossain, C. Standing, and C. Chan, "The development and validation of a two-staged adoption model of RFID technology in livestock businesses," Information Technology & People, 2017.

[86] J. Wei, P. B. Lowry, and S. Seedorf, "The assimilation of RFID technology by Chinese companies: A technology diffusion perspective," Information & Management, vol. 52, no. 6, pp. 628-642, 2015.

[87] B. Sivathanu, "Adoption of industrial IoT (IIoT) in auto-component manufacturing SMEs in India," Information Resources Management Journal (IRMJ), vol. 32, no. 2, pp. 52-75, 2019.

[88] T. P. Liang, C. W. Huang, Y. H. Yeh, and B. Lin, "Adoption of mobile technology in business: a fit - viability model," Industrial management & data systems, 2007.

[89] R. Sandu and E. Gide, "Technological, organisational and environmental (TOE) factors that influence the adoption of cloud based service SMEs in India," in 2018 IEEE 11th International Conference on Cloud Computing (CLOUD), 2018, pp. 866-870: IEEE.

[90] T. Kim and W. Chiu, "Consumer acceptance of sports wearable technology: The role of technology readiness," International Journal of Sports Marketing and Sponsorship, vol. 20, no. 1, pp. 109-126, 2019.

[91] M. R. Lynn, "Determination and quantification of content validity," Nursing research, vol. 35, no. 6, pp. 382-386, 1986.

[92] A. Bhattacherjee, Social science research: Principles, methods, and practices. University of South Florida, 2012.

[93] R. Kumar, "Research methodology: A step-by-step guide for beginners," 2018.

[94] M. A. Hertzog, "Considerations in determining sample size for pilot studies," Research in nursing & health, vol. 31, no. 2, pp. 180-191, 2008.

[95] J. F. Hair, C. M. Ringle, and M. Sarstedt, "Partial least squares structural equation modeling: Rigorous applications, better results and higher acceptance," Long range planning, vol. 46, no. 1-2, pp. 1-12, 2013.

[96] J. F. Hair Jr, G. T. M. Hult, C. M. Ringle, M. Sarstedt, N. P. Danks, and S. Ray, Partial least squares structural equation modeling (PLS-SEM) using R: A workbook. Springer Nature, 2021.

[97] J. F. Hair, C. M. Ringle, and M. Sarstedt, "PLS-SEM: Indeed a silver bullet," Journal of Marketing theory and Practice, vol. 19, no. 2, pp. 139-152, 2011.

[98] C. Fornell and F. L. Bookstein, "Two structural equation models: LISREL and PLS applied to consumer exit-voice theory," Journal of Marketing research, vol. 19, no. 4, pp. 440-452, 1982.

Research on Network Flow Based on Statistical Analysis Methods

Maruf Juraev[1], Inomjon Yarashov[2], Adilbay Kudaybergenov[3], Alimdzhan Babadzhanov[4], Zilolaxon Mamatova[5]
Tashkent Institute of Technology, Management and Communication, Tashkent, Uzbekistan[1]
Jizzakh branch of the National University of Uzbekistan Named after Mirzo Ulugbek, Jizzakh, Uzbekistan[1]
Tashkent International University of Education, Department of Information Technology, Tashkent, Uzbekistan[2]
University of World Economy and Diplomacy, Tashkent, Uzbekistan[2]
Karakalpak State University, Nukus, Uzbekistan[3]
National University of Uzbekistan Named After Mirzo Ulugbek, Tashkent, Uzbekistan[3]
Engineering Federation of Uzbekistan, Tashkent, Uzbekistan[4]
Fergana State University, Fergana, Uzbekistan[5]

***Abstract*—To solve the problem of detecting unexpected situations in network traffic, it is proposed to determine the normal, unexpected states and behavior of the system using statistical methods. The statistical analysis methods used are the mean, variance, asymmetry coefficient, kurtosis coefficient, opposition coefficient and entropy coefficient. All statistical analysis methods allow for a deeper understanding of the uniqueness of the data. Each coefficient provides different measurement values, with which it is possible to determine the general characteristics of the data. The application of statistical analysis methods in network traffic is one of the most common methods for implementing the technology of detecting unexpected situations. For this, network traffic recorded in real time in laboratory conditions is used. Packet features of the network traffic are extracted from the recorded data, and then statistical analysis is performed using the extracted features. Markov chain is one of the most effective tools for analyzing situations and events occurring in network traffic. Markov chain is formed using the results of statistical analysis, and a Markov chain is constructed. This approach represents the state of the network flow in a probabilistic model and serves as an effective tool in monitoring the network flow.**

***Keywords*—Unexpected situation; mean; variance; asymmetry coefficient; kurtosis coefficient; opposition coefficient; entropy coefficient; packet analysis; markov chain; packet features**

I. Introduction

Today, ensuring information protection is one of the most pressing issues. The increase in the volume of data also creates new problems. This means that network security officers need to strengthen measures to eliminate or prevent problems in the proper organization of network management [32]. Traditional methods do not provide sufficient results in the face of new attacks [7]. Therefore, it is necessary to use modern analysis methods. In particular, traffic analysis, anomaly detection [13], and deep analysis of packet features are effective tools [15]. These methods are widely used by many researchers due to the fact that they can be implemented in real time [16].

Analysis through packet features is a time-saving method for detecting unexpected events in network operations [33]. Researchers use neural network, machine learning [6], deep learning and analytical analysis methods [31] to analyze packet features [3]. Statistical analysis methods have been little studied. Therefore, statistical analysis methods were used in this research work.

The research study investigated unexpected network traffic patterns. To do this, first, the network traffic was recorded, and packet features were extracted. The data was recorded in a laboratory environment using the Wireshark program, and packet features were extracted based on the developed algorithm.

Statistical analysis methods [28], including mean, variance, asymmetry coefficient, kurtosis coefficient, opposition coefficient and entropy coefficient, were used to analyze the extracted packet features [1], [2]. Markov chain is the most effective method for in-depth analysis of unexpected situations [21]. As a result of the research work, statistical analysis methods and Markov chain architecture were developed to identify unexpected situations in network flow [14]. Based on the results obtained, an assessment was made of unexpected situations in network flow.

The novelty of the study is that it proposes a Markov chain model, combining statistical indicators based on real traffic data. This allows not only to identify unexpected situations in the network, but also to predict their development trends in advance.

The study consists of six sections: Introduction, Related works, Problem statement, Proposed methodology, Experimental results and discussions, and Conclusion.

II. Related Works

Ramachandran et al. [1] have studied probability theory and mathematical analysis in detail in statistical analysis. In their book, they have justified the analysis of theoretical knowledge and practical skills necessary for solving real-life statistical problems [26].

Hu et al. [2], performed quantitative data analysis and sample variance analysis. Based on the experiments, they proposed new solutions for variance analysis [35].

Alasmar and Zakhleniuk used variance from statistical analysis algorithms to extract link data from the incoming data.

The proposed link size estimation approach measures the statistical parameters and then measures the rate of change in traffic. The heavy length distributions performed better than the Gaussian model [3].

Chydzinski [4] studies the effect of interdependence of packet arrival time and service delivery on the degradation of queuing system performance in network operation.

Kasse et al. [5] present an approach using the Path-scan model to identify paths that exhibit anomalous behavior in network communications. They developed a Markov model consisting of three states to model the states in the network.

Li and Wu used a Markov chain model to calculate the probability of agents forwarding packets in the idle or busy state and analyze the packet forwarding behavior of agents. On this basis, they proposed a game model for packet forwarding strategies [6].

Zhou and Nikaein [7] proposed an adaptive packet collection method to improve the reliability of random access. Random access [30] is not checked for each packet, but instead, random access is checked when the number of collected packets reaches a certain threshold. During packet collection, there may be delays or packet loss. To overcome this problem, a Markov model was used.

Sun and Zhao proposed a Markov chain method for modeling photovoltaic array noise and studied the accuracy of the modeling [8].

Doncel described the average information age of a system [24] using the stochastic hybrid system method. The study investigated the average information age of a queuing system formed by two queues, the first queue storing the system updates and the second one storing the energy required to deliver the state update to the monitor [34]. The queues process one packet [9].

Ott and Swanson study Markov chain models derived from protocols in the TCP (Transmission Control Protocol) paradigm. The researchers study the problems of packet loss, latency, and network congestion [10].

Even studied the convergence rate of Markov chains and optimized functions [23] under the least restrictive assumptions. He showed a theoretical lower bound for methods for sampling stochastic gradients along a Markov chain path and a dependence on the beat time of the Markov chain [11].

Smiesko et al. [12] used statistical parameters to detect DDoS attacks in real time. Statistical analysis methods such as kurtosis coefficient, entropy, correlation coefficient, and Hurst exponent were used to analyze network flow data [17].

Lu uses data transformation analysis to reduce risks in online shopping and ensure online security [34]. Opposition coefficient, correlation coefficient and other methods of statistical analysis were used [13].

Kenyon et al. [14] analyzed the payload of packets using entropy and found that entropy is an effective method for anomaly detection. They compared real-time entropy values with baseline values to detect anomalies in the flow [18].

In calculating the maximum entropy of network traffic, Gu et al. [15] proposed an effective method for detecting behavior-based anomalies, which is commonly applied to TCP and UDP (User Datagram Protocol) traffic data. The calculation used the protocol information and the sending port number features.

Doriguzzi-Corin et al. [16] propose to detect DDoS (Distributed Denial-of-Service) attacks by extracting packet features and exploiting the features.

In the analyzed research works, it was mainly studied on the basis of analysis based on packet data. However, the data in the packets were considered as the basis for analysis, but their interdependence was rarely analyzed. To solve this problem, the research study used a Markov chain in combination with statistical analysis methods, which allows for accurate evaluation of the results obtained.

III. Problem Statement

Developing effective methods for detecting and responding to threats by analyzing large networks has always been relevant [19]. One of the biggest challenges is to detect attacks that exhibit anomalous behavior in selected packet features of network traffic [20]. Detection and response methods must be adaptable to a wide variety of network environments, and these methods should be implemented without human intervention as much as possible [25], [27].

Attack detection should be as accurate as possible. Because a falsely detected attack can lead to a denial of service for legitimate users in network operations. If the detection results in the detection of attacks that are not occurring, the attack will not be detected, meaning the detection process will not work properly [26], [28]. The method used in the detection process should be effective and fast in relation to the various attacks that currently exist.

Research shows that statistical features and statistical processing are a useful approach for detecting anomalous intrusions.

Developing useful methods for detecting unexpected events that occur as a result of device failures or unforeseen events observed in network operations is an important task today. The main requirement for the developed methods is the ability to detect any type of unexpected situation in real time, including data that is unknown in advance and is also scattered over time. For this, the recorded network flow characteristics are used. When recording a network flow, the network flow is recorded in pcap format using the Wireshark program and the flow characteristics are extracted for analysis. The extracted characteristics are stored in csv file format. The flow characteristics are evaluated using statistical analysis methods and Markov chains.

IV. The Proposed Methodology

In the proposed research method, network traffic is recorded, and statistical analysis is performed based on the collected data. Fig. 1 depicts the network traffic data collection and analysis. First, the incoming network traffic data was recorded using the Wireshark program in a laboratory setting. The recorded data is saved in the pacp file format. Second, the data stored in the pcap

file is read based on the developed algorithm, 89 different features are extracted from the read data and saved in the csv file format. This process is carried out in three stages. Reading, feature extraction, and saving the features in the csv file format. The data in the csv file is processed and checked using statistical analysis algorithms [1]. A Markov chain is formed using the results of statistical analysis. The results of statistical analysis are important in determining the set of Markov chain states.

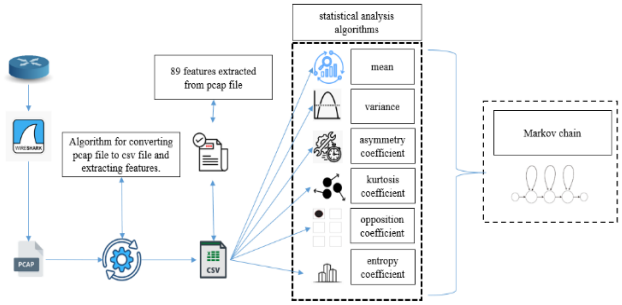


Fig. 1. Statistical analysis structure of network flow feature extraction.

A. Extracting Network Flow Characteristics

Network traffic was captured in a laboratory environment using Wireshark. During the recording, data was transmitted through 6 computers and various network activity conditions were captured. Fig. 1 shows a model of the network traffic capture process. Eighty-nine features were extracted from the traffic data recorded in a laboratory environment. The extracted features allow us to solve the research problem.

In the algorithm in Fig. 2, the data recorded in the pcap file is read through Wireshark and the initial properties of the packet are extracted: *srcIP*, *dstIP*, *srcPort*, *dstPort*, *protocol*, *timestamp*, *headBytes*, *payloadBytes*, *flags*, *windowSize*, *payload*. These properties are the main information for forming the flow data. Based on this information, eighty-nine different properties are determined.

```
1: function nextPacket()
2: packetInfo ← None
3: while pcapPtr < pcapLen do ▷ Packet reading continues
4:    totGen.nextId() ▷ Create a new ID for each package
5:    timeHigh ← pcapData[pcapPtr:pcapPtr + 4] ▷ High timestamp
6:    timeHigh ← struct.unpack(typeI, timeHigh)[0] ▷ Unpacking time
7:    timeLow ← pcapData[pcapPtr + 4:pcapPtr + 8] ▷ Lower timestamp
8:    timeLow ← struct.unpack(typeI, timeLow)[0] ▷ Unpacking time
9:    timeStamp ← 1000000 * timeHigh + timeLow ▷ Timestamp calculation
10:   caplen ← pcapData[pcapPtr + 8:pcapPtr + 12] ▷ Package length
11:   caplen ← struct.unpack(typeI, caplen)[0] ▷ Unpacking the packet length
12:   pcapPtr += 16 ▷ Go to package information
13:   packetData ← pcapData[pcapPtr:pcapPtr + caplen] ▷ Get package information
14:   pcapPtr += caplen ▷ Move the pointer to the next packet.
15:   if isIpv4TCP(packetData) then
16:      packetInfo ← getIpv4Info(packetData, timeStamp) ▷ IPv4 TCP packet
17:      break
18:   else if isIpv6TCP(packetData) then
19:      packetInfo ← getIpv6Info(packetData, timeStamp) ▷ IPv6 TCP packet
20:      break
21:   end if
22: end while
23: return packetInfo ▷ Return packet data
24: end function
```

Fig. 2. Pseudocode for reading packet data from a pcap file.

In Fig. 3, timestamps and flags from each packet of data are used to create a stream from the data. When the stream is complete, the data is written to a csv file.

```
1: function addPacket(packet: BasicPacketInfo)
2:   if packet == None then
3:     return
4:   end if
5:   currentTS ← packet.getTimeStamp() ▷ Packet timestamp
6:   pktFwdFlowId ← packet.getFwdFlowId() ▷ Packet transmission flow identifier
7:   pktBwdFlowId ← packet.getBwdFlowId() ▷ Packet reverse flow identifier
8:   if pktFwdFlowId in self.currentFlows or pktBwdFlowId in self.currentFlows then
9:     if pktFwdFlowId in self.currentFlows then
10:       flowId ← pktFwdFlowId ▷ Flow ID
.    .    .    .    .    .    .    .    .    .    .    .
.    .    .    .    .    .    .    .    .    .    .    .
61: function dumpPayloadToCSV()
62:   open "payload.csv" for writing as csvfile
63:   writer ← csv.writer(csvfile) ▷ Opening and writing a CSV file
64:   for each flow in self.finishedFlows do
65:     flow ← flowType(flow) ▷ Getting the flow
66:     output ← flow.generateFlowFeatures() ▷ Getting the features
67:     writer.writerow(output) ▷ Writing features to CSV
68:   end for
69: end function
```

Fig. 3. Pseudocode for capturing stream data based on packet data from a pcap file.

In Fig. 4, flag features are also extracted to perform the calculations required for the research results. This data is written to a csv file after the data in the pcap file is completely read using the dumpFeatureToCSV function [29]. This process is performed each time a pcap file is read, and the data from each pcap file is saved to a separate csv file.

```
1: function init(generator, srcIP, dstIP, srcPort, dstPort, protocol, timeStamp, headBytes, payloadBytes, flags,
TCPWindow, payload)
2: id ← generator.nextId() ▷ Create a unique ID for the package
3: srcIP ← srcIP ▷ Setting the source IP address
4: dstIP ← dstIP ▷ Setting the destination IP address
5: srcPort ← srcPort ▷ Setting the source port
6: dstPort ← dstPort ▷ Setting the destination port
7: protocol ← protocol ▷ Protocol setup
8: timeStamp ← timeStamp ▷ Set timestamp
9: headBytes ← headBytes ▷ Setting the packet header length
10: payloadBytes ← payloadBytes ▷ Load length setting
.    .    .    .    .    .    .    .    .
.    .    .    .    .    .    .    .    .
80: return (flags >> 5) & 1 ▷ Check the URG flag
81: end function
82: function hasFlagECE()
83: return (flags >> 6) & 1 ▷ Check the ECE flag
84: end function
85: function hasFlagCWR()
86: return (flags >> 7) & 1 ▷ Check the CWR flag
87: end function
```

Fig. 4. Pseudocode for extracting features needed in a calculation.

B. Mathematical Representation of Markov Chain

Markov chains were first used by Andrey Markov in 1906 to show that the law of large numbers does not require that random variables be independent. Since then, Markov chains have been used in biology, finance, computer science, mathematics, and other fields [5, 21]. A Markov chain is a model that helps us visualize a sequence of random variables, states, and the states in which each of them can take on values from a given set. Markov chains can also be used to make predictions, predict future events, or understand a process that has occurred [22].

Consider a simple Markov chain with states and transition matrix $\|p_{ij}\|_1^n$. Given $r_1, \dots, r_n$, the terminal probabilities of the chain have the property of termination. The set of all states $\{n\}$ of the chain is decomposed into two disjoint subsets, that is, $\{n\} = \{n_1\} \cup \{n_2\}$. The transition from the initial set to the next one through $R_{1,2}$ and $R_{2,1}$ and vice versa from the final to the initial state is described by the Formula (1) below:

$$R_{1,2} = \sum_{\substack{i\in\{n_1\} \\ j\in\{n_2\}}} r_i p_{ij}\,,\ R_{2,1} = \sum_{\substack{i\in\{n_2\} \\ j\in\{n_1\}}} r_i p_{ij} \qquad (1)$$

Lemma. $R_{1,2} = R_{2,1}$ probability flows are balanced [36].

In [4], in particular, all the states in the first subset $\{n_1\}$ are numbered $1, \dots, k$, and all the positions of the Markov chain are numbered according to the chosen decomposition as $\{n_1\}, \{n_2\}$ (where k is the number of items in this subset). The items in the complement of the set $\{n\}$ belonging to the set $\{n_2\}$ are numbered $k+1, \dots, n$.

The final probabilities $r_1, \dots, r_n$ of the complete chain give rise to the relationship [Formula (2)]:

$$r_i = \sum_{j=1}^{n} r_j p_{ji}, all\ i = 1, \dots, n) \quad (2)$$

the i-number position is extracted in the following Formula (3):

$$-r_i(1 - p_{ii}) = \sum_{\substack{j=1 \\ j \neq i}}^{n} r_j p_{ji} = 0, i = 1, \dots, n) \quad (3)$$

The stochastic transition matrix $\|p_{ij}\|_1^n$ is Formula (4):

$$\sum_{j=1}^{n} p_{ij} = 1, i = 1, \dots, n \quad (4)$$

$(1 - p_{ii})$ expression is expanded as in Formula (5):

$$-r_i \sum_{\substack{j=1 \\ j \neq i}}^{n} p_{ji} + \sum_{\substack{j=1 \\ j \neq i}}^{n} r_j p_{ji} = 0, i = 1, \dots, n) \quad (5)$$

When passing to probability values, the first k of the above formula is added [Formula (6)]:

$$-\sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} r_i p_{ij} - \sum_{i=1}^{k} \sum_{j=k+1}^{n} r_j p_{ij} + \\ + \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} r_i p_{ji} + \sum_{i=1}^{k} \sum_{j=k+1}^{n} r_j p_{ji} = 0 \quad (6)$$

The first term of the above Formula (6) is equal to the third term [Formula (7)].

$$\sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} r_i p_{ij} = \sum_{i=1}^{k} \sum_{j=1}^{k} r_j p_{ij} - \sum_{i=1}^{k} r_j p_{ii} = \\ = \sum_{j=1}^{k} \sum_{i=1}^{k} r_j p_{ji} - \sum_{i=1}^{k} r_j p_{ii} = \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} r_j p_{ji} \quad (7)$$

(Here the sum indices are replaced equivalently by $i \leftrightarrow j$, and the order of the sums is changed.)

Using the above formulas, a generalized formula is obtained as in Formula (8):

$$-\sum_{i=1}^{k} \sum_{j=k+1}^{k} r_i p_{ij} + \sum_{i=1}^{k} \sum_{j=k+1}^{n} r_j p_{ij} = 0 \quad (8)$$

A sequence of random variables $X_0, X_1, X_2, \dots$ is called a Markov chain if the values of the variables are taken in the state space $\{1,2, \dots N\}$ and there exists a matrix $P = (p_{ij})$ of size $N \times N$ for any $n \geq 0$. $X_0, X_1, X_2, \dots$ denotes the set of states [Formula (9)].

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \\ = P(X_{n+1} = j | X_n = i) = p_{ij} \quad (9)$$

P is called the transition matrix of the matrix chain, and P_{ij} is the probability of transition from i to j.

The transition probability $p_{ij} = P(X_{n+1} = j | X_n = i)$ is written in the form of a matrix of p_{ij} and is as follows [Formula (10)]:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdot & \cdot & \cdot & p_{1n} \\ p_{21} & p_{22} & \cdot & \cdot & \cdot & p_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ p_{n1} & p_{n2} & \cdot & \cdot & \cdot & p_{nn} \end{bmatrix} \quad (10)$$

The sum of the transition probabilities $p_{11}, p_{12}, \dots, p_{1n}$ in each row must be equal to $p_{11} + p_{12} + \dots + p_{1n} = 1$ [Formula (11)].

$$\sum_{j=1}^{N} p_{ij} = \sum_{j=1}^{N} P(X_{n+1} = j | X_n = i) = 1 \quad (11)$$

where, the sum of all the possibilities in the row together is 100%.

So, if $X_0, X_1, X_2, \dots$ forms a Markov chain, then the transition probability of a matrix P of size $N \times N$ at step t is equal to $P(X_t = j | X_0 = i) = (P^t)_{ij}$.

If $X_0, X_1, X_2, \dots$ forms a Markov chain [11], then the probability distribution of X_0 of a matrix P of size $N \times N$ is equal to 1, and the probability distribution of a row vector π^T of size N is expressed as $\pi^T P^t$ [Formula (12)].

$$X_0 = \pi^T \Rightarrow X_t = \pi^T P^t \quad (12)$$

Stationary probability as in Formula (13):

$$\pi P = \pi \quad (13)$$

The condition $\sum \pi_i = 1$ must also be met, which indicates the probability of the process remaining in a given state for how long.

C. Formulation of Statistical Features of Network Flow with Unexpected Situation

Statistical features are used to detect unexpected situations in network traffic. The following statistical features are calculated for network traffic during the analysis:

- Mean as in Formula (14):

$$m_i = \frac{1}{n} \sum_{j=i}^{i+n} S_j \quad (14)$$

where, m_i is the mean value, n is the number of packets in the selected time interval, $S_i - j$ is the sample value of the traffic intensity at time, and the sum of values from $j = i$ to $i + n$ is obtained [1].

- Variance as in Formula (15):

$$D = \frac{1}{n-1} \sum_{j=i}^{i+n} \left(S_j - m_j\right)^2 \quad (15)$$

where, D is the variance, n is the length of the window under study (how many values were obtained), $S_i - j$ is the intensity at time (number of packets, traffic volume), m_j is the mean value [2].

- asymmetry coefficient as in Formula (16):

$$K_a = \frac{\frac{1}{n}\sum_{j=i}^{i+n}(S_j - m_j)^3}{D^3} \quad (16)$$

where, $S_i - j$ is the sample value at time (e.g., the number of packets), m_i is the mean value, D is the variance, and n is the number of values. The asymmetry coefficient represents the shift in the center of gravity of the statistical distribution [4].

- kurtosis coefficient as in Formula (17):

$$K_k = \frac{\frac{1}{n}\sum_{j=i}^{i+n}(S_j - m_j)^4}{D^4} - 3 \quad (17)$$

where, $S_i - j$ is the value at time, m_i is the mean, D is the variance, n is the number of values. The kurtosis or asymmetry coefficient measures the height or sharpness of the peak of a statistical distribution. In other words, it tells how far the values in the distribution are not far from or clustered around the mean [12].

- opposition coefficient as in Formula (18):

$$K_{oc} = \frac{1}{\sqrt{\eta}} \quad (18)$$

η is the excess parameter, defined as Formula (19):

$$\eta = K_k = \frac{\mu_4}{\sigma^4} \quad (19)$$

where, σ is the standard deviation (i.e. the square root of the variance), μ_4 is the fourth central moment (i.e. the average of the 4th order deviations of all values from the mean). The parameter K_{oc} varies from 0 (according to the Cauchy distribution) to 1 (discrete binomial distribution). For a normal distribution, $K_{oc} = \frac{1}{\sqrt{3}} \approx 0.577$.

Another parameter that characterizes the shape of the distribution is the entropy coefficient.

- entropy coefficient as in Formula (20):

$$K_{en} = \frac{\Delta_E}{\sigma} \quad (20)$$

where, Δ_E is the entropy value of the error, which means the level of uncertainty of the distribution, σ means the standard deviation, that is, the degree of deviation from the mean value [Formula (21)].

$$\Delta_E = \frac{1}{2} e^{H(\frac{X}{X_n})} \quad (21)$$

where, $H(\frac{X}{X_n})$ is the entropy (uncertainty) change introduced by measuring the random variable X, and e is a mathematical constant. The entropic error shows that the greater the uncertainty in the distribution, the greater the entropic error [Formula (22)].

$$H\left(\frac{X}{X_n}\right) = \int_{+\infty}^{-\infty} p(X) \ln\big(p(X)\big) dX \quad (22)$$

where, $p(X)$ is the probability density of the random variable [14]. Table I presents the entropic error distribution type.

TABLE I. ENTROPY COEFFICIENT DISTRIBUTION

Distribution type	K_{en} value	Note
Normal distribution	$K_{en} = \frac{\sqrt{2\pi e}}{2} \approx 2.066$	Maximum entropy
Uniform distribution	$K_{en} = \sqrt{3} \approx 1.73$	Average entropy
Cauchy distribution and discrete binomial distribution	$K_{en} = 0$	Very low entropy

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

Statistical analysis methods are used to record network traffic in a laboratory environment and identify unexpected events.

For verification, the recorded data was recorded at different times. The files in csv format were merged. Statistical analysis is performed based on this data.

A. Using Statistical Analysis Method

Fig. 5 shows the sample average value graph. In it, the sample average value is calculated based on Formula (14), $m_i = 1646.5714$, and according to the obtained result, the packet size in 9 flows was higher than the mean. This means that an unexpected situation occurred in the network flow.

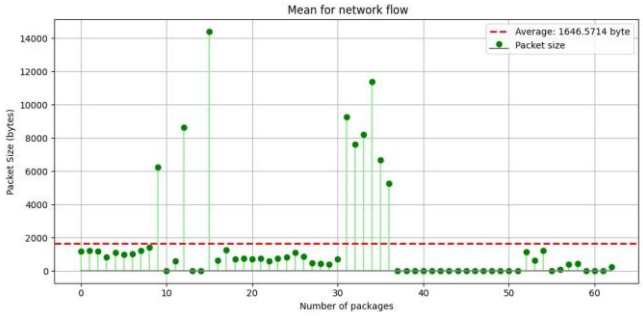


Fig. 5. Mean value graph of network flow.

In Fig. 6, the variance value was calculated based on Formula (15) and was equal to $D = 9474632.51^2$. In this statistical analysis, the square root of the variance of 9 network flows did not fall within the standard deviation range.

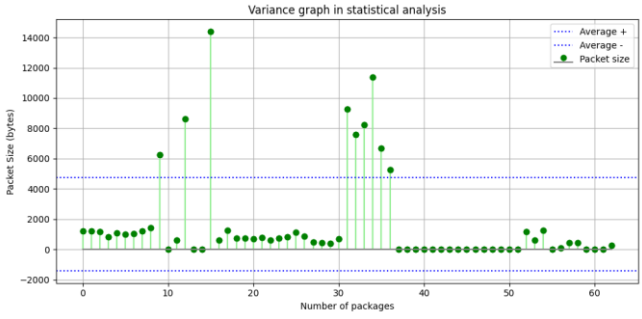


Fig. 6. Variance value graph of network flow.

The result of the asymmetry coefficient is presented in Fig. 7. The asymmetry coefficient is effective in determining how the data flow in the network is distributed, in which cases the packet size is small or large. This information is very important for

network optimization or security verification. The asymmetry coefficient helps to determine the presence of large packets in the network, which may indicate network congestion or problems. According to Formula (16), the asymmetry coefficient is equal to $K_a = 2.463632$. If the asymmetry coefficient is positive, this indicates a large number of large packets in the network and some uncertainties.

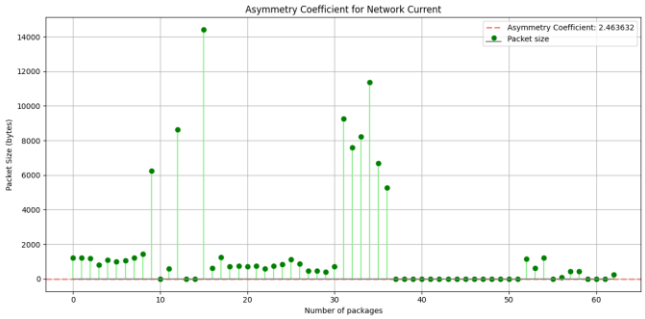


Fig. 7. Graph of the asymmetry coefficient of network flow.

The statistical distribution of the kurtosis coefficient can take very small or very large values. These results are useful in identifying unexpected situations in the network. In Fig. 8, the kurtosis coefficient is equal to $K_k = 8.352490$ according to Formula (17), which means that the value of the kurtosis coefficient is large. This indicates that there is an unexpected situation in the network.

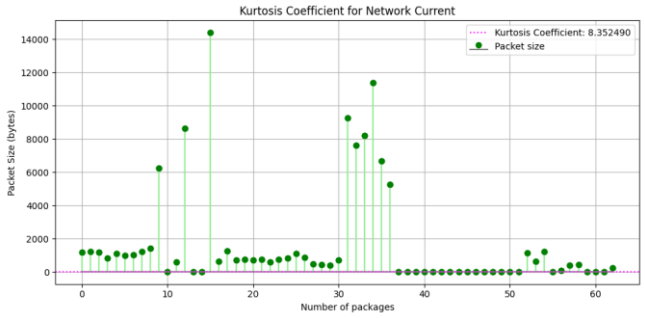


Fig. 8. Kurtosis coefficient graph of network flow.

The opposition coefficient is an indicator that provides deep statistical information about the symmetry and degree of deviation of the distribution in the network flow. The opposition coefficient is a very effective tool for identifying states in the network. The states that can be identified using K_{oc} are listed in Table II.

TABLE II. OPPOSITION COEFFICIENT CASES

Opposition coefficient	Distribution type	Note
$K_{oc} \approx 1.0$	Discrete, two-valued	Repeated and binary stream, 0 and 1500 byte packets, packets sent by test process or Botnet
$K_{oc} \approx 0.577$	Normal (Gaussian)	Stable network packet sizes are normally distributed
$K_{oc} < 0.4$	Heavy-tail	There are unexpected flows. For example, DoS or DDoS attacks, Botnets, etc.
$K_{oc} \approx 0$	Widely deviant distribution	Extremely dangerous, anomalous or exploitable condition

According to the result of the statistical analysis by recording the network current, the opposition coefficient is equal to $K_{oc} = 0.3516$. This indicates that there are unexpected situations in the network flow. In general, the combined use of statistical analysis results provides more accurate results. Fig. 9 presents the opposition coefficient result in graphical form.

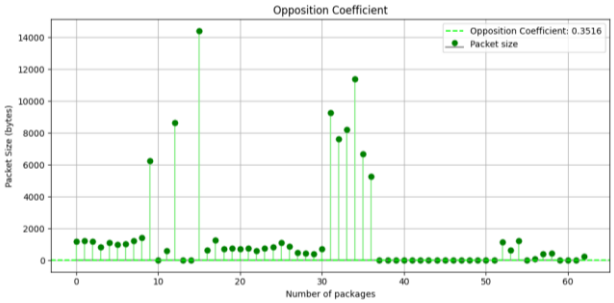


Fig. 9. Opposition coefficient in statistical analysis of network flow.

The entropy coefficient can be used to distinguish between normal and unexpected situations in a network. The entropy coefficient is an important statistical indicator for traffic analysis, attack detection, and security monitoring. As a result of the statistical analysis in Fig. 10, $K_{en} = 1.6231$. It indicates that there are different packets in the network flow, and there is an unexpected situation.

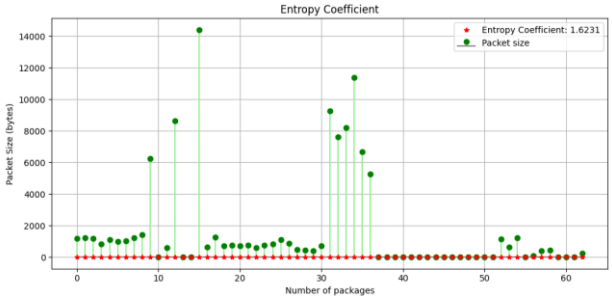


Fig. 10. Entropy coefficient in statistical analysis of network flow.

The analysis of the presented characteristics shows that the mean, variance, asymmetry coefficient, kurtosis coefficient, opposition coefficient, and entropy coefficient indicate the presence of unexpected situations in the network flow. Fig. 11 shows an overview of the research results.

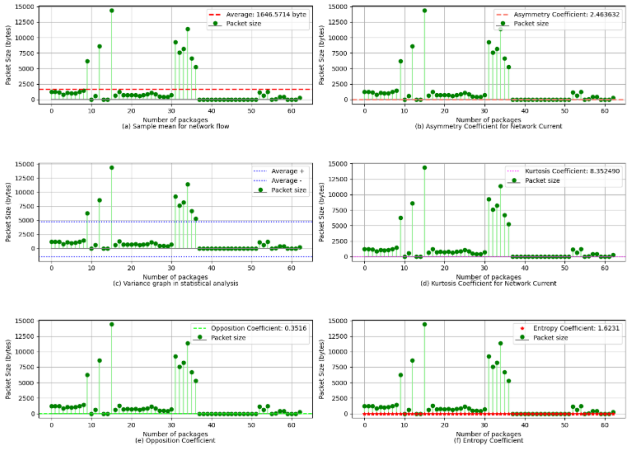


Fig. 11. General graph of statistical analysis results.

B. Markov Chain Construction Based on Statistical Analysis Results

Based on the results, statistical analysis methods are used to identify unexpected situations. The analysis of the obtained statistical characteristics shows that when an unexpected situation occurs in the network flow, the results of the mean and variance distinguish the unexpected situation, and the detection result may have an informative nature.

The values of the asymmetry coefficients and the kurtosis coefficient were large, and the values of the opposition coefficient and the entropy coefficient were low. According to the results of the study, these indicators are informative, since an unexpected situation occurs in the network flow. Table III shows the values of all statistical analysis results. The results obtained in the construction of the set of states of the Markov chain are the main value.

TABLE III. STATISTICAL ANALYSIS OF UNEXPECTED SITUATIONS IN NETWORK FLOW

№	Statistical analysis	The value of statistical analysis
1	Mean	$m_i = 1646.5714$
2	Variance	$D = 9474632.51^2$
3	Asymmetry coefficient	$K_a = 2.463632$
4	Kurtosis coefficient	$K_k = 8.352490$
5	Opposition coefficient	$K_{oc} = 0.3516$
6	Entropy coefficient	$K_{en} = 1.6231$

The Markov chain is evaluated based on the obtained statistical results. In the evaluation based on the Markov chain, based on the statistical analysis result, the first set of cases $X_0, X_1, X_2, \ldots$ is constructed, and the set of cases is presented in Table IV.

TABLE IV. SET OF STATES FOR A MARKOV CHAIN

№	Status name	Statistical analysis method
1	low load	variance, mean
2	average load	entropy coefficient, variance
3	high load	asymmetry coefficient, kurtosis coefficient

These three states were considered as states of the Markov chain, and based on the information obtained from the statistical analysis, the transition probability was calculated and a transition matrix was constructed according to Formula (1).

$$P = \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.2 & 0.4 & 0.4 \\ 0.1 & 0.2 & 0.7 \end{bmatrix}$$

The stationary probability was calculated using Formula (13). It was equal to $\pi = [0.1852 \quad 0.2963 \quad 0.5185]$.

Based on the transition probability matrix, Fig. 12 shows a graph of a Markov chain for network flow analysis. The graph shows the relationship between the three states.

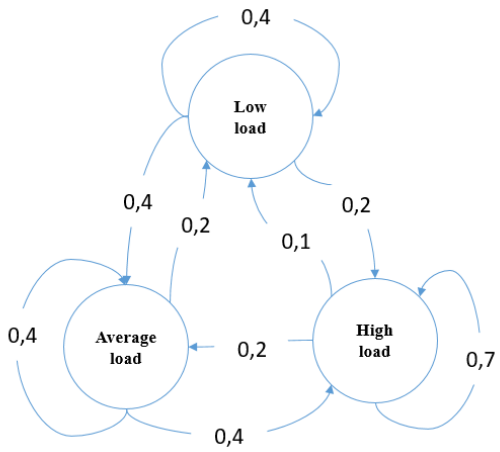


Fig. 12. Markov chain of network flow analysis.

VI. CONCLUSION

In solving the problem of detecting unexpected situations in network traffic, the use of statistical methods such as mean, variance, asymmetry coefficient, kurtosis coefficient, opposition coefficient and entropy coefficient supports the use of increasing efficiency and real-time operation. In detecting unexpected situations, packet features were extracted from the network traffic recorded in real time under laboratory conditions. Statistical analysis was performed using the extracted features. Markov chain is the most effective tool for analyzing situations and events occurring in network traffic and Markov chain was formed using the results of statistical analysis and Markov chain was constructed.

As a result, the low load stationary probability was 18.52%. This situation is rare in the system. This means that the network is in a low load state for a short time. The results of the statistical analysis correspond to the variance and mean, respectively. The average load stationary probability was 29.63%. This situation occurs moderately in the network flow. This means that there are changes, and the average load is frequent in the network. The high load stationary probability was 51.85%. The network flow time is more in this state. Therefore, the results of the statistical analysis are fully consistent with the results of the mean, variance, asymmetry coefficient and kurtosis coefficient. The Markov chain representing the probabilities of network states gave the same result as the statistical analysis.

In future work, using Markov chains in conjunction with statistical analysis methods will be effective. Markov chains are effective in showing the interaction between states when moving from one state to another. When used in conjunction with statistical analysis, the results of evaluating the object being analyzed will be more accurate.

REFERENCES

[1] Ramachandran K, Tsokos C. Mathematical Statistics with Applications in R, 2nd Edition. Elsevier, Inc.; 2015.

[2] Hu, C., & Liangping, H. (2022). How to use analysis of variance correctly—an analysis of variance for the univariate quantitative data collected from the orthogonal design. Sichuan Mental Health, 35(3), 201-206. doi:https://doi.org/10.11886/scjsws20220510004

[3] Alasmar, M., & Zakhleniuk, N. (2017). Network link dimensioning based on statistical analysis and modeling of real internet traffic. Ithaca: Retrieved from https://www.proquest.com/working-papers/network-link-dimensioning-based-on-statistical/docview/2076751109/se-2

[4] Chydzinski, A. (2025). Asymmetry between queues with correlated service and correlated arrivals. Symmetry, 17(1), 106. doi:https://doi.org/10.3390/sym17010106

[5] Kasse, M., Charrier, R., Berred, A., Bertelle, C., & Delpierre, C. (2024). An approach for anomaly detection in network communications using k-path analysis. Journal of Cybersecurity and Privacy, 4(3), 449. doi:https://doi.org/10.3390/jcp4030022

[6] Li, Y., & Wu, X. (2018). Packet forwarding strategies in multiagent systems: An evolutionary game approach. Wireless Communications & Mobile Computing (Online), 2018 doi:https://doi.org/10.1155/2018/4965343

[7] Zhou, K., & Nikaein, N. (2016). Random access with adaptive packet aggregation in LTE/LTE-A. EURASIP Journal on Wireless Communications and Networking, 2016, 1-15. doi:https://doi.org/10.1186/s13638-016-0531-6

[8] Sun, F., & Zhao, C. (2019). Research and modeling of photovoltaic array channel noise characteristics. Energies, 12(7), 1257. doi:https://doi.org/10.3390/en12071257

[9] Doncel, J. (2021). Age of information of a server with energy requirements. PeerJ Computer Science, doi:https://doi.org/10.7717/peerj-cs.354

[10] Ott, T. J., & Swanson, J. (2006). Asymptotic behavior of a generalized TCP congestion avoidance algorithm. Ithaca: Retrieved from https://www.proquest.com/working-papers/asymptotic-behavior-generalized-tcp-congestion/docview/2091799438/se-2

[11] Even, M. (2023). Stochastic gradient descent under markovian sampling schemes. Ithaca: Retrieved from https://www.proquest.com/working-papers/stochastic-gradient-descent-under-markovian/docview/2781016529/se-2

[12] Smiesko, J., Segec, P., & Kontsek, M. (2024). Machine recognition of DDoS attacks using statistical parameters. Mathematics, 12(1), 142. doi:https://doi.org/10.3390/math12010142

[13] Lu, F. (2024). Online shopping consumer perception analysis and future network security service technology using logistic regression model. PeerJ Computer Science, doi:https://doi.org/10.7717/peerj-cs.1777

[14] Kenyon, A., Deka, L., & Elizondo, D. (2024). Characterising payload entropy in packet Flows—Baseline entropy analysis for network anomaly detection. Future Internet, 16(12), 470. doi:https://doi.org/10.3390/fi16120470

[15] Gu, Y.; McCallum, A.; Towsley, D. Detecting anomalies in network traffic using maximum entropy estimation. In Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement, Berkeley, CA, USA, 19–21 October 2005; p. 32

[16] Doriguzzi-Corin, R., Luis Augusto, D. K., Mendozzi, L., Siracusa, D., & Savi, M. (2024). Introducing packet-level analysis in programmable data planes to advance network intrusion detection. Ithaca: Retrieved from https://www.proquest.com/working-papers/introducing-packet-level-analysis-programmable/docview/2836658859/se-2

[17] Abdulganiyu, O.H.; Ait Tchakoucht, T.; Saheed, Y.K. A systematic literature review for network intrusion detection system (IDS). Int. J. Inf. Secur. 2023, 22, 1125–1162. [CrossRef]

[18] Adrian-Tiberiu Costin, Zinca, D., & Dobrota, V. (2021). LAN traffic capture applications using the libtins library. Electronics, 10(24), 3084. doi:https://doi.org/10.3390/electronics10243084

[19] Taramit, H., José Jaime Camacho-Escoto, Gomez, J., Orozco-Barbosa, L., & Haqiq, A. (2022). Accurate analytical model and evaluation of wi-fi halow based IoT networks under a rayleigh-fading channel with capture. Mathematics, 10(6), 952. doi:https://doi.org/10.3390/math10060952

[20] Sembiring, I., Suharyadi, Iriani, A., Jenni Veronika, B. G., & Jusia, A. G. (2023). A novel approach to network forensic analysis: Combining packet capture data and social network analysis. International Journal of Advanced Computer Science and Applications, 14(3) doi:https://doi.org/10.14569/IJACSA.2023.0140353

[21] Kabulov, A.; Baizhumanov, A.; Saymanov, I. Synthesis of Optimal Correction Functions in the Class of Disjunctive Normal Forms. Mathematics 2024, 12, 2120. https://doi.org/10.3390/math12132120

[22] Kabulov, A.; Normatov, I.; Saymanov, I.; Baizhumanov, A. On the Completeness of Classes of Correcting Functions of Heuristic Algorithms, Azerbaijan Journal of Mathematics, 2025, vol 15, no. 2. (accepted)

[23] Kabulov, A., Baizhumanov, A., Berdimurodov, M. (2024). On the minimization of k-valued logic functions in the class of disjunctive normal forms. Journal of Mathematics, Mechanics and Computer Science, 121(1), 37–45. https://doi.org/10.26577/JMMCS202412114

[24] Kabulov, A.; Saymanov, I.; Babadjanov, A.; Babadzhanov, A. Algebraic Recognition Approach in IoT Ecosystem. Mathematics 2024, 12, 1086. https://doi.org/10.3390/math12071086

[25] Saymanov, I. . (2024). Logical automatic implementation of steganographic coding algorithms . Journal of Mathematics, Mechanics and Computer Science, 121(1), 122–131. https://doi.org/10.26577/JMMCS2024121112

[26] Islambek Saymanov, et al. Numerical Methods of Synthesis of a Correct Algorithm for Solving Recognition Problems. Advances in Artificial Intelligence and Machine Learning. 2025;5(1):202. https://dx.doi.org/10.54364/AAIML.2025.51202

[27] A. Kabulov, I. Normatov, E. Urunbaev and F. Muhammadiev, "Invariant Continuation of Discrete Multi-Valued Functions and Their Implementation," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 2021, pp. 1-6, doi: 10.1109/IEMTRONICS52119.2021.9422486.

[28] A. Kabulov, I. Yarashov and A. Otakhonov, "Algorithmic Analysis of the System Based on the Functioning Table and Information Security," 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 2022, pp. 1-5, doi: 10.1109/IEMTRONICS55184.2022.9795746.

[29] Kabulov, A. V., Normatov, I. H. About problems of decoding and searching for the maximum upper zero of discrete monotone functions. Journal of Physics: Conference Series, 1260(10), 102006, 2019. doi:10.1088/1742-6596/1260/10/102006.

[30] A. Kabulov, I. Saymanov, I. Yarashov and A. Karimov, "Using Algorithmic Modeling to Control User Access Based on Functioning Table," 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 2022, pp. 1-5, doi: 10.1109/IEMTRONICS55184.2022.9795850.

[31] Kabulov, A. V., Normatov, I. H., Ashurov A.O. Computational methods of minimization of multiple functions. Journal of Physics: Conference Series, 1260(10), 10200, 2019. doi:10.1088/1742-6596/1260/10/102007.

[32] A. Kabulov, I. Normatov, A. Seytov and A. Kudaybergenov, "Optimal Management of Water Resources in Large Main Canals with Cascade Pumping Stations," 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Vancouver, BC, Canada, 2020, pp. 1-4, doi: 10.1109/IEMTRONICS51293.2020.9216402.

[33] E. Navruzov and A. Kabulov, "Detection and analysis types of DDoS attack," 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 2022, pp. 1-7, doi: 10.1109/IEMTRONICS55184.2022.9795729.

[34] A. Kabulov, I. Saymanov, I. Yarashov and F. Muxammadiev, "Algorithmic method of security of the Internet of Things based on steganographic coding," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 2021, pp. 1-5, doi: 10.1109/IEMTRONICS52119.2021.9422588.

[35] A. Babadzhanov, E. Urunbaev and I. Saymanov, "Problem of Synthesis of Minimal Forms of Logical Functions," 2022 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2022, pp. 1-5, doi: 10.1109/ICISCT55600.2022.10146903.

[36] Stefaniuk, V. L. (2011). Determinirovannye tsepi Markov. Information processes, 11(4), 423-427

A Steel Surface Defect Detection Method Based on Lightweight Convolution Optimization

Cong Chen[1], Ming Chen[2], Hoileong Lee[3], Yan Li[4], and Jiyang YU[5]*
School of Marine Information Engineering, Hainan Tropical Ocean University, Sanya 572022, China[1]
School of Information and Intelligent Engineering, University of Sanya, Sanya 572022, China[2]
Faculty of Electronic Engineering & Technology, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia[3]
Thalesgroup, Ottawa, ON, K1K 4Z9, Canada[4]
Graduate School of Management of Technology, Pukyong National University, Busan 48547, Korea[5]

***Abstract*—Surface defect detection of steel, especially the recognition of multi-scale defects, has always been a major challenge in industrial manufacturing. Steel surfaces not only have defects of various sizes and shapes, which limit the accuracy of traditional image processing and detection methods in complex environments. However, traditional defect detection methods face issues of insufficient accuracy and high miss-detection rates when dealing with small target defects. To address this issue, this study proposes a detection framework based on deep learning, specifically YOLOv9s, combined with the C3Ghost module, SCConv module, and CARAFE upsampling operator, to improve detection accuracy and model performance. First, the SCConv module is used to reduce feature redundancy and optimize feature representation by reconstructing the spatial and channel dimensions. Second, the C3Ghost module is introduced to enhance the model's feature extraction ability by reducing redundant computations and parameter volume, thereby improving model efficiency. Finally, the CARAFE upsampling operator, which can more finely reorganize feature maps in a content-aware manner, optimizes the upsampling process and ensures detailed restoration of high-resolution defect regions. Experimental results demonstrate that the proposed model achieves higher accuracy and robustness in steel surface defect detection tasks compared to other methods, effectively addressing defect detection problems.**

***Keywords*—YOLOv9s; steel surface defect detection; C3Ghost module; SCConv module; CARAFE upsampling operator**

I. INTRODUCTION

The surface quality of steel is one of the key indicators for measuring its performance and reliability, and it is widely applied in high-precision fields such as aerospace, automotive manufacturing, and building structures. Surface defects not only compromise the aesthetic quality of steel but also jeopardize its structural integrity, potentially causing safety hazards. Therefore, steel surface defect detection has become an important task in modern manufacturing. Traditional defect detection methods, such as manual visual inspection, ultrasonic testing, and optical scanning, although still of certain value in specific situations, face issues such as low efficiency, poor accuracy, and difficulty in identifying complex defects due to their reliance on manual operation or being limited to specific detection conditions [1]. In particular, in large-scale production environments, manual inspection cannot meet the requirements for high speed, high precision, and high reliability [2].

In recent years, with the rapid development of deep learning technology, particularly breakthroughs in the field of computer vision, automated defect detection technology has seen significant improvements. Deep learning algorithms, especially Convolutional Neural Networks (CNNs), have demonstrated powerful capabilities in image recognition, enabling efficient and accurate identification and classification of various defects on steel surfaces. Compared to traditional methods, deep learning-based detection methods not only improve detection efficiency but also overcome the limitations of manual inspection, adapting to the complex and variable production environment. In particular, the YOLO (You Only Look Once) [3-6] object detection algorithm, due to its efficient real-time performance and accuracy, has been widely adopted for steel defect detection.

Many researchers have made significant contributions to the field of defect detection. For instance, Gao et al. [7] integrated the attention mechanism and weighted bi-directional feature pyramid network (BiFPN) into the YOLOv5 architecture, achieving good results. However, there are still certain shortcomings when handling small-sized defects. Yu et al. [8] proposed the introduction of structural reparameterization, context transformation modules, and simplified generalized feature pyramid networks to improve the model's accuracy. Although they performed excellently in terms of mAP, complex background noise affected defect detection performance, indicating that there is still room for improvement in the model's robustness to interference. Yang et al. [9] proposed a detection method combined with a supervised spatial attention module (SSAM) to improve the detection accuracy of defects such as surface cracks and rolled-in scale on steel. However, this method increases the computational complexity of the model and raises hardware requirements. Subburaj et al. [10] proposed the DBCW-YOLO model, which integrates attention mechanisms and enhanced feature extraction techniques to improve defect detection accuracy. However, integrating multiple advanced technologies increases the model's complexity, affecting real-time performance. Wu et al. [11] proposed the Hyper-YOLO model, which improves detection performance by replacing the CSP module in YOLOv5 with the Ghost module, the PAFPN module with the Hyper FPN module, and introducing α-CIoU and α-DIoU loss functions. However, these improvements increase the model's computational cost. Xu et al. [12] improved the YOLOv5 algorithm by introducing a coordinate attention

mechanism and multi-scale feature fusion, significantly improving the model's detection accuracy. However, these improvements increase the model's complexity, leading to longer training and inference times. Kong et al. [13] proposed an improved steel surface defect detection algorithm based on YOLOv8, enhancing small-target detection accuracy by introducing a non-attention mechanism and improving the SPPF module. However, the algorithm requires substantial computational resources, limiting its deployment in resource-constrained environments. Mi et al. [14] proposed a surface defect detection method for hot-rolled steel strips based on multi-scale feature perception and adaptive feature fusion, effectively improving the model's detection performance. However, it performs poorly when handling highly complex or noisy images and requires longer training times.

In summary, although significant progress has been made in existing steel surface defect detection algorithms, there are still numerous challenges and areas for optimization, especially regarding the detection accuracy of small-sized defects. To address this, this paper proposes a steel surface defect detection method based on deep learning with YOLOv9s, aiming to solve the problems encountered by current methods when handling small-sized defects through a series of optimization designs, thereby improving detection accuracy and efficiency. The core contributions of this study are mainly reflected in the following key aspects:

1) The SCConv module is adopted, and through the aforementioned reconstruction of spatial and channel dimensions, the model can reduce redundant information without losing important features, making the network more efficient and improving feature representation accuracy.

2) The C3Ghost module is introduced, which helps the model focus on the most distinguishable defect features by reducing redundant feature channels, while significantly decreasing the computational and memory load.

3) The CARAFE (Content-Aware Re-Assembly of Feature Maps) upsampling operator is incorporated, which, through content-based feature reorganization, dynamically adjusts the resolution of the feature map, thereby better preserving the spatial information of the image during the upsampling process.

II. Related Work

A. Depthwise Separable Convolution

Depthwise separable convolution [15] is an efficient convolutional operation that significantly reduces computational cost and parameter count by decomposing the traditional convolution operation into two simpler operations. This approach enhances the efficiency of neural networks. The process consists of the following steps: In a standard convolution, the input feature map has a size of $H \times W \times M$, the convolution kernel has a size of $K \times K \times M$, and the number of convolution kernels is N, resulting in an output feature map of size $H \times W \times N$. Consequently, the parameter count of standard convolution is $P_1 = K \times K \times M \times N$, and the computational cost is $C_1 = H \times W \times K \times K \times M \times N$.

Depthwise separable convolution decomposes the standard convolution kernel into two parts: depthwise convolution and pointwise convolution. Depthwise convolution applies $K \times K \times 1$ convolution kernels to each input channel independently, with a total of M such kernels. Pointwise convolution then uses $1 \times 1 \times M$ convolution kernels to combine all input channels, ultimately generating N output channels. This approach significantly reduces computational cost and parameter count, enhancing network efficiency. Consequently, the total parameter count of depthwise separable convolution is $P_2 = K \times K \times M + M \times N$, with a computational cost of $C_2 = K \times K \times M \times H \times W + M \times N \times H \times W$. Specifically, the parameter count for depthwise convolution is $K \times K \times M$, and for pointwise convolution, it is $M \times N$. The computational cost for depthwise convolution is $K \times K \times M \times H \times W$, while for pointwise convolution, it is $M \times N \times H \times W$.

From the perspective of parameter count and computational cost, depthwise separable convolution significantly reduces both by decomposing the operations of standard convolution. The ratios of the parameter count and computational cost between depthwise separable convolution and standard convolution are $\frac{P_2}{P_1} = \frac{K \times K \times M + M \times N}{K \times K \times M \times N} = \frac{1}{N} + \frac{1}{K^2}$ and, $\frac{C_2}{C_1} = \frac{K \times K \times M \times H \times W + M \times N \times H \times W}{H \times W \times K \times K \times M \times N} = \frac{1}{N} + \frac{1}{K^2}$, respectively. It can be observed that depthwise separable convolution effectively lowers computational overhead, especially when processing high-dimensional inputs. Therefore, compared to standard convolution, depthwise separable convolution offers higher computational efficiency.

B. YOLOv9 Algorithm

The YOLOv9 model features a unique design in its three key components: the backbone network, the neck network, and the head network. The backbone network adopts the General Efficient Layer Aggregation Network (GELAN), which ingeniously integrates CSPNet and ELAN. This fusion increases the network's width, facilitating smoother gradient flow throughout the network and enhancing its feature extraction capability. The neck network, serving as a crucial bridge between the backbone and head, employs the PGI-based multi-level auxiliary information module. This design enables each level of the feature pyramid to receive information on all object sizes, integrating gradient information from different prediction branches to facilitate parameter updates. As a result, it effectively achieves multi-scale feature fusion while incorporating both deep and shallow feature information without significantly increasing computational cost. The head network adopts a decoupled head design, which consists of two branches: a classification branch and a bounding box (box) branch. Additionally, it divides detection boxes into three scales—large, medium, and small—targeting objects of different sizes. The classification branch is responsible for predicting the object's category, while the bounding box branch focuses on determining the object's location. Each scale of detection boxes plays a distinct role in improving detection accuracy across different object sizes.

III. METHOD DESIGN

This study improves YOLOv9s [16], as shown in Fig. 1, by incorporating the SCConv module. The SCConv module dynamically adjusts the weights of different channels to more effectively capture important spatial information while suppressing irrelevant or redundant features. The C3Ghost module is introduced, and through an optimized design, it maintains a low computational cost, allowing the model to sustain fast inference speed even when handling complex tasks. Additionally, the CARAFE upsampling operator is used to reorganize the feature map in a content-aware manner, enhancing the fine-grained reconstruction of steel surface defect regions, making defect boundaries clearer.

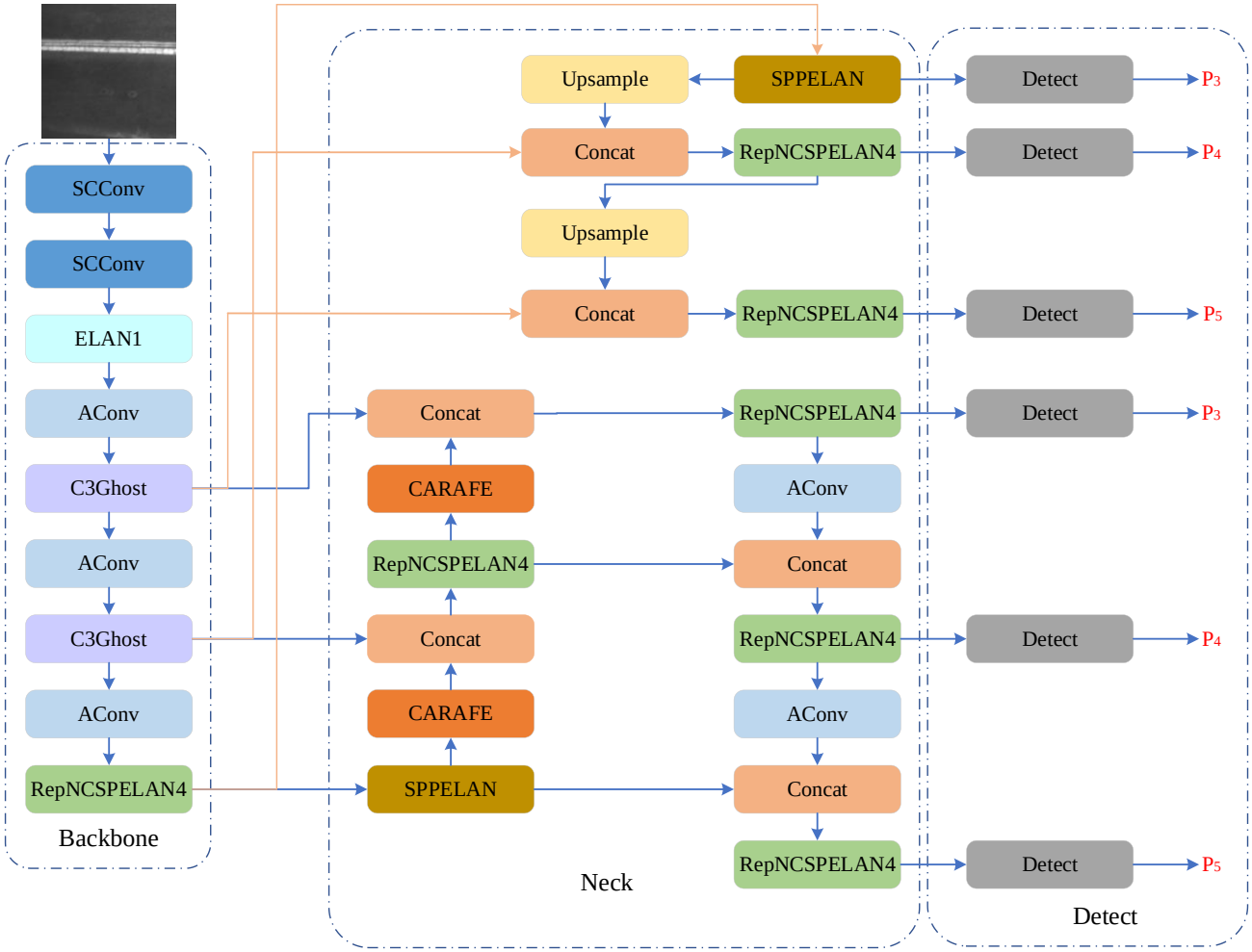


Fig. 1. Improved YOLOv9s network architecture diagram.

A. SCConv Module

YOLOv9s, as an efficient object detection model, demonstrates outstanding performance in various complex scenarios. However, its convolutional layers still exhibit redundancy in both spatial and channel dimensions during feature extraction, which not only increases the computational burden but also limits the efficiency of feature representation to some extent. To address this issue, we introduce the SCConv [17] (Spatial and Channel Reconstruction Convolution) module, which optimizes the features through the Spatial Reconstruction Unit (SRU) and Channel Reconstruction Unit (CRU), thereby reducing redundant computations and enhancing feature representation capability.

As shown in Fig. 2, to further elaborate on the principles and process of the SRU structure, we can delve deeper into the details of the separation and reconstruction operations. The core objective of the separation operation is to distinguish between feature maps with rich information and those with less information. In this way, the SRU can extract useful information while reducing reliance on irrelevant data.

First, the input feature map $X \in R(N \times C \times H \times W)$ is processed through normalization. This is done using the Group Normalization (GN) layer. This normalization operation helps to eliminate biases between different batches of data in convolutional neural networks, making the model training more stable. The normalization formula is [Formula (1)]:

$$X_{out} = GN(X) = \frac{\gamma(X-\mu)}{\sqrt{\sigma^2+\varepsilon}} + \beta \quad (1)$$

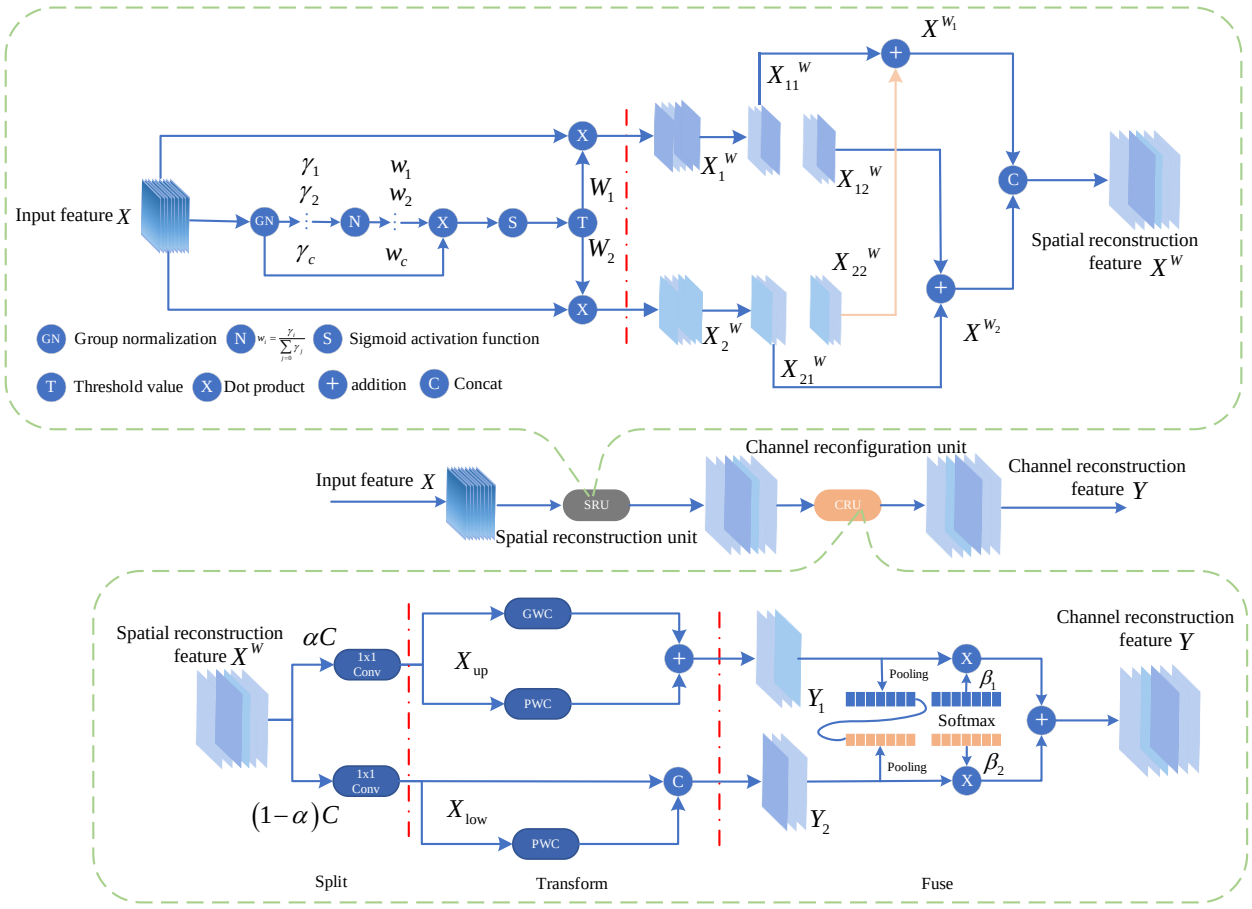


Fig. 2. SCConv module.

Here, μ and σ are the mean and standard deviation of the input feature map X along a specific dimension (usually the channel dimension). γ and β are the trainable affine transformation parameters, and ε is a small constant used to prevent division by zero when the standard deviation is zero.

Next, we calculate the weight W_γ based on each channel feature, which aims to measure the importance of each channel in the overall feature map. The calculation of weight W_γ is as follows [see Formula (2)]:

$$W_\gamma = \{w_i\} = \frac{\gamma_i}{\sum_{j=1}^{C}\gamma_j}, \quad i, j = 1, 2, \ldots, C \tag{2}$$

Here, γ_i represents the affine transformation parameters of each channel in group normalization. The normalization of the weights is performed by dividing each channel's γ_i by the sum of the affine parameters of all channels, ensuring that the weight values remain within a reasonable range.

Next, the normalized weight W_γ is gated through the Sigmoid activation function, yielding a value within the range (0, 1), which represents the "importance" of each channel feature. Specifically as in Formula (3):

$$W = \mathrm{Gate}(\sigma(W_\gamma(GN(X)))) \tag{3}$$

Here, σ (·) is the Sigmoid activation function, which restricts the output to the range (0, 1). Then, a threshold is set (usually 0.5), and the weights greater than this threshold are set to 1 to obtain information-rich weights W_1; weights smaller than the threshold are set to 0 to obtain information-poor weights W_2.

Subsequently, the input feature X is mapped through two different weight matrices W_1 and W_2, resulting in information-rich feature X_1^w and information-poor feature X_2^w. Then, these two types of features are fused through a cross-addition method to further enhance the expressive power of feature information and reduce spatial redundancy. Finally, the fused features X^{w1} and X^{w2} are concatenated together to obtain the final spatially reconstructed feature map X^w, which is represented by the following Formula (4):

$$\begin{cases} X_1^W = W_1 \otimes X, \\ X_2^W = W_2 \otimes X, \\ X_{11}^w \oplus X_{22}^w = X^{w1}, \\ X_{21}^w \oplus X_{12}^w = X^{w2}, \\ X^{w1} \bigcup X^{w2} = X^w. \end{cases} \tag{4}$$

In the formula, $\otimes$ represents element-wise multiplication, $\oplus$ represents element-wise addition, and $\bigcup$ represents the Concat module.

As shown in Fig. 2, the CRU (Channel Reduction Unit) structure is discussed in terms of how it optimizes feature maps through a series of operations to reduce redundancy and improve computational efficiency. The three main stages of CRU include Split, Transform, and Fuse.

In the Split stage, the feature map X^w is divided along the channel dimension into two parts: the upper part contains αC channels, and the lower part contains $(1-\alpha)C$ channels, where α is a tunable parameter, typically set to 0.5, indicating that the feature map is split into two equal parts. The purpose of this step is to reduce the number of channels in each part, thereby reducing the computational burden. Next, a 1x1 convolution operation is applied to compress the number of channels in the feature map. The 1x1 convolution has the ability to reduce the number of channels while also serving as a tool for feature transformation. The compression ratio r controls the number of channels in the output feature map, balancing the trade-off between computational cost and performance. The compressed feature map is then divided into the upper part X_{up} and the lower part X_{low}, preparing for the next stage of operations.

The Transform stage is the core part of the CRU, primarily aimed at extracting features and enhancing computational efficiency through two operations: In the upper transformation stage, the input is the upper part of the feature map X_{up} obtained from the split. Groupwise Convolution (GWC) and Pointwise Convolution (PWC) are used for feature extraction. Groupwise Convolution divides the input channels into several small groups for convolution operations, thereby reducing computational complexity. Pointwise Convolution performs a 1×1 convolution at each pixel location to further extract features. The formula for the upper transformation is [Formula (5)]:

$$Y_1 = M^G X_{up} + M^{P_1} X_{up} \tag{5}$$

Here, M^G and M^{P_1} are the learnable weights of GWC and PWC, respectively, and Y_1 is the output feature map of the upper transformation stage.

In the lower transformation stage, the input is the lower part of the split feature map X_{low}. PWC is used to extract features, and the result is concatenated with the original X_{low} feature map. The concatenation operation is denoted by the symbol $\bigcup$. The formula for the lower transformation is [Formula (6)]:

$$Y_2 = M^{P_2} X_{low} \bigcup X_{low} \tag{6}$$

Here, M^{P_2} is the learnable weight of PWC, and Y_2 is the output of the lower transformation stage.

In the Fusion stage, the simplified SKNet method is used to merge the output features Y_1 and Y_2 from the upper and lower transformation stages. Global spatial information is obtained through global average pooling. The formula for the global channel descriptor S_m is as follows [Formula (7)]:

$$S_m = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Y_c(i, j), m = 1, 2 \tag{7}$$

Subsequently, the two global channel descriptors S_1 and S_2 are stacked together.

Channel attention operations are used to generate importance weights β_1 and β_2, which represent the importance of the corresponding channels. The calculation formula is as follows [Formula (8)]:

$$\beta_1 = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}, \beta_2 = \frac{e^{s_2}}{e^{s_1} + e^{s_2}} \tag{8}$$

Finally, the reconstructed feature $\beta_1 Y_1 + \beta_2 Y_2$ is obtained through the weighted calculation.

B. C3Ghost Module

In the YOLOv9s model, the main goal of replacing the RepNCSPELAN4 module with the C3Ghost [15,18] module is to enhance the model's computational efficiency and detection performance. The core idea behind the design of the C3Ghost module is to optimize the network structure, reduce computational load and the number of parameters, while not sacrificing the model's detection accuracy, thereby achieving efficient object detection. Compared to traditional convolutional modules, the C3Ghost module uses the GhostConv operation, which can extract more effective features at a lower computational cost. Therefore, we introduced the C3Ghost module to significantly improve computational efficiency while maintaining detection accuracy.

As shown in Fig. 3, the workflow of the C3Ghost module is as follows: First, the input feature map X is processed through a convolution layer (Conv) for preliminary processing. Then, it passes through multiple GhostConvBottleneck modules, which improve efficiency by reducing computational load and the number of parameters, while enhancing feature extraction capabilities. After that, a Concat operation is performed, followed by a Conv operation from another path. Finally, the concatenated feature map undergoes a convolution layer (Conv) to generate the final output X.

C. CARAFE Upsampling Operator

The CARAFE [19] upsampling operator is introduced in the YOLOv9s model to replace traditional upsampling methods, primarily due to its unique advantages. CARAFE is a content-aware upsampling mechanism that dynamically adjusts the upsampling process based on the input features, generating higher-quality feature maps, especially excelling in small object detection. Therefore, using the CARAFE upsampling operator can improve the model's detection accuracy for small objects and overall performance.

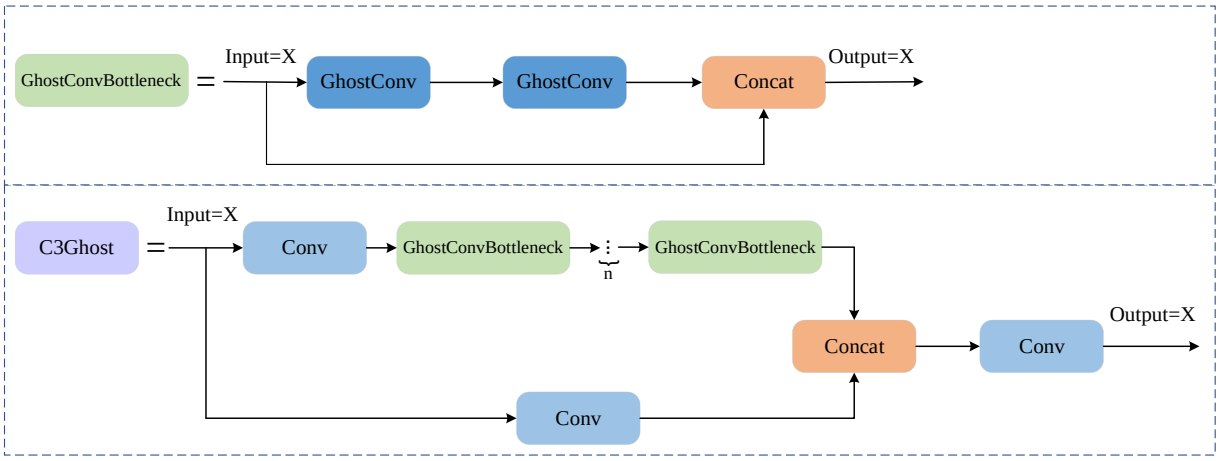


Fig. 3. C3Ghost module.

As shown in Fig. 4, in the upsampling kernel prediction section, first, the channel compression part compresses the input feature map X using a weight matrix W, reducing the number of channels from C to C_m, resulting in a compressed feature representation. Next, the feature map undergoes content encoding, converting the input feature map into a more processable form. Then, a spatial dimension expansion operation is performed to obtain an upsampling kernel of the form $\sigma H \times \sigma W \times k_{up}^2$, where the upsampling factor σ is set. Finally, a kernel normalization operation is applied to ensure that the sum of the kernel weights equals 1, resulting in the output feature map O.

In the feature reorganization section, the feature map undergoes spatial dimension expansion, converting into feature blocks of $K_{up} \times K_{up}$. These feature blocks will prepare for the final upsampling operation. Finally, a pointwise operation is performed, where the expanded feature blocks are corresponded with the points in the original feature map, generating the final upsampled feature map X'.

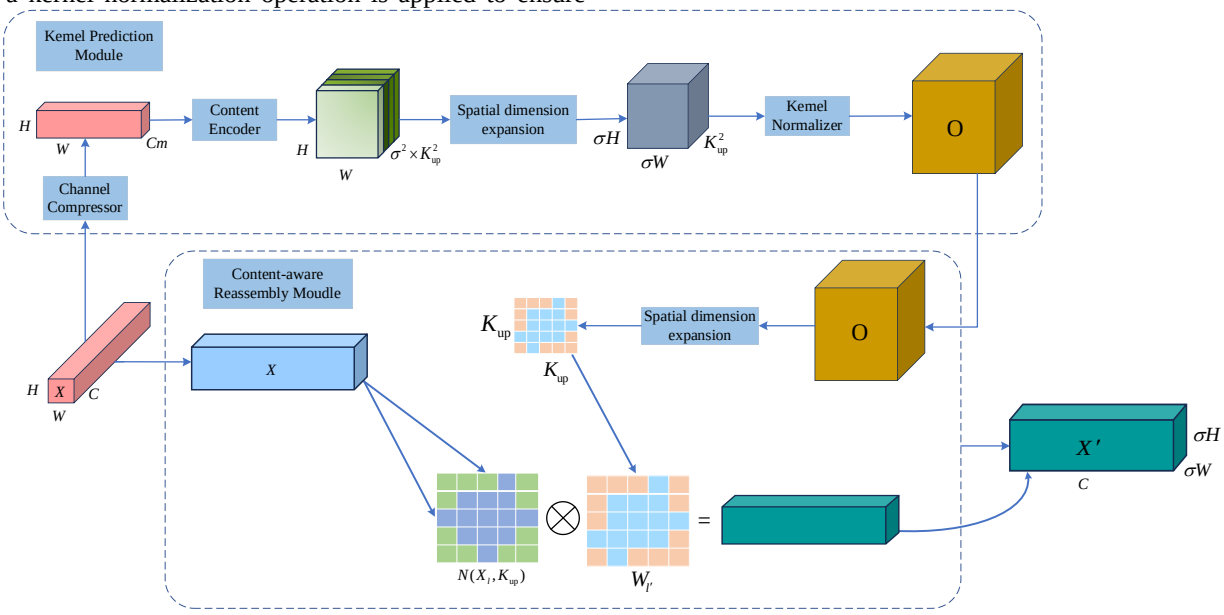


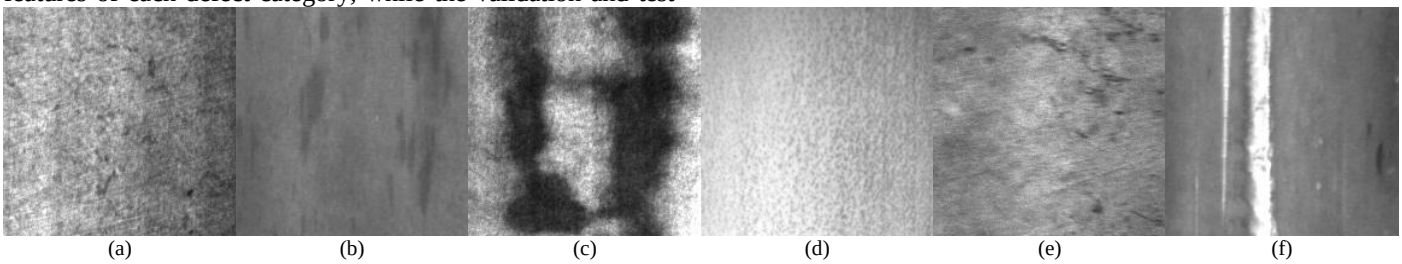
Fig. 4. CARAFE Upsampling operator.

IV. Experimental Setup

A. Dataset

The steel surface defect dataset used in this study, provided by Northeastern University, contains six typical defect categories: a) Crazing, b) Inclusion, c) Patches, d) Pitted Surface, e) Rolled-in Scale, and f) Scratches. The dataset consists of 1,800 images, with 300 samples for each category, ensuring a balanced distribution that comprehensively covers common defect types and features found on steel surfaces. These images were captured using industrial equipment and authentically reproduce the visual characteristics and complexity of surface defects in real-world production processes. Each image is annotated with bounding boxes, a common format used in object detection tasks, which include the defect location and corresponding category information. To reasonably allocate the dataset for training and testing purposes, it is divided into training, validation, and test sets in a 7:1:2 ratio, containing

1,260, 180, and 360 images, respectively. This division ensures that the training phase has sufficient data to effectively learn the features of each defect category, while the validation and test sets provide a scientific evaluation of the model's performance. The six types of steel surface defect images are shown in Fig. 5.



(a) (b) (c) (d) (e) (f)

Fig. 5. Six types of steel surface defects.

B. Experimental Platform

The experiments in this study were conducted on a high-performance computing platform with comprehensive software and hardware configurations, which fully meet the computational resource requirements of deep learning tasks. The software environment uses the Ubuntu 20.04 operating system, with model development and training based on Python 3.8 and PyTorch 1.11.0, and it supports CUDA 11.3 acceleration, significantly improving the efficiency of large-scale data processing during training. The hardware environment includes an NVIDIA RTX 4090 GPU (24GB memory), providing powerful support for deep learning computations; a 22-core AMD EPYC 7T83 64-Core Processor (vCPU), suitable for data preprocessing and training task scheduling; 90GB of RAM, ensuring stable loading of high-resolution images and large-scale model training; as well as a 30GB system disk and a 50GB data disk, used for operating system operation and experimental data storage, respectively. This platform adopts GPU and CPU parallel acceleration to fully utilize computational resources, improving experimental efficiency and task processing capabilities. Additionally, its software and hardware combination offer high stability and scalability, supporting the optimization and expansion needs of future experiments.

C. Hyperparameter Settings

This study made reasonable settings for the training parameters to maximize model performance while ensuring efficient training. During the training process, the pre-trained model "yolov9-s.pt" was loaded, and the image resolution was set to 640×640. The batch size was set to 8 to adapt to hardware resources and ensure the stability of gradient updates. The number of epochs was set to 300 to ensure that the model fully learns the feature patterns in the data. The number of workers was set to 4 to parallelize the data loading process, thereby improving training efficiency. In terms of optimizer parameters, the learning rate was set to 0.01, which provides a good balance between model convergence speed and training stability. The momentum parameter was set to 0.937 to reduce gradient oscillation and accelerate the convergence process. The weight decay coefficient was set to 0.0005 to limit the model's complexity and prevent overfitting. With these fine-tuned parameter settings, this study ensured both the efficiency of the training process and the stability of model performance.

D. Evaluation Metrics

The experiments in this study use F1 score, Precision (P), Recall (R), Average Precision (AP), and mean Average Precision (mAP) as evaluation metrics [20], and also consider the number of parameters (Parameters). The calculation formulas for these metrics are as follows [Formula (9) to (13)]:

$$\text{Precision} = \frac{T_{\text{p}}}{T_{\text{p}} + F_{\text{p}}} \tag{9}$$

$$\text{Recall} = \frac{T_{\text{p}}}{T_{\text{p}} + F_{\text{N}}} \tag{10}$$

$$\text{AP} = \int_0^1 P(R)\text{d}R \tag{11}$$

$$\text{mAP} = \frac{1}{n}\sum_{\text{i}=0}^{n} AP(i) \tag{12}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

Here, T_{p} represents the number of correctly detected defect targets; F_p represents the number of incorrectly detected defect targets; F_N represents the number of missed defect targets; n represents the number of defect categories; and $AP(i)$ represents the average precision for the i-th target class.

V. Experimental Analysis

A. Algorithm Comparison Experiment

To validate the proposed performance improvements, we compared the improved YOLOv9s model (Ours) to RT-DETR-18, Faster RCNN, SSD, and seven other mainstream YOLO target detection models, all tested on the same data set and training rounds. The results are shown in Table I. The comparison metrics include Precision, Recall, mean Average Precision (mAP@0.5), Frames Per Second (FPS), computational complexity (GFLOPS), and model parameter count (Params). These metrics enable us to evaluate the performance of each algorithm and its efficiency under different computational resources. Among all the algorithms, the proposed improved model demonstrates a superior balance between precision and recall, while also achieving significant improvements in mAP and inference speed.

TABLE I. EXPERIMENTAL COMPARISON RESULTS OF DIFFERENT ALGORITHMS

Algorithm	Precision/%	Recall/%	mAP@0.5/%	FPS (f/s)	GFLOPS	Params (M)
Faster R-CNN[21]	79.9	—	77.5	—	—	—
SSD[22]	70.2	—	68.6	—	—	—
YOLOv5s[23]	71.7	71.2	75.2	208.3	15.8	7.02
YOLOv7[24]	63.3	74.8	71.7	109.8	103.2	36.50
YOLOv8s[25]	75.7	67.1	71.7	312.5	28.4	11.12
YOLOv9s	70.7	77.3	78.1	94.3	38.7	9.60
YOLOv10s[26]	77.5	67.4	74.4	125.0	24.5	8.03
YOLOv11s[27]	74.1	74.7	78.4	86.2	21.3	9.41
RT-DETR-18[28]	79.3	67.4	73.4	68.0	57.0	19.87
Ours	**77.0**	**72.4**	**79.6**	**109.9**	**36.9**	**9.07**

Table I presents the comparison experiment results of various object detection algorithms, including Precision, Recall, mean Average Precision (mAP@0.5), Frames Per Second (FPS), computational complexity (GFLOPS), and model parameter count (Params). Among these algorithms, The Precision of Faster R-CNN reached 79.9%, and the mAP@0.5 achieved 77.5%. The Precision of SSD reached 70.2%, and the mAP@0.5 achieved 68.6%. YOLOv5s has a high inference speed (208.3 FPS) and low computational complexity (15.8 GFLOPS), but its Precision (71.7%) and Recall (71.2%) are relatively moderate. YOLOv7 performs well in Recall (74.8%), but its Precision is lower (63.3%) and computational complexity is higher. YOLOv8s achieves a high Precision (75.7%) but has a lower Recall (67.1%) and an extremely fast inference speed (312.5 FPS). YOLOv9s performs well in Recall and mAP (78.1%), but its inference speed is slower (94.3 FPS). YOLOv10s has high Precision (77.5%) with moderate inference speed (125.0 FPS). YOLOv11s shows a high Recall (74.7%) but slower inference speed (86.2 FPS). RT-DETR-18, although having high Precision (79.3%), has slower inference speed and lower mAP (73.4%). Our model achieves a good balance in both Precision (77.0%) and Recall (72.4%), with an mAP of 79.6%, inference speed of 109.9 FPS, moderate computational complexity (36.9 GFLOPS), and a model parameter count of 9.07 M. It demonstrates high overall performance, making it an efficient and balanced choice.

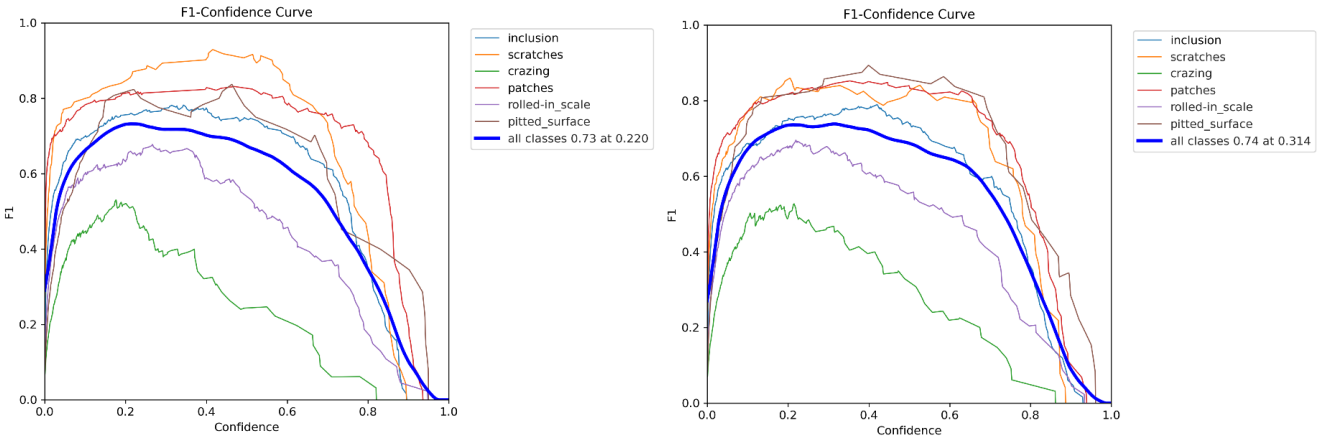
TABLE II. COMPARISON OF VARIOUS ALGORITHMS IN TERMS OF AVERAGE PRECISION (AP%)

Algorithm / Category	Inclusion	Scratches	Crazing	Patches	Rolled-in_Scale	Pitted_Surface
YOLOv5s	78.6	91.1	41.3	89.1	70.9	80.3
YOLOv7	75.6	80.5	38.9	89.9	66.6	78.9
YOLOv8s	70.2	89.0	32.8	86.4	66.1	85.7
YOLOv9s	81.2	94.3	48.0	89.1	70.8	85.5
YOLOv10s	75.8	90.5	43.8	81.7	74.0	80.7
YOLOv11s	82.1	92.3	49.0	89.9	69.5	87.8
RT-DETR-18	78.6	93.1	48.9	88.0	52.5	79.5
Ours	**84.1**	**92.5**	**52.0**	**88.8**	**75.0**	**85.1**

Based on the Average Precision (AP) comparison results in Table II, the "Ours" model performs the best across multiple defect categories, particularly excelling in the inclusion and crazing categories, with precision rates of 84.1% and 52.0%, respectively, significantly higher than all other algorithms. YOLOv9s and YOLOv11s also perform well in most categories, especially in scratches, where they achieve precision rates of 94.3% and 92.3%, respectively, although they are slightly lacking in the Rolled-in_scale category. YOLOv5s, YOLOv7, and YOLOv10s perform well in the scratches and patches categories, but their precision in crazing and Rolled-in_scale is lower. YOLOv8s shows strong performance in scratches and Pitted_surface, but lags significantly in crazing. RT-DETR-18 also demonstrates good performance in scratches and crazing. In summary, the "Ours" model outperforms other algorithms across multiple defect categories, showing significant performance improvements.

B. Result Visualization

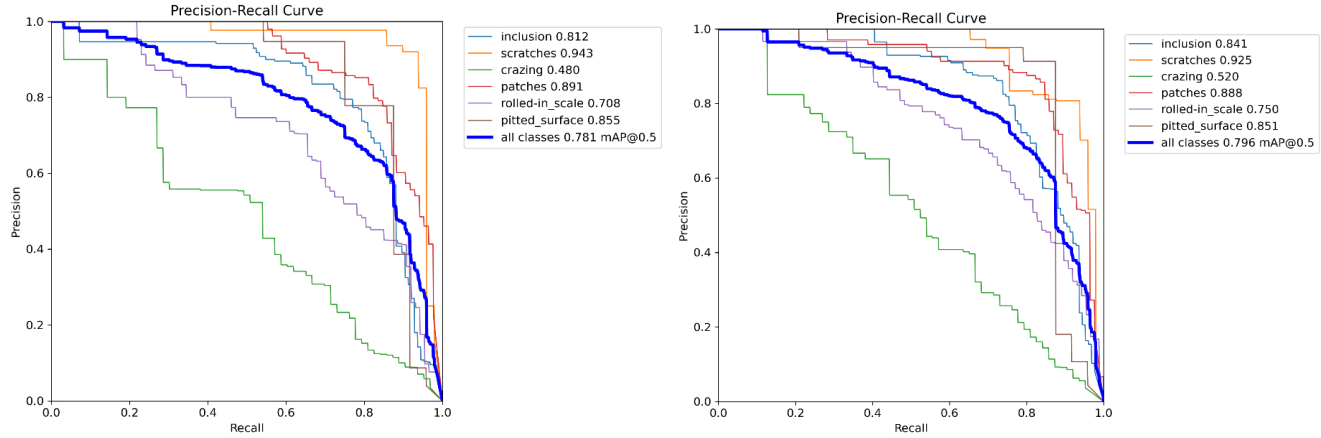
To evaluate the effectiveness of the improved algorithm, it needs to be compared with the original algorithm. YOLOv9s, as a classic object detection algorithm, is widely used in various tasks. With continuous optimization of the algorithm, the improved YOLOv9s has shown better performance in both precision and recall. The F1 score, which considers both precision and recall, provides an effective way to assess the algorithm's performance in detection tasks. By comparing the F1 scores of YOLOv9s and the improved YOLOv9s, we can visually understand the improvements in the balance between precision and recall, thereby evaluating the performance optimization effect of the improved algorithm.



(a) (b)

Fig. 6. F1-Confidence Curve: (a) YOLOv9s Algorithm; (b) Improved YOLOv9s Algorithm.

Fig. 6 displays the F1-Confidence curves for both YOLOv9s and the improved YOLOv9s. The blue curve represents the F1 score for all categories. From the comparison, it is evident that the improved YOLOv9s shows a significant enhancement in F1 scores across categories, particularly in the higher confidence regions, indicating a clear optimization in precision. Specifically, in the left-side YOLOv9s plot, the blue curve reaches an F1 score of 0.73 at a confidence of 0.220, while in the right-side improved YOLOv9s plot, the blue curve reaches an F1 score of 0.74 at a confidence of 0.314. This demonstrates that the improved YOLOv9s has an overall increase in F1 score across all categories, with better performance at higher confidence levels.

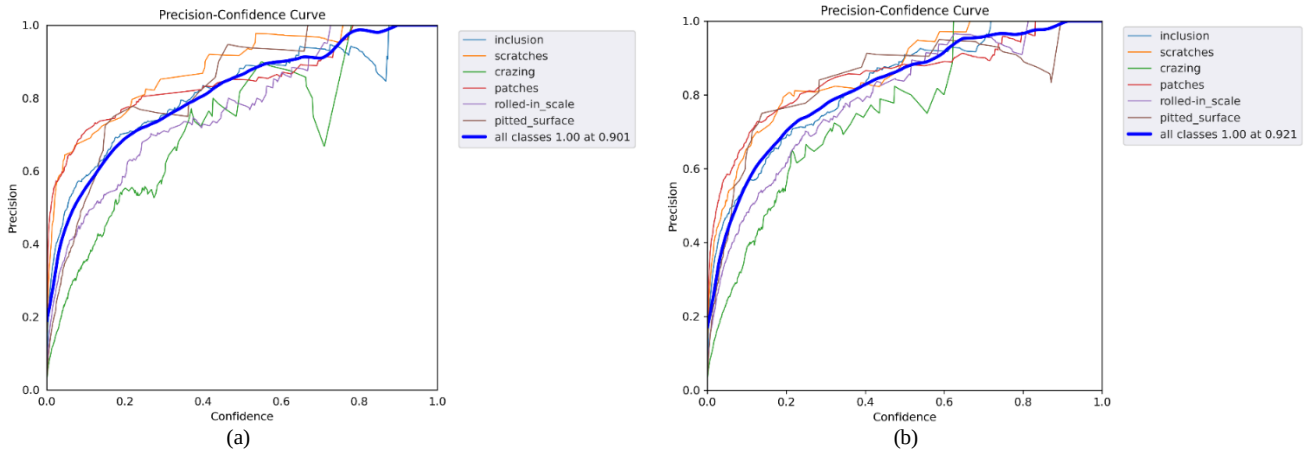


(a) (b)

Fig. 7. Precision-Recall Curve: (a) YOLOv9s Algorithm; (b) Improved YOLOv9s Algorithm.

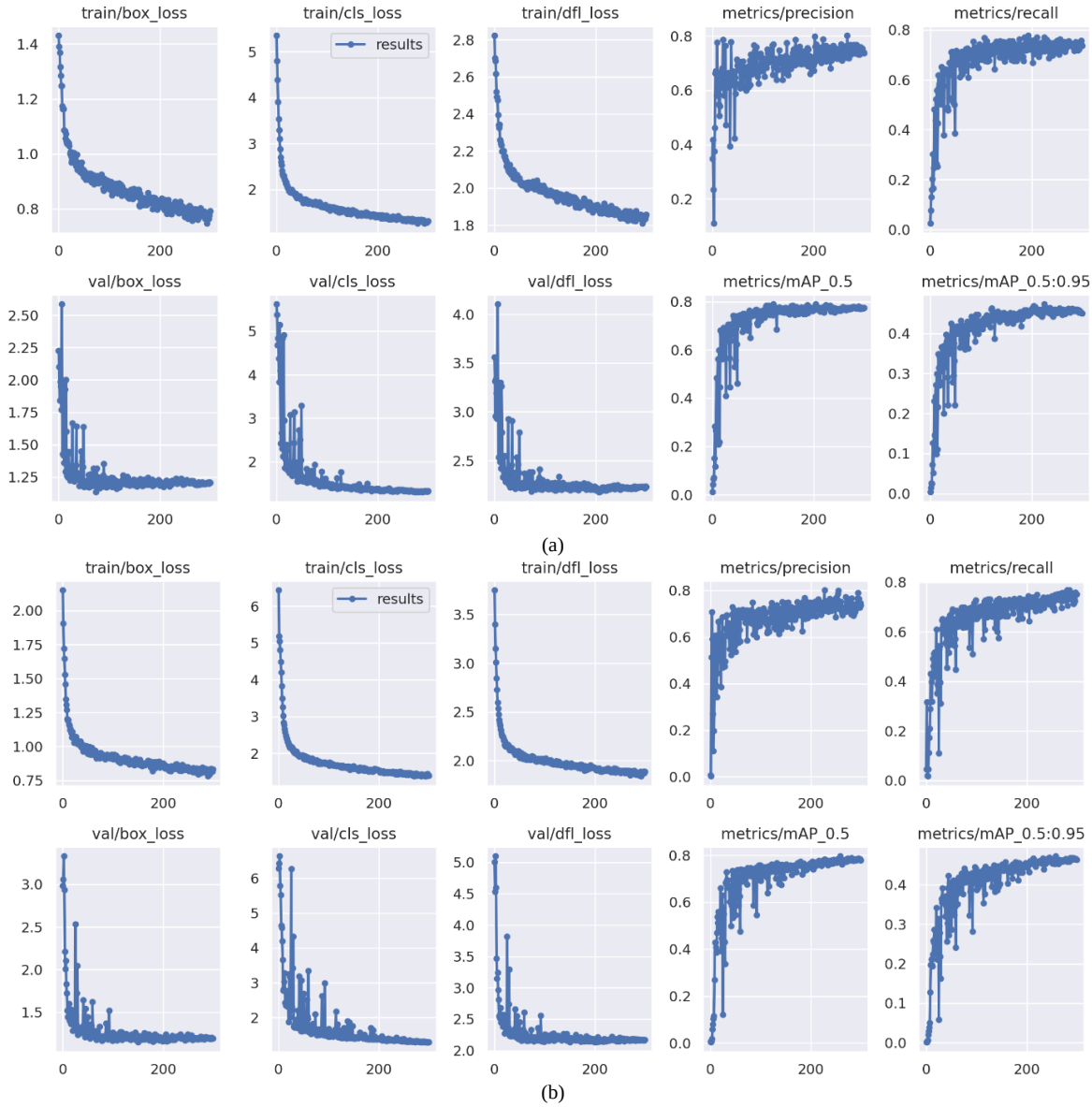
Fig. 7 shows the Precision-Recall (PR) curves for YOLOv9s and the improved YOLOv9s across different categories. In the left-side YOLOv9s plot, the blue curve represents the Precision-Recall curve for all categories, with an mAP value of 0.781. In the right-side plot of the improved YOLOv9s, the blue curve has an mAP value of 0.796, which shows an improvement over YOLOv9s, indicating that the improved algorithm performs better in balancing precision and recall. Overall, the improved YOLOv9s shows enhanced precision across multiple categories, particularly in the "inclusion" and "crazing" categories, where the improved algorithm demonstrates superior performance.

Fig. 8 shows the Precision-Confidence curves for YOLOv9s and the improved YOLOv9s across different categories. In the left-side YOLOv9s plot, the blue curve represents the Precision-Confidence curve for all categories, with a precision of 1.00 at a confidence of 0.901. In the right-side plot of the improved YOLOv9s, the blue curve also reaches a precision of 1.00, but at a confidence of 0.921. A comparison reveals that the improved YOLOv9s shows an overall increase in precision at higher confidence levels and outperforms YOLOv9s in multiple categories, especially demonstrating more stable precision at lower confidence levels.



(a) (b)

Fig. 8. Precision-Confidence Curve: (a) YOLOv9s Algorithm; (b) Improved YOLOv9s Algorithm.



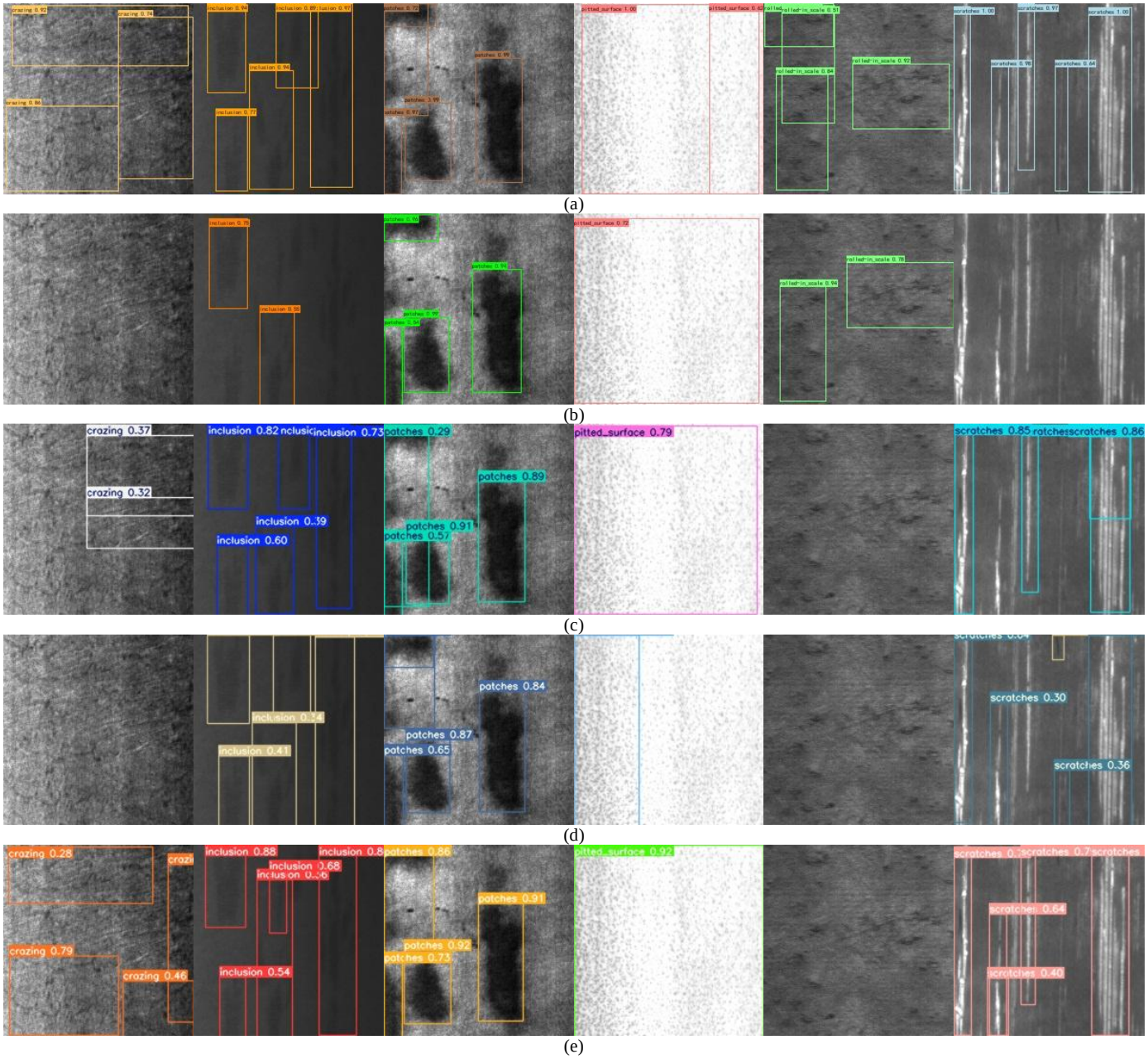
(a)

(b)

Fig. 9. Change Curves of Various Loss Functions and Evaluation Metrics During Training and Validation: (a) YOLOv9s Algorithm; (b) Improved YOLOv9s Algorithm.

Fig. 9 shows the change curves of various loss functions and evaluation metrics during the training and validation process for YOLOv9s and the improved YOLOv9s. These include box loss, classification loss, depth regression loss, precision, recall, and mean average precision (mAP). It can be observed that as training progresses, the losses for both the training and validation sets gradually decrease, indicating continuous improvement in the model's performance in box prediction, classification, and depth regression. Meanwhile, the precision, recall, and mAP metrics also increase, showing a significant improvement in the detection capabilities of the improved YOLOv9s. Overall, the improved YOLOv9s outperforms YOLOv9s across all metrics, demonstrating effective optimization and stronger detection performance during training.

As shown in Fig. 10 below, we compare the detection results of various algorithms on the steel surface defect dataset, including a) Faster R-CNN, b) SSD, c) YOLOv5s, d) YOLOv7, e) YOLOv8s, f) YOLOv9s, g) YOLOv10s, h) YOLOv11s, i) RT-DETR-18, and j) Ours. Through intuitive visualization, we can clearly observe the recognition performance of each algorithm on different targets, including localization accuracy and classification confidence. The improved algorithm (Ours) demonstrates higher localization accuracy across multiple defect categories, showing a distinct advantage over the other algorithms. These comparisons allow us to more comprehensively assess the performance improvements of the improved algorithm in practical applications.



(a)

(b)

(c)

(d)

(e)

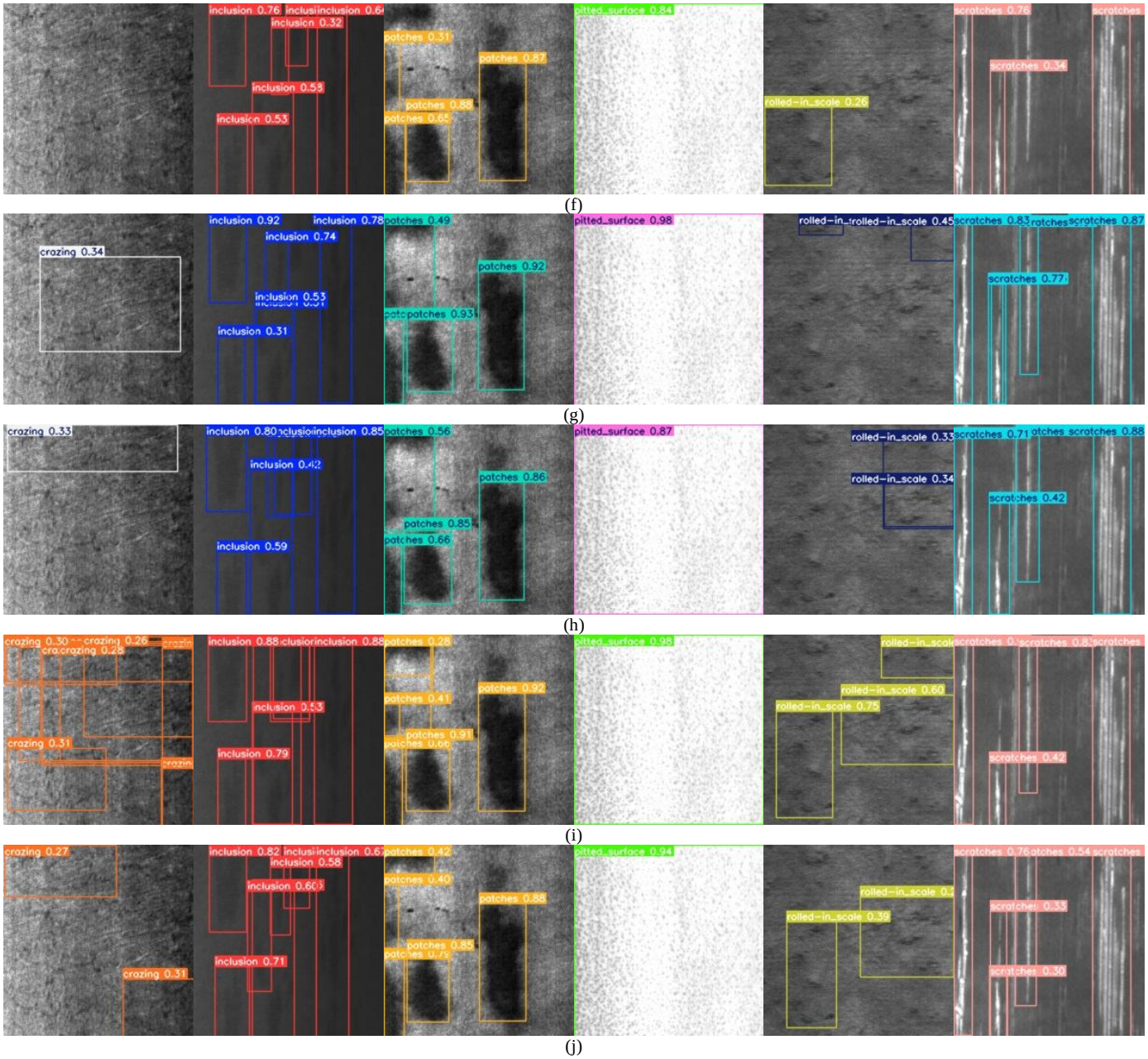


Fig. 10. Detection Results on the Steel Surface Defect Dataset: (a) Faster RCNN; (b) SSD; (c) YOLOv5s; (d) YOLOv7; (e) YOLOv8s; (f) YOLOv9s; (g) YOLOv10s; (h) YOLOv11s; (i) RT-DETR-18; (j) Ours.

Based on the visualization results of steel surface defect detection shown in Fig. 10, different models exhibit varying advantages and disadvantages when detecting defects such as crazing, inclusion, patches, pitted surface, rolled-in scale, and scratches. Overall, traditional object detection methods like Faster R-CNN and SSD have certain limitations. Although Faster R-CNN produces more detection boxes, it suffers from severe target overlapping, making small targets prone to being missed. SSD performs well in detecting large targets but has weaker recognition capabilities for small targets, resulting in a relatively low overall recall rate. In comparison, YOLOv5s demonstrates a noticeable improvement in both recall rate and precision over SSD, although some instances of missed detection still occur. YOLOv7 performs slightly worse than YOLOv5s, with a higher false detection rate and an unremarkable recall rate. YOLOv8s performs well in detecting large and medium-sized targets but still struggles with certain small targets. YOLOv9s and YOLOv10s further enhance detection confidence and recall rates, showing improved detection performance for rolled-in scale and scratches while reducing false detections. YOLOv11s achieves the best overall performance among all categories, offering higher recall rates and confidence scores, more precise detection boxes, and the lowest false detection rate. Additionally, RT-DETR-18 achieves a well-balanced trade-off between precision and recall. Compared to the YOLO series, it provides higher detection confidence and a lower false detection rate, particularly demonstrating stable performance in detecting pitted surfaces

and patches. However, there is still room for improvement in detecting certain small targets, such as scratches. Ultimately, the proposed model (Ours) outperforms all other detection models, delivering the most precise detection boxes, the highest recall rate, the most stable confidence scores, and the lowest false detection rate. Compared with other models, it accurately identifies all defect categories, effectively avoiding common issues of missed and false detections. Notably, it shows significant advantages in detecting small targets such as scratches and inclusions, highlighting its high applicability and reliability in steel surface defect detection tasks. Future optimizations could focus on enhancing detection capabilities for extremely small targets (e.g., fine scratches) to improve detection sensitivity, making it even more efficient and precise for industrial inspection applications.

C. Ablation Study

To further validate the impact of different modules on model performance, we conducted ablation experiments. The ablation study compares the performance of different model combinations across multiple performance metrics, including Precision, Recall, mean Average Precision (mAP), Frames Per Second (FPS), GFLOPS, and parameter count (Params). These experiments aim to assess the contribution of each module to the model's performance, helping us understand which optimizations effectively improve the detection results.

TABLE III. ABLATION STUDY RESULTS

Number	Experiment	Precision/%	Recall/%	mAP@0.5/%	FPS (f/s)	GFLOPS	Params (M)
1	YOLOv9s	70.7	77.3	78.1	94.3	38.7	9.60
2	YOLOv9s+CARAFE	76.4	71.3	78.3	82.6	39.0	9.74
3	YOLOv9s+C3Ghost	78.4	70.3	79.5	131.5	34.5	8.91
4	YOLOv9s+SCConv	73.9	75.5	79.1	75.7	40.8	9.62
5	YOLOv9s+CARAFE+C3Ghost+SCConv	**77.0**	**72.4**	**79.6**	**109.9**	**36.9**	**9.07**

According to the ablation experiment results in Table III, each model shows different performance across various metrics such as precision, recall, mAP, FPS, GFLOPS, and parameter count. Experiment 1 uses YOLOv9s as the baseline model, with a precision of 70.7%, recall of 77.3%, mAP of 78.1%, and FPS of 94.3, demonstrating good performance. Experiment 2 introduces the CARAFE module based on Experiment 1, increasing precision to 76.4%, but recall drops to 71.3%, and FPS significantly decreases to 82.6. Experiment 3 adds the C3Ghost module to Experiment 1, resulting in a further improvement in precision to 78.4% and a slight decrease in recall to 70.3%. Additionally, the FPS notably increases to 131.5, reflecting significant speed optimization. Experiment 4 introduces the SCConv module based on Experiment 1, improving both precision and recall, but FPS drops to 75.7. Finally, Experiment 5 combines the CARAFE, C3Ghost, and SCConv modules, showing improvements in precision, recall, and mAP, with precision at 77.0%, recall at 72.4%, mAP at 79.6%, and FPS rising to 109.9, despite a slight increase in parameter count. Overall, Experiment 5 (YOLOv9s + CARAFE + C3Ghost + SCConv) achieves a better balance between precision and speed, showing significant improvements over the baseline model.

VI. DISCUSSION

Although the improved YOLOv9s model achieves significant enhancements in defect detection performance, several limitations were observed during experimental testing. First, the model's detection accuracy slightly declines when dealing with extremely small-sized defects, such as fine scratches or micro-inclusions, particularly under conditions of strong illumination interference or highly complex surface textures. Second, while the introduction of the SCConv, C3Ghost, and CARAFE modules enhances feature representation and reconstruction capabilities, the complexity of the combined modules leads to a marginal increase in inference latency compared to the original YOLOv9s. Although the overall computational overhead remains acceptable, this factor could pose potential challenges for deployment in ultra-real-time industrial detection systems where strict latency requirements exist. In future work, optimizing the lightweight structure of the improved model to further reduce inference time without compromising detection accuracy will be explored. Additionally, strategies such as multi-scale feature enhancement, adaptive noise suppression, and transfer learning under domain adaptation settings will be investigated to improve the model's performance in detecting extremely small defects and generalizing across varied production environments.

VII. CONCLUSION

This study aims to improve the YOLOv9s model to enhance its performance in steel surface defect detection tasks. To address the shortcomings of the existing YOLOv9s model in detecting small-sized defects on complex steel surface scenes, this study proposes a deep learning-based YOLOv9s method for steel surface defect detection by introducing the SCConv module for optimization, the design optimization of the C3Ghost module, and the CARAFE upsampling operator. Detailed experimental validation of the proposed method is provided. The main conclusions of this study are as follows: First, the adoption of the SCConv module significantly improved the feature extraction efficiency and automatically suppressed redundant and unimportant features. The introduction of the C3Ghost module, by streamlining the computation process and reducing redundant convolution calculations, effectively reduced the model's computational complexity while maintaining a high feature representation capability. To address the issues of small defects and blurred boundaries in steel surface defect images, the CARAFE (Content-Aware Receptive Field) upsampling operator was incorporated. CARAFE finely reconstructs feature maps in a content-aware manner, improving the detail reconstruction accuracy of defect regions, and showing significant advantages in the clarity of defect boundaries and the localization accuracy of small defects. Experimental results

show that the improved model achieved mean average precision (mAP) of 79.6% and precision of 77.0%, which represent improvements of 1.5% and 6.3%, respectively, compared to the baseline model. The findings of this study are of significant practical value for improving the quality and production efficiency of industrial products.

REFERENCES

[1] Wang, Sen, et al. "Lightweight tomato ripeness detection algorithm based on the improved RT-DETR." Frontiers in Plant Science 15 (2024): 1415297.

[2] Ma, Yuxin, et al. "Surface defect inspection of industrial products with object detection deep networks: a systematic review." Artificial Intelligence Review 57.12 (2024): 1-48.

[3] Redmon J. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[4] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[5] Farhadi A, Redmon J. Yolov3: An incremental improvement[C]//Computer vision and pattern recognition. Berlin/Heidelberg, Germany: Springer, 2018, 1804: 1-6.

[6] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." arXiv preprint arXiv:2004.10934 (2020).

[7] Gao, Yang, et al. "Research on steel surface defect classification method based on deep learning." Scientific Reports 14.1 (2024): 8254.

[8] Yu, Tao, et al. "CRGF-YOLO: An Optimized Multi-Scale Feature Fusion Model Based on YOLOv5 for Detection of Steel Surface Defects." International Journal of Computational Intelligence Systems 17.1 (2024): 154.

[9] Yang, Tianle, and Jinghui Li. "Steel Surface Defect Detection Based on SSAM-YOLO." International Journal of Information Technologies and Systems Approach (IJITSA) 16.3 (2023): 1-13.

[10] Subburaj, Kannan, et al. "DBCW-YOLO: an advanced yolov5 framework for precision detection of surface defects in steel." Matéria (Rio de Janeiro) 29.4 (2024): e20240549.

[11] Wu, Guinan, and Qinghong Wu. "Enhancing Steel Surface Defect Detection: A Hyper-YOLO Approach with Ghost Modules and Hyper FPN." IAENG International Journal of Computer Science 9 (2024).

[12] Xu, Yiming, et al. "The steel surface multiple defect detection and size measurement system based on improved yolov5." Journal of Electrical and Computer Engineering 2023.1 (2023): 5399616.

[13] Kong, Haozheng, and Congzhe You. "Improved steel surface defect detection algorithm based on YOLOv8." IEEE Access (2024).

[14] Mi, Zengzhen, et al. "Steel strip surface defect detection based on multiscale feature sensing and adaptive feature fusion." AIP Advances 14.4 (2024).

[15] Kou, Renke, et al. "LW-IRSTNet: Lightweight infrared small target segmentation network and application deployment." IEEE Transactions on Geoscience and Remote Sensing (2023).

[16] Xu, Keyu, et al. "RMT-YOLOv9s: An Infrared Small Target Detection Method Based on UAV Remote Sensing Images." IEEE Geoscience and Remote Sensing Letters (2024).

[17] Li, Jiafeng, Ying Wen, and Lianghua He. "Scconv: Spatial and channel reconstruction convolution for feature redundancy." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[18] Sonawane, Sandip, and Nitin N. Patil. "Performance Evaluation of Modified YOLOv5 Object Detectors for Crop-Weed Classification and Detection in Agriculture Images." SN Computer Science 6.2 (2025): 1-12.

[19] Meng, Bo, and Weida Shi. "Small traffic sign recognition method based on improved YOLOv7." Scientific Reports 15.1 (2025): 5482.

[20] Kou R, Wang C, Peng Z, et al. Infrared small target segmentation networks: A survey[J]. Pattern Recognition, 2023, 143: 109788.

[21] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[22] Liu, Wei, et al. "Ssd: Single shot multibox detector." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016.

[23] Chen, Chen, et al. "Improved YOLOv5s model for key components detection of power transmission lines." arxiv preprint arxiv:2502.06127 (2025).

[24] Liu, Le, et al. "Improved YOLOv7 Based on Reparameterization and Attention Mechanism for Continuous Casting Slab Detection." IEEE Transactions on Instrumentation and Measurement (2025).

[25] Zhong, Jiaqi, et al. "Improved real-time object detection method based on YOLOv8: a refined approach." Journal of Real-Time Image Processing 22.1 (2025): 1-13.

[26] Meng, Zhichao, et al. "YOLOv10-pose and YOLOv9-pose: Real-time strawberry stalk pose detection models." Computers in Industry 165 (2025): 104231.

[27] Alkhammash, Eman H. "Multi-Classification Using YOLOv11 and Hybrid YOLO11n-MobileNet Models: A Fire Classes Case Study." Fire 8.1 (2025): 17.

[28] Zhou, Chuncheng, et al. "MSRT-DETR: A novel RT-DETR model with multi-scale feature sequence for cell detection." Biomedical Signal Processing and Control 103 (2025): 107378.

Landslide Detection Method Based on Lightweight Convolution and Attention Mechanisms

Cong Chen[1], Chengyang Zhang[2], Ran Chen[3], Jiyang YU[4]*
School of Marine Information Engineering, Hainan Tropical Ocean University, Sanya 572022, China[1, 3]
Xiamen City, Fujian Province, 361004, China[2]
Graduate School of Management of Technology, Pukyong National University, Busan 48547, Korea[4]

***Abstract*—Landslide monitoring is a crucial component of geological disaster early warning systems. Traditional landslide detection methods often suffer from insufficient accuracy or low efficiency. To address these issues, this study proposes an improved landslide detection algorithm based on YOLOv11n, aiming to enhance both detection accuracy and efficiency by optimizing the model structure. First, the GhostConv module is introduced to reduce redundant computations, thereby improving computational efficiency. Additionally, the C3K2-SCConv optimization module is incorporated, which enhances feature extraction capability and improves the recognition of landslides at different scales by integrating multi-scale information and a weighted convolution strategy. Furthermore, the SimAM attention mechanism is implemented to adaptively adjust feature map weights, strengthening key features in landslide regions and improving detection accuracy. Experimental results demonstrate that the improved model achieves a mean average precision (mAP@0.5) of 83.3%, a precision of 85.5%, and a recall of 78.1%, representing increases of 2.0%, 3.2%, and 2.8%, respectively, compared to the baseline model. The proposed improvements provide a more accurate and efficient landslide detection method, contributing to the precision of geological disaster early warnings and enhancing the reliability of disaster prevention and mitigation efforts.**

Keywords*—YOLOv11n; landslide detection; GhostConv module; C3K2-SCConv module; SimAM attention mechanism

I. INTRODUCTION

Landslides, as a common and severe geological disaster, typically occur with the sudden collapse or movement of mountain masses, directly threatening the safety of local residents, infrastructure, and the ecological environment [1]. Particularly in mountainous and coastal regions with high rainfall, the frequency of landslide occurrences continues to rise due to triggering factors such as precipitation, earthquakes, and engineering construction. Traditional landslide monitoring methods primarily rely on ground surveys and physical measurement techniques, such as leveling measurements [2], ground-based radar (GB-InSAR) [3], and GPS positioning [4]. These methods can provide high-precision monitoring data and offer significant advantages, especially in small-scale landslide monitoring and early warning applications. However, traditional methods often suffer from limitations in data acquisition, high labor intensity, and difficulties in real-time monitoring. These challenges become particularly prominent in large-scale, complex terrains or post-disaster scenarios, where timely and accurate monitoring results may not be readily available. In recent years, the integration of deep learning technology has introduced a novel solution for landslide monitoring. Deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) have made remarkable progress in fields such as image processing [5-7], pattern recognition, and big data analysis. Specifically, in landslide detection, deep learning models can automatically extract features from large-scale remote sensing imagery, identify potential landslide-prone areas, and detect changes over time. This approach overcomes the reliance of traditional methods on expert experience and manual feature extraction, offering a more efficient and automated solution.

Although deep learning technology has made significant progress in landslide disaster monitoring, several challenges and limitations remain in current research. For example, Han et al. [8] proposed a novel neural network model called Dynahead-YOLO, which integrates scale-aware, spatial-aware, and task-aware attention mechanisms to enhance landslide decoding in complex backgrounds. However, small-sample landslide detection remains a challenge. Yang et al. [9] introduced a lightweight attention-guided YOLO model (LA-YOLO-LLL), which incorporates MobileNetv3 as the backbone network and designs a lightweight pyramid feature reuse and fusion attention mechanism while integrating the level set layer to generate precise landslide boundaries. However, at high resolutions, the model still struggles with detecting small-scale landslides. Liu et al. [10] proposed the SE-YOLOv7 landslide detection algorithm, which integrates an attention mechanism and an improved loss function, significantly improving detection accuracy. Nevertheless, the algorithm is still prone to false detections in areas with complex terrain features. Liu et al. [11] introduced a multi-scale sample-based complex background enhancement method (MSSCBE) to improve sample quality in landslide detection. However, the model exhibited poor generalization ability. Zhang et al. [12] designed a multi-scale feature extraction module that combines efficient channel attention, average pooling, and spatially separable convolution. Despite its effectiveness, challenges persist in processing small-scale landslides in complex backgrounds. Yu et al. [13] proposed an automatic loess landslide detection method based on Google Earth Engine (GEE) and an improved YOLOX algorithm, achieving an average accuracy of 95.43%. However, the model's performance still needs improvement in detecting small-scale landslides, and it is prone to false detections at the edges of remote sensing images. Hou et al. [14] developed a landslide detection method based on an improved YOLOX

object detection model, achieving promising detection performance. However, the accuracy and robustness of the model in detecting landslides across multiple categories, different terrains, and landscapes remain relatively low. Cheng et al. [15] reviewed deep learning technology and its fundamental principles, along with the current status of landslide remote sensing databases. They introduced classic deep learning models for landslide detection and analyzed the strengths and limitations of various models. Chandra et al. [16] proposed a novel generalized efficient layer aggregation network (GELAN) based on an attention mechanism, which significantly improved detection performance. However, the model's complexity remains a major drawback.

Based on the existing research, despite the significant progress made by deep learning techniques in landslide detection tasks, there are still several key challenges that need to be addressed. To this end, this study proposes a landslide detection algorithm based on the improved YOLOv11, with its core contributions mainly reflected in the following aspects:

1) By introducing the GhostConv module, the model's parameter count and computational complexity are reduced while maintaining high performance.

2) The integration of the C3K2-SCConv module enhances the model's ability to recognize multi-scale features, improving the detection of landslides and other complex features.

3) The incorporation of the SimAM attention mechanism strengthens the representation of key features, further improving the overall model performance.

II. Related Work

A. Vision Transformer (ViT)

Vision Transformer (ViT) [17] is a neural network architecture designed for image recognition tasks, inspired by the Transformer model originally developed for natural language processing (NLP). The core idea of ViT is to divide an image into fixed-size patches and process them in a manner similar to how words are handled in NLP tasks. Specifically, an image is first segmented into small patches, each of which is flattened and mapped to a high-dimensional space, akin to word embeddings in NLP. To help the model understand the spatial position of these patches in the original image, positional encoding is added. The patch embeddings, combined with positional encoding, are then fed into a Transformer encoder, which leverages the self-attention mechanism to capture relationships between different patches. This allows the model to grasp the global context of the image rather than just local features. After processing through the encoder, a classification token (similar to the [CLS] token in BERT) is added to the sequence, and its output is used for image classification. Like other deep learning models, ViT is trained in an end-to-end manner using labeled datasets. One of the key advantages of ViT is its strong performance on large-scale datasets, where it often outperforms traditional convolutional neural networks (CNNs, especially when the model size is large. Additionally, due to the self-attention mechanism, ViT can capture the global context of an image, enabling it to model long-range dependencies between different image patches. However, ViT also has some limitations. First, it requires a large amount of training data to achieve optimal performance, and on small datasets, it often underperforms compared to CNNs. Second, due to the high computational complexity of the Transformer architecture, ViT requires significant computational resources when processing high-resolution images. Despite these challenges, ViT represents a major breakthrough in computer vision and has inspired numerous improved versions, such as DeiT (Data-efficient Image Transformer), which is designed to enhance performance when training data is limited.

B. SE Attention Mechanism

The Squeeze-and-Excitation (SE) attention mechanism module [18] is a method to enhance the performance of neural networks by adaptively recalibrating the inter-channel feature responses. Its core idea is to learn the "importance" of each channel, thereby improving the network's feature representation capability. The working process of the SE module can be divided into three main steps: First, global average pooling (GAP) is applied to compress the spatial information of the input feature map, resulting in a global feature vector of size 1×1×C, where C represents the number of channels. Next, the SE module processes these global features through an "excitation" process, which consists of a two-layer fully connected neural network. The first layer uses the ReLU activation function to introduce non-linearity, while the second layer applies the Sigmoid activation function to output the "weight" or "importance" values for each channel, which lie between 0 and 1. These values represent the contribution of each channel to the final decision. Finally, the SE module performs a "recalibration" by multiplying the obtained channel weights with the original feature map on a per-channel basis, enhancing the response of important channels while suppressing less important ones. The advantages of the SE module lie in its strong adaptability, as it automatically learns the significance of each channel, and its low computational cost, making it an effective method to improve model performance in visual tasks.

C. YOLOv11 Algorithm

The network architecture of YOLOv11 [19] is primarily divided into three parts: the backbone, the neck, and the head. In the backbone section, YOLOv11 introduces innovative convolution mechanisms such as C3K2, which allow for more effective extraction of key features in the shallow layers of the network. This creates a more optimized feature extraction structure. Compared to traditional convolution structures, it enables the network to capture essential feature information crucial for object detection with greater precision, especially in complex scenarios. This sets a solid foundation for subsequent detection tasks and addresses the issue of incomplete or inaccurate feature extraction in previous models. In the neck section, the C2PSA mechanism is employed, which integrates a multi-head attention mechanism within C2. This allows the model to better capture spatial context information, improving feature fusion and information transmission efficiency. Compared to conventional designs, this mechanism makes fuller use of contextual information and comprehensively integrates features from different layers. It resolves issues such as reduced detection accuracy and insufficient feature fusion

when dealing with complex spatial relationships in objects, which had previously affected model performance. In the head section, two depthwise separable convolutions are added, reducing computational load and parameter count, significantly improving computational efficiency and inference speed. This is a strategy that has been less commonly adopted in earlier models.

III. METHOD DESIGN

This study presents improvements to the YOLOv11n model, as shown in Fig. 1. The GhostConv module replaces standard convolution to reduce redundancy and computational cost during the early stages of feature extraction, providing a lightweight yet effective foundation for subsequent processing. The C3K2-SCConv module enhances multi-scale feature representation by embedding a spatial-channel attention mechanism, thereby strengthening the model's feature extraction capability. The SimAM attention mechanism is introduced after feature fusion, adaptively adjusting neuron response weights based on energy differences. This not only suppresses background noise but also enhances focus on critical regions. These modules work in a complementary manner: GhostConv ensures computational efficiency, C3K2-SCConv enriches feature expressiveness, and SimAM refines attention. Their integration forms a cohesive structure that balances speed, accuracy, and robustness, ultimately improving the model's ability to accurately detect landslides.

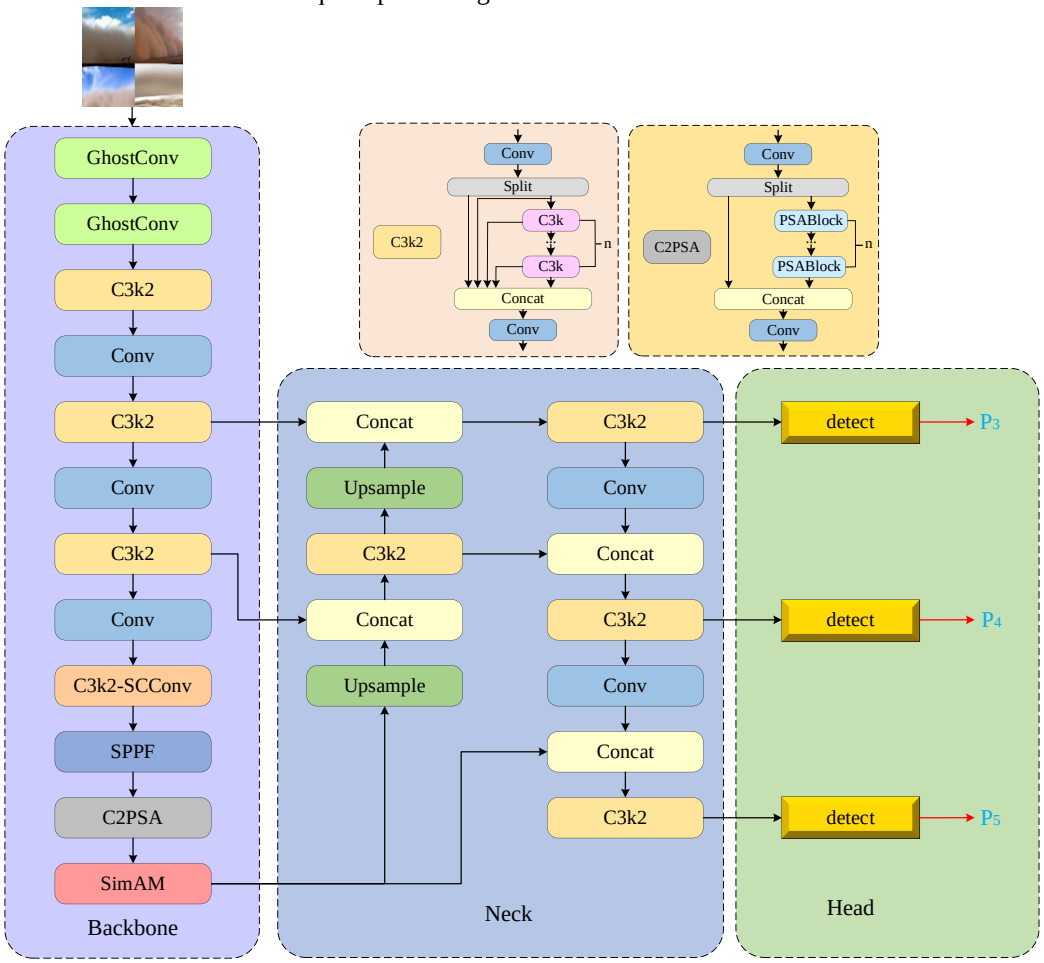


Fig. 1. Network structure of the improved YOLOv11n model.

A. GhostConv Module

YOLOv11n, as a lightweight object detection model, although having a fast inference speed, still faces the issue of computational redundancy, particularly during the shallow feature extraction phase. Traditional convolution operations involve high computational cost and parameter redundancy. To address this problem, this study adopts GhostConv to replace standard convolutions, thereby reducing the computational load and improving computational efficiency.

GhostConv [20] is a technique designed to enhance the efficiency of Convolutional Neural Networks (CNNs) by optimizing convolution operations to reduce computational and storage costs while maintaining model performance. As shown in Fig. 2, the core idea is to reduce the computational load and parameter count by using low-complexity convolution operations during feature map generation. Traditional convolution operations involve sliding a convolution kernel over the input feature map and performing weighted summation, which is computationally expensive, especially when dealing with high-dimensional inputs. In contrast,

GhostConv decomposes the input feature map by generating high-quality feature maps using standard convolutions, followed by simple operations (such as linear transformations) to generate low-complexity "Ghost" feature maps. These Ghost feature maps, although requiring fewer computations and parameters, retain most of the input information and are fused with the feature maps generated by the standard convolution. This design significantly reduces the model's computational and storage requirements, leading to improved computation speed.

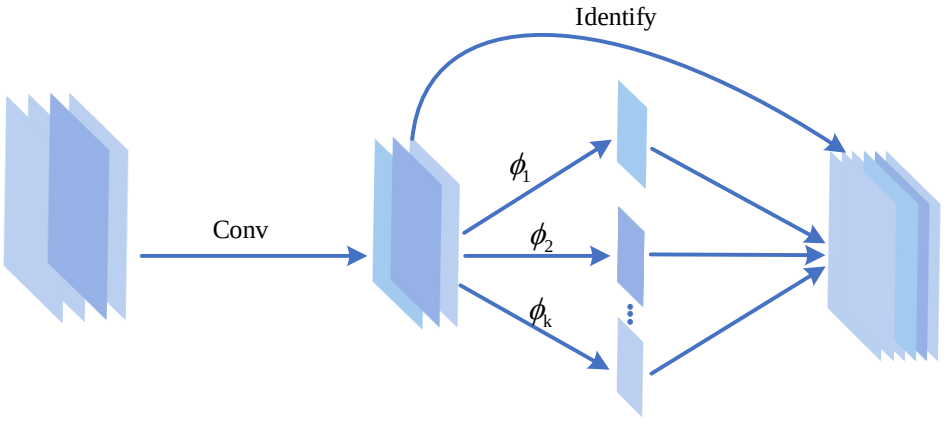


Fig. 2. GhostConv module.

B. C3k2-SCConv Module

The replacement of C3k2 with C3k2-SCConv is aimed at enhancing feature extraction capabilities and improving object detection accuracy while maintaining computational efficiency. C3k2-SCConv inherits the multi-branch structure of C3k2 and introduces SCConv [21] to strengthen feature representation. SCConv potentially enhances the model's focus on key information through spatial and channel attention mechanisms, leading to better performance in complex backgrounds or small object detection tasks. Additionally, SCConv increases the local receptive field, enhancing the robustness of feature representation, allowing the model to maintain high detection capability even in cases of object deformation or occlusion. Compared to traditional convolution, C3k2-SCConv reduces redundant computation through more efficient feature extraction, optimizes cross-layer feature fusion, and reduces information loss, thus achieving a better balance between accuracy and speed.

As shown in Fig. 3, the BottleneckSCConv module is an optimization of the feature extraction unit based on the classic Bottleneck structure, achieved by replacing one of the standard convolution layers (Conv) with SCConv, thereby enhancing feature representation capabilities. The process begins with initial convolution processing, followed by the SCConv module, which may incorporate local attention or cross-channel interaction mechanisms to strengthen key information extraction. Subsequently, the features from the SCConv branch are fused with the original input through a concatenation (Concat) operation, thereby improving the model's feature representation ability. Building on this, C3k2-SCConv inherits the C3k2 structure and introduces the BottleneckSCConv in its Bottleneck structure to further enhance feature extraction. C3k2-SCConv constructs deep feature extraction capabilities by stacking multiple BottleneckSCConv blocks (n times), and combines multi-branch information through feature concatenation (Concat) operations. Finally, the features are integrated through convolutional layers. This improvement enhances the model's feature extraction ability while maintaining its efficiency, making it more suitable for high-performance detection tasks.

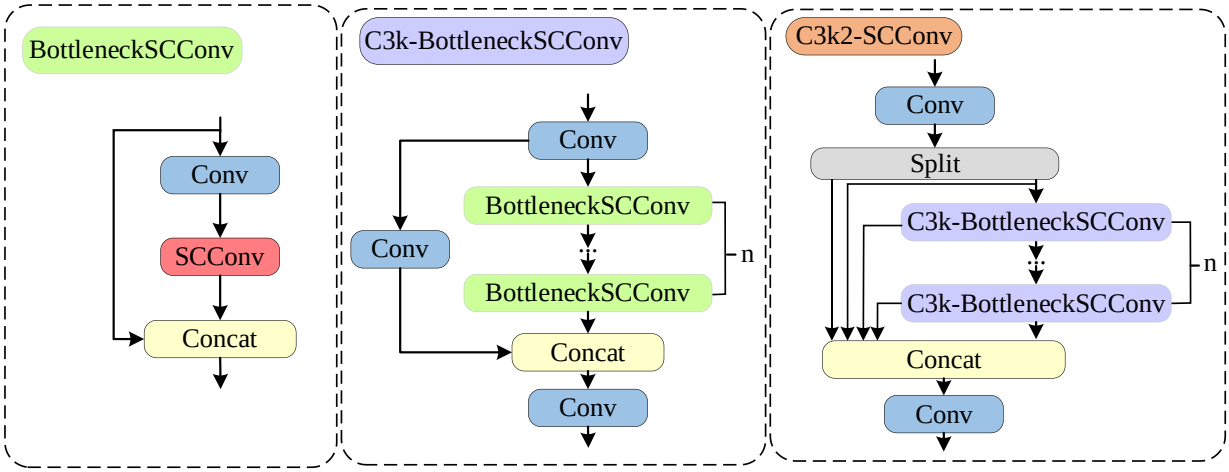


Fig. 3. C3k2-SCConv module.

C. SimAM Attention Mechanism

In this study, we optimized the YOLOv11n object detection algorithm to enhance feature extraction capabilities and detection accuracy while maintaining high computational efficiency. We strengthened the model's attention mechanism to further optimize the expression of target features and the suppression of background information. Specifically, we introduced the SimAM (Simple Attention Module) attention mechanism after the C2PSA module to increase the model's focus on important features, reduce interference from irrelevant information, and thereby improve the accuracy of object detection.

As shown in Fig. 4, the core idea of the SimAM [22] mechanism is to evaluate the differences between neurons through a physical model in order to determine their importance. The energy function $E(x_i)$ calculates the difference between the target neuron x_i and its neighboring neurons $\{x_j \mid j \in N\}$. A lower energy value indicates a greater difference between the target neuron and its surrounding neurons, implying higher importance. The formula for the energy function is:

$$E(x_i) = \frac{4(\sigma^2 + \lambda)}{(x_i - \mu)^2 + 2\sigma^2 + 2\lambda} \tag{1}$$

In the formula, x_i represents the target neuron, and λ is a hyperparameter

SimAM also requires the calculation of the mean and variance of the neurons within the neighborhood. This helps to more comprehensively assess the relationship between each neuron and its surrounding neurons. The formulas for calculating the mean and variance are as follows:

$$\mu = \frac{1}{N}\sum_{j=1}^{N} x_j \tag{2}$$

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N} (x_j - \mu)^2 \tag{3}$$

In the formulas, N represents the number of neurons within the neighborhood, and x_j denotes the value of each adjacent neuron.

Finally, SimAM adjusts the output of each neuron based on its weight, enhancing the contribution of important neurons, thereby improving the model's performance. The formula for the weighted output is as follows:

$$\tilde{X} = sigmoid(\frac{1}{E}) \odot X \tag{4}$$

where, E groups are all $E(x_i)$ across the channel and spatial dimensions. “sigmoid” is added to restrict too large value in E. It will not influence the relative importance of each neuron because a sigmoid is a monofonic function.

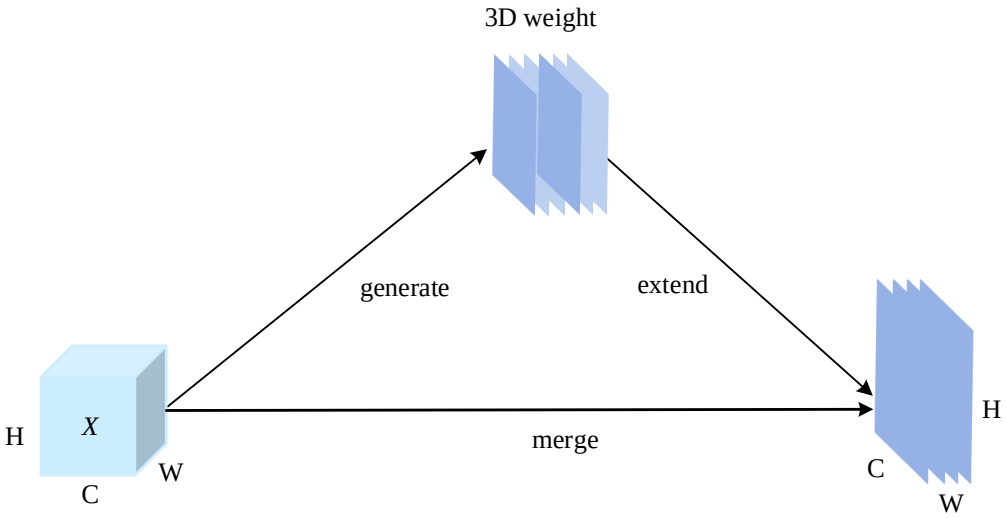


Fig. 4. SimAM attention mechanism.

IV. Experimental Setup

A. Dataset

The dataset used in this study consists of two main categories: landslides and storms, with all images manually annotated to ensure accuracy and consistency. Landslides refer to the phenomenon of soil, rocks, and other materials sliding or collapsing due to natural factors such as rainfall and earthquakes. Storms include severe meteorological events such as thunderstorms and hurricanes, typically accompanied by strong winds and heavy rainfall. Existing landslide recognition datasets primarily rely on high-resolution remote sensing images; however, such data fail to accurately reflect the complex environments present during landslides, such as dust clouds during dry conditions (as shown in Fig. 5). Additionally, processing remote sensing images demands significant computational resources, limiting the feasibility of real-time applications. To address these challenges, this study constructed a multi-domain landslide dataset consisting of 4,735 images, with 77.8% of the data sourced from news and

live video frame extraction, and 22.2% from drone remote sensing data, aimed at improving the model's recognition ability in complex scenarios. To enhance the dynamic features of the data, frame extraction was employed. Given the high-speed motion characteristics of landslides, the video frame extraction interval was set to 1, 1.5, or 2 seconds. Since some images were obstructed by watermarks, blurred, or blocked by buildings, offline data augmentation was applied to part of the news and live video data, resulting in 1,125 images, or 24% of the total dataset. The data augmentation methods combined the few-shot learning strategies proposed by Wang et al. [23], Lu et al. [24], and the random erasure and noise addition techniques proposed by Zhong et al. [25] and Zhang et al. [26], aiming to enhance the system's robustness.



Fig. 5. Comparison of the dust phenomena caused by sandstorms and landslides.

B. Hyperparameter Settings

During training, the image resolution was set to 640×640, with a batch size of 64 to fully utilize the GPU's computational power and accelerate model training. The number of training epochs was set to 200, and the initial learning rate was set to 0.01. To improve data loading efficiency, 4 worker threads were used. Additionally, the momentum parameter was set to 0.937, and the weight decay coefficient was set to 0.0005 to optimize the training process and prevent overfitting.

C. Experimental Platform

The experiments in this study were conducted on the following experimental platform: PyTorch 1.10.0 was used as the deep learning framework, with Python version 3.8 running on the Ubuntu 20.04 operating system, and CUDA version 11.3 to leverage GPU acceleration. The experiments were performed using a single RTX 4090 (24GB) GPU for training, equipped with an AMD EPYC 7T83 64-Core Processor (22 vCPUs) and 90GB of memory. This platform provided powerful computational capabilities for model training, effectively accelerating the training process.

D. Evaluation Metrics

In this study, the evaluation metrics used include F1 score, precision (P), recall (R), average precision (AP), and mean average precision (mAP) [27]. Additionally, the number of parameters (Parameters) was also considered. The formulas for these metrics are as follows:

$$\text{Precision} = \frac{T_p}{T_p + F_p} \tag{4}$$

$$\text{Recall} = \frac{T_p}{T_p + F_N} \tag{5}$$

$$\text{AP} = \int_0^1 P(R)\mathrm{d}R \tag{6}$$

$$\text{mAP} = \frac{1}{n}\sum_{i=0}^{n} AP(i) \tag{7}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

where, T_p represents the number of correctly detected targets; F_p represents the number of falsely detected targets; F_N represents the number of missed targets; n denotes the number of categories; and $AP(i)$ is the average precision of the i-th target class.

V. Experimental Analysis

A. Algorithm Comparison Experiment

To evaluate the advantages of our improved algorithm, we conducted a comparison with other algorithms. In the field of object detection, algorithm performance is typically assessed using several key metrics, such as precision, recall, mAP@0.5, FPS (frames per second), GFLOPS (computational complexity), and the number of parameters (Params). Precision and recall primarily reflect the model's detection accuracy and recall capability, while mAP@0.5 provides a comprehensive measure of the model's performance across different detection thresholds. FPS and GFLOPS assess the model's real-time performance and computational efficiency, respectively, while the number of parameters reflects the model's complexity. By comparing these metrics, we can gain a deeper understanding of the strengths and weaknesses of different algorithms in terms of accuracy, real-time performance, and computational efficiency.

From the experimental comparison results in Table I, different object detection algorithms exhibit varying performances in metrics such as precision, recall, mAP@0.5, FPS, GFLOPS, and the number of parameters (Params). The Ours algorithm achieved the best results in precision (85.5%), recall (78.1%), and mAP (83.3%), while maintaining high real-

time performance (128.5 FPS), low computational complexity (6.3 GFLOPS), and a relatively low number of parameters (2.72M), demonstrating the optimal overall performance. YOLOv11n excelled in real-time performance (147.0 FPS) and computational efficiency (6.3 GFLOPS), but its precision (82.3%) and mAP (81.3%) were slightly lower than those of Ours. YOLOv5s performed well in real-time performance (137.0 FPS), with moderate precision (78.6%) and mAP (76.5%), but its computational complexity (15.8 GFLOPS) and parameter count (7.01M) were relatively high. YOLOv10n showed moderate performance in precision (81.6%), mAP (77.5%), and real-time performance (122.6 FPS), with lower computational complexity (8.3 GFLOPS) and parameter count (2.73M). YOLOv9-gelan-s demonstrated excellent precision (85.3%) and mAP (82.4%), but its computational complexity (26.2 GFLOPS) and parameter count (7.07M) were higher. In comparison, RT-DETR-18 and YOLOv7-tiny performed poorly in terms of precision, recall, and mAP, with RT-DETR-18 having a higher computational complexity (56.9 GFLOPS) and parameter count (19.87M). Overall, the Ours algorithm generally outperforms traditional algorithms in terms of performance.

TABLE I. EXPERIMENTAL COMPARISON RESULTS OF DIFFERENT ALGORITHMS

Algorithm	Precision/%	Recall/%	mAP@0.5/%	FPS (f/s)	GFLOPS	Params (M)
YOLOv5s[28]	78.6	76.5	76.5	137.0	15.8	7.01
YOLOv7-tiny[29]	76.5	66.8	71.3	120.8	13.0	6.01
YOLOv8n[30]	83.2	70.7	78.9	106.3	8.1	3.00
YOLOv9-gelan-s[31]	85.3	74.5	82.4	81.3	26.2	7.07
YOLOv10n[32]	81.6	72.1	77.5	122.6	8.3	2.73
RT-DETR-18[33]	76.6	67.4	70.5	68.4	56.9	19.87
YOLOv11n	82.3	75.3	81.3	147.0	6.3	2.58
Ours	**85.5**	**78.1**	**83.3**	**128.5**	**6.3**	**2.72**

TABLE II. COMPARISON RESULTS OF VARIOUS ALGORITHMS IN AVERAGE PRECISION (AP/%)

Algorithms / Classes	YOLOv5s	YOLOv7-tiny	YOLOv8n	YOLOv9-gelan-s	YOLOv10n	RT-DETR-18	YOLOv11n	Ours
landslide	65.8	56.9	66.8	74.3	66.8	62.7	72.0	**74.3**
storm	87.1	85.7	90.9	90.5	88.1	78.4	90.6	**92.3**

From the average precision (AP) comparison results in Table II, it is evident that different algorithms exhibit significant performance differences in the landslide and storm tasks. YOLOv9-gelan-s and Ours tied for the best performance in the landslide category, both achieving an average precision of 74.3%. In the storm category, Ours achieved the best performance with an average precision of 92.3%, significantly outperforming other algorithms. YOLOv8n and YOLOv11n also performed well in both tasks, with YOLOv8n achieving an average precision of 90.9% in the storm category. YOLOv10n achieved average precisions of 66.8% and 88.1% for the landslide and storm categories, respectively, slightly outperforming YOLOv5s, which scored 65.8% and 87.1%. However, neither of these two algorithms reached the optimal performance. In contrast, YOLOv7-tiny and RT-DETR-18 performed poorly, particularly in the storm category, where RT-DETR-18's average precision was only 78.4%, the lowest among all algorithms. Overall, the Ours algorithm excelled in both tasks, especially in complex scenarios (such as storms), where it demonstrated higher detection accuracy, offering the best overall performance. YOLOv10n and YOLOv5s showed moderate performance, with room for further optimization.

TABLE III. COMPARISON RESULTS OF VARIOUS ALGORITHMS IN PRECISION (P/%)

Algorithms / Classes	YOLOv5s	YOLOv7-tiny	YOLOv8n	YOLOv9-gelan-s	YOLOv10n	RT-DETR-18	YOLOv11n	Ours
landslide	69.4	66.5	73.9	79.3	73.7	72.8	79.7	**77.1**
storm	87.8	86.5	92.6	91.3	89.4	80.4	84.9	**93.8**

From the data in Table III, it can be observed that different algorithms exhibit varying precision in the landslide and storm categories. In the landslide category, YOLOv11n performed the best with a precision of 79.7%, outperforming YOLOv7-tiny (66.5%). In the storm category, Ours demonstrated the best performance, with a precision of 93.8%, significantly surpassing other algorithms, especially RT-DETR-18 (80.4%). YOLO series algorithms (such as YOLOv5s, YOLOv8n, etc.) also performed well in the storm category, but their precision varied greatly in the landslide category. Overall, Ours performed best in storm detection, while YOLOv11n and Ours were the most outstanding in landslide detection. RT-DETR-18 performed relatively poorly in both tasks.

From the recall (R) data in Table IV, Ours demonstrated the best performance in both the landslide and storm categories, with a landslide recall of 67.2% and a storm recall of 89.0%, significantly outperforming other algorithms. YOLOv11n

followed closely, performing well in the landslide category with a recall of 63.7%, and also showing good performance in the storm category. YOLOv5s achieved a recall of 87.7% in the storm category, which was quite notable, but in the landslide category, its recall was 65.3%, slightly lower than that of Ours. In contrast, YOLOv7-tiny and RT-DETR-18 generally had lower recall rates, with YOLOv7-tiny's recall in the landslide category being only 51.4%. Overall, Ours has a significant advantage in recall, making it suitable for applications that require high sensitivity to missed detections.

TABLE IV. COMPARISON RESULTS OF VARIOUS ALGORITHMS IN RECALL (R/%)

Algorithms / Classes	YOLOv5s	YOLOv7-tiny	YOLOv8n	YOLOv9-gelan-s	YOLOv10n	RT-DETR-18	YOLOv11n	Ours
landslide	65.3	51.4	56.4	63.0	63.0	63.4	63.7	**67.2**
storm	87.7	82.2	84.9	86.1	81.3	71.4	87.0	**89.0**

B. Result Visualization

To comprehensively evaluate the superiority of the Ours algorithm in object detection tasks, we conducted a visual comparison with several existing detection algorithms. Although the existing algorithms, such as YOLOv5s, YOLOv7-tiny, YOLOv8n, YOLOv9-gelan-s, YOLOv10n, RT-DETR-18, and YOLOv11n, exhibit certain detection capabilities in different scenarios, they still have limitations in terms of detection accuracy, target confidence, and missed detection phenomena. Therefore, we visually compared the performance of different algorithms in the landslide (landslide) and storm (storm) detection tasks.

Fig. 6 shows the performance comparison of different models in landslide and storm detection. YOLOv5s exhibits basic detection capabilities, but with lower confidence, resulting in some targets being misdetected or missed. YOLOv7-tiny has even lower confidence, with some landslide areas not detected and a higher number of false detections, indicating that it is not robust enough for small targets or complex backgrounds. YOLOv8n shows improvements in landslide detection, but the confidence in the storm category remains relatively low. Overall, it is more stable than YOLOv7-tiny, but there are still some missed detections. In contrast, YOLOv9-gelan-s, YOLOv10n, RT-DETR-18, and YOLOv11n show improvements in detection accuracy, but all exhibit overlapping detection boxes, indicating that these models may have redundant predictions when processing landslide areas, generating multiple detection boxes for the same target region, which affects the final detection results. Additionally, YOLOv11n also experiences missed detections, with some landslide areas not being recognized. Among all models, Ours achieves the best detection performance, with the highest confidence, the most stable detection boxes, a wider detection range for landslide targets, and effectively reduces overlapping detection boxes while avoiding missed detections, resulting in the best overall performance.

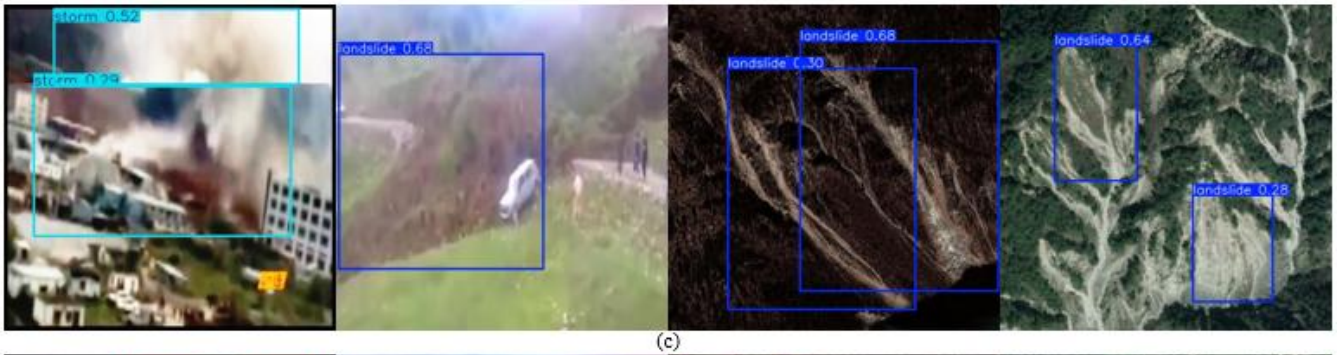
In object detection tasks, the performance of the model is typically evaluated using precision, recall, and F1 score. The F1-Confidence curve reflects the variation in the F1 score at different confidence thresholds, providing an intuitive representation of the model's overall detection capability at different decision boundaries. To compare the performance of different object detection models, we plotted the F1-Confidence curve and conducted an analysis.



(a)



(b)


storm 0.52
landslide 0.68
landslide 0.68
landslide 0.30
landslide 0.64
landslide 0.28

(c)


storm 0.62
landslide 0.56
landslide 0.58
landslide 0.44

(d)


landslide 0.61

(e)


landslide 0.30
landslide 0.79
landslide 0.88

(f)

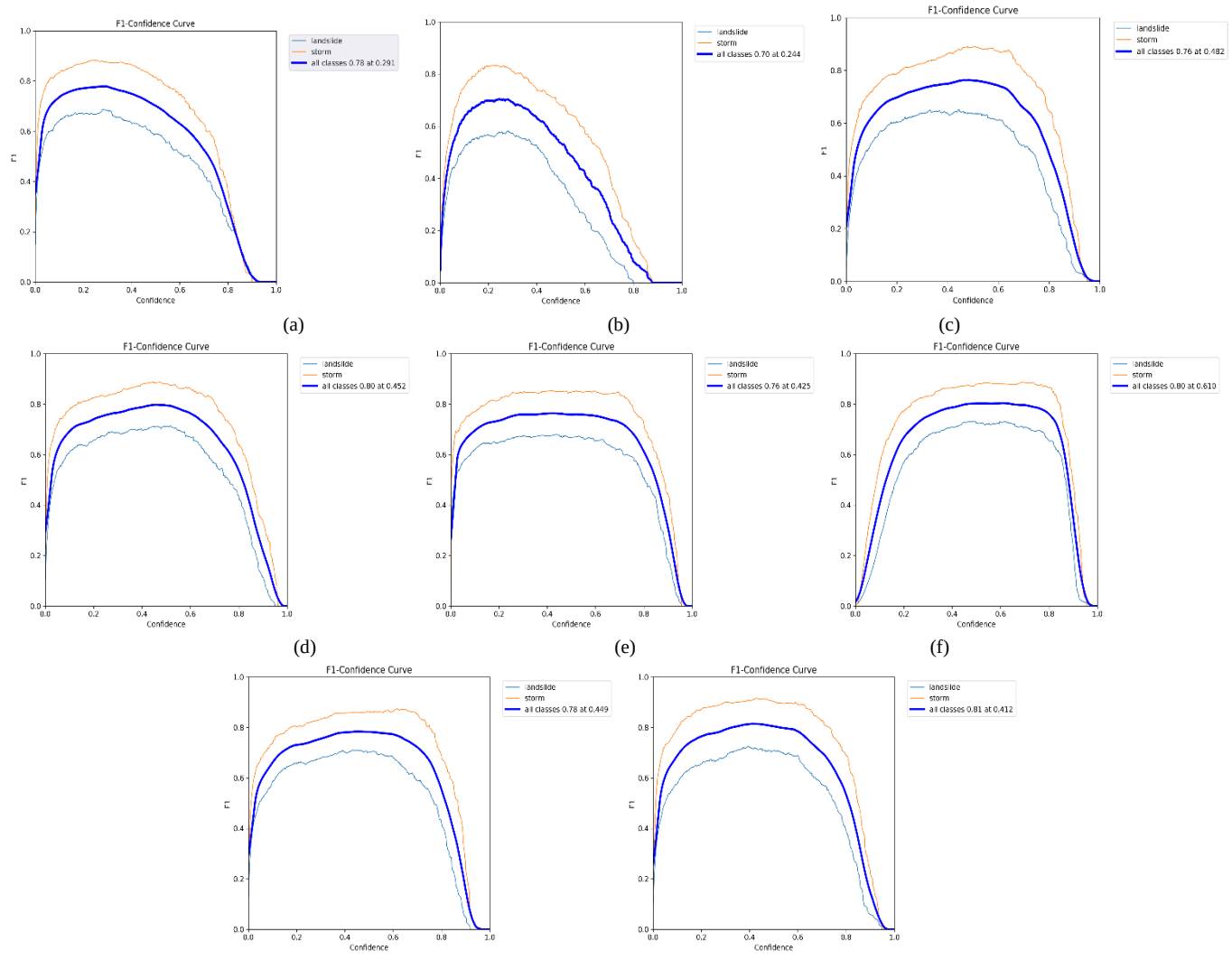

storm 0.61
storm 0.38
landslide 0.79
landslide 0.43

(g)



(h)

Fig. 6. Detection results of different algorithms on the dataset: (a) YOLOv5s; (b) YOLOv7-tiny; (c) YOLOv8n; (d) YOLOv9-gelan-s; (e) YOLOv10n; (f) RT-DETR-18; (g) YOLOv11n; (h) Ours.



(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 7. F1-Confidence Curve: (a) YOLOv5s; (b) YOLOv7-tiny; (c) YOLOv8n; (d) YOLOv9-gelan-s; (e) YOLOv10n; (f) RT-DETR-18; (g) YOLOv11n; (h) Ours.

To comprehensively evaluate the detection capabilities of different object detection models at various confidence thresholds, we plotted the F1-Confidence curve to visually demonstrate the variation in F1 scores across models. Fig. 7(a) to Fig. 7(h) correspond to YOLOv5s, YOLOv7-tiny, YOLOv8n, YOLOv9-gelan-s, YOLOv10n, RT-DETR-18, YOLOv11n, and our model (Ours), used to compare their performance at different decision boundaries. The results in Fig. 7 show that Ours achieves the highest F1 score of 0.81 at a confidence threshold of 0.412, outperforming all other models

and demonstrating the best performance. YOLOv9-gelan-s, RT-DETR-18, and YOLOv11n also performed well, but still fell short of Ours. YOLOv5s and YOLOv10n achieved F1 scores of 0.78 and 0.76, respectively, showing acceptable overall performance, better than YOLOv7-tiny (0.70) and YOLOv8n (0.76), but still not surpassing Ours. YOLOv7-tiny had the lowest F1 score at only 0.70, indicating weaker detection capabilities. Overall, Ours maintained a high F1 score across different confidence ranges, with a low false detection rate, indicating stronger object detection capabilities.

This study improved the YOLOv11n object detection algorithm and evaluated its training and validation processes. To analyze the model's convergence and detection performance, we plotted the loss curve and the trends of key evaluation metrics. The loss curve illustrates the optimization of the model during the training process, while metrics such as precision, recall, and mAP reflect the model's detection capabilities. Through these curves, we validated the stability and performance improvement of the modified model.

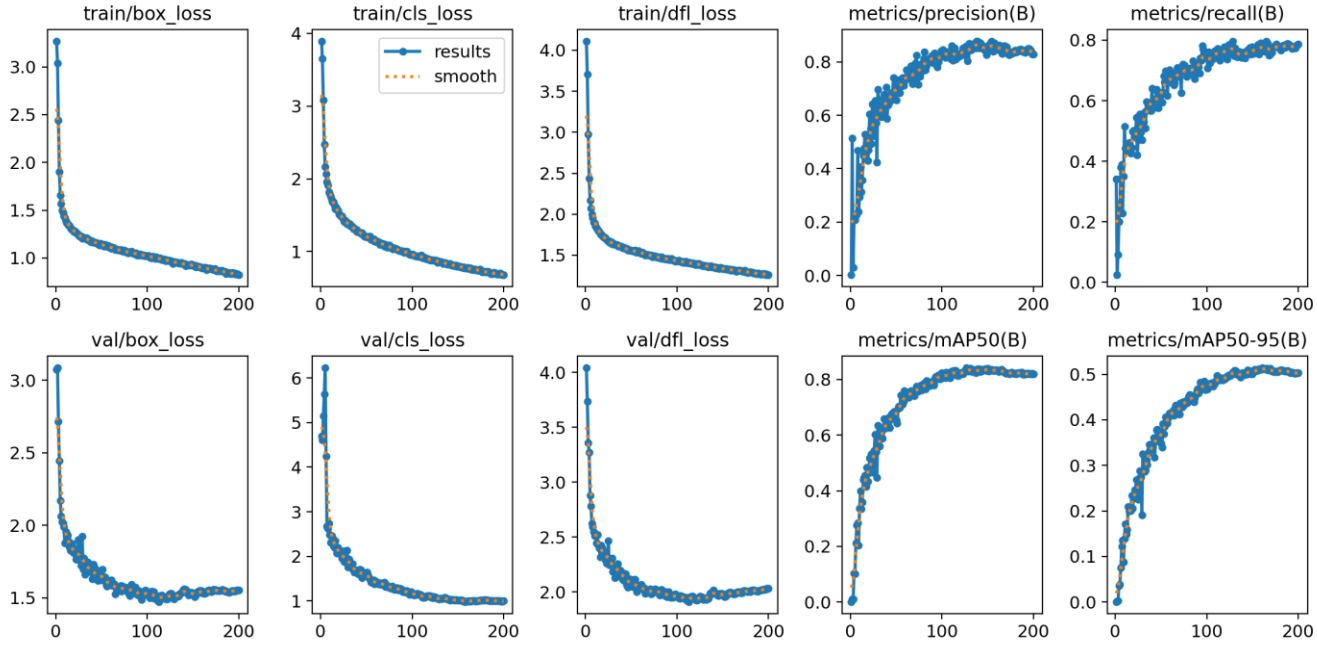
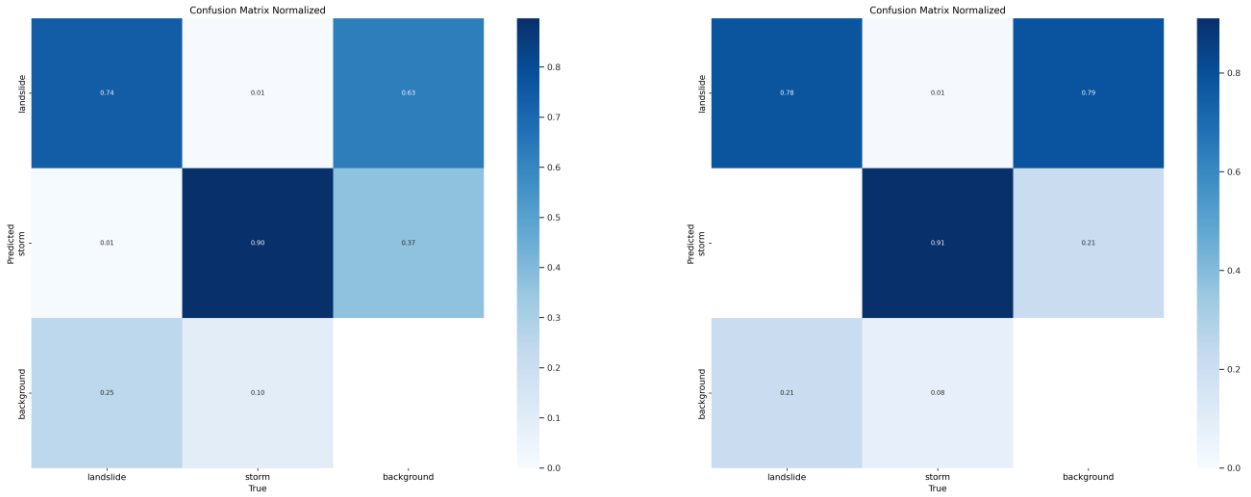


Fig. 8. Loss reduction trends and performance metric variations of the improved YOLOv11n algorithm during training and validation.

Fig. 8 illustrates the loss reduction trend and changes in key evaluation metrics during the training and validation process of the improved YOLOv11n algorithm, used to assess the model's convergence and detection performance. From the loss curves, both training losses (box_loss, cls_loss, dfl_loss) and validation losses (val/box_loss, val/cls_loss, val/dfl_loss) show a steady downward trend, indicating that the model continuously optimizes its object detection capabilities during training without significant overfitting or underfitting. In terms of model evaluation metrics, precision and recall steadily increase and stabilize, suggesting that the improved model gradually enhances its target classification and detection abilities. The mAP50 is close to 0.833, and the mAP50-95 is approximately 0.516, demonstrating the model's good detection performance and strong generalization ability at different IoU thresholds. Overall, the improved YOLOv11n algorithm outperforms the original version, but its small target detection capabilities and mAP50-95 metric can still be further improved by adjusting the learning rate, optimizing data augmentation strategies, or enhancing non-maximum suppression (NMS) processing.

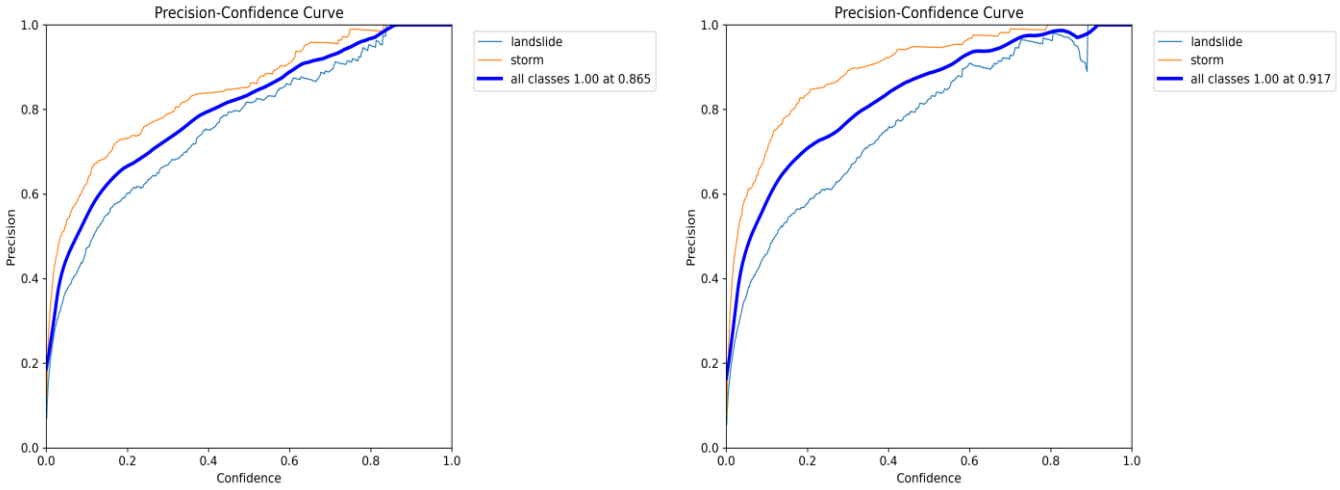
To more clearly evaluate the model's classification ability, the confusion matrix is used to visually quantify the model's prediction accuracy and misclassification rates for different categories. To assess the detection performance of the YOLOv11n algorithm and its improved version in the landslide and storm categories, we plotted the normalized confusion matrix. This allows for a comparative analysis of the classification accuracy and misclassification rates of the two models, in order to verify the performance improvement of the modified model.

Fig. 9 shows the confusion matrices of the YOLOv11n algorithm and the improved YOLOv11n algorithm, used to evaluate the model's classification performance on landslide and storm targets. From the results, the improved YOLOv11n algorithm shows improvements in both landslide and storm detection. The accuracy of landslide detection increased from 0.74 to 0.78, and the accuracy of storm detection increased from 0.90 to 0.91, while the false detection rate decreased. In particular, the proportion of storms misclassified as background reduced from 0.37 to 0.21, indicating that the improved model is more stable in storm detection. Additionally, the background misclassification rate also decreased, with the probability of landslides being misidentified as background dropping from 0.25 to 0.21, suggesting that the improved algorithm has optimized false detection suppression.



(a) (b)

Fig. 9. Normalized confusion matrix: (a) YOLOv11n algorithm; (b) Improved YOLOv11n algorithm.

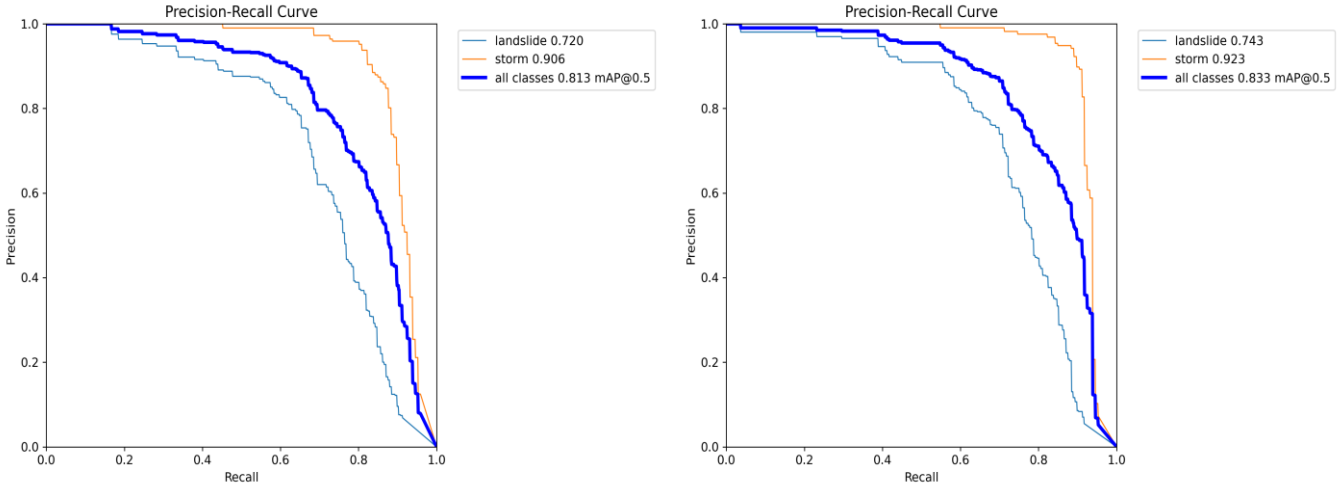


(a) (b)

Fig. 10. Precision-confidence curves: (a) YOLOv11n algorithm; (b) Improved YOLOv11n algorithm.

Fig. 10 presents the Precision-Confidence curves for the YOLOv11n algorithm [Fig. 10(a)] and the improved YOLOv11n algorithm [Fig. 10(b)], used to evaluate the variation in precision at different confidence thresholds. The left panel [Fig. 10(a)] shows the original YOLOv11n, while the right panel [Fig. 10(b)] shows the improved YOLOv11n. The curves correspond to landslide, storm, and all classes. Overall, as the confidence threshold increases, precision gradually improves, indicating a reduction in false positives at higher confidence levels and more accurate predictions. Compared to the original model, the improved YOLOv11n shows an increase in precision across all confidence ranges, with the confidence threshold corresponding to the highest precision increasing from 0.865 to 0.917. This suggests that the improved model maintains a higher level of precision at higher confidence thresholds, thereby reducing false positives caused by low-confidence predictions.

Fig. 11 presents the Precision-Recall (P-R) curves for the YOLOv11n algorithm [Fig. 11(a)] and the improved YOLOv11n algorithm [Fig. 11(b)], used to assess the variation in precision at different recall rates. The left panel [Fig. 11(a)] shows the original YOLOv11n, while the right panel [Fig. 11(b)] shows the improved YOLOv11n. The P-R curves for landslide, storm, and all classes are plotted, with the performance at mAP@0.5 indicated for each class. Overall, the mAP@0.5 of the improved model increased from 0.813 to 0.833 across all classes, with the mAP of the storm class rising from 0.906 to 0.923 and the mAP of the landslide class increasing from 0.720 to 0.743, indicating an improvement in detection performance. The P-R curves show that the storm class consistently outperforms the landslide class in detection, while the improved YOLOv11n maintains a slower decline in precision at higher recall rates, suggesting enhanced detection ability at higher recall levels and a reduction in false negatives. Overall, the improved model achieves better performance in balancing precision and recall.



(a) (b)

Fig. 11. Precision-recall curves: (a) YOLOv11n algorithm; (b) Improved YOLOv11n algorithm.

C. Ablation Study

To evaluate the impact of different modules on the YOLOv11n object detection performance, we conducted ablation experiments and performed a comparative analysis of detection accuracy, recall rate, mAP@0.5, inference speed (FPS), computational complexity (GFLOPS), and parameter count (Params) for each improvement. By progressively introducing modules such as C3k2-SCConv, GhostConv, and SimAM, we assess their contributions to model performance and whether the improvements can maintain high detection accuracy while also considering inference speed and computational efficiency. The experimental results, as shown in Table V, present the performance of each configuration across various metrics.

TABLE V. ABLATION STUDY RESULTS

number	Experiment	Precision/%	Recall/%	mAP@0.5/%	FPS (f/s)	GFLOPS	Params (M)
1	YOLOv11n	82.3	75.3	81.3	147.0	6.3	2.58
2	YOLOv11n+C3k2-SCConv	85.6	74.6	82.2	102.8	6.4	2.73
3	YOLOv11n+C3k2-SCConv+GhostConv	85.9	75.3	82.8	130.3	6.3	2.72
4	YOLOv11n+C3k2-SCConv+GhostConv+SimAM	**85.5**	**78.1**	**83.3**	**128.5**	**6.3**	**2.72**

Table V presents the ablation experiment results for YOLOv11n and its various improvements, evaluating the impact of each configuration on precision, recall, mAP@0.5, frame rate (FPS), computational complexity (GFLOPS), and parameter count (Params). The baseline model, YOLOv11n, achieves the fastest inference speed with an mAP@0.5 of 81.3% and FPS of 147.0 f/s, but there is still room for improvement in detection accuracy. After adding C3k2-SCConv, precision increases to 85.6%, and mAP@0.5 slightly improves, but recall decreases, and FPS drops to 102.8 f/s, indicating that the increased computational load impacts inference efficiency. The further introduction of GhostConv increases precision to 85.9%, recall recovers by 5.3%, and mAP@0.5 rises to 82.8%, while FPS reaches 130.3 f/s, suggesting that GhostConv effectively reduces computational redundancy and enhances inference speed. Finally, with the addition of SimAM, recall further improves to 78.1%, and mAP@0.5 increases to 83.3%. Overall, the gradual inclusion of C3k2-SCConv, GhostConv, and SimAM significantly enhances detection accuracy while maintaining high inference speed, with minimal impact on computational complexity and parameter count.

VI. DISCUSSION

There are several limitations in the current study. First, the dataset lacks images captured under extreme weather conditions such as heavy rain or snow, which may weaken the model's generalization ability in complex real-world environments. Second, the sample distribution of landslide and storm categories may be imbalanced, potentially leading to suboptimal detection performance for underrepresented classes. Moreover, the current research primarily focuses on the detection of landslides and storms, while the model's ability to identify other types of geological disasters, such as debris flows or earthquakes, has not been sufficiently validated. Additionally, the model is relatively complex, which not only increases the time cost for training and inference but also poses challenges for deployment and maintenance. Future research can address these issues in two main directions. One is the expansion and augmentation of the dataset, by incorporating a larger number of images from various scenarios and weather conditions to enhance the model's robustness and generalization. The other is the optimization of the model architecture and its real-time performance, by exploring more efficient network designs and optimization algorithms to reduce computational complexity and parameter size.

VII. CONCLUSION

This study optimizes the YOLOv11n object detection algorithm by introducing the GhostConv module to optimize the convolutional computations, reducing redundant calculations while significantly improving the model's operational efficiency without compromising feature extraction capability. The introduction of the C3k2-SCConv module further optimizes the feature extraction process, effectively enhancing the representation of key features in landslide images, while maintaining optimal computational performance. The incorporation of the SimAM attention mechanism adapts the model's focus on different region features, allowing the network to more effectively perceive critical information in landslide images, thereby further improving detection accuracy. Experimental results show that the improved model outperforms other algorithms in terms of mAP@0.5, precision, and recall, demonstrating superior detection performance and computational efficiency.

REFERENCES

[1] Casagli N, Intrieri E, Tofani V, et al. Landslide detection, monitoring and prediction with remote-sensing techniques[J]. Nature Reviews Earth & Environment, 2023, 4(1): 51-64.

[2] Savvaidis P D. Existing landslide monitoring systems and techniques[J]. From stars to earth and culture, 2003: 242-258.

[3] Ferrigno F, Gigli G, Fanti R, et al. GB-InSAR monitoring and observational method for landslide emergency management: the Montaguto earthflow (AV, Italy)[J]. Natural Hazards and Earth System Sciences, 2017, 17(6): 845-860.

[4] Qiu D, Wang L, Luo D, et al. Landslide monitoring analysis of single-frequency BDS/GPS combined positioning with constraints on deformation characteristics[J]. Survey Review, 2019, 51(367): 364-372.

[5] Zhang, Shengdong, et al. "Generative adversarial and self-supervised dehazing network." IEEE Transactions on Industrial Informatics 20.3 (2023): 4187-4197.

[6] Zhang, Shengdong, et al. "Semantic-aware dehazing network with adaptive feature fusion." IEEE Transactions on Cybernetics 53.1 (2021): 454-467.

[7] Liu, Yun, et al. "Single nighttime image dehazing based on unified variational decomposition model and multi-scale contrast enhancement." Engineering Applications of Artificial Intelligence 116 (2022): 105373.

[8] Han, Zheng, et al. "A novel Dynahead-Yolo neural network for the detection of landslides with variable proportions using remote sensing images." Frontiers in Earth Science 10 (2023): 1077153.

[9] Yang, Yueheng, et al. "Lightweight attention-guided YOLO with level set layer for landslide detection from optical satellite images." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17 (2024): 3543-3559.

[10] Liu, Qing, et al. "Se-yolov7 landslide detection algorithm based on attention mechanism and improved loss function." Land 12.8 (2023): 1522.

[11] Liu, Xiaohui, Ling Xu, and Jinyu Zhang. "Landslide detection with Mask R-CNN using complex background enhancement based on multi-scale samples." Geomatics, Natural Hazards and Risk 15.1 (2024): 2300823.

[12] Zhang, Wenjie, et al. "LS-YOLO: A novel model for detecting multiscale landslides with remote sensing images." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17 (2024): 4952-4965.

[13] Yu, Zhengbo, Ruichun Chang, and Zhe Chen. "Automatic detection method for loess landslides based on GEE and an improved YOLOX algorithm." Remote Sensing 14.18 (2022): 4599.

[14] Hou, Heyi, et al. "A universal landslide detection method in optical remote sensing images based on improved YOLOX." Remote Sensing 14.19 (2022): 4939.

[15] Cheng, Gong, et al. "Advances in deep learning recognition of landslides based on remote sensing images." Remote Sensing 16.10 (2024): 1787.

[16] Chandra, Naveen, et al. "A Novel Attention-Based Generalized Efficient Layer Aggregation Network for Landslide Detection from Satellite Data in the Higher Himalayas, Nepal." Remote Sensing 16.14 (2024): 2598.

[17] Arnab, Anurag, et al. "Vivit: A video vision transformer." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[18] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[19] Wang, Dingran, et al. "SDS-YOLO: An improved vibratory position detection algorithm based on YOLOv11." Measurement 244 (2025): 116518.

[20] Wang, Tao, and Shiqing Zhang. "DSC-Ghost-Conv: A compact convolution module for building efficient neural network architectures." Multimedia Tools and Applications 83.12 (2024): 36767-36795.

[21] Li, Jiafeng, Ying Wen, and Lianghua He. "Scconv: Spatial and channel reconstruction convolution for feature redundancy." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[22] Yang, Lingxiao, et al. "Simam: A simple, parameter-free attention module for convolutional neural networks." International conference on machine learning. PMLR, 2021.

[23] Wang, Yaqing, et al. "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53.3 (2020): 1-34.

[24] Lu, Jiang, et al. "A survey on machine learning from few samples." Pattern Recognition 139 (2023): 109480.

[25] Zhong, Zhun, et al. "Random erasing data augmentation." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.

[26] Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).

[27] Kou R, Wang C, Peng Z, et al. Infrared small target segmentation networks: A survey[J]. Pattern Recognition, 2023, 143: 109788.

[28] Chen, Chen, et al. "Improved YOLOv5s model for key components detection of power transmission lines." arxiv preprint arxiv:2502.06127 (2025).

[29] Liu, Le, et al. "Improved YOLOv7 Based on Reparameterization and Attention Mechanism for Continuous Casting Slab Detection." IEEE Transactions on Instrumentation and Measurement (2025).

[30] Zhong, Jiaqi, et al. "Improved real-time object detection method based on YOLOv8: a refined approach." Journal of Real-Time Image Processing 22.1 (2025): 1-13.

[31] Anandakrishnan, Jayakrishnan, et al. "Precise Spatial Prediction of Rice Seedlings From Large Scale Airborne Remote Sensing Data Using Optimized Li-YOLOv9." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024).

[32] Meng, Zhichao, et al. "YOLOv10-pose and YOLOv9-pose: Real-time strawberry stalk pose detection models." Computers in Industry 165 (2025): 104231.

[33] Zhou, Chuncheng, et al. "MSRT-DETR: A novel RT-DETR model with multi-scale feature sequence for cell detection." Biomedical Signal Processing and Control 103 (2025): 107378.

Integration of 2D-CNN and LSTM Networks for Enhanced Image Processing and Prediction in Alzheimer's Disease

Aya Mohamed Abd El-Hamed, Mohamed Aborizka
Department of Computer Science-College of Computing and Information Technology,
Arab Academy for Science, Technology, and Maritime Transport, Sheraton, 11757, Cairo, Egypt

***Abstract*—The early diagnosis of Alzheimer's disease remains a major challenge due to the complexity of magnetic resonance image interpretation and the limitations of existing diagnostic models. The slow memory loss associated with the gradual loss of thinking abilities, known as Alzheimer's disease, is the most common element of the illness. Effective early diagnosis is therefore essential to treatment; unfortunately, the traditional diagnostic procedure, which involves analyzing magnetic resonance images, is a complex process and prone to mistakes. This study aims to successfully merge these cognitive models with advanced deep learning techniques to enhance the diagnostic capabilities of Alzheimer's disease using a fusion model with 3-dimensional convolutional neural networks and long short-term memory networks. The proposed approach uses three-dimensional convolutional neural networks to extract intricate features from volumetric magnetic resonance images, while long short-term memory networks analyze sequential data to identify key temporal patterns that indicate the progression of Alzheimer's disease. The dataset used in this study is the Alzheimer's Disease Neuroimaging Initiative dataset, which contains magnetic resonance images labeled into four categories: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The dataset consists of 6,400 magnetic resonance images in total, split into training (70%), validation (15%), and testing (15%) sets. These outcomes demonstrate that the hybrid model improves predictive accuracy significantly over current benchmarks on this topic. This study highlights the importance of introducing deep learning models into clinical practice, thereby providing an efficient tool for early-stage Alzheimer's disease diagnosis, ultimately improving patient outcomes through early and accurate intervention.**

Keywords*—*Alzheimer's disease; magnetic resonance imaging; two-dimensional convolutional neural network; long short-term memory; deep learning; early detection

I. Introduction

One well-known neurological disorder that affects memory and cognitive function is Alzheimer's disease (AD). The disease is named after Dr. Alois Alzheimer, a well-known German expert who initially discovered the basic pathophysiology of the underlying pathology in 1906 [1].

The prevalence of Alzheimer's disease increases significantly with age, altogether influencing the elderly. The trademark of this condition is the amassing of anomalous protein totals inside the brain, which continuously disables cognitive work and creates genuine challenges with day-by-day tasks [2].

Our understanding of AD has advanced impressively over the past century, especially with respect to the part that the amyloid-beta protein plays within the early stages of the infection. However, there are critical concerns related to the worldwide rise in AD predominance. In recent years, machine learning has opened up unused roads for the early discovery and administration of AD. By leveraging cutting-edge algorithms and large-scale medical imaging datasets, AI techniques have significantly enhanced diagnostic workflows and accuracy [28]. Machine-learning models can distinguish subtle patterns in neuroimaging, genetic data, and cognitive assessments that may indicate early signs of this insidious disease [3].

The novelty of this study lies in the integration of two-dimensional convolutional neural networks (2D-CNNs) and long short-term memory (LSTM) networks to enhance both spatial feature extraction and temporal pattern recognition. This combined approach provides significant improvement over prior research that typically employs CNNs or LSTMs separately, leading to more robust Alzheimer's disease classification. Additionally, we introduce a comprehensive dataset preprocessing pipeline and apply SMOTE to handle class imbalance, an aspect not fully explored in existing literature. This model integrates 2D-CNN for efficient feature selection from MRI scans and LSTM networks for the temporal analysis of sequential MRI data. A thorough performance evaluation is conducted using metrics such as accuracy, recall, and F1-score, demonstrating the superiority of our method over conventional techniques. Through the application of these cutting-edge strategies, our objective is to supply healthcare experts with a robust diagnostic tool that not only improves the accuracy of Alzheimer's disease detection but also enhances the understanding of disease progression. Additionally, the physicians will be motivated to take good decisions about care, ultimately leading to improved patient outcomes.

While several studies have applied CNNs or RNNs individually for medical image classification, their effectiveness in capturing both spatial and temporal dependencies remains limited. This study addresses this gap by combining 2D-CNN with LSTM to leverage both spatial features and sequential context from MRI slices.

The structure of this study is as follows: Section II presents a comprehensive review of related work on deep learning models used for Alzheimer's disease diagnosis. Section III outlines the proposed methodology, including model architecture, dataset details, and preprocessing steps. Section IV provides the experimental results, and Section V discusses the model's performance using standard evaluation metrics. Finally, Section VI concludes the study and highlights potential directions for future work.

II. Literature Review

This section discusses more complicated deep learning techniques, specifically Convolutional Neural Networks (CNNs) and LSTM networks, for the early diagnosis and detection of AD using neuroimaging data. The literature review section explores various studies that have integrated CNNs, LSTMs, and Transformers for AD diagnosis. However, most prior works have either focused on CNNs for spatial feature extraction or LSTMs for temporal sequence analysis, but have not effectively combined them.

Recently, many studies highlighted the important role of merging CNNs and LSTMs when analyzing MRI scans for AD diagnosis. A multimodal fusion model was created that combined a CNN for spatial feature extraction and an LSTM for temporal sequence analysis, achieving an accuracy of 92.3% using the Alzheimer's Disease Neuroimaging Initiative dataset [4].

Similarly, the research done by Hu et al. [8] introduced a Transformer-based model, VGG-TSwinformer, for predicting the progression from Mild Cognitive Impairment (MCI) to Alzheimer's Disease. Although, retrieving a lower accuracy of 77.2%, this model was the first to merge CNNs with Transformer models, showing how important it is to control the temporal aspects of longitudinal medical data. Despite its drawbacks, this innovative approach indicates a promising avenue for further research in AD diagnosis. Additionally, recent AI-driven methodologies have shown substantial promise in predicting Alzheimer's progression using advanced deep learning frameworks [14].

In recent research, explainability in AI models has become increasingly important. The XGBoost model was combined with the SHAP (SHapley Additive exPlanations) method in the study by Bogdanovic et al. [6] to create clear conclusions about the model's predictions. This approach not only achieved a high F1-score of 0.84, but it also increased the clarity of the diagnostic process, making it more dependable and accessible for clinical application. In a recent study, an explainable AI-based model for Alzheimer's disease prediction was developed using a multimodal dataset [7].

This outcome shows the efficiency of combining clinical, neuroimaging, and psychological data in AD prediction. Using a different data modality and focusing on explainability shows how the complexity and accuracy of the current AD diagnostic techniques have enhanced [7]. Another contribution has been made by Qiu et al. [8] who developed a significant multimodal deep learning framework that combined MRI data with clinical features using CNN and CatBoost models. This study shows the importance of merging various data sources to enable an extensive diagnosis of AD. Additionally, Ngnamsie Njimbouom et al. [9] has created a multi-modal model by using artificial neural networks (ANN) and CNN to predict dental caries.

Although they mainly focused on dentistry applications, their methodology is applied to AD research as it shows the value of multi-modality in improving prediction accuracy [9].

Hoang et al. [10] applied Vision Transformer models to predict the transition from mild cognitive impairment to Alzheimer's disease. Their findings revealed an accuracy of 83.27% and provided visual representations of key brain regions associated with the progression of AD, consequently improving the interpretability of these predictive models [10]. Pang et al. [11] have provided a detailed analysis of deep learning-based medical imaging report production, which included AD diagnostic reports. Their research shows two vital components in the development of effective tools for AD detection: the challenges associated with imbalanced datasets and the significance of interdisciplinary collaboration.

Javeed et al. [12] conducted a systematic evaluation of machine learning techniques for dementia prediction, focusing on data modalities such as voice, clinical features, and picture data. Their results showed the effectiveness of image-based machine learning models for dementia prediction, but they also identified limitations such as model overfitting and the need for larger datasets.

Finally, Khatri and Kwon [13] used PET scans to predict the progression of AD through an explainable Vision Transformer model that includes self-supervised learning. Their approach achieved a significant accuracy of 92.31%. It highlights the value of integrating robust predictive capabilities with interpretability in clinical environments.

The deep learning techniques indicate their importance in all this research, especially the integration of CNNs, LSTMs, and Transformers, which have improved the ability to diagnose AD with greater accuracy and transparency. Future advancements in the early detection and management of AD are well-founded in this field of study by handling the challenges of data diversity and model explainability.

Table I lists the main techniques and associated accuracy of many studies related to the early detection and classification of AD using several deep learning models.

TABLE I. Related Works with Methodologies and Classes

Study	Methodology	No. of Class	Accuracy
Haq et al., 2024 [5]	CNN-LSTM (Multimodal Fusion)	Three Class (AD, MCI, NC)	92.30%
Hu et al., 2023 [8]	VGG-TSwinformer (Transformer + CNN)	Binary Class (MCI vs pMCI)	77.20%
Bogdanovic et al., 2022 [6]	XGBoost + SHAP (Explainable ML)	Multi-Class (AD, MCI, NC, etc.)	84.00%

Jahan et al., 2023) [3]	Random Forest (Multimodal, Explainable AI)	Four Class (AD, MCI, NC, LMCI)	98.81%
(Atito et al., 2021) [22]	Vision Transformer with Self-Supervised Learning	Binary Class (MCI vs AD)	92.31%
Qiu et al., 2022) [8]	Multimodal Deep Learning Framework (CNN + CatBoost)	Four Class (NC, MCI, AD, nADD)	AUC 0.971
(Ngnamsie Njimbouom et al., 2022 [9]	Multi-Modal Dental Caries Prediction (CNN + ANN)	Binary Class (Caries vs No Caries)	High Accuracy
Hoang et al., 2023 [10]	Vision Transformers (ViT)	Binary Class (MCI vs AD)	83.27%
Pang et al., 2023 [7]	Deep Learning-based Report Generation	Multi-Class (Various)	Survey
Javeed et al., 2023 [6]	Explainable Vision Transformer (Self-Supervised)	Binary Class (MCI vs AD)	Systematic Review
Khatri & Kwon, 2023) [13]	Systematic Review of ML for Dementia Prediction	Various (AD, MCI, NC, Voice Data)	92.31%

III. METHODOLOGY

To improve clarity, a flowchart illustrating the workflow of the proposed model is shown in Fig. 1. Additionally, Algorithm 1 provides a structured representation of the key steps in the model's training and classification process.



Fig. 1. Workflow of the hybrid 2D-CNN + LSTM model for AD classification.

Algorithm 1: Hybrid 2D-CNN + LSTM Model for AD Classification

1. Data Preprocessing:
   - Load MRI images from the ADNI dataset.
   - Resize images to 176x176 pixels.
   - Normalize pixel values to [0,1].
2. Model Training:
   - Pass images through 2D-CNN layers for feature extraction.
   - Flatten extracted features and feed them into an LSTM layer.
3. Evaluation:

-Compute accuracy, precision, recall, F1-score. Evaluate the model using standard classification metrics: accuracy, precision, recall, and F1-score.

-Generate confusion matrix and Receiver Operating Characteristic (ROC) curve.

The following workflow diagram illustrates the steps involved in the Hybrid two-dimensional convolutional neural network (2D-CNN) + long short-term memory (LSTM) model for AD classification. It details the data preprocessing, model training, and evaluation phases.

A. Adopted Base Model Architecture

2D-CNN or Conv2D: A 2D Convolutional Neural Network is a type of deep learning model specifically designed for processing two-dimensional data, such as medical imaging scans. In a 2D-CNN, the convolutional layer applies an arrangement of channels (or filters) over the input information to capture spatial chains of command and patterns, such as edges, surfaces, and more complex highlights, by sliding the channels over the width and height measurements of the input. This makes a difference in identifying critical highlights inside the information while protecting spatial connections, making 2D-CNNs especially successful for assignments like image recognition [15].

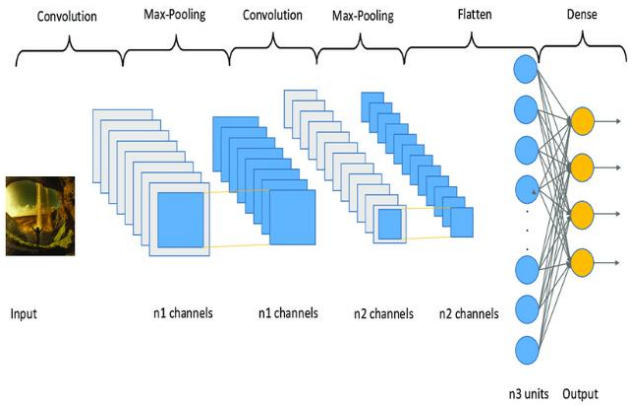


Fig. 2. 2D-CNN architecture.

Fig. 2 illustrates the architecture of a 2D-CNN used for image classification. The process begins with an input image, which undergoes multiple convolutional layers to extract features, followed by max-pooling layers that reduce spatial dimensions while preserving key information. The extracted features are then flattened into a 1D vector and passed through a fully connected (dense) layer for classification. The final output layer produces the classification results, making the architecture suitable for tasks such as object detection and image recognition.

LSTM Architecture design: LSTM is built with a feedback mechanism that allows it to preserve information for long durations. LSTM essentially forms the current input by considering the past yield, hence putting away the future yield in its short-term memory. The cell state is essential to the functioning of LSTM, as it dictates the data that must be preserved or eliminated [16].

The LSTM network functions using three main gates and two essential mechanisms. The Forget Gate decides which information to keep or eliminate from long-term memory. The Input Gate gathers new data and evaluates its relevance relative to existing stored information. Lastly, the Output Gate produces the final output based on the decisions of the other two gates. These gates function using the sigmoid and hyperbolic tangent (tanh) activation functions. A schematic representation of the LSTM architecture is illustrated in Fig. 3.

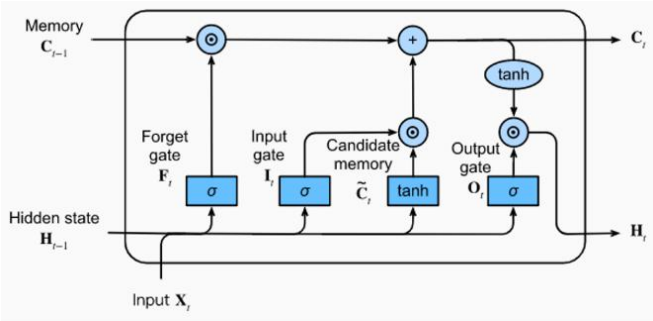


Fig. 3. LSTM architecture.

B. Proposed Hybrid Model

In this study, CNNs and LSTM networks were combined, creating a hybrid deep learning model that enhances image classification as shown in Fig. 4. The first three convolutional blocks each consisted of a convolutional layer followed by a max-pooling layer.

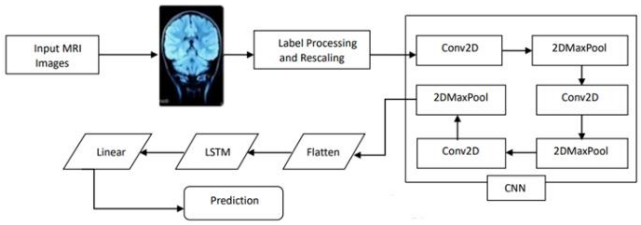


Fig. 4. Proposed hybrid model for Alzheimer's detection.

ReLU activation and 32, 64, and 128 filters with a 3x3 kernel size are used in the convolutional layers. These layers are crucial for the extraction of low-level to high-level features from the input images, such as edges, textures, and complex patterns. Each convolutional layer is followed by a max-pooling operation that uses a 2x2 pooling window with 'same' padding is applied to efficiently reduce the spatial dimensions of the feature maps and reduce the possibility of overfitting.

To further mitigate overfitting, a dropout layer with a rate of 0.3 is applied after the convolutional and pooling layers. To prevent the model from becoming overly dependent on a single feature, this regularization method randomly disables a segment of the input units during the training phase. After the convolutional layers, the model's output is converted into a one-dimensional vector, which is subsequently reshaped to enable sequential processing by the LSTM layer.

The 128-unit LSTM layer is important for identifying temporal patterns and dependencies in the sequential data extracted from MRI images. By enabling the model to understand the relationships and feature development throughout the input data, this layer improves the model's ability to provide accurate predictions. The flattened and reshaped feature vectors are processed by the LSTM layer, which efficiently preserves and makes use of this temporal information for additional analysis.

Finally, the model divides into multiple dense layers, starting with a fully linked layer with 1024 units that uses the ReLU activation function. This layer uses the features retrieved from the previous convolutional and LSTM layers to perform high-level reasoning.

An additional dropout layer is applied to maintain model flexibility. The architecture has been completed with a four units output layer that represents the number of classes in the classification operation. The model's final predictions are generated by this layer using the softmax activation function, producing a probability distribution over the classes.

C. Dataset

This study organized the dataset into two main directories, training and testing, each containing subfolders corresponding to numerous stages of AD. The total number of MRI images were divided across the four classes as follows:

- Mild Demented: 717 train + 179 test
- Moderate Demented: 52 train + 12 test
- Non-Demented: 2,560 train + 640 test
- Very Mild Demented: 1,792 train + 448 test

This structure supports effective model training and evaluation, ensuring that it can accurately classify images corresponding to various stages of AD.

D. Data Preprocessing and Parameters

To enhance model generalization and prevent overfitting, data augmentation techniques were applied, including rotation, flipping, brightness adjustment, and contrast normalization. These transformations ensured that the model learns robust features independent of variations in MRI images. Additionally, The Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic samples for underrepresented classes, mitigating class imbalance and improving classification performance. In this study, the hybrid 2D-CNN + LSTM architecture was selected to leverage the strengths of both models. CNNs excel at extracting spatial features from medical images, while LSTMs are effective in capturing temporal dependencies. Given that Alzheimer's disease progresses over time, integrating LSTM with CNN allows for improved detection of subtle changes in MRI images that might indicate early stages of the disease. Unlike traditional CNNs that process static images, the LSTM component enabled our model to analyze sequential MRI scans, capturing temporal variations that are crucial for accurate diagnosis. Furthermore, this hybrid approach outperforms standalone CNN and RNN models, as demonstrated in our experimental results.

Several preprocessing steps were required before using the dataset into the machine learning model. First, all images were resized to a consistent 176x176 resolution to maintain

consistency across the dataset. In order to improve the model's performance during training, normalization is applied by rescaling pixel values from the range [0, 255] to [0, 1].

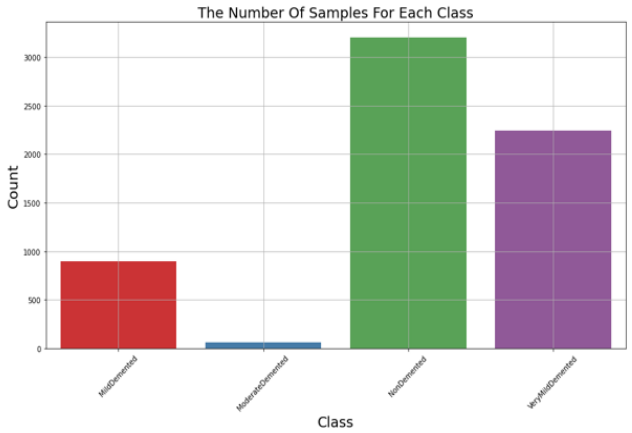


Fig. 5. The number of samples for each class before addressing the imbalance.

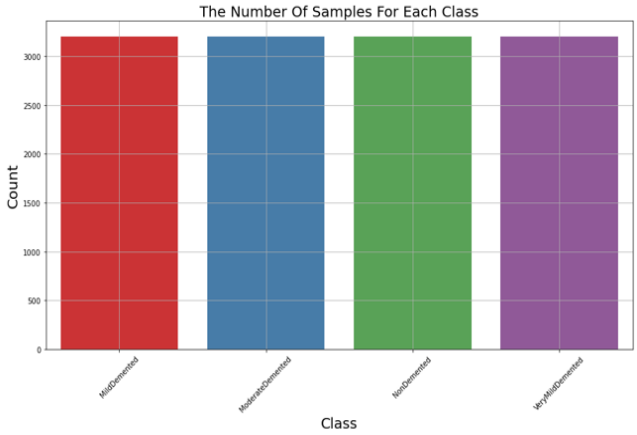


Fig. 6. The number of samples for each class after addressing the imbalance.

The distribution of samples across the four classes before applying SMOTE is shown in Fig. 5. The Synthetic Minority Over-Sampling Technique (SMOTE) has addressed the dataset's inherent class imbalance. By generating synthetic examples for underrepresented classes, SMOTE balances the dataset and enhances the model's ability to generalize across all categories. The class distribution after applying SMOTE is shown in Fig. 6. According to Joloudari et al., an effective class imbalance learning method based on SMOTE and CNNs was developed [17].

70% of the dataset was used for training, with the remaining 30% evenly divided between validation and testing. This balanced split ensured that the model can be thoroughly evaluated and validated during training. The model took input in the form of 176x176 pixel images with three color channels corresponding to the RGB color space. Training was performed with a batch size of 32 over 50 epochs, with a learning rate of 0.001. The loss function employed was categorical cross-entropy, while the Adam optimizer was used to efficiently handle sparse gradients.

These preprocessing steps, combined with the balanced dataset, provide a strong foundation for training a deep learning model capable of accurately classifying the different stages of AD.

IV. Results

A. Environmental Setup

The TensorFlow Keras package is used for classification, with the model executed on an AMD Ryzen 7 and 32 GB RAM. All computations are performed in Python. The proposed technique is evaluated using accuracy, precision, recall, F1-score, specificity, and confusion matrix.

The data visualization of classification results is shown in Fig. 7, illustrating the distribution of predictions across categories.

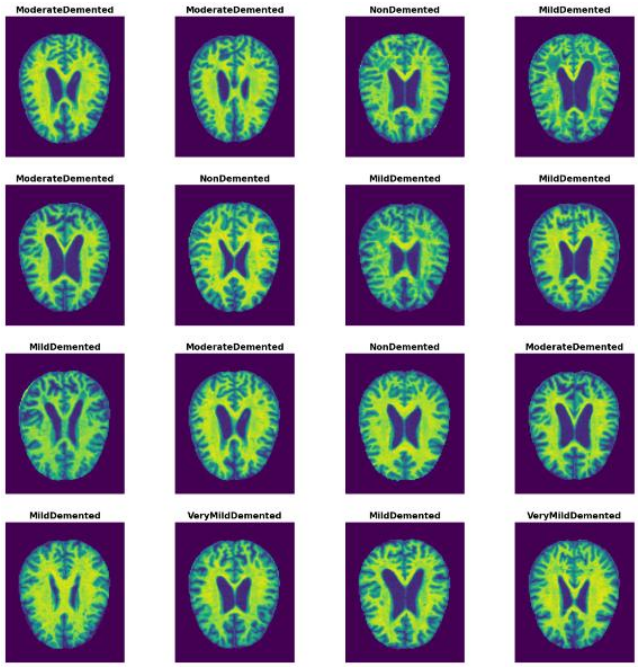


Fig. 7. Data show.

True Negatives (TN), True Positives (TP), False Negatives (FN), and False Positives (FP) indicate the model's performance as measured by the confusion matrix. The performance metrics utilized in this study are listed below:

Accuracy:

$$Accuray = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Sensitivity:

$$\text{Recall} = \text{T P F N} + \text{T P} \quad (3)$$

Specificity:

$$\text{Recall} = \text{T N T N} + \text{F P} \quad (4)$$

F1 measure:

$$F1 = 2 \times \text{Precision} \times \text{Recall Precision} + \text{Recall} \quad (5)$$

Precision measures the accuracy of positive identifications, while accuracy indicates how well the classifier predicts all categories correctly. Specificity evaluates how accurately the model predicts negative instances, and sensitivity assesses its ability to detect positives. The F1-score, which factors in both false negatives and positives, is the harmonic mean of precision and sensitivity.

B. Result Analysis

The main contribution of this study is the development of a hybrid 2D-CNN and LSTM model for efficient Alzheimer's disease classification.

The training and validation loss curve is presented in Fig. 8, while the training and validation accuracy curve is shown in Fig. 9.

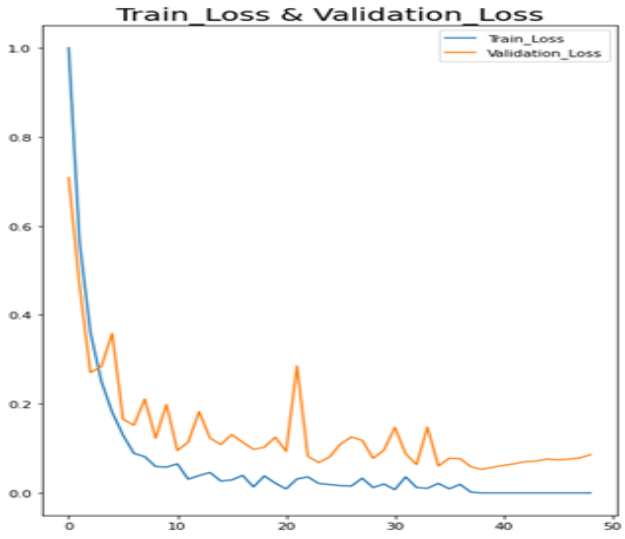


Fig. 8. Training and validation loss curve.

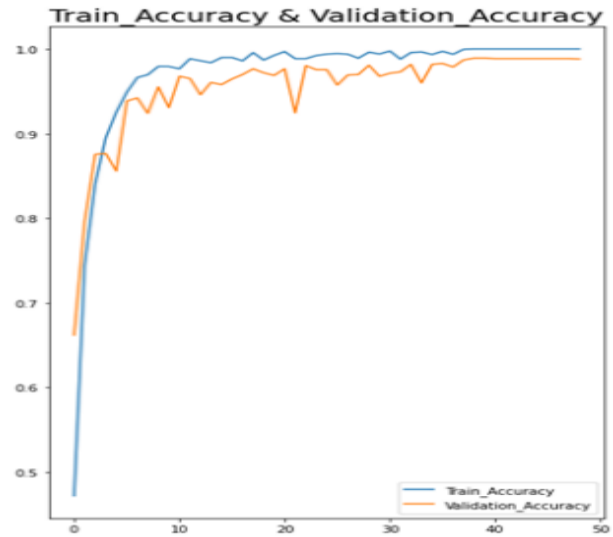


Fig. 9. Training and validation accuracy curve.

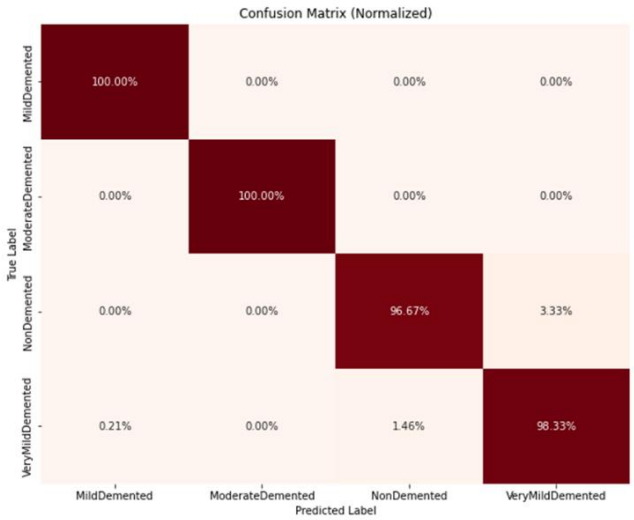


Fig. 10. Confusion matrix.

Fig. 10 presents the confusion matrix of the hybrid model, where rows represent the actual classes and columns represent the predicted classes. The number of cases assigned to each category is shown in the matrix cells. All Mild Demented instances were correctly classified with 100% accuracy, and similarly, all Moderate Demented cases were perfectly classified. For Non-Demented individuals, 96.67% were correctly identified, with only 3.33% misclassified as Very Mild Demented. Finally, 98.33% of Very Mild Demented cases were accurately classified, with 1.46% misclassified as Non-Demented and 0.21% as Mild Demented.

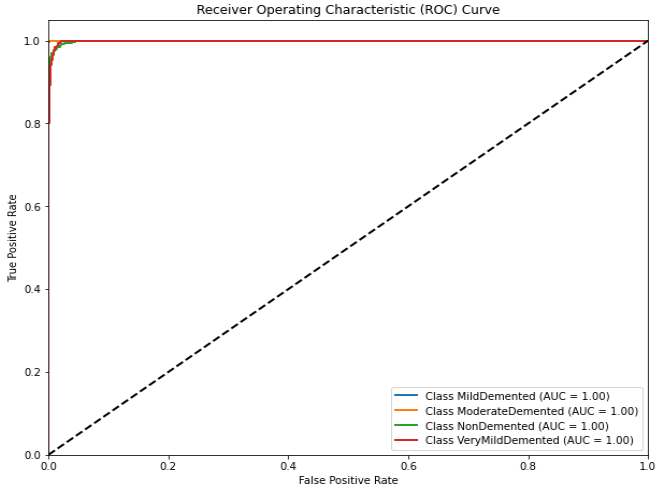


Fig. 11. ROC-AUC curve.

Fig. 11 shows the ROC curve, highlighting the model's performance by plotting the true positive rate against the false positive rate at different decision thresholds. The ROC curve illustrates the model's reliability, with all classes—Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented—achieving an Area Under the Curve (AUC) score of 1.00. The perfect AUC scores for all classes emphasize the model's accuracy in distinguishing between categories.

The performance of our hybrid model achieves an accuracy of 99%, with sensitivity at 100%, a precision rate of 98% and up to 100%, a specificity of 76%, and an F1-score ranging from 98% to 100%.

The final output of the proposed model's classification is presented in Fig. 12, showing the predicted labels for test samples.

The output of the proposed model:

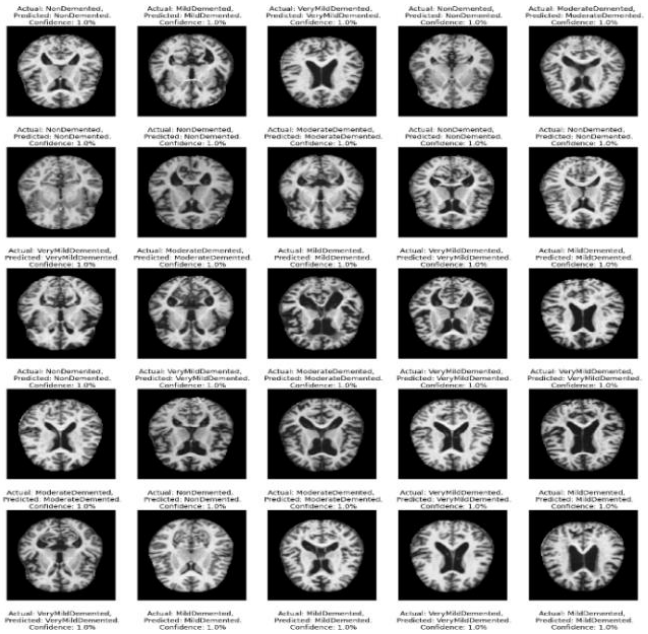


Fig. 12. The output of the proposed model.

V. Discussion

The combination of 2D-CNN and LSTM layers was found to be effective for the image classification task. Despite extensive experimentation with various architectures, including ResNet, DenseNet, VGG, and CNN with SVM, the results varied significantly. The CNN with LSTM approach, along with SMOTE and normalization, achieved the best performance among the tested methods.

Table II summarizes a comparison of different models proposed by researchers for classifying Alzheimer's disease using MRI images. Using the ADNI dataset, Samhan et al. [18] used the VGG16 model and achieved a 97% accuracy rate, while Yang et al. [19] applied the VGG19 model to 3,210 images across four classes and achieved a 97.8% accuracy rate. Using 6,000 images across four classes, Pradhan et al. [26] applied VGG16 technique and achieved an accuracy of 94%. In contrast, using the Inception V4 model with the OASIS dataset, Mohammed et al. [22] achieved an accuracy of 73.75%. Also, Feng et al. [27] combined 3D-CNN and FSBi-LSTM models on the ADNI dataset and has achieved an impressive accuracy of 94.82%. This study's proposed model, integrating 2D-CNN and LSTM, achieved a peak accuracy of 99% on 6,400 images across four classes. This model exceeds the performance of the existing models and highlights its significant potential for effective diagnosis of AD on a larger dataset.

TABLE II. A Comparison of Various Existing Models Along Side the Proposed Model

Authors	Number of Classes	Methodology	Accuracy
L. F. Samhan et al. [15]	ADNI dataset	VGG16	97%
K. Yang et al. [16]	3210 images + 4 classes	VGG19	97.8%
Z. Cui et al. [29]	2400 images + 2 classes	Inception V3	85.7%
V. Patil et al. [30]	223 images	DenseNet	96.4%
E. A. Mohammed et al. [19]	OASIS dataset	Inception V4	73.75%
J. Venugopalan et al. [20]	503 images	CNN	78%
A. Pradhan et al. [21]	6000 images + 4 classes	VGG16	94%
F. Razavi et al. [22]	51 images	S Filter + Regression	98.3%
T. J. Saleem et al. [23]	ADNI dataset	DNN	67%
M. Zaabi et al. [24]	4870 images + 2 classes	CNN + Transfer Learning	92.81%
G. Folego et al. [25]	ADNI dataset	LeNet-5	52.3%
Feng C et al. [31]	ADNI dataset	3D-CNN and FSBi-LSTM	94.82%
Dua M et al. [32]	OASIS dataset	CNN + RNN + LSTM	92.22%
Proposed Model	**6400 images + 4 classes**	**2D-CNN + LSTM**	**99%**

VI. Conclusion and Future Work

The experimental evaluation demonstrates the superiority of our hybrid 2D-CNN + LSTM model, achieving 99% accuracy, 100% sensitivity, and an F1-score of 98%. Compared to existing models, our approach offers a significant improvement in early-stage AD detection. The use of augmentation and class balancing techniques contributed to enhancing model robustness. Future work will focus on extending the dataset, incorporating additional neuroimaging modalities, and exploring transformer-based architectures to further enhance diagnostic precision.

Despite promising results, the study has some limitations. First, the dataset used was relatively small, which may impact the generalizability of the model. Second, the model does not integrate clinical metadata such as age or cognitive scores, which could potentially enhance classification accuracy.

Acknowledgment

This research, "Integration of 2D Convolutional Neural Networks and Long Short-Term Memory Networks for Enhanced Image Processing and Prediction in Alzheimer's Disease", was supported by the department of Computer Science at the Arab Academy for Science, Technology, and Maritime Transport's Computer Science Department during 2023–2024. We are sincerely grateful to our colleagues and collaborators for their invaluable knowledge and assistance in completing this research. Special thanks are also due to the medical professionals and data providers who supplied the necessary datasets that made this research possible.

FUNDING INFORMATION

Self-funded.

CONFLICT OF INTEREST

None

REFERENCES

[1] Alzheimer's Disease Neuroimaging Initiative (ADNI), 'ADNI Data Repository,' Available: https://ida.loni.usc.edu/login.jsp, Accessed: Aug. 2024.

[2] U. Haq, Q. Zhou, Y. Huang, and J. Rizwan, 'Multimodal Fusion Diagnosis of Alzheimer's Disease via Lightweight CNN-LSTM Model Using Magnetic Resonance Imaging (MRI),' in Proc. IEEE Conf. Med. Imaging, 2024.

[3] S. Jahan et al., 'Explainable AI-based Alzheimer's Prediction and Management using Multimodal Data,' IEEE Trans. Biomed. Eng., vol. 70, no. 4, pp. 1223-1234, Nov. 2023.

[4] G. M. Hoang, U.-H. Kim, and J. G. Kim, 'Vision Transformers for the Prediction of Mild Cognitive Impairment to Alzheimer's Disease Progression using Mid-Sagittal MRI,' IEEE Access, vol. 11, pp. 4545-4560, 2023.

[5] E. U. Haq, Q. Y. Zhou, Y. Huang, and J. Rizwan, 'Multimodal Fusion Diagnosis of Alzheimer's Disease via Lightweight CNN-LSTM Model Using Magnetic Resonance Imaging (MRI),' in Proc. Int. Conf. AI Med. Imaging, 2024.

[6] A. Javeed et al., 'Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions,' J. Med. AI, vol. 15, no. 3, pp. 203-215, 2023.

[7] T. Pang, P. Li, and L. Zhao, 'A Survey on Automatic Generation of Medical Imaging Reports Based on Deep Learning,' IEEE Rev. Biomed. Eng., vol. 16, pp. 78-102, 2023.

[8] Z. Hu et al., 'VGG-TSwinformer: Transformer-Based Deep Learning Model for Early Alzheimer's Disease Prediction,' in Proc. Int. Conf. AI Healthcare, 2023.

[9] O. Hansson et al., 'The Alzheimer's Association Appropriate Use Recommendations for Blood Biomarkers in Alzheimer's Disease,' Alzheimers Dementia, vol. 18, no. 6, pp. 2314-2325, 2022.

[10] 'Statistics of Alzheimer Disease,' Available: https://cfah.org/alzheimers-dementia-statistics/, Accessed: 2023.

[11] 'What is Alzheimer?' Available: https://www.diva-portal.org/smash/get/diva2:1735569/FULLTEXT01.pdf, Accessed: 2023.

[12] 'A robust and clinically applicable deep learning model for early detection of Alzheimer's,' Available: https://www.dovepress.com/getfile.php?fileID=87924, Accessed: 2023.

[13] 'Exploring the Role of Advanced AI Models in Alzheimer's Prediction,' Nat. Sci. Rep., vol. 13, no. 42, pp. 1224-1239, 2023.

[14] 'Predicting Alzheimer's Progression with AI-driven Methodologies,' Front. Digit. Health, vol. 4, 2022, doi: 10.3389/fdgth.2022.875895.

[15] 'Alzheimer's Disease Statistics,' Available: https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.13390, Accessed: 2023.

[16] 'Medical Report Prediction Using Deep Learning,' Available: https://tankmitesh29.medium.com/medical-report-predication-with-deep-learning-2e3e40289fc3, Accessed: 2023.

[17] 'AI in Medical Imaging: Insights and Predictions,' Available: https://ar5iv.labs.arxiv.org/html/2107.07045, Accessed: 2023.

[18] 'Current Trends in Medical AI,' Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4719918, Accessed: 2023.

[19] M. M. Rana et al., 'A Robust and Clinically Applicable Deep Learning Model for Early Detection of Alzheimer's,' IET Image Process., vol. 17, no. 14, pp. 3959–3975, 2023.

[20] M. Khatun et al., 'Hybridized Convolutional Neural Networks and Long Short-Term Memory for Improved Alzheimer's Disease Diagnosis from MRI Scans,' in Proc. Int. Conf. Comput. Inf. Technol., 2023.

[21] '2023 Alzheimer's Disease Facts and Figures,' Alzheimer's Dementia, vol. 19, no. 4, pp. 1598–1695, 2023, doi: 10.1002/alz.13016.

[22] T. Akan, S. Alp, and M. A. N. Bhuiyan, 'Vision Transformers and Bi-LSTM for Alzheimer's Disease Diagnosis from 3D MRI (Version 1),' arXiv, 2024, doi: 10.48550/ARXIV.2401.03132.

[23] T. J. Saleem et al., "Deep Neural Network for Alzheimer's Diagnosis Using the ADNI Dataset," IEEE J. Biomed. Health Inform., vol. 27, no. 1, pp. 1–10, 2023.

[24] M. Zaabi et al., "Alzheimer's Detection via Transfer Learning using CNN," Comput. Biol. Med., vol. 145, 105464, 2022.

[25] G. Folego et al., "Diagnosis of Alzheimer's Disease Using LeNet-5 on MRI Scans," J. Med. Imaging Health Inform., vol. 10, no. 3, pp. 500–507, 2020.

[26] A. Pradhan et al., "Improved Alzheimer's Classification using VGG16 with Class Balancing Techniques," Biomed. Signal Process. Control, vol. 78, 103936, 2023.

[27] C. Feng et al., "Hybrid 3D-CNN and FSBi-LSTM for Alzheimer's Diagnosis," Neurocomputing, vol. 456, pp. 224–234, 2022.

[28] Yang, Y., Ye, L., & Feng, Z. (2025). Application of Artificial Intelligence in Medical Imaging: Current Status and Future Directions. iRADIOLOGY, 3(2), 144–151. https://doi.org/10.1002/ird3.70008

[29] Z. Cui et al., "Diagnosis of Alzheimer's Disease with Inception V3," Brain Imaging Behav., vol. 15, no. 2, pp. 445–452, 2021.

[30] V. Patil et al., "MRI-Based Alzheimer's Classification Using DenseNet," ICT Express, vol. 8, no. 4, pp. 531–537, 2022.

[31] Feng C et al., "Enhanced Deep Learning for AD Detection using FSBi-LSTM," IEEE Access, vol. 10, pp. 18244–18255, 2023.

[32] Dua M et al., "Hybrid CNN-RNN-LSTM Network for Alzheimer's Diagnosis from MRI," Neural Comput. Appl., vol. 35, pp. 857–869, 2023.

Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition

Alaa Alsaeedi[1], Amal AlMansour[2], Amani Jamal[3]
Computer Science Department, King Abdulaziz University, Jeddah, Saudi Arabia[1, 2, 3]
Computer Science Department, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia[1]

***Abstract*—Recently, there has been a significant reliance on the Internet. This creates a fertile environment for various risks, including fraud, privacy violations, and theft. The most common and dangerous risks at present are known as deepfakes. The development of deepfake technologies relies on advancements in artificial intelligence. Deepfake content can greatly affect privacy and security, posing a significant risk to many fields. Therefore, recent research has focused on mechanisms to detect real content from fake content. These mechanisms are classified into two main types: single-modal and multimodal detection. It is worth noting that the widespread deepfake technology has recently become more complex. This may hinder traditional single-mode detection methods in detecting video clips. In this study, we designed an effective multimodal fusion mechanism that integrates pre-trained audio, visual, and textual features. Our framework is based on three considerations: audio features, visual features, and emotion recognition. Emotion recognition focuses on three considerations: audio emotion, facial emotion, and sentiment of speech. We take advantage of the sentiment of speech to ensure there is consistency between audio and visual emotion with the meaning of words. As we achieved, the sentiment of speech makes our model more accurate and robust than when we used the audio-visual emotion inconsistency measures only. In our experiment, we used the FakeAVCeleb dataset, and we achieved 95.24% accuracy, proving our assumption of the impact of the sentiment of speech, the emotion of audio tone, and facial expressions to detect deepfakes.**

***Keywords*—Machine learning; deepfake; multimodal; sentiment of speech; emotion recognition**

I. Introduction

Recently, artificial intelligence technologies (AI) have witnessed significant development. This development has substantial uses, as AI technologies have become part of most areas of our daily lives [1]. Like any major advancement, AI development has both positive and negative aspects, leading to its use being categorized into these two sides [1]. Among the positive uses of AI, some have been used to enhance innovation and progress in various fields such as medicine, education, and industry. For example, AI has contributed to the development of medical technologies that help diagnose diseases more accurately, improve education by providing smart educational methods, and increase production efficiency in manufacturers [1]. On the contrary, some have exploited AI unethically to achieve harmful goals such as espionage or violating privacy, spreading misleading information, or manipulating public opinions by creating fake content, which is called deepfake [2], [1]. Deepfakes are one of the most concerning uses of these advancements. That involves digital manipulation of media such as text, audio, and visuals [2]. According to Spector [3], it is estimated that a substantial portion, around 50% of the billions of audio files, images, and videos uploaded daily on social and professional platforms, is manipulated. The term deepfake combines "deep learning" and "fake", indicating manipulation based on AI [2]. AI-based manipulations such as Variational Auto-Encoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models (DMs) aim to reach unprecedented levels of realism [4]. Notably, deepfakes are utilized across various fields, including hyper-realistic computer-generated imagery (CGI), virtual reality (VR), augmented reality (AR), education, animation, art, and cinema. However, their deceptive nature also makes them susceptible to malicious use [5]. The major theft incident that occurred in 2021 sheds light on the severe financial risks associated with deepfake technology. Thieves employed AI-based voice cloning to mimic a bank official. That enabled them to fraudulently obtain $35 million from a UAE bank [6]. Facial expressions and speech are essential to human interaction, and it is crucial in biometric-based identity recognition. That means altered faces and voices present significant challenges to the integrity of online information and security systems [2].

Deepfakes can manifest in various media, all aiming to alter current content for the benefit of the manipulator. A deep fake can occur on a single modal such as text, audio, or visual, as shown in Fig. 1 and 2.

Otherwise, multimodal, typically shown with videos, such as audio-visual or text-visual manipulation, which is a more complicated deepfake technology, which is shown in Fig. 3 and 4, respectively [4].

Text-based deepfake means manipulating unprecedented volumes of online misinformation, such as fake news and rumors circulating across the internet, influencing public perceptions of significant social events [11]. Recent advances in neural generative models, like GPT-2, have exacerbated this issue, as these models can generate highly fluent and coherent text, potentially allowing adversaries to produce convincing fake news [11]. Audio-based deepfakes are artificially generated or altered audio recordings that convincingly mimic real speech [12]. Lastly, a visual-based deepfake is the replacement of a targeted individual's face in a video with that of another person by splicing a synthesized facial region into the original footage [5].



Fig. 1. Video deepfake from Celeb-DF dataset [7].

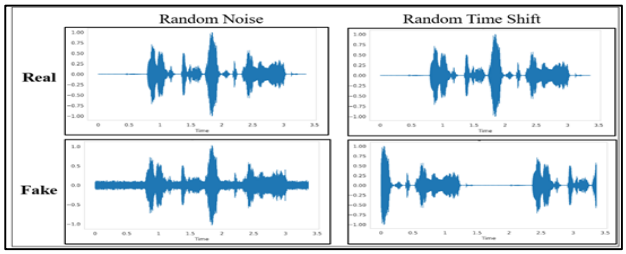


Fig. 2. Audio deepfake from ASVspoof 2019 TFRecord dataset[8]. Random noise means the background noise, and random time shift means changing the audio order.

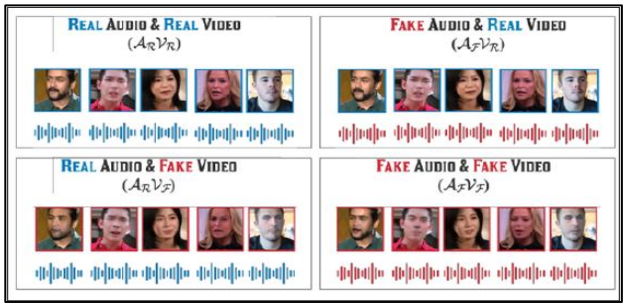


Fig. 3. Audio and visual multimodal manipulation [9].

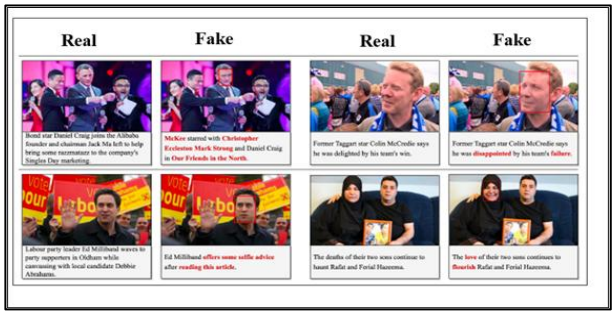


Fig. 4. Text and visual multimodal manipulation[10] .

Deepfake detection has become substantial in the digital age for various reasons. First, it helps stop misinformation, which can convince public opinion and have significant social and political effects [13]. Second, the integrity of news, social media, and legal evidence to support digital media authenticity [14]. Finally, since deepfakes can impersonate people for harmful purposes, it protects against financial fraud and identity theft [15]. To deepfake detection, researchers are developing and evaluating algorithms that can effectively identify manipulated media [4]. These detection models are categorized into two main groups based on the manipulation type [4]. Single-modal detection models focus on analyzing one specific type of data, such as text, image, or audio. Multimodal detection models integrate the analysis of many types of manipulations, combining images, audio, and text. Multi-modal better handles real-world deepfakes [4].

As noted above, multimodal detection is superior to single-modal detection due to the employment of many levels of detection. Thus, studies indicate that integrating emotion recognition, which means increasing detection levels, into deepfake detection provides a promising enhancement in the fight against advanced AI-generated media [16]. Emotion recognition is an aspect of artificial intelligence designed to interpret human emotions through nonverbal analysis [17]. Emotional recognition can be accomplished through several forms, including text, visual, and audio forms [17]. Text emotion recognition systems assess written text for emotions [18]. Classifying emotions like happiness, rage, and sadness includes processing textual signals, including word choice, syntax, and language patterns [18]. Visual emotion recognition employs computer vision to examine facial features in images and videos [17]. In audio emotion identification, AI algorithms use tone, pitch, and rhythm to infer emotions [17]. Conventional detection techniques frequently depend on perceptual discrepancies or technical defects. As the quality of deepfakes enhances, these indicators become increasingly difficult to discern. Consequently, examining emotional authenticity, including the intensity and genuineness of expressions, will enhance the distinction between authentic and altered information [16], [19], and [20].

Recent studies have shown that multimodal detection strategies, either visual-audio or text-audio, can give the deep learning model complementary information. However, the current strategies are still struggling with detecting fake content, especially with the datasets that use complex manipulation tools. The related works that used an emotional indicator achieved better accuracy than those without emotion detection. Nevertheless, there is no investigation about the concatenation of emotions with the sentiment of speech, "what I say and how I say it". Thus, we will get better information than the traditional emotion detection when we integrate the sentiment of speech—what I say—with audio emotion and facial emotion—how I say it.

Motivated by this gap, we introduce our novel framework, titled "Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition". A framework for developing an efficient multimodal deepfake detection system. Our method detects deepfakes based on emotional mismatches in three considerations.

The main contributions are as follows:

- Compute the mismatch between the audio tone and the facial emotion.
- Compute the mismatch between the audio tone and the sentiment of speech.
- Compute the mismatch between the facial emotion and the sentiment of speech.

- Perform an ablation study to analyze the impact of each field in our model with five experiments (audiovisual emotion model, sentiment of speech only model, audio emotion with sentiment of speech model, facial emotion with sentiment of speech model, and our approach that includes all of above model).

The structure of this study is as follows: Section II presents a review of the literature. Section III describes the proposed framework and methodology. Section IV details the experiments and reports the results. Section V provides a discussion of the findings. Finally, Section VI concludes the study and outlines future works.

II. Literature Review

As we mentioned earlier, deepfake techniques have recently become more accurate and sophisticated due to advances in artificial intelligence. Therefore, it is important to keep up with this development with effective detection tools. We categorize the literature reviews according to the approaches to detecting deepfake content. We break it into three subsections: single-modal, multimodal, and emotional deepfake detection approaches. Then we will present the existing datasets. This section will conclude with research gaps and our contribution.

A. Single-Modal Detection Approach

As we know, in single-modal detection, there are three types of methods for detecting deepfakes: textual, audio, or visual detection. In this subsection, we will present the literature reviews that utilize a single type of deepfake detection. Divided into sub-subsections are text-based, audio-based, and visual-based. This subsection concludes with an analytical table of the literature mentioned there.

1) Text-based approach. The following studies aim to propose a deepfake detection model based on textual content. Saravani et al. [21] examined a technique for identifying social media bots through the analysis of tweets. Utilizing the deepfake dataset comprising 25,572 tweets from bots mimicking human accounts, the model processed data and employed a classification architecture that included two fully connected layers. The authors attained a two per cent enhancement in accuracy over prior models by incorporating bidirectional encoder representations from transformers (BERT) for text representation that used natural language processing (NLP), a bidirectional long-short-term memory (BiLSTM) layer is a type of neural network used to preserve word order, and the neural network expand vector of locally aggregated descriptors (NeXtVLAD) layer for effective information summarization. Li et al. [22] used large language models (LLMs) that closely resemble human writing to study the challenges of distinguishing human-authored from machine-generated writing. The authors built their identification system using human-authored texts from Reddit comments, news headlines, and academic works. To ensure diversity, they used 27 LLMs and three prompt types to generate machine-generated writings. The detection system provided a probability score for classifications and used the Longformer neural network, which outperformed related models. Jensen-Shannon distance revealed that text variety inhibited the detection of linguistic trends. The detection method had an 86.54% recall rate on GPT- 4 texts. Zhong et al. [11] examined the growing problem of online deception, focusing on GPT-2-generated fake news. The researchers developed FAST, a graph-based method for analyzing literary compositions' factual architecture to distinguish human-generated from machine-generated material. They examined the news-style (GROVER) and web text-style (GPT-2) datasets for binary classification. We found that human-authored texts cited the same elements in their sentences, but machine-generated texts incorporated unnecessary sections, making differentiation easier.

2) Audio-based approach. The following research efforts aim to develop a deepfake detection model based on voice analysis. A study by Mcuba et al. [23] used deep learning techniques, especially convolutional neural networks (CNNs), to make it easier to spot deepfake sounds. They used the Baidu Silicon Valley AI Lab's VCTK and LibriSpeech speaker datasets for research. During preprocessing, they converted audio recordings into images and extracted features like mel-spectrograms and MFCCs for analysis. Hamza et al. [12] studied deepfake audio identification. The authors used advanced machine learning techniques, specifically MFCCs for audio evaluation, to improve detection approaches. To ensure high-quality preprocessing for model training, the authors used the Fake-or-Real (FoR) dataset with approximately 195,000 samples of real and synthetic speech. This study focused on feature extraction and classification models, with the Support Vector Machine (SVM) achieving the greatest accuracy of 98.83%. Shorter audio segments increased classification performance, with SVM outperforming in clean and loud circumstances. Pianese et al. [24] focused on speaker biometrics and trained the algorithm on real data to ensure real-world performance. Centroid-based (CB) and maximum similarity (MS) were the study's key approaches. The study tested advanced models such as RawNet2- antispoofing on ASVSpoof2019, FakeAVCelebV2, and In-The-Wild Audio Deepfake. Centroid-based (CB) testing compares test audio to the real audio average. Second, maximum-similarity testing (MS) compares test audio to real audio with the highest similarity score. The MS method performed well across datasets, especially in complicated, real-world circumstances. Muller et al. [25] examined text-to-speech synthesis advances and their effects on audio deepfakes, which may impersonate human voices. The authors used ASVspoof 2019 and IWA. They created 37.9 hours of audio samples from real and fictitious prominent people to test the models. ADAM optimizes cross-entropy loss-trained deep learning models.

TABLE I. ANALYTICAL REVIEW OF SINGLE-MODAL APPROACH

Study	Approach	Model	Datasets	Performance Metrics	Evaluation
Saravani et al. [21]	Text- based	BERT, BiLSTM, and NeXtVLAD	Deepfake	Accuracy	92%
Li et al. [22]	Text- based	Longformer	Customized (Texts produced by GPT-4)	Accuracy	86.54%
Zhong et al. [11]	Text- based	Graph Convolutional Network (GCN), LSTM, and Next Sentence Prediction (NSP)	GROVER and GPT-2	Accuracy	GROVER =87.97% GPT-2= 93.1%
Mcuba et al [23]	Audio-based	CNN with Adam, SDG, and Adadelta optimizers	VCTK with the LibriSpeech speakers	Accuracy	Adadelta =85.9% SDG = 83.6% Adam = 72.2%
Hamza et al. [12]	Audio-based	SVM	FoR	Accuracy	98.83%
Pianese et al. [24]	Audio-based	Supervised method, CB, and MS	ASVSpoof20 19, FakeAVCele bV2, and IWA	Calculate the average area under the curve (AUC) for all datasets	Supervised method = 79.6% CB = 88.4% MS= 91.3%
Muller et al. [25]	Audio-based	Multiple neural network	ASVspoof20 19 and IWA dataset	Equal Error Rate (EER)	ASVspoof2019 =9.85% IWA= 60.10%
Wodajo and Atnafu [5]	Visual-based	CViT (CNN and Vision Transformer (ViT))	DFDC	Accuracy	91.5%
Cozzolino et al. [10]	Visual-based	ID-Reveal which uses temporal ID network	VoxCeleb2 dataset	Accuracy	80.4%
Li et al. [26]	Visual-based	DSP-FWA by using CNNs	Own dataset (Celeb-DF) and others as UADFV and DFDC	AUC	Celeb-DF=64.6% UADFV=97.7% DFDC=75.5%

Comparing neural network topologies requires feature extraction, according to the authors. They found that raw audio features outperformed processed features and that longer audio segments reduced model error rates from 19.89% to 9.85%.

3) Visual-based approach. The studies below focus on developing a deepfake detection model based on image analysis. Wodajo and Atnafu [5] developed CViT to detect false content. The study required three steps for preparation: we recognized faces from video frames, scaled them to 224 x 224 pixels in RGB format for consistent input, and optimized content to reduce background interference. Rotation, flipping, and color changes boosted dataset variety and model generalization. The CViT used convolutional layers and transformer processes to record spatial hierarchy in images and apply self-attention to detect essential parts. The CViT had 91.5% accuracy on the enlarged dataset, although dataset heterogeneity and visual artifacts affected it. Cozzolino et al. [10] used the Temporal ID Network and 3D Morphable Model (3DMM) Generative Network to create the ID-Reveal method for detecting deepfake movies. The temporal ID network developed an embedded vector that identifies an individual using convolutional layers and similarity score computations, while the 3DMM generative network created authentic 3D characteristics. Test videos were compared to pristine reference videos for authenticity. The ADAM optimizer gave ID-Reveal 80.4% accuracy and 0.91 AUC for high-quality videos. Li et al. [26] proposed detection methods and the Celeb-DF dataset of high-quality deepfake videos of celebrities. They examined nine detection methods, including CNNs and head movement assessment. Deep learning-based spatio-temporal features for video forgery detection with attention (DSP-FWA) performed best. DSP-FWA looks at videos' spatial and temporal data using an attention method to bring out important parts and make detection more accurate.. Table I shows that deepfake detection techniques are either in text, audio, or visual domains. With each approach demonstrating unique strengths. Text-based models, Saravani et al. [21], used BERT, BiLSTM, and the NeXtVLADGCN, showed good accuracy in detecting manipulated texts, equal to 92%. While audio-based detection, Hamza et al. [12], used SVM, which outperformed other methods with an accuracy of 98.83%. On the other side, visual-based models, Wodajo and Atnafu [5] used the CViT and achieved good accuracy around 91%. Overall, selecting the right detection method depends on the type of deepfake being analyzed.

B. Multimodal Detection Approach

In multimodal detection, the detection is achieved by analyzing more than one type of manipulation. As we mentioned, there are two main combination types of methods for multimodal detecting deepfake: textual-visual or audio-visual detection. In this subsection, we will present the literature reviews that utilize multimodal deepfake detection. We divided this into subsections: text-visual based models and audio-visual-based models. This subsection concludes with an analytical table of the literature mentioned there.

1) Text-visual-based approach. Text-visual multimodal detection provides a framework for deepfake detection that integrates textual and visual analysis. Shao et al.[27] studied system localized modifications in coupled photos and text, with a specific focus on news stories. They have created the multimodal media manipulation dataset called Detecting and Grounding MultiModal Media Manipulation (DGM4). This dataset of 230,000 samples, including 152,574 modified pairs, examines the impact of simultaneous picture and text alterations. The detection model hierarchical multimodal manipulation reasoning transformer (HAMMER) and its enhanced version, HAMMER++, have proposed approaches to align and analyze image-text interactions for more accurate

modification detection. Liu et al.[4] developed the Unified Frequency-Assisted Trans Former Framework (UFAFormer) to discover and evaluate media modifications in images and text. It includes picture and text encoders, a frequency encoder that leverages DWT, and a Forgery-Aware Mutual Module (FAMM) to promote feature integration. UFAFormer improved accuracy on DGM4. Change recognition and classification improved as the authors devised components that incorporate multiple sources of information. Wang et al. [28] suggested a framework that used unimodal feature extractors to extract unique photo and text properties. The model used dual-branch cross-attention (DCA). It improved manipulation identification by analyzing both modalities simultaneously. The study focused on manipulation grounding, which involved identifying altered visual parts and text phrases. The data preparation included standardizing text to fifty tokens and scaling pictures to 256x256. Their method reached 95.11%.

2) Audio-visual-based approach. The following efforts provide a framework for deepfake detection that integrates audio and visual analysis. Khalid et al. [29] introduced the evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. They used the FakeAVCeleb dataset, which had 600 clips starring various celebrities. The authors employed three evaluative methodologies: unimodal, ensemble, and multimodal. Unimodal techniques examined audio or video separately, employing baseline models such as VGG16 and Xception; nevertheless, they encountered difficulties when one modality was authentic and the other was fabricated. Ensemble approaches amalgamate predictions from distinct classifiers for audio and video, enhancing accuracy. Nonetheless, the multimodal strategy, which concurrently assessed both data kinds, demonstrated diminished efficacy, indicating that existing techniques were insufficient for intricate deepfake identification. Lewis et al. [30] used multimodal deep learning to detect false videos utilizing spatial, spectral, and temporal anomalies. NOLANet uses multiple detection methods, including audio and visual attributes to classify content as real or false to identify deepfakes. They used Facebook's Deepfake Detection Challenge dataset (FDDC). Multi-stage detection used LSTM networks for categorization. Preprocessing the films involved extracting frames at 30 frames per second while maintaining the audio. They used BlazeFace to trim and recognize faces at 128x128 pixels. Face landmarks were extracted using FANet. Multiple LSTM networks processed integrated visual and auditory data after DCT analysis. Aligning aural and visual data helped discover deepfake discrepancies. Raza and Malik [31] proposed a multimodal representation learning multimodal trace for deepfake detection. The framework classified audio and video inputs as authentic or fake using a structured architecture. Essential components include audio and video feature extractors, mixers to combine them, and a classification layer. Data is prepared for analysis after normalization and feature extraction by Fast Fourier transform and ResNet. IntrAmodality Mixer Layers (IAML) evaluated audio and video inputs separately to discover patterns. Multi-label categorization using IEML unifies audio and video input modalities. Muppalla et al. [32] presented an integrated audio-visual feature for multimodal deepfake detection. That aimed to improve the identification of multimodal deepfakes by amalgamating audio and visual data via a distinct classification methodology. This entailed compressing video frames to a standardized dimension. Which is about 300x300 pixels and transforming audio into Mel spectrograms. The approach applied deep learning networks for feature extraction and implemented a multi-task learning technique. The assessment utilized two principal datasets. The FakeAVCeleb and the TMC datasets. They tested the framework using two models, the capsule network and the Swin transformer; the capsule network has the best accuracy. Cheng et al. [7] introduced a voice-face matching detection (VFD) system by looking at how well the voices and faces in deepfake videos match each other. They developed a method that first trains the model on regular data. Then fine-tune it for deepfake detection. The authors trained their model with the Voxceleb2 dataset, which includes over 17 million audio utterances from more than 6,000 celebrities. In the testing, they used many datasets, such as DFDC and FakeAVCeleb. Feng et al. [33] proposed self-supervised video forensics by audio-visual anomaly detection. The model was trained on real, unlabeled speech videos from the Lip-Reading Sentences 2 (LRS2) and Lip-Reading Sentences 3 (LRS3) datasets. By learning the normal patterns of audio-visual synchronization.

This led to the classification of anomalies by flagging videos that deviated from the established distribution of features representing the relationship between audio and visual elements. A key component of the method involved calculating synchronization scores for audio-video pairs using a specialized network. The model maximized synchronization scores for real pairs and identified fake ones by analyzing their probability distributions. The evaluation of the model was conducted using two datasets. The FakeAVCeleb and the large-scale Korean-language deepfake detection (KoDF) datasets. Oorloff et al. [34] proposed an audio-visual feature fusion (AVFF) framework for detecting video deepfakes. By employing a self-supervised learning strategy to capture the relationship between audio and visual elements. A trained classifier differentiated between real and fake videos by examining characteristics from both audio and visual elements. The training procedure used the LRS3 dataset for feature extraction. The FakeAVCeleb and KoDF datasets are used to evaluate the model. Table II shows multimodal deepfake detection methods which shows promising results by combining different inputs like text, audio, or visuals. Wang et al. [28] achieved the highest result for text-visual models at 91.42%. While Muppalla et al. [32] reported an impressive 99.20% multimodal result for audiovisual models. The ensemble based on the voting boosts detection robustness in Khalid et al. [29]. Overall, integrating multiple modalities enhances performance. However, effectiveness depends on the datasets, algorithms used, and the integration method used to merge two single models.

TABLE II. ANALYTICAL REVIEW OF MULTIMODAL APPROACH

Study	Approach	Model	Datasets	Performance metrics	Evaluation	
Shao et al. [27]	Text- visual based	Image detection by BBox Detector	DGM4	Intersection over Union mean (IoUmean)	Image = 76.51%	
		Text detection by Token Detector		Recall	Text= 72.14%	
		Multimodal by Multi-Label Classification		Mean Average Precision (mAP)	Multi-label =86.29%	
Liu et al. [4]	Text- visual based	Image encoder by vision transformer (ViT-B/16)	DGM4	Intersection over Union mean (IoUmean)	Image = 78.33%	
		Text encoder by BETR base model		Recall	Text= 70.61%	
		Multimodal by unified decoder to manipulate detection and grounding across image, text, and image-text pairs		Mean Average Precision (mAP)	Multi-label =87.85%	
Wang et al. [28]	Text- visual based	Image encoder by vision transformer (ViT-B/16)	DGM4	Intersection over Union mean (IoUmean)	Image = 80.83%	
		Text encoder by RoBERTa base model		Recall	Text= 70.73%	
		Binary classifier to determine whether manipulation is present. Fine-grained classifier to identify specific types of manipulation.		Mean Average Precision (mAP)	Multi-label =91.42%	
Khalid et al. [29]	Audio-visual based	Audio using many deep learning methods	FakeAV Celeb	Accuracy	Audio = 76%	
		Image using many methods such as VGG16			Image = 81%	
		Ensemble approach based on voting method. Multimodal using CNN			Ensemble = 78% Multimodal= 67.3%	
Lewis et al. [30]	Audio-visual based	Audio using LipSpeech sub-network	FDDC	Accuracy	Audio = 59.21%	
		Image using VSNet sub- network			Image = 61.59%	
		Multimodal using LSTM network detector			Multimodal= 65.18%	
Raza and Malik [31]	Audio-visual based	Audio IAML sub- network	FakeAVCeleb	Accuracy	Audio = NA	
		Image IAML sub- network			Image = NA	
		Multimodal using mixer layer with IEML			Multimodal= 92.9%	
Muppall a et al. [32]	Audio-visual based	Audio details not found	FakeAV Celeb and TMC	Accuracy of Capsule network	FakeAVCeleb	TMC
		Image details not found			Audio =99.80%	Audio =99.68%
		Multimodal using capsule network and the swin transformer			Image =61.59%	Image = 96.43%
					Multimodal= 65.18%	Multimodal= 99.20%
Cheng et al.[7]	Audio-visual based	VFD uses a dual-stream network (audio and visual)	DFDC and FakeAVCeleb	Accuracy	DFDC	FakeAV Celeb
		Multimodal using a matching function			Audio = NA	Audio = NA
					Image = NA	Image = NA
					Multimodal= 80.96%	Multimodal= 81.52%

Feng et al. [33]	Audio-visual based	Audio using VGG-M for encoding Multimodal using a matching function. Image using ResNet-18 for encoding Multimodal using autoregressive method	FakeAVCeleb and KoDF	AUC	FakeAVCeleb Audio = NA Image = NA Multimodal= 95.8%	KoDF Audio = NA Image = NA Multimodal= 86.9%
Oorloff et al. [34]	Audio-visual based	Unimodal encoders (audio, visual) Multimodal use: Cross-modal fusion: (audio-to-visual) and (visual-to-audio) Complementary masking while one modality is masked, the other is visible.	FakeAVCeleb and KoDF	AUC	FakeAVCeleb Audio = NA Image = NA Multimodal= 94%	KoDF Audio = NA Image = NA Multimodal= 95.5%

C. Emotion-Based Detection Approach

This subsection will focus on the related deepfake detection models that utilize emotional recognition features to detect fake content. Mittal et al. [19] proposed affective cue-based audio-visual deepfake detection. They found inconsistencies in actual and fake videos of the same person using a Siamese neural network and embedding vectors. Deepfakes are detected by evaluating emotional variations between speech and facial expressions. Their model was evaluated on DF-TIMIT and DFDC, two well-known deepfake detection datasets. Hosler et al. [20] suggested that using emotional inconsistencies to detect deepfakes is meaningful. Two dimensions were utilized to model emotion: valence (positive/ negative effect) and arousal. Real and fake videos have different emotional patterns, according to the authors. Deepfake technology struggled to recreate real emotions. First, video emotional features were extracted for detection. To express the subject's emotions across time, use valence and arousal. Finalizing the video's deepfake detection utilizing these emotional signals. The tests used two datasets. An emotion prediction model is trained on SEMAINE, and deepfake detection on DFDC. An LSTM model predicted emotional states and captured video emotional dynamics using extracted features. Conti et al. [35] presented a semantic approach for deepfake speech detection through emotion recognition (SER). The model identified emotional inconsistencies in synthetic audio by extracting emotional features and feeding them into a classifier. The system identified four primary types of emotions in speech, which are angry, happy, sad, and neutral. The system effectively distinguished reality from synthetic speech.

TABLE III. ANALYTICAL REVIEW OF EMOTIONAL APPROACH

Study	Approach	Model	Datasets	Performance Metrics	Evaluation	
Mittal et al. [19]	Audio-visual based	Emotions extraction Memory fusion network (MFN) to extract emotions from audio and visual modalities Detection by Siamese network to detect inconsistencies.	DF-TIMIT and DFDC	AUC	Without emotion indicator DFTIMIT= 94.8% DFDC=8 2.8%	With emotion indicator DFTIMIT= 96.3% DFDC= 84.4%
Hosler et al. [20]	Audio-visual based	Emotions extraction by Extracting low-level features for face and voice. Detection by Uses LSTM models to predict continuous emotional states (valence and arousal) over time.	DFDC	Accuracy	Without emotion indicator = 87.5%	With emotion indicator = 99.5%
Conti et al. [35]	Audio based	Emotions extraction by Speech Emotion Recognition (SER) component that extracts emotional features from speech using a 3D- Convolutional Recurrent Detection by Neural Network (CRNN) Synthetic Speech Detector (SSD) classifies the input as real or deepfake using a Random Forest classifier.	ASVspoof 2019	AUC	Without emotion indicator = NA	With emotion indicator = 98%

Testing showed high accuracy, especially in clean audio, and demonstrated better performance than other models. The core idea was that synthetic speech often failed to replicate authentic emotional patterns, making SER-based features valuable for detection. Table III shows that incorporating emotional recognition significantly enhances the performance of deepfake detection models. Mittal et al. [19] and Hosler et al. [20] proposed detection models using audio-visual deepfake detection. Demonstrated that including emotional features improved the performance. Hosler et al. [20] achieved impressive growth from 87% to around 99%. Conti et al. [35] also showed high performance with emotion extraction by speech emotion recognition. These findings suggest that emotion extraction can be a crucial factor in improving the performance of deepfake detection methods. However, there

are some considerations to enhance the performance, including selecting datasets, algorithms used, and merge methods.

The literature analysis presented above shows that several studies have been conducted to detect deepfake content with different types of content. As we mentioned, as deepfake technology advances, methods for detecting these intricate alterations must also improve. As we observed in Table II, multimodal detection in most cases has improved the performance of deepfake detection, while the model in Muppalla et al. [30] achieved an accuracy of around 99%, the poor detection accuracy of the visual-based model affected the overall multimodal detection accuracy when they used the FakeAVCeleb dataset. They reached an accuracy of around 99% on audio and 61% on image, leading to the overall detection of the multimodal dropping to around 66%. This situation necessitates further investigation into the methods of merging the individual models. Furthermore, notice in Table III that when emotion analysis is added to the model, it enhances the performance. Looking at Mittal et al. [19] and Hosler et al. [20] on how the emotion-based analysis increased performance results, however, there are limited studies that utilize this feature. In this study, we will aim to design audio-visual multimodal deepfake detection leveraging emotional recognition. The emotion extraction will be in three points: audio, image, and sentiment of speech. And the inconsistency between them. Considering the side effects of the multimodal dropped performance. That may occur due to the ineffective selection of the algorithm of audio detection and visual detection, as well as the integration methods.

III. Framework and Methodology

This section explores the architecture of our proposed framework, as shown in Fig. 5, which is designed to detect deepfake content by leveraging emotion recognition in multimodality.

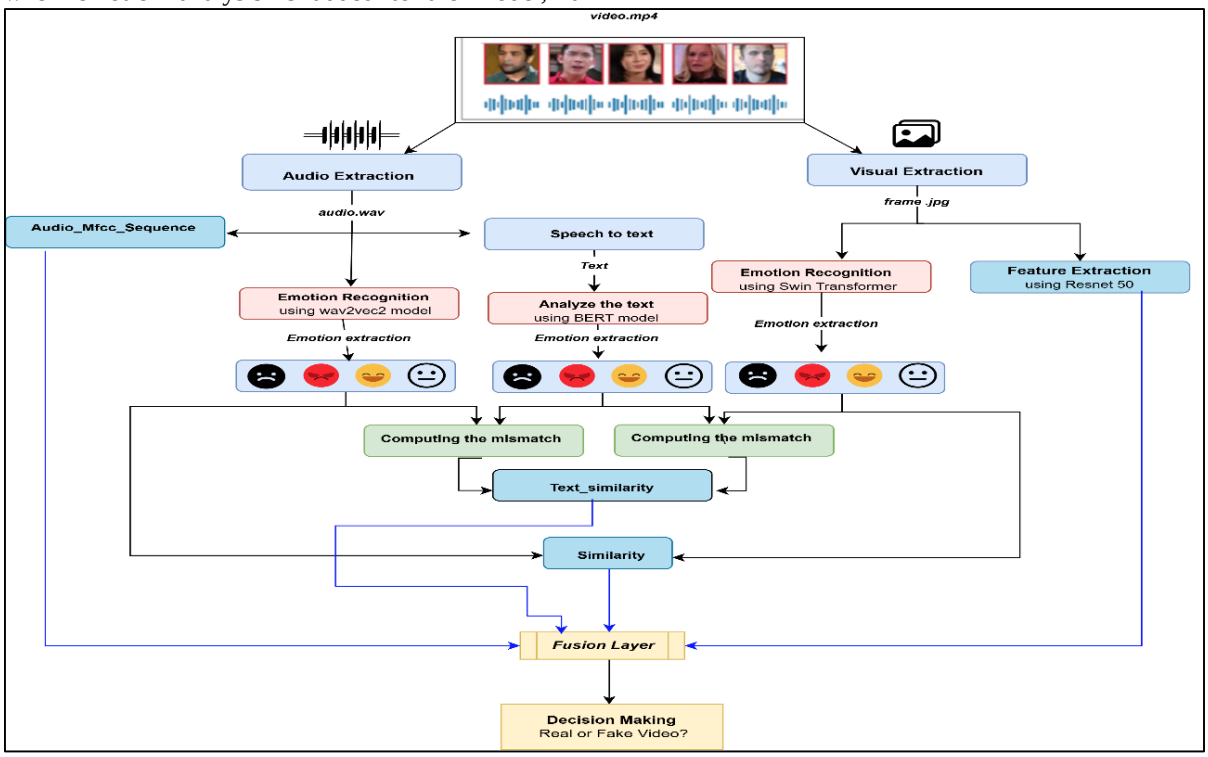


Fig. 5. The model framework.

A. Audio Features Extraction

The first component of our framework is the audio extraction and analysis, containing two main models. The first model aims to extract the main features of (.wav) audio, which includes thirteen Mel-Frequency Cepstral Coefficients (MFCC). The second model seeks to convert the speech to text, aka speech recognition, by using the pre-trained model called wav2vec2 [refer Appendix A] [36]. After that, the audio section involves the emotion extraction stage of each audio and text, aka the sentiment of speech. For the emotion recognition of the audio tones, we used the wav2vec2 pretrained model [37]. For sentiment analysis, we used a pretrained model using the bidirectional encoder representations from transformer (BERT) [38].

B. Visual Features Extraction

The second component is a visual analysis part that extracts the frames (.jpg) with image size 224*244 pixels and then extracts the visual features by a pretrained model using Residual Network - 50 layers (ResNet50) model to extract the noise, sharpness, and edge density. After that, the visual emotion extraction using the pretrained model used the shifted window transformer (Swin transformer) [39].

C. Cross-Model Emotional Consistency Stage

The consistency computation will be between the three modalities: text, audio, and visual. As follows, between the sentiment of speech and the emotion of the audio tone. Between the sentiment of speech and the emotion of the visual

part. Last, compute the mismatch or consistency between the emotion of the audio tone and the emotion of the visual part.

D. The Fusion Layer

Our framework's fusion technique is based on concatenates modality-specific characteristics from the text, audio, and visual streams before passing them through a classification. In particular, a Long Short-Term Memory (LSTM) network with a hidden size of 256 is used to process the temporal data from the video frames in order to identify dynamic patterns over time. After that, this output is coupled with emotion embeddings from each modality, visual quality metrics (noise, edge density, sharpness), and high-level features taken from the audio signal (MFCC and spectral contrast). The final classification output (real or false) is obtained by passing the resultant feature vector through two fully connected layers using ReLU activations.

This straightforward yet effective fusion method enables the model to integrate emotional signals and quality-related features from multiple modalities into a unified decision space. Although more sophisticated approaches—such as attention-based mechanisms—could further enhance performance, we intentionally opted for a lightweight architecture. Our goal was to ensure reproducibility and maintain interpretability, providing a clear and stable baseline for multimodal deepfake detection.

We implemented a modular pipeline to extract and align multimodal features from the FakeAVCeleb dataset. For each video, seven evenly spaced frames were extracted, alongside full audio and transcribed text. Visual features were obtained using a pretrained ResNet50 model, and temporal dependencies were captured using an LSTM module. Additional visual quality metrics (sharpness, edge density, and noise) were computed per frame. Audio features included MFCCs and spectral contrast, while emotional vectors for audio, text, and visual modalities were parsed from pre-extracted sentiment scores. Cosine similarity was used to measure the alignment between emotion vectors. A full architecture diagram is shown in Fig. 5.

This fusion strategy is theoretically grounded in the concept of cross-modal emotional coherence, which hypothesizes that authentic content exhibits natural alignment between facial expressions, vocal tone, and verbal sentiment. By explicitly modeling emotional alignment across modalities—rather than fusing only raw or semantic features—the framework captures inconsistencies that are typical in manipulated content. When compared to partially fused techniques, the ablation studies empirically show that integrating emotional consistency results in significant performance benefits. Our method employs emotional incompatibility as deception cues, providing a new layer of semantic analysis in fake content detection, in contrast to classic multimodal fusion, which only considers spatial or temporal correlations.

IV. Experiments and Results

In this section, we will explore our experiments and results and how we adopt a structured training approach that ensures stability and generalizability to train our deepfake detection model successfully.

A. Experiments

In this subsection, we will present the experiment setup, the dataset that has been used and the ablation study that was conducted in our experiments

1) Experimental setup. The model was trained for 10 epochs using the Adam optimizer with a learning rate of 1e-4. We applied cross-entropy loss and set the batch size to 4, with a fixed randomization seed of 101 in all experiments. The training dataset consisted of 80% of the total data, while the remaining 20% was used 10% for validation and 10% for testing.

2) Dataset. The dataset that has been used in our study is called the FakeAVCeleb dataset [9]. That contains 20,500 video samples, which are divided into the following categories: 500 real videos from the VoxCeleb2 dataset, featuring real audio. 500 videos that combine synthetic (fake) audio with actual footage. 9,000 videos with real audio and fake video. 10,000 videos containing fake audio and video. The average length of each video is roughly 7.8 seconds. The dataset offers a variety of actual and altered audiovisual material combinations to facilitate multimodal deepfake detection research. Fig. 6 below shows the categories of the FakeAVCeleb dataset. The original FakeAVCeleb dataset contains 20,500 samples. For this study, we selected a representative and balanced subset consisting of 998 video samples, equally divided into 499 real and 499 fake videos. The samples were randomly selected from the full dataset while ensuring class balance. To streamline the training and evaluation process, we manually created a CSV file that includes the file paths and corresponding labels for each video sample. This setup allows for efficient data loading and consistent processing across experiments.



Fig. 6. The FakeAVCeleb dataset [9].

3) Ablation study. To ensure the productivity of our contribution, we conducted the ablation study for each emotion feature in our multimodal deepfake detection model as follows:

a) Audio-visual emotion only (No sentiment): In this experiment, we focused on the features extracted from audio, which are thirteen MFCC, and from visual, which are noise,

sharpness, and edge density. Furthermore, the emotion consistency computation between audio and visual modalities. Without any sentiment information. Table IV shows the detailed results of the experiment.

TABLE IV. AUDIO-VISUAL WITH EMOTION RECOGNITION RESULT

Model	Performance		
	Accuracy	*F1-Score*	*Recall*
Audio-visual with emotion recognition	76.19%	72.15%	73.3%

b) Audio emotion and text sentiment (No visual Emotion): In this experiment, we focused on the features extracted from audio, which are thirteen MFCC, and from visual, which are noise, sharpness, and edge density. Furthermore, the emotion consistency computation is between audio and text modalities without any computation between audio and visual emotion. Table V shows the detailed results of the experiment.

TABLE V. AUDIO-VISUAL WITH AUDIO EMOTION AND TEXT SENTIMENT

Model	Performance		
	Accuracy	*F1-Score*	*Recall*
Audio-visual with audio emotion and text sentiment	76.19%	69.21%	68.33%

c) Visual emotion and text sentiment (No audio emotion): In this experiment, we focused on the features extracted from audio, which are thirteen MFCC, and from visual, which are noise, sharpness, and edge density. Furthermore, the emotion consistency computation is between visual and text modalities without any computation between audio and visual emotion. Table VI shows the detailed results of the experiment.

TABLE VI. AUDIO-VISUAL WITH VISUAL EMOTION AND TEXT SENTIMENT

Model	Performance		
	Accuracy	*F1-Score*	*Recall*
Audio-visual with visual emotion and text sentiment	76.19%	69.21%	68.33%

d) Sentiment fusion only (No cross-audio-visual emotion): In this experiment, we focused on the features extracted from audio, which are thirteen MFCC, and from visual, which are noise, sharpness, and edge density. Furthermore, the emotion consistency computation between visual and text modalities and between audio and text modalities without any computation between audio and visual emotion. Table VII shows the detailed result of the experiment.

e) Full multimodal emotion and sentiment fusion (Our approach): In this experiment, it was conducted on the complete framework features, which are the features extracted from audio, which are thirteen MFCC, and from visual which are noise, sharpness, and edge density, the emotion consistency computation between visual and text modalities, and between audio and text modalities. Furthermore, the computation between audio and visual emotion. Table VIII shows the detailed results of the experiment.

TABLE VII. AUDIO- VISUAL WITH SENTIMENT ANALYSIS

Model	Performance		
	Accuracy	*F1-Score*	*Recall*
Audio-visual with sentiment analysis	80.95%	78.57%	81.67%

TABLE VIII. OUR APPROACH

Model	Performance		
	Accuracy	*F1-Score*	*Recall*
Our Approach	95.24 %	95.24%	95.45%

These results empirically validate the contribution of sentiment indicators in multimodal fusion, showing that their absence significantly degrades performance. This supports our hypothesis that emotional coherence between modalities is a strong signal for deepfake detection. These ablation results are further discussed and interpreted in Section V, where we reflect on their implications and underlying patterns.

B. Results

In this subsection, we present the performance evaluation of our proposed model through three comparative perspectives. First, we compare our results with existing studies that utilized the same dataset (FakeAVCeleb) to provide a fair benchmark under similar data conditions. Second, we examine studies that incorporated emotion-related features, particularly those focusing on emotional or affective cues from audio or visual streams, to highlight the novelty of integrating the sentiment of speech into the detection process. Finally, we compare our model against state-of-the-art baseline methods, including both visual-only and multimodal approaches such as FaceForensics++[40] and LipForensics[41]. These comparisons aim to position our work within the broader deepfake detection landscape and demonstrate the effectiveness of our multimodal sentiment-aware framework.

1) Comparative results with the related used same dataset. In this part, we will explore the related works that use the FakeAVCeleb dataset to detect the deepfake content. Table IX shows the comparison details.

2) Comparative results with emotion-based approach. In this part, we will present the related works that focused on emotion recognition. Table X below shows the details of the comparison.

Although the model results were not the highest compared to previous studies, it opened a new area of research in how the sentiment of speech can affect the detection of deepfakes if considered along with the emotions of images and audio. Thus, when all the experimental strategies were fixed and the ablation study was conducted, it was found that the sentiment of speech indicator increased the productivity of the model in detecting deepfakes. The model accuracy without sentiment of speech reached 76%, however, when adding the sentiments, the accuracy increased to 95.24%.

3) Comparative results with state-of-the-art baseline models. We also compare our model with well-known baseline methods like FaceForensics++[40] and LipForensics [41].

Including both visual-only and multimodal approaches. This helps to place our work in context and shows how our use of audio, video, and sentiment features adds value to deepfake detection. Table XI shows the details of the comparison.

As shown in Table XI, FaceForensics++ [40] is a visual-only approach that uses XceptionNet and achieved 95.7% AUC on the FF++ dataset. LipForensics [41], on the other hand, detect a deepfake based on inconsistencies in lip movements and performs even better, reaching 97.1% AUC. While both are strong and widely used methods, they focus only on the visual aspect of the content and don't take audio or emotional signals into account. This is where our approach brings something new, by combining audio, video, and sentiment features to better capture manipulations that might be subtle or spread across different modalities. On the other hand, our approach looks at the problem a bit differently. Instead of relying on just the visual side, we combine both audio and video inputs — and go a step further by including the sentiment behind the speech. This gives the model a better chance at catching those subtle manipulations that might slip past visual-only methods. When we tested it on a balanced sample from the FakeAVCeleb dataset, the model reached 95.24% accuracy, which shows how adding emotional cues from speech can really make a difference in spotting deepfakes, especially when the fakeness is hard to notice.

TABLE IX. COMPARING WITH RELATED WORKS ON THE SAME DATASET

Study	Deepfake detection	Emotion recognition	Sentiment of speech	Result
Pianese et al. [24]	Audio-based	No	No	91%
Khalid et al. [29]	Audio-visual based	No	No	81%
Raza and Malik [31]	Audio-visual based	No	No	92.9%
Muppalla et al. [32]	Audio-visual based	No	No	65.18%
Cheng et al. [7]	Audio-visual based	No	No	81.52%
Feng et al. [33]	Audio-visual based	No	No	95.8%
Oorloff et al. [34]	Audio-visual based	No	No	94%
Our study	**Audio-visual based**	**YES**	**YES**	**95.24%**

TABLE X. COMPARING WITH RELATED USED EMOTION RECOGNITION

Study	Deepfake detection	Dataset	Emotions based	Sentiment of speech	Result
Mittal et al. [19]	Audio-visual	DF-TIMIT	Yes	NO	94.8
Hosler et al. [20]	Audio-visual	DFDC	Yes	NO	99.5%
Conti et al. [35]	Audio	ASVspoof2019	Yes	NO	98%
Our study	**Audio-visual**	**FakeAVCeleb**	**Yes**	**Yes**	**95.24%**

TABLE XI. COMPARING WITH BASELINE MODELS

Study	Deepfake detection	Dataset	Result
FaceForensics [40]	Visual	FaceForensics++	AUC = 95.7%
LipForensics[41]	Visual (Lip-based)	FaceForensics++/ DFDC	AUC= 97.1%
Our study	**Audio, Visual, Sentiment**	**FakeAVCeleb**	**Accuracy= 95.24%**

V. DISCUSSION

As mentioned before, the proposed framework depends on the extraction and analysis of emotions from visual, textual, and audio modalities. BERT is used for speech sentiment, Swin Transformer is used for visual emotion, and wav2vec2 is used for audio tones. An LSTM-based fusion layer is then utilized to fuse the consistency calculated by comparing these emotional cues across modalities. To ensure the accuracy of the results and emotional studies separately, we conducted the ablation study. Table XII shows the results from the ablation study.

TABLE XII. RESULTS FROM THE ABLATION STUDY

Experiment	Accuracy	F1-score	Recall
Audio-visual with emotion recognition	76.19%	72.15%	73.3%
Audio- visual with audio emotion and text sentiment	76.19%	69.21%	68.33%
Audio- visual with visual emotion and text sentiment	76.19%	69.21%	68.33%
Audio- visual with sentiment analysis	80.95%	78.57%	81.67%
Our Approach	95.24 %	95.24%	95.45%

As we noted above, we conducted a study that focused on emotions in five experiments. The first experiment contained models of audio, image features, and emotional mismatch between audio and image, which is what came in previous studies. It's got modest results in terms of accuracy, F1-score, and recall. Moreover, in the second and third experiments, we observe that both F1-score and recall decreased compared to the first experiment. This indicates that partial modality fusion is insufficient to enhance the distinction between real and fake content. This justifies how the fourth experiment improved performance significantly, as we removed the partial modality fusion. Now the model measures the mismatch from three aspects: first, between the text and the audio, let's call it A. Second, between the text and the image, let's call it B, and third, between A and B. In the last experiment, which is our approach, we achieved the highest results, with a consistency across all metrics, which are accuracy, F1-score, and recall. Fig. 7 below shows the consistency across all metrics between all experiments. The final experiment demonstrates a strong balance across all evaluation metrics, with accuracy (95.24%), F1-score (95.24%), and recall (95.45%) closely aligned. This consistency indicates that the model maintains a reliable tradeoff between accuracy and recall. In contrast, earlier experiments show a notable gap between accuracy and other metrics. That suggests that partial fusion strategies (emotion only, sentiment-audio only, or sentiment-visual only) may lead to biased predictions or limited generalization. The balanced results in our approach highlight their robustness and reliability in detecting deepfakes across various modalities.

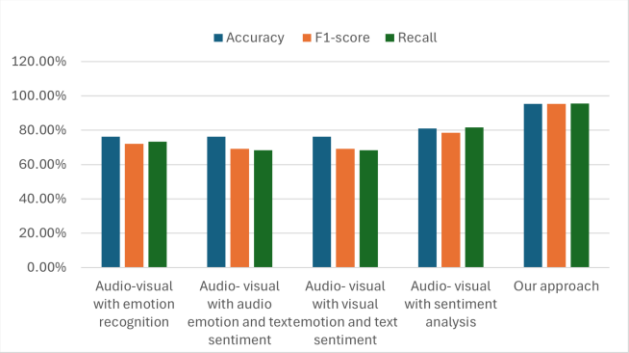


Fig. 7. Evaluation metrics of experiments.

We also explore additional enhancements which is not adding significant improvements. Table XIII shows the detailed results.

TABLE XIII. ENHANCEMENTS

Experiment	Accuracy	F1-score	Recall
The approach with gating mechanism	90.48%	90.45%	90.45%
The approach with the FGSM	95.24 %	95.24%	95.45%
The approach without enhancement	95.24 %	95.24%	95.45%

Finally, using the FakeAVCeleb dataset, this study examines how speech sentiment and emotion recognition affect detection accuracy. When compared to previous studies utilizing the same dataset, our model performed effectively, achieving an accuracy of 95.24%. Opening a new viewpoint of the effects of sentiment of speech on deepfake detection productively. To the best of our knowledge, no prior research using this dataset has examined this emotional aspect. Furthermore, in the domain of deepfake detection, no previous study has considered both speech sentiment and emotion recognition simultaneously. Our research presents this combination on the FakeAVCeleb dataset that provides a novel and useful viewpoint for developing multimodal deepfake analysis with the sentiment of speech.

VI. CONCLUSION

People's reliance on the Internet and social media has increased significantly in this era. With the development of artificial intelligence technologies, data manipulation has become accessible to everyone, and there are even free manipulation tools. In this study, we explored the framework for detecting deepfakes in video by using emotion recognition in three modalities: audio, images, and the sentiment of transcribed speech.

Our research contributed to this combination of emotions, achieved a high result, which is 95.24%, and provided a novel and useful viewpoint for developing multimodal deepfake analysis with the sentiment of speech. However, this study has certain limitations. The experiments were conducted solely on the FakeAVCeleb dataset, which may limit the generalizability of the results. This choice was made due to the dataset's unique structure, which provides audio and visual components—making it highly suitable for emotion- and sentiment-based fusion analysis. Moreover, while we relied on pre-trained models such as wav2vec2, BERT, and Swin Transformer for feature extraction, this decision was made deliberately to isolate and focus on the core contribution of this work: emotional consistency fusion across modalities. In future work, we aim to build our fine-tuned domain-specific models tailored for deepfake detection tasks, allowing for deeper integration of modality-aware learning and potentially improving detection performance and adaptability across datasets. Furthermore, we will conduct a detailed failure case analysis to address issues like speaker mismatches and low-quality audio. We also plan to integrate explainability methods such as SHAP and Grad-CAM to better interpret model decisions and enhance transparency.

REFERENCES

[1] Tableau [n.d.], 'What are the advantages and disadvantages of artificial intelligence?', https: //www.tableau.com/data-insights/ai/advantages-disadvantages. Accessed: November 29, 2024.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Akhtar, Z., Pendyala, T. L. and Athmakuri, V. S. [2024], 'Video and audio deepfake datasets and open issues in deepfake technology: being ahead of the curve', Forensic Sciences 4(3), 289–377.

[3] Spector, N. [2017], 'So it's fine if you edit your selfies, right? think again', https://www.nbcnews.com/business/consumer/ so-it-s-fine-if-you-edit-your-selfies-not-n766186.

[4] Liu, P., Tao, Q. and Zhou, J. T. [2024], 'Evolving from single-modal to multi-modal facial deepfake detection: A survey', arXiv preprint arXiv:2406.06965 .

[5] Wodajo, D. and Atnafu, S. [2021], 'Deepfake video detection using convolutional vision transformer', arXiv preprint arXiv:2102.11126.

[6] Limer, E. [2021], 'Bank robbers in the middle east reportedly cloned someone's voice to steal $35 million', https://gizmodo.com/ bank-

robbers-in-the-middle-east-reportedly-cloned-someo-1847863805. Accessed: 2024-11-24.

[7] Shah, S. [n.d.], 'Sih31st [kaggle notebook]', https://www.kaggle.com/code/sohamshah03/ sih31st/notebook. Retrieved:2024-11-12.

[8] Awsaf49 [n.d.], 'Asvspoof 2019 tfrecord dataset [data set]', https://www.kaggle.com/ datasets/awsaf49/asvspoof-2019-tfrecord-dataset. Retrieved: 2024-11-12.

[9] DASH Lab, FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. [Online]. Available: https://sites.google.com/view/fakeavcelebdash-lab/. [Accessed: Apr. 28, 2025].

[10] Cozzolino, D., Rössler, A., Thies, J., Nießner, M. and Verdoliva, L. [2021], Id-reveal: Identityaware deepfake video detection, in 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 15108–15117.

[11] Zhong, W., Tang, D., Xu, Z., Wang, R., Duan, N., Zhou, M. and Yin, J. [2020], 'Neural deepfake detection with factual structure of text', arXiv preprint arXiv:2010.07475 .

[12] Hamza, A., Javed, A. R. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z. and Borghol, R. [2022], 'Deepfake audio detection via mfcc features using machine learning', IEEE Access 10, 134018–134028.

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. [2020], 'Generative adversarial networks', Communications of the ACM 63(11), 139–144.

[14] Chesney, R. and Citron, D. [2019], 'Deepfakes and the new disinformation war: The coming age of post-truth geopolitics', Foreign Affairs 98, 147.

[15] Maras, M.-H. and Alexandrou, A. [2019], 'Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos', The International Journal of Evidence & Proof 23(3), 255–262.

[16] Becker, C., Conduit, R., Chouinard, P. A. and Laycock, R. [2024], 'Can deepfakes be used to study emotion perception? a comparison of dynamic face stimuli', Behavior Research Methods pp. 1–17.

[17] VISO.ai [n.d.], 'Visual emotion recognition with ai', https://viso.ai/deep-learning/ visual-emotion-ai-recognition/. Accessed: 2024-11-24.

[18] Abdallah, A., Al-Sharif, M. H. and Al-Zoubi, A. [2022], 'A systematic review of emotion recognition from video: Opportunities and challenges', Computational Intelligence and Neuroscience. https://doi.org/10.1155/2022/2645381.

[19] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A. and Manocha, D. [2020], Emotions don't lie: An audio-visual deepfake detection method using affective cues, in 'Proceedings of the 28th ACM International Conference on Multimedia', pp. 2823–2832.

[20] Hosler, B., Salvi, D., Murray, A., Antonacci, F., Bestagini, P., Tubaro, S. and Stamm, M. C. [2021], Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotionalinconsistencies, in 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 1013–1022.

[21] Saravani, S. M., Ray, I. and Ray, I. [2021], Automated identification of social media bots using deepfake text detection, in 'International Conference on Information Systems Security', pp. 111–123.

[22] Li, Y., Li, Q., Cui, L., Bi, W., Wang, Z., Wang, L. and Zhang, Y. [2024], 'Mage: Machinegenerated text detection in the wild', Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics 1, 36–53.

[23] Mcuba, M., Singh, A., Ikuesan, R. A. and Venter, H. [2023], 'The effect of deep learning methods on deepfake audio detection for digital investigation', Procedia Computer Science 219, 211–219.

[24] Pianese, A., Cozzolino, D., Poggi, G. and Verdoliva, L. [2022], Deepfake audio detection by speaker verification, in '2022 IEEE International Workshop on Information Forensics and Security (WIFS)', IEEE, pp. 1–6.

[25] Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A. and Böttinger, K. [2022], 'Does audio deepfake detection generalize?', arXiv preprint arXiv:2203.16263 .

[26] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3207-3216).

[27] Shao, R., Wu, T., Wu, J., Nie, L. and Liu, Z. [2023], Detecting and grounding multi-modal media manipulation, in 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. Accessed: 2024-12-18.

[28] Wang, J., Liu, B., Miao, C., Zhao, Z., Zhuang, W., Chu, Q. and Yu, N. [2024], Exploiting modality-specific features for multi-modal manipulation detection and grounding, in 'ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 4935–4939.

[29] Khalid, H., Kim, M., Tariq, S. and Woo, S. S. [2021], Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors, in 'Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection', pp. 7–15.

[30] Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z. and Palaniappan, K. [2020], Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning, in '2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)', IEEE, pp. 1–9.

[31] Raza, M. A. and Malik, K. M. [2023], Multimodaltrace: Deepfake detection using audiovisual representation learning, in 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 993–1000.

[32] Muppalla, S., Jia, S. and Lyu, S. [2023], Integrating audio-visual features for multimodal deepfake detection, in '2023 IEEE MIT Undergraduate Research Technology Conference (URTC)', pp. 1–5.

[33] Feng, C., Chen, Z. and Owens, A. [2023], 'Self-supervised video forensics by audio-visual anomaly detection', Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 10491–10503.

[34] Oorloff, T., Koppisetti, S., Bonettini, N., Solanki, D., Colman, B., Yacoob, Y. and Bharaj, G. [2024], Avff: Audio-visual feature fusion for video deepfake detection, in 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 27102–27112.

[35] Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F. and Tubaro, S. [2022], Deepfake speech detection through emotion recognition: a semantic approach, in 'ICASSP 2022- 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 8962–8966.

[36] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460.

[37] Yang, S., Chi, P., Huang, Y., Lim, C., Lee, H., Sung, D. L., Lin, J., & Yi, H. (2021). SUPERB: Speech processing universal PERformance benchmark. Interspeech 2021.

[38] Kirouane, A. (2022). BERT-Emotions-Classifier [Computer software]. Hugging Face. https://huggingface.co/ayoubkirouane/BERT-Emotions-Classifier

[39] Segni, M. W. (2023). swin-tiny-patch4-window7-224-finetuned-face-emotion-v12 [Computer software]. Hugging Face. https://huggingface.co/MahmoudWSegni/swin-tiny-patch4-window7-224-finetuned-face-emotion-v12

[40] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1-11).

[41] Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5039-5049).

APPENDIX A: GLOSSARY OF TERMS AND ACRONYMS

- Sentiment-Emotion Conflict: This is a discrepancy between a person's tone of voice or facial expression and what they say, or the sentiment or meaning of their words.
- MFCC (Mel-Frequency Cepstral Coefficients): These are characteristics extracted from audio signals that replicate the way sound is perceived by the human ear. They aid in capturing the speaker's voice's emotional cues in our investigation.
- SwinT (Swin Transformer): A smart image-processing model that breaks an image into windows and shifts them to capture visual details more accurately. We use it to analyze facial emotions frame by frame.
- wav2vec2: A speech model that picks up knowledge straight from unprocessed audio. In our situation, it is helpful to identify the sentiment expressed as well as the words themselves.
- BERT: A language model that understands context in text. We use it to analyze the emotional meaning of the transcribed speech.
- FGSM (Fast Gradient Sign Method): A method for determining if the model can still produce accurate predictions by significantly altering the input data. It enables us to assess the robustness of the model.
- Gating Mechanism: When merging several data types, a clever layer that learns to "turn on" or "off" specific properties helps the model concentrate on what really matters.
- Emotion-Sentiment Fusion: This is fusing a person's tone of voice or facial expression with their words to convey their feelings. This combination can highlight minute indications of video tampering.

A Novel Multi-Modal Deep Learning Approach for Real-Time Live Event Detection Using Video and Audio Signals

Integrating Audio-Visual Fusion for Robust Real-Time Event Recognition

Pavadareni R[1], A. Prasina[2], Samuthira Pandi V[3]*, Ibrahim Mohammad Khrais[4], Alok Jain[5], Karthikeyan[6]
Department of AI&DS, Chennai Institute of Technology, Chennai, India[1]
Department of ECE, Chennai Institute of Technology, Chennai, India[2]
Centre for Advanced Wireless Integrated Technology, Chennai Institute of Technology, Chennai, India[3]
Faculty of Economics and Administrative Sciences-Islamic Banks, Zarqa University, Zarqa, Jordan[4]
School of Electronics and Electrical Engineering, Lovely Professional University, Phagwara, India[5]
Department of Electronics and Communication Engineering, Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi - Vel Tech Road, Avadi, Chennai[6]

***Abstract*—Recent developments in live event detection have primarily focused on single-modal systems, where most applications are based on audio signals. Such methods normally rely on classification approaches involving the Mel-spectrogram. Single-modal systems, though effective in some applications, suffer from severe disadvantages in capturing the complexities of a real-world event, which thereby reduces their reliability in dynamically changing environments. This research study presents a novel multi-modal deep learning approach that combines audio and visual signals in order to enhance the accuracy and robustness of live event detection. The innovation lies in the use of two-stream LSTM pipelines, allowing for temporally consistent modeling of both input modalities while keeping a real-time processing pace through feature-level fusion. Unlike many of the recent transformer models, we are utilizing proven techniques (MFCC, 2D CNN, ResNet and LSTM) in a latency-aware and deployment-friendly architecture suitable for embedded and edge-level event detection. The AVE (Audio Video Events) dataset, consisting of 28 categories, has been used. For the visual modality, video frames undergo feature extraction through a 2D CNN ResNet and temporal analysis through an LSTM. Simultaneously, the audio modality employs MFCC (Mel Frequency Cepstral Coefficients) for feature extraction and LSTM to capture temporal dependencies. The features extracted from both audio and video modalities are concatenated for fusion. The proposed integration leverages the complementary nature of audio and visual inputs to create a more comprehensive framework. The outcome yields 85.19% accuracy in audio and video-based events due to the effective fusion of spatial and temporal cues from diverse modalities, outperforming single-modal baselines (audio-only or video-only models).**

Keywords*—Multi-modality; feature fusion; early fusion; concatenation audio-video signals; convolutional neural network (CNN); Long Short-Term Memory (LSTM); Mel Frequency Cepstral Coefficients (MFCC); ResNet (Residual Network)

I. INTRODUCTION

Event detection has become a pervasive application of artificial intelligence, found in applications as diverse as video surveillance, sports analysis, public security systems, and even entertainment. The identification and classification of significant events in time based on the analysis of multimedia data—mainly audio and video. However, traditional unimodal approaches based solely on audio or video often fail to record very well the complex interaction often occurring between vision and hearing in ordinary situations. Most events in real-world scenarios do not occur in isolation. These events are often characterized by noise, temporal overlap, and a complex interplay of audio and visual stimuli. A barking dog following a car, a crowd cheering a goal, or shattered glass during the act of burglary—all are clearer when we can hear and see them at the same time. This is why modern event detection systems need multi-modal solutions that can effectively fuse both audio and visual data and make better, more knowledgeable decisions. Researchers have made encouraging strides in this direction. For instance, Convolutional Neural Networks (CNNs) have done well in extracting meaningful patterns from audio waveforms, while 3D CNNs are able to handle spatiotemporal data from video streams. Yet, with all their promise, these methods hit some tough bottlenecks: synchronizing asynchronous data streams, reducing computational overhead, and building fusion mechanisms through which audio and video can actually complement each other, rather than merely coexist. To address these challenges, recent studies such as [1] and [2] have also explored the appropriateness of multi-modal deep learning architectures.

These studies experiment with new architectures such as light vision transformers and advanced fusion techniques in the hope of improving both the accuracy and versatility of event detection models in the wild [3]. The fusion process of deciding how and when to merge features between modalities has turned into a very critical aspect that can significantly impact the performance of the model. This is where AVFusion comes in—our proposed multi-modal architecture designed specifically for live event recognition. It was experimented with over the well-liked Audio-Visual Event (AVE) benchmark dataset, including 4143 annotated videos across 28 real-world categories of events.

AVFusion leverages principles from works such as [4], with the feature-level fusion approach being targeted towards learning information about how an audio and video stream correlate from one time period to another.

In the AVFusion framework, we extract audio features as Mel Frequency Cepstral Coefficients (MFCCs), which is a representation that mimics the human ear's perception.

These features are then passed through Long Short-Term Memory (LSTM) networks to capture the temporal development of the sounds. Visually, the video frames are processed through ResNet to extract spatial features, again followed by LSTMs to capture motion and continuity over time. By fusing the encoded information of both modalities rather than at decision time, we have a richer, more integrated representation of the event. This combined audio-visual processing, while enhancing classification performance, remains computationally efficient and enables the system to operate in near real-time without heavy hardware. Our approach draws inspiration from recent developments like [5] and [6], highlighting the importance of intelligent fusion in multi-modal tasks.

The rest of the study is organized as follows: Section II: Literature Survey gives a concise review of the shift from unimodal solutions towards advanced multi-modal systems with accomplishments and ongoing difficulties. Section III: AVFusion Framework explains the structure in terms of preprocessing tasks, audio and video feature extraction, fusion pipeline, and training techniques like early stopping to prevent overfitting. Section IV: Experimental Setup. Section V: Results and Discussion reports empirical results on the AVE dataset, such as performance measures, visual plots, and comparative analysis to show the robustness of AVFusion. Finally, Section VI concludes the study.

While each of the elements used in AVFusion - MFCCs, 2D CNNs, ResNet, LSTMs, and concatenation at feature-level - is very well established, it is this focus on integration that makes this work novel. Most of the state-of-the-art models emphasized architectural complexity at the cost of runtime efficiency, while AVFusion was built for real-time event detection in a complex, very real-world setting. This framework not only focused on accuracy, but our aim was to find a balance among latency, accuracy, and modularity for implementation into surveillance, fault detection and embedded systems. This perspective challenges the idea that only architectural innovation leads to novelty, but also affirms how smart design for systems can produce feasible, scalable solutions to complex deep multi-modal challenges.

II. Literature Survey

Over the past decade, event detection has undergone a significant transformation. It began with simple unimodal approaches and evolved into complex multi-modal frameworks integrating auditory, visual, and contextual cues. This historical progression highlights how foundational research laid the groundwork for more integrated and adaptive systems. Each stage rounds a rather particular response to emerging real-world events that become progressively more complex.

In the onset research, most methods for event detection focus specifically on audio or visual sources, rarely both. Audio-only systems were among the first to make breakthroughs. One example is a 2018 study that used some spectrogram-based recognition of speech commands in conjunction with its own dataset and achieved an impressive 95% score on it [2]. The effectiveness was impressive during specific time periods, but it had no modeling of time-it could not know how the sounds evolve in time, leaving it deaf to dynamic auditory contexts.

Advancements soon followed. A 2023 study applied Electrodermal Activity (EDA) features in CNNs and LSTMs, achieving 96.6% accuracy on a Kaggle dataset. Though promising for real-time applications, it remained restricted to audio input without visual data. This model could not distinguish between acoustically similar but visually distinct events or understand spatial dynamics, a critical shortcoming in nuanced scenarios.

The field of video-based event detection has developed significantly over time. In 2017, researchers developed a method for detecting visual events (fights, crashes, and falls) using CNNs and LSTMs via spatiotemporal analysis [9]. The method was successful in recognizing visual events temporally, but did not include the auditory modality, and was relatively slow and computationally heavy, especially when it came to utilizing its predictive capabilities for real-time video-based decision making. By 2024, researchers began applying transformers in video classification pipelines [3], but effectiveness in modeling spatial and temporal dependencies was uneven at best. Despite advances made over the years, most video-based approaches still do not capture the auditory modality when optimizing for video classification tasks, and there are not typically any major increases in classifier accuracy for these tasks.

Having been trained and tested in narrow or conformed settings, those single-modal pipelines had difficulty capturing the complex and multi-sensory aspects of the scenarios outside the laboratory walls. Further studies led the research to develop modalities whereby multiple sensory streams are integrated to create a more complete representation of the world. The early attempts in this domain were made during the early 2010s. A concerted effort was made by one team to bring audio and video into a single platform using CNNs in the year 2014 [6]. Although pioneering, the model lacked mechanisms for encoding temporal relationships. The same year, the concept of grouplets was introduced to temporally align audio and video characteristics [7]. It was found to have an extension to innovations, but noise sensitivity and lack of a quantitative basis hold it back severely. At the same time, photofusion for visuals through CNNs and RNNs would prove to be a more organized exercise in temporal modeling by 2015. Once again, however, audio remained unincorporated in many early visual models [8].

In 2020, the state-of-the-art complete survey on feature fusion techniques was published. It enumerated and classified strategies as early, late, and hybrid fusion and created a conceptual map for future research [4]. The resulting taxonomy was to bring unity to the field and inspire a diversely methodological and application-specific solution scheme.

This momentum has accelerated rapidly. In 2019, a milestone was defined in the detection of audio-visual inputs along contextual metadata by the combination of deep learning model to recognize aggressive behavior on trains. Although

specific accuracy metrics were not reported, it was among the pioneering efforts to integrate real-world context into multi-modal models. A dynamic fusion model implemented in 2023 for micro-video recommendations introduced visual, auditory, and textual cues into fusion via meta-learning; so far, it achieves an accuracy of 87.9% on the UCF51 dataset. This opened significant new opportunities, showing that flexible fusion models may be better than static, single-modal systems [24].

By 2023, the most sophisticated multi-modal models were audio-visual speech recognition systems. Although high in computational cost, a model that used 3D CNNs together with BiLSTMs reached a phenomenal score of 98.56% on the LRW dataset [11]. Nevertheless, such accuracy boils down to a very high computational cost. However, it once reflected the power of multi-modal integration when resources permitted.

These advances were supported by the development of relevant datasets. The AVE data set, established in 2018, was an example. It collated 4,143 real-world recorded samples concerning 28 categories, such as from musical performances to machinery failures [23]. Today, it remains a significant benchmark for testing audio-visual fusion models. Other datasets like UCF51 and Kinetics-Sounds caught importance in 2024, scaling up generalizability to broad action recognition tasks and helping models achieve more robust results, such as 87.9% accuracy [22].

New research is steadily pushing the boundaries. The 2020 study utilized CNNs and LSTMs for visual temporal modeling [12]. Though excluding audio inputs, it nevertheless advanced some techniques for fusion. A parallel development in 2021 saw multi-modal analysis move into environmental monitoring, equipping sensor data and deep learning for urban water quality predictions-a powerful demonstration of multi-modalism beyond the traditional event detection pathway [13][10].

The fusion strategies have reached maturity. The early fusion integrates features before the learning process begins, as observed in a 2015 video event classification study [14]. Late fusion is applied in cases such as newborn pain detection in 2021; it combines outputs after separate models work on each modality, allowing modularity and interpretability [15]. Hybrid fusion is a sophisticated technique: it went to work in a 2023 medical imaging study where CNNs and transformers were integrated in the middle of the stream to preserve a variety of feature representations while enhancing model simplicity [16].

Attention mechanisms are receiving extended focus lately. The visual relationship model in the year 2022 used attention layers to isolate certain key interaction scenes, such as people making physical contact [17]. On the other hand, the affect detection system in 2024 was trained on long video-audio sequences using stacked transformers for picking up subtle emotional cues over time [18].

However, challenges still remain. In the survey on anomaly detection in 2019, it was highlighted that datasets presently available do not capture the messiness of the real world-the presence of interference, occlusion, and noise is rarely represented [19]. This notion was echoed in a healthcare review of 2023, which was calling for more "imperfect" multi-modal datasets that increasingly reflect lived environments [20]. Nevertheless, while transformers have greatly transformed fusion, a 2024 survey pointed out their failure due to sizable computations, restricting their applications in real-time or edge settings [21].

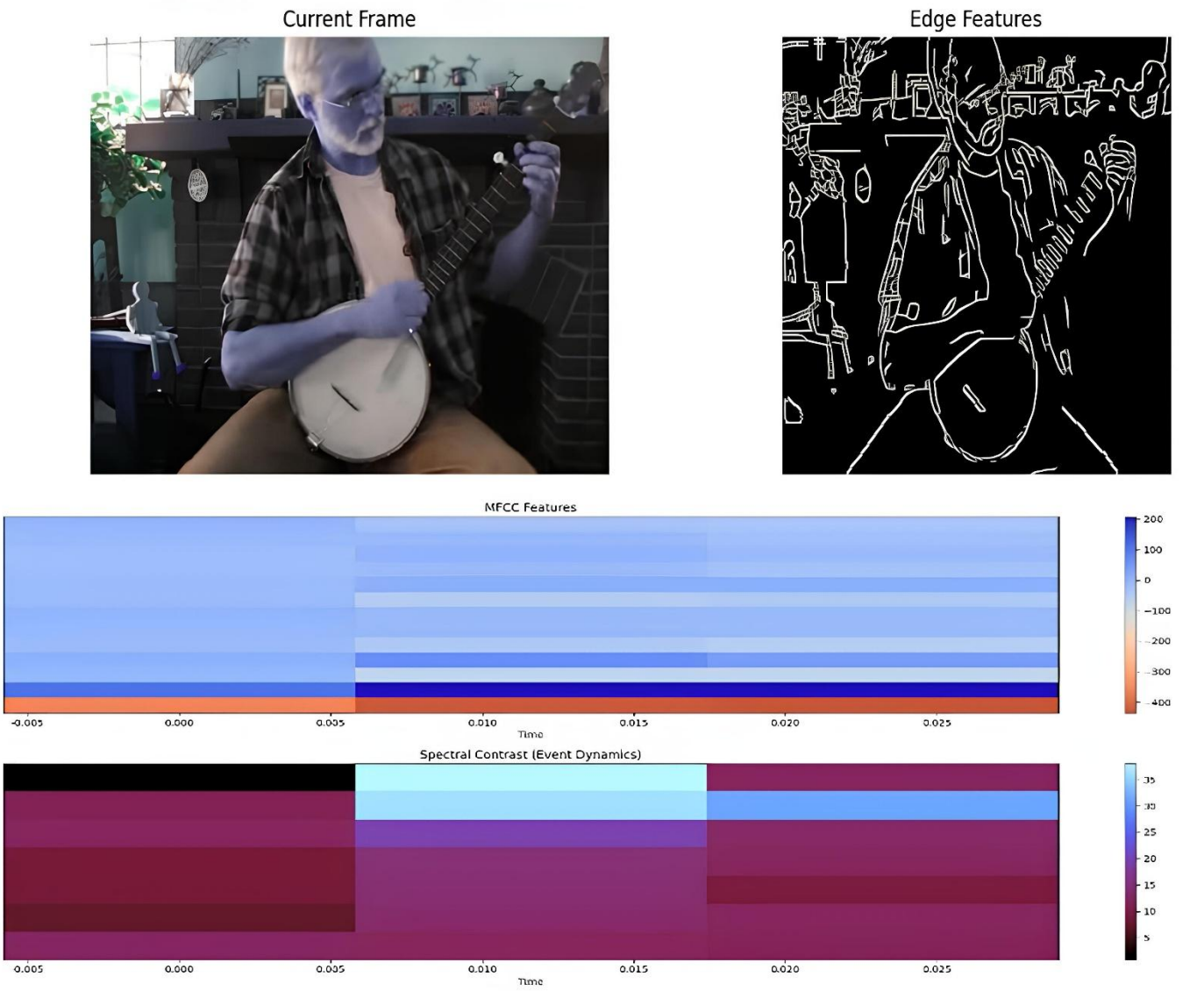
Text-based modalities also acted as a complement. A 2014 study on social event detection applied NLP techniques, such as TF-IDF, without any audio or video [5]. While such work appeared simplistic according to today's standards, it paved the way for the convergence of language, vision, and sound.

All influences dovetail them into AVFusion: our multi-modality framework, which integrates lessons from the previous decade into a tangential practice: with MFCCs and LSTMs to extract temporal patterns from audio, ResNet, and LSTMs to model spatial and sequential video dynamics, followed by feature-level concatenation of these streams into a unified representation space. Although not the most computationally intensive solution, it strikes a good balance between reliability and efficiency. Achieves 85.19% accuracy on AVE. It shows that good fusion could be as competitive as more heavyweight fusion setups. It does not require the most advanced state-of-the-art systems like 2023 precision health fusion model [20] or 2024 AVE specific architectures [22], AVFusion proves intelligent design, not plain raw complexities, could realize a world impact.

Our multi-modal framework, the modification from the last decade into an operational, streamlined system: MFCCs and LSTMs tap temporal patterns from audio, while ResNet and LSTMs model spatial and sequential dynamics of video. Feature-level concatenation binds these streams into a united representation space. Although it might not be the most computationally intensive solution, AVFusion achieves a balance between reliability and efficiency. It achieved an accuracy of 85.19% on the AVE dataset, showing that thoughtful fusion can vie with the more complicated setups. Not even calling for the state-of-the-art systems such as the 2023 precision health fusion model [20] or 2024 AVE specifications architectures [22], AVFusion proves that effective system design and practical impact are not solely a function of model complexity.

III. AVFUSION FRAMEWORK

The AVFusion method offers a high probability of detecting live events by merging audio and video in an intelligent way. In a four-fold setting: audio preprocessing, video preprocessing, fusion, and classification, it could generally be viewed as a pipeline shown in Fig. 1.



(a) Multi-modal feature extraction during a Musical event



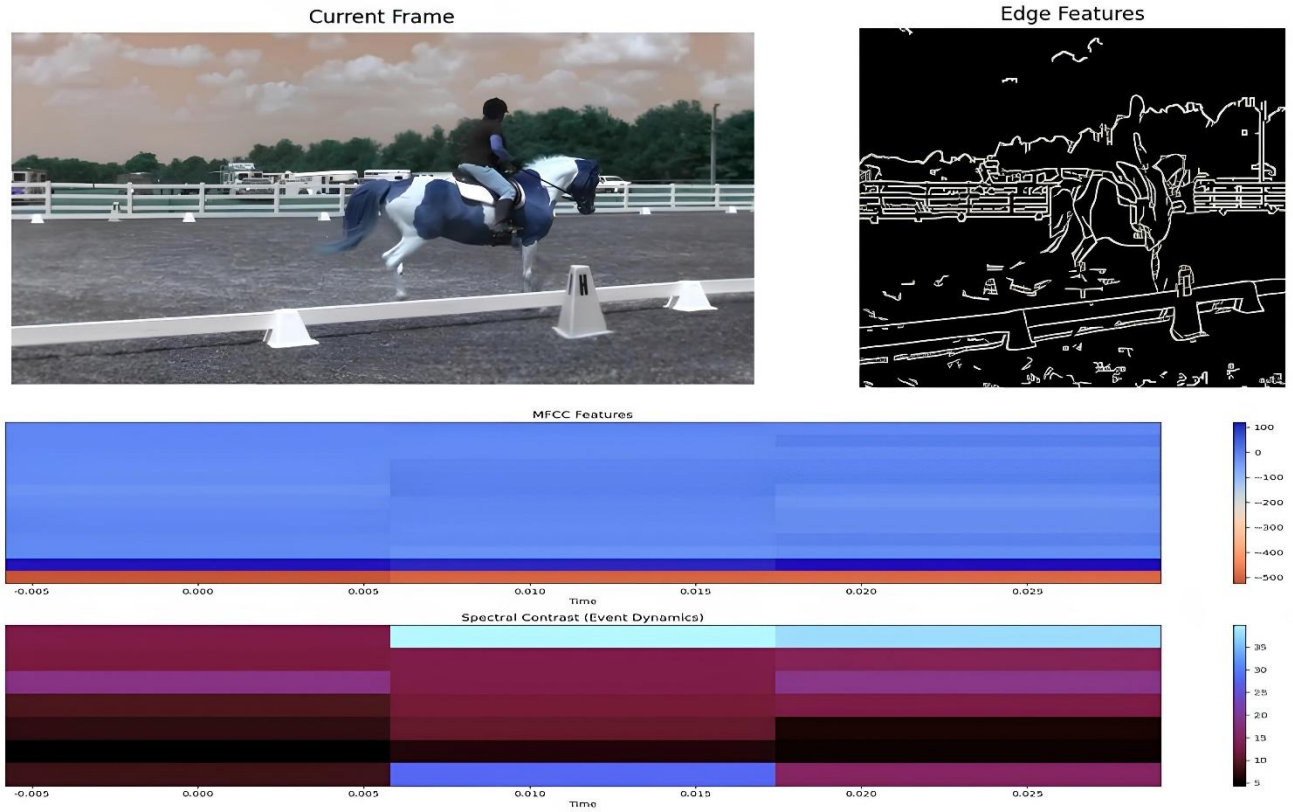
(b)



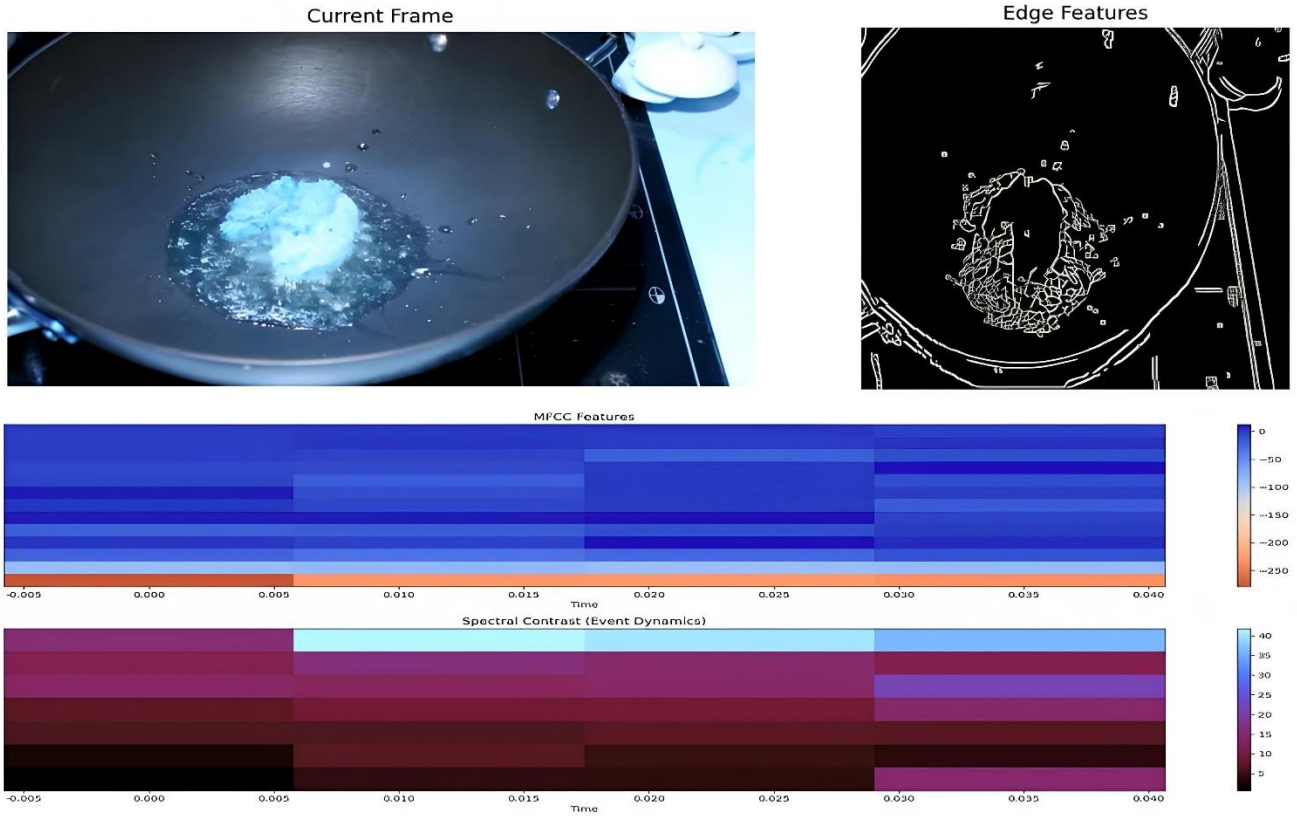
(c)



(d)



(e) Multi-modal feature extraction for an event of horse-riding.



(f) Multi-modal feature extraction for a cooking-related audio-visual.

Fig. 1. (a-f): Three AVE events with raw frames, audio features, and processed outputs [25].

A. Audio Preprocessing

In this context, audio refers to the temporal signal representation of sound events, such as the sharp twang of a banjo or the continuous sizzle of a frying pan. It is the process of converting an audio stream into Mel-Frequency Cepstral Coefficients (MFCCs)-a signal representation that resembles human hearing. It starts with a filtering step known as a pre-emphasis, which shifts the overall gain towards the high frequencies.

$$y(t) = x(t) - 0.97\,x(t-1) \quad (1)$$

Here, x(t) represents a raw audio signal, whereas y(t) stands for its filtered version. Afterwards, an audio signal must be segmented into short frames, and a Fast Fourier Transform (FFT) is computed to extract the power spectrum:

$$P(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn} \quad (2)$$

where, P(k) is a complex-valued frequency component, x[n] is the time-domain signal at sample n, N is frame size, k is frequency bin index, $e^{-j\frac{2\pi}{N}kn}$ is the complex exponential kernel. Next, a Mel-scale filter bank is used to filter the signal to match human audition:

$$m(f) = 2595 log_{10}(1 + \frac{f}{700}) \quad (3)$$

where, f represents frequency in Hertz (Hz). We then take the spectrum's log to match loudness perception, apply a Discrete Cosine Transform (DCT), and achieve MFCCs:

$$C_k = \sum_{m=1}^{M} log(S_m) cos(\frac{\pi k}{M}(2m - 0.5)) \quad (4)$$

where, Ck are the k-th MFCC coefficients, Sm energy of the m-th Mel filter bank, and M is the number of Mel filter banks. This DCT decorrelates the log-Mel energies and compacts the signal's most important spectral information into a small number of coefficients. Fig. 2 shows this whole audio preprocessing flow, from raw signal to MFCCs. You can see the MFCCs and spectral contrast in action for three AVE events—Banjo, Horse Riding, and Frying—in Fig. 1(a), Fig. 1(e), and Fig. 1(f).

Those figures show the raw video frame on top, with the audio features (MFCCs and spectral contrast) below, capturing the sound patterns over time—like the twang of a banjo or the sizzle of frying. After getting those features, they are fed into an LSTM to track how the audio evolves, which is crucial for events that unfold over a few seconds.

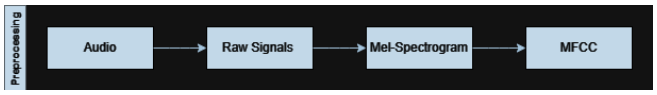


Fig. 2. Audio preprocessing flowchart.

B. Video Preprocessing

Video is about rapidly transforming shapes; for example, a horse in a gallop or the shape of a frying pan. We sample visual frames at 4 f/s, resize the frames to square 224×224 pixels, and normalize them using mean subtraction. A 2D-CNN on ResNet-50 extracts spatial features:

$$y(i,j) = \sum_{m=1}^{M-1}\sum_{n=0}^{N-1} x(i+m, j+n) \cdot W(m,n) + b \quad (5)$$

Here, x is the input image, W is the convolutional filter, b is the bias, and y (i, j) represents the activation at position (i, j) in the output feature map. This convolution captures spatial features such as edges, textures, and object contours in the input frame.

Temporal Modelling: The features are passed through an LSTM network that encodes the temporal evolution of video events. This LSTM allows the model to see how visual patterns evolve in time and is a requirement to identify the occurrence of events that last over a sequence of frames, for example, a fight or a car crash. An LSTM comes next to catch how those features change over time:

$$h_t = LSTM(h_{t-1}, F_t) \quad (6)$$

- ht is the hidden state at time step t,
- $ht-1$is the hidden state at the previous time step.

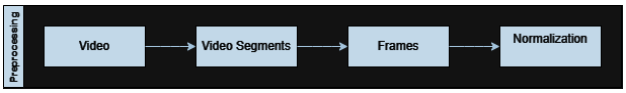


Fig. 3. Video preprocessing flowchart.

Fig. 3 maps out this video preprocessing flow, from frames to LSTM output. For a real look at it, check out Fig. 1. Fig. 1(b), (c), and (d) show the raw frames for Banjo, Horse Riding, and Frying—those are the starting points. Then, Fig. 1(a), (e), and (f) show the edge features after the CNN processes them, highlighting key shapes like the banjo's strings, the horse's legs, or the pan's edges. Those processed outputs help the model zero in on what makes each event tick.

C. Model Architecture

The architecture of AVFusion is a parallel architecture that consists of both an audio part and a video part, which are ultimately fused for joint event prediction. The audio part of the AVFusion model consists of a 3-layer LSTM with 128 hidden units taking in 40-dimensional MFCC features for each frame to extract the temporal dynamics of sound. The video part of the model contains spatial feature extraction, using ResNet-50, then a 2-layer LSTM with 256 hidden units to learn motion and temporal dependencies across video frames.

Additionally, both pipelines of the AVFusion model are trained end-to-end from their associated loss function with a shared loss function that allows the model to optimize both modalities at once without any spatial or temporal lag. Batch normalization is applied to each layer of the AVFusion model to stabilize the training of the model. In additional, a dropout rate of 0.5 is also used to combat overfitting. The final stage of the AVFusion model is to align the outputs of both audio and video pipelines over the frame, take the concatenation of the features from both modalities, and treat as a single representation with a prediction head for classification.

D. Fusion and Classification

Bringing both modalities together to bring forth a more accurate and holistic event detection model after feature generation by both audio and video models separately predicted is the last process. The fusion process includes Concatenation: Audio features and video features are concatenated into one

feature vector. This allows the model to take into account both visual as well as audio information of an event at once. The concatenated features will pass through fully connected networks, which can also be referred to as dense layers so that it can learn the relation between audio and video features. The combined feature vector is calculated as follows:

$$x_{fusion} = Concat(f_{audio}(x_{audio}), f_{video}(x_{video})) \quad (7)$$

- where, f_{audio} and f_{video} are temporally aligned feature sequences with matching time steps. The fusion is performed along the feature dimension after temporal alignment, and $Concat$ denotes the concatenation of the audio and video feature vectors. x_{audio} and x_{video} represent the feature vectors obtained from the audio and video models, respectively

Temporal Alignment: Before combining the audio and video features to be fused, the audio and video features need to be temporally aligned so that both content modalities (e.g., audio and video) are in synch. MFCCs are calculated with a window and a hop size so that an MFCC vector is calculated when each frame of the video is presented (e.g., one MFCC vector per frame). Consequently, f_{audio} and f_{video} are in temporal alignment at the same time step. This allows for concatenation by element (or frame) so that the fused (concatenated) feature vector is from the same time or time context from the audio and video modality at the same time steps.

Dropout is applied to prevent overfitting, and the output layer uses a sigmoid activation function to perform binary classification, determining whether an input belongs to the target class or not. The final classification is performed by passing the combined feature vector through a fully connected layer (dense layer) and a sigmoid activation function:

$$y_{pred} = \sigma(W x_{fusion} + b) \quad (8)$$

- W is the weight matrix,
- b is the bias term,
- $xfusion$ is the concatenated feature vector,
- σ is the function that outputs a value between 0 and 1, representing the sigmoid activation function, the probability of event detection.

AVFusion is shown in Fig. 4, with audio pipeline (3-layer LSTM, 128 units, 40 MFCCs) and video pipeline (ResNet-50, 2-layer LSTM, 256 units) that merge into a single decision. Batch Normalization along with dropout (0.5) is applied during all training processes to make the training process stable and reliable. AVFusion's design paradigms emphasize operational efficiency, data synchronization and modular scalability; rather than layers of abstraction or multi-layered fusion. Preprocessed and temporally synchronized audio and video streams assist the model to integrate asynchronous cues, without resorting to slower and variable time-frames of attention-based processing. Class imbalance is addressed through dynamic loss weighting and PCA is used for dimensionality reduction; such that the model is able to retain functionality even if hardware constraints are imposed. Although AVFusion employs a feature-level concatenation approach for fusion, a relatively simple fusion approach, this is intentional and allows the model to maintain nearly real-time inference speeds (30 FPS) with interpretability and generalizability; therefore, AVFusion's original contribution as a system is the coordination of many effective approaches into a single operationalizable.

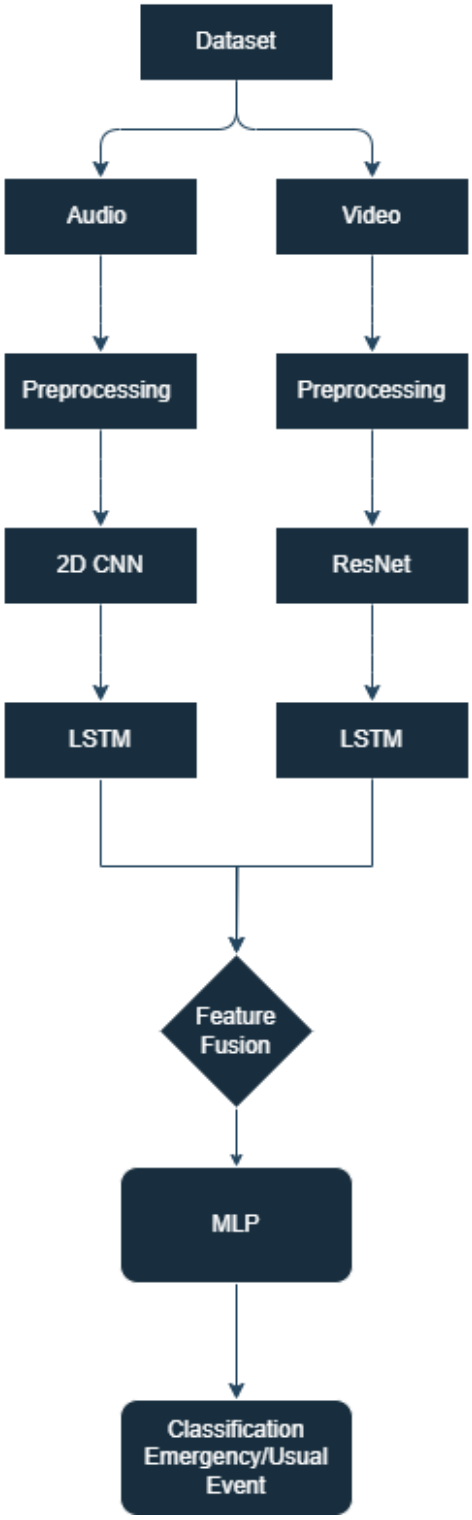


Fig. 4. Complete AVFusion architecture.

E. Evaluation

We examined the robustness and real-world applicability of AVFusion beyond training and validation, evaluating a new and unseen test split of the AVE dataset. This step was necessary to evaluate how well the model generalizes to different audio-visual scenarios previously unknown to it.

A key part of the training scheme that became a problem was the inherent class imbalance in the AR-Videos dataset. While the dataset contained common events (for example, general background noise or silent scenes), it also contained events that are extremely rare events (for example, frying or blowing out candles). If we are to ignore this imbalance, it is likely that the model would bias its predictions towards the majority classes - this results in inflated overall accuracy at the expense of poorly represented categories. As a technique to improve the model's performance overall, we introduced class weighting to the training scheme so that the underrepresented events in the dataset were weighted more heavily in the loss function for the training model. As a result, the underrepresented events in the model were encouraged to be formally attended to, therefore uncovering more subtle patterns, which leads to better generalizabililty and less likelihood of certain high-frequency classes.

To reduce the class imbalance that exists in the AVE dataset, we used class weights that were determined based on the inverse of the class frequencies. For every class c of the form c (wc), will be equal to $= \frac{1}{f_{c,}}$, is the relative frequency of class c in the training data. These class weights were applied to the loss function in the training process, which ensured that rare events were contributing relatively more to the optimization than the frequent events.

The second measure we took was early stopping and ReduceLROnPlateau to avoid overfitting, which is crucial in the presence of fewer high-signal events. These methods also monitored validation loss in real time, stopping training once the improvements plateaued, while also dynamically reducing the learning rate to refine optimization in the last few epochs.

While speed is a crucial variable in detection, other methods may be just as worthy. In live deployment scenarios such as surveillance or smart monitoring, the model's inference time on the test videos was observed. The AVFusion model stood up to the demands of a balanced approach to speed and accuracy; therefore, the model's performance supports its potential for real-time application.

The confusion matrix shown in Fig. 5 gives a detailed account of AVFusion's performance on 28 AVE classes. Good performance is noted for instances with obvious audio-visual signatures (like "instrument playing"), while slips are also noted for events with overlapping cues or which occur in complex environments. This frank profiling not only maps out the successes but also gives pointers toward future refinements.

The thorough benchmarking against single-modal baselines set forth in Section V, leveraging AVFusion to excel above audio-only or video-only methodologies, makes a statement in itself. It is thus established that the system is integrated: the whole is greater than its parts.

F. Integration with External Systems

It was created as more than just a proof-of-concept or academic experiment; it was made for actual deployment. One interesting application concerns the potential incorporation of this Tech-Scopes Data Connectivity's Connector system integrates AVFusion with power monitoring units (PMUs) to detect faults using embedded device data.

Tech-Scopes collects information regarding faults from embedded devices placed in critical infrastructures. These data are then collected into a centralized point of data collection, which can be identified as a nerve center in that ecosystem of monitoring. That point comprises a network of servers consolidated with display monitors and external hardware interfaces, and then a cluster of 14 PMUs, with each assigned to monitor voltage, current, and frequency across the grid.

AVFusion would sit in this very framework with audiovisual contextualization along those existing electrical information streams. With the connection of cameras and microphones placed close to monitored equipment, that is, transformers, switches, and circuit breakers, it will be capable of streaming environmental data in real-time, thus increasing the possibility of fault detection using multi-sensory analysis.

Consider the following: a power transformer starts to give off a sizzle-like sound, and as the flash hits a fuse, brief sparks are visible to the camera. By themselves, given the context mentioned, these signs would hardly excite suspicion within the walls of any classical PMU, heavily relying on electrical parameters as first indicators. However, an AVFusion trained to detect such Arabic patterns could have quickly associated these events and promptly generated an alert in context that hints at a potential equipment failure or safety hazard before it even goes down.

This real-time, sensory-rich detection embodies a paradigm shift in how we view power grid safety. Instead of being based on numerical anomalies in electrical signals, AVFusion now introduces the capability of visual and auditory intelligence into the mix, greatly enhancing its ability to recognize early-stage faults or irregularities proceeding serious failures.

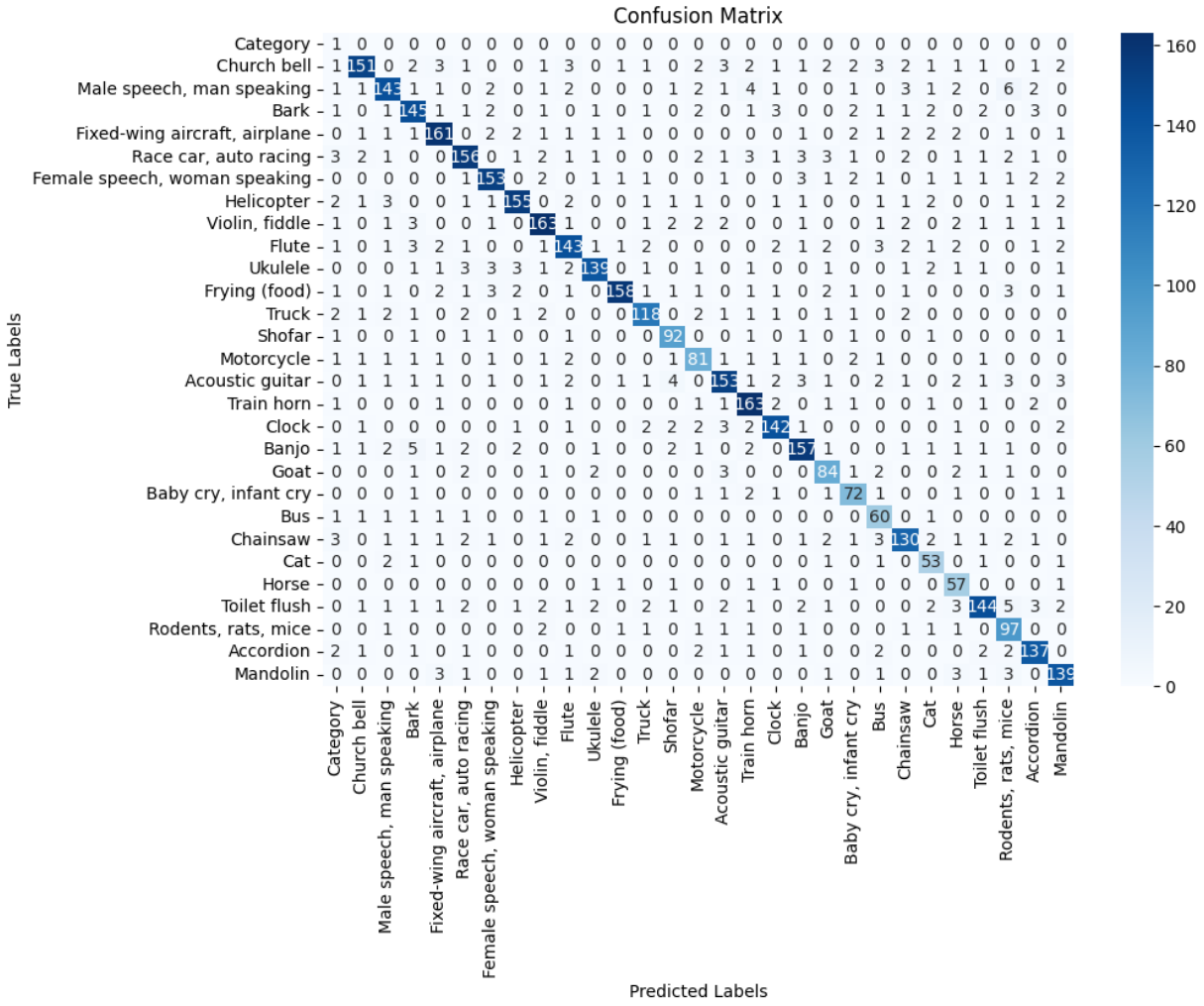


Fig. 5. Confusion matix.

Still conceptual at this stage, the collaboration has a strong foundation upon which next-generation predictive maintenance tools capable of multi-modal fault detection would truly become a reality by integrating AVFusion and PMU-based systems like those from Tech-Scopes. This is pivotal towards resilient and intelligent energy infrastructure.

G. Optimizations

But large deep-learning models are associated with great computation requirements, something that makes them completely unfit for scenarios which cannot compromise on quickness and responsiveness. In such situations, AVFusion has been designed, and many deliberate optimizations have been employed to ensure that AVFusion runs optimally, without sacrificing results.

In practical applications, it starts with the dimensionality reduction. Indeed, MFCCs are good for the audio features; however, in their raw form, they tend to be computationally heavy. Therefore, take 20 dimensions of the dimensionality of the MFCC features using Principal Component Analysis (PCA); this cuts down majorly on the memory requirement while speeding up the processing, as it still retains most of the key informative characteristics of the audio signal.

Batch processing then comes. The batch size of 32 counts will give a good trade-off between memory consumption and computational throughput, hence allowing the model to process data efficiently in parallel using the breadth of GPU processing capabilities.

The other major pipeline involves a convolutional neural network (CNN) that has been tuned for stability and performance. The most wondrous thing about it was gradient clipping, in which case the gradients do not explode during back propagation. The gradients have been clipped on a threshold, thereby ensuring convergence of the algorithm and definitely stable training.

Though the model AVFusion emulates ResNet-50 with LSTM layers, it will be even more practical in the future as there are optimizations to support the model as a PCA dimensionality reduction step, batching the output from openVINO, and the limited NLP dimension with features fusion. The "lightweight" has to do with the model's ability in real-time inferring with commercially available GPUs (RTX 3090) without the addition of multiple large transformer stacks or excessive amount of multi-head attention. There will need to be some reduction in

model size in the future by pruning or through the use of backbones such as MobileNet or EfficientNet-lite.

Although the model is not light in the absolute sense, it is relatively computationally efficient in comparison to transformer-based architectures or other hybrid architectures that leverage cross-modal attention. Furthermore, AVFusion has modular design considerations that are easy to implement and deploy, as well as make it responsive to real-time applications such as embedded fault detection and surveillance. Next steps include investigating backbone replacement (e.g. MobileNetV3, EfficientNet-lite) or exploring lightweight attention-based hybrid fusion modules.

AVFusion executes at about 30 frames per second (FPS) on a single NVIDIA RTX 3090 GPU with an average response time of ~ 32 ms per video segment (10-seconds long with matched audio). The total parameters in the model are about 38.5M, and the model processes at under 7.5 GFLOPs per inference time, which makes AVFusion fit for real-time applications in surveillance and monitoring systems.

AVFusion is capable of running video data through any pipeline at 30 frames per second, which is ideal for real-time requirements. This makes the system not only able to detect events accurately but also quick enough to be applied as a practical system for situations like surveillance, fault detection, or emergency response.

IV. Experimental Setup

The two widely practiced interventions were incorporated in that regard to train stability and generalization. First came early stopping with a patience threshold set at 10 epochs, which means once the validation performance of the model did not improve for 10 consecutive epochs, training would cease so as not to overfit. The second one we had included was ReduceLROnPlateau, variable learning rate scheduling, monitoring validation loss; if any improvements would stagnate, it would reduce the learning rate by half, making the model adjust weights finely toward the end of epochs and perfecting its learning without much retraining.

All training and evaluations were performed using PyTorch on an NVIDIA RTX 3090 GPU, capable of the throughput required to process high-resolution video streams and dense audio spectrograms simultaneously. Since real-time event detection was among the core requirements for deploying AVFusion, we made processing speed a priority during the study. The system operated at an impressive rate of 30 frames per second (FPS) throughout, thus meeting the demands of online video analysis and real-time applications.

In our evaluation of AVFusion, we benchmarked it against single-modal baselines (i.e., those trained on audio or video inputs only) and the best competing multi-modal systems available in the literature. These comparisons explicitly showed the strength of AVFusion's feature extraction capacity.

V. Results and Discussion

This section dives into how AVFusion performs on audio-visual event detection, using the AVE dataset. It's built on a mix of 2D-CNNs and LSTMs for both audio and video streams, tied together with a fusion mechanism. This section presents the model's performance, explains the fusion process, and outlines the most important metrics used to evaluate it.

A. Model Performance

The AVFusion has thus exhibited tremendous results by attaining an accuracy of 85.19% on the AVE test set, which strongly corroborates the model architecture and the fusion strategy. The accuracy level speaks to the successful integration of audio and visual features, and the accuracy reflects successful integration and strong generalization across event types, to see how well the system generalizes to varying event types.

In dealing with sound, AVFusion takes advantage of Mel-Frequency Cepstral Coefficients (MFCCs) to characterize the rich texture and pattern of sounds. These features are then fed into Long Short-Term Memory (LSTM) layers, highly competent for modeling temporal dependencies-very critical for identifying time-varying sound events, such as the gradual twang of a strummed banjo string or the faint shifting of pitch during frying. Such sounds would almost be inaudible for static audio encoders, which is why the LSTMs represent a critical part of the audio pipeline in AVFusion.

As for the visual stream, it employs the 2D Convolutional Neural Network (CNN) to extract rich spatial features-the silhouette of a person brushing his or her teeth or a flame coming out of a gas stove-and LSTMs for sequencing in time, helping the model learn the pattern of occurrence, such as a horse moving or someone playing a drum. The video modality helps disambiguate sounds with their visual counterpart, thus increasing reliability and context-awareness in event detection.

Fusing both modalities is a key design element in itself: fusion by feature level, which brings together the temporal intelligence of audio and the spatial-temporal richness of video. Merging both perspectives allows AVFusion to perform just about two orders of magnitude better than a unimodal baselines, which typically miss classifying or entirely loses an event that is subtle.

Fig. 6 shows the accuracy-and-loss epochs training dynamics: This x-axis holds the epochs of training (50, in total), and the y-axis shows the performance metrics. With training, accuracy keeps climbing until it reaches a sprawl above 85.19%, while loss keeps on decreasing because there is effective learning without overfit. This convergence is smooth, indicating that the architecture and training strategy, including the fusion mechanics and optimizer settings and regularization techniques, were all well aligned.

From here, we see that such a potent training curve, combined with such high final performance, demonstrates the robustness and adaptability of AVFusion, setting a solid foundation for future real-time real-world audio-visual event detection systems. Fusion of Audio and Video Modalities. AVFusion's fusion is a two-step process: It handle audio and video features separately, then merge them for the final prediction. For video, it use a 2D-CNN to pull spatial features—like the outline of a frying pan—then a Time Distributed wrapper and LSTM layers to track how those frames change over time. For audio, it extracts spatial features with MFCCs and use LSTMs to follow the sound's temporal flow—like a frying

sizzle coming and going. The fusion takes place at the decision level. It combines the audio and video features into a single vector and passes it through a fully connected layer with a sigmoid activation. Doing so allows the model to make a better-informed decision by reconciling the spatial information of the video with the time patterns of the audio.

B. Evaluation Metrics

AVFusion's mechanism of fusion is deliberately and carefully conceived in two stages, whereby audio and video features are first processed separately and merged later for a joint final prediction. This separation allows each modality to exploit its capability before passing on both members' insights for a complementary fusion.

From the video perspective, the model begins with a 2D template CNN to extract useful spatial representations from individual frames - recognize a frying pan shape or the flickering of a flame or perhaps the motion of an instrument. In order to maintain the temporal order of visual cues, these spatial features are wrapped into a TimeDistributed layer, which simply treats each frame as a time step. The entire sequence is then dumped into an LSTM layer so that the model can witness how these spatial features evolve over time and thus learn motion dynamics, such as the forward-and-back swing of a golf club or the slow tilt of a falling object.

The same care was taken with input audio. Starting with audio signal processing, audio signals are converted first to MFCCs, considerably compact spatial representations of sound textures. These MFCCs that encode variations of pitch, tonality, and energy are then input to the LSTM layers, which forge the capability of tracking sound patterns as they change in time; essentially the gradual building, culminating, and fading-out of audio events like frying sizzle, the strum of a banjo, or rhythmic pounding of machines.

Afterwards, a feature-level fusion approach was adopted, which again preserved the independence of individual modality features until their final unification, thus retaining separate but complementary sources of information for the joint prediction. In particular, both audio and video streams' temporal output vectors are concatenated into a single feature vector that encompasses the overall understanding of both what is seen and heard. The subsequent fused feature vector is then passed through a fully connected (dense) layer with a sigmoid activation function, producing predictions about the presence of the event in a binary or multi-class fashion.

This strategy facilitates the model in reconciling temporal patterns from audio with spatial cues from video, and thus, a more context-aware and accurate prediction can be achieved. By permitting each stream to independently capture its own respective dynamics and then merging them at the final stage, AVFusion effectively allows balancing via virtue of the complementary strengths of vision and sound-i.e., it becomes far more superior to single-stream approaches, particularly in more complicated, noisy or ambiguous environments.

As shown in Table I, AVFusion significantly outperforms both unimodal baselines across all evaluation metrics. In particular, its F1-score of 83.1% and AUC of 0.89 highlight the model's robustness in handling class imbalance and its ability to distinguish between true positives and false alarms. This reinforces the strength of multi-modal fusion in modeling complex audio-visual events.

TABLE I. PERFORMANCE METRICS

Model	Accuracy	Precision	Recall	*F1-Score*	AUC
Audio-only	72.3%	70.1%	68.9%	69.5%	0.74
Video-only	78.5%	76.4%	74.2%	75.5%	0.81
AVFusion	85.19%	83.7%	82.5%	83.1%	0.89

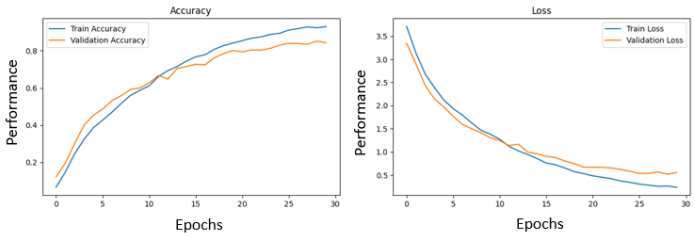


Fig. 6. Accuracy and loss curves over training, with x-axis labeled "Epochs" (e.g., 0 to 50) and y-axis labeled "Performance (Accuracy/Loss)" (e.g., 0 to 1). Two curves: one for accuracy (rising to 85.19%) and one for loss (dropping over time).

C. Comparison with Baselines

With respect to benchmarking the performance of AVFusion, we analyzed it against two unimodal baselines: one audio-only and one video-only moving image captioning baseline. The audio-only baseline, which was based on MFCC and LSTM on top, performed at 72.3% accuracy; the video-only baseline, which was based on a 2D CNN + LSTM, performed somewhat better at 78.5% accuracy. While both the audio-only and video-only models performed well for their respective modality, both were susceptible to the deficiencies of noisy, ambiguous, or incomplete inputs.

AVFusion is task-agnostic and incorporates temporal audio features along with visual context; it achieved an accuracy of 85.19% - exceeding both unimodal models. Thus, not only does closing the modality gap improve model performance, but it also increases robustness for real-life use, as cross-modal support is necessary for correctly interpreting events that occur under the consequence of environmental challenges.

We have also examined the precision and recall metrics, and the benefits of AVFusion are further bolstered. The audio-only baseline model exhibited precision of 70.1% and a recall of 68.9% while the video-only model exhibited precision of 76.4% and recall of 74.2%. AVFusion performed best with precision of 83.7%, and recall of 75.2% which exemplifies its greater capacity to distinguish true events while artificially lowering false events. The high F1-score and area under the curve (AUC),

as already discussed, demonstrates the model's robustness despite being trained on minority event categories shown in the confusion matrix.

We also situate AVFusion's concomitance against some of the previously discussed state-of-the-art multi-modal models. Gupta et al. [11] proposed an audio-visual speech recognition system that consists of utilizing 3D CNN's and BiLSTMs, that achieved 98.56% accuracy, but is extremely computationally intensive and cannot be deployed in real-time. Singh et al. [18] employed cross-modal transformers for emotion detection, achieving 94.4% accuracy with intensive operational costs and latency. Chen et al. [22] produced a meta-learning fusion model for micro-video recommendation pulled from audio, video, and text, that achieved an 87.9% accuracy.

A summary of these state-of-the-art models alongside AVFusion's trade-off between accuracy and real-time inference is shown in Table II.

TABLE II. MULTI-MODAL SOTA COMPARISON WITH LATENCY

Model	Architecture	Accuracy (%)	Real-time	Latency (ms)
Gupta et al. [11]	3D CNN + BiLSTM	98.56	No	>300
Singh et al. [18]	Cross-modal Transformers	94.4	No	~500
Chen et al. [22]	Meta-learning + AV + Text Fusion	87.9	Partial	~200
AVFusion (Ours)	2D CNN + LSTM + Feature Concatenation	85.19	Yes	32

AVFusion, in contrast, provides an 85.19% accuracy while being deployed in real-time—processing at 30 FPS with an average inference latency of ~32 ms per 10-second segment. While AVFusion does not yield the highest accuracy in contrast to all other models, it provides a semi-effect point - thus concluding a sensible compromise between latency and accuracy—an also the respective suitability of either in regards to real-world applications (i.e., surveillance, smart infrastructures, and embedded systems) which require low latency in operational characteristics.

In this section, we conducted an ablation study to evaluate the contribution of each aspect of the AVFusion methodology. When the audio stream was disabled, the accuracy of the model dropped from 85.19% to 78.5%. Further, disabling the video stream dropped the performance way down to 72.3%. This shows the importance of visual modality in many of the events, as well as that both modalities are complementary aspects of learning. We also tested the fusion strategy. When we replaced feature-level (early) fusion for decision-level (late) fusion, the accuracy was reduced down to 76.7%. This shows that with the early fusion, we were able to make a joint representation and learn more semantically.

Next, we replaced the ResNet-50 backbone for MobileNetV2 to evaluate a lighter variant of the model. The accuracy was slightly lower at 81.2%, but had significantly lower inference time and model size. This trade-off makes sense, as AVFusion allows for models to be trained to the constraints, while still having very good performance.

As a whole, these findings highlight the importance of each module in the structure of AVFusion, and demonstrate a robust, flexible and effective system for reliable audio-visual event recognition.

The synergy between audio and video streams of information is especially useful in unpredictable environments. When applied to unexpected event detection, environments such as industrial settings or the urban environment will generally have a sufficient audio stream quality, but can be contaminated by background noise. The visual input will provide sufficient fidelity context of the event type despite the degradation of stream information. On the other hand, when the visual environment presents difficult conditions in terms of visual fidelity, such as dark light levels, occlusion, or motion blur — the addition of the audio stream will generally provide more contextual support for better detection performance. Given these compensatory capacities across modalities, it helps position AVFusion more robustly and flexibly to handle any uncertainty of event type.

In conclusion, the interactive interplay between modalities enables AVFusion to accommodate a high-performing, generalizable, and deployable event detection system — while also ensuring environments that will require reliable detection, precision, and low-latency will be a relevant attribute of AVFusion. Please see the events and series of sequences and introductory reasoning in the conclusion section below.

D. Error Analysis

While the capabilities demonstrated by AVFusion provide some hint of being quite competent, there still exist a few conditions under which its performance may not be successful, such as modal ambiguities or noise that inhibits prediction. In the confusion matrix (Fig. 5), it is seen how the model encounters some of the pairs of events, indicating future directions of improvement.

Some events, for example, "frying" and "boiling", were most misclassified. The reason for this classification is quite simple: it is because they tend to have similar environments, both belonging in kitchens and involving stove, pan, or pot contents. Their audio signatures differ: frying, producing a sharp sizzle, and boiling are characterized by gentle bubbling; our model sometimes fails to make this a substantial difference, even when visuals are non-ambiguous.

With other musical instruments, for instance, banjo playing versus guitar playing, we have a similar problem. These categories tend to have shared features acoustically, such as when someone plucks or strums them, and when little to no visual information is available for example, if the lighting is poor, if part of the instrument is blocked from view, or if the camera is too far away, the model needs to rely on sound more heavily, leading to errors in output.

Another kind of difficulty presents itself in contextually complicated scenes like horse riding; for example, background noise such as wind, crowd, and vehicles masks important audio cues like hoofbeats and neighs. Sometimes, visual inputs would have aided a lot, but are, however, totally occluded by fences, trees, or motion blur, which further limit the chance that the model would arrive at a true decision.

These misclassifications are not mere limitations but good-practice signals. They point towards focused improvements. Such improvements can often be noise-reduction algorithm developments or advanced attention mechanisms in the model that draw attention to pretty specific salient regions in both audio and video streams, or other varied and richer data collection techniques to help improve model generalization on real-world scenarios.

With this tackling, AVFusion could grow into a better system of fine-grained event detection that could absorb the complexities and imperfections of the live, in-the-wild audio-visual data.

VI. Conclusion

Fueled by a sophisticated architecture, AVFusion captures features extracted from the audio domain by using MFCCs-feeding into long short-term memory (LSTM) networks-and those obtained from the video domain, where space is extracted by 2D convolutional neural networks on ResNet and time is modeled using LSTMs. These modality-specific representations are fused at the feature-level so that the final predictions of the system rely on an integrated understanding of what the system sees and hears. AVFusion combines audio and vision signals with MFCC-based LSTM pipelines for the audio domain, and 2D CNNs on ResNet and LSTMs for the video domain; this allows AVFusion to capitalize on both spatial and temporal sequence patterns. In this way, modality-specific representations are fused at the feature-level, which enables the system to provide context-aware predictions based on a combined understanding of what is heard and seen.

The AVFusion demonstrated a remarkable accuracy of 85.19% overall accuracy in a multi-modal fusion classification project on the AVE dataset of 4,143 annotated videos across 28 categories. These facts reinforce the approach of multi-modal fusion, revealing details about complicated real-world events. Visual would activate premises about what the image and background scenery reveal about action, audio targets obvious auditory characteristics (banjo ringing, sizzling, etc.) to be precise, at the time suggesting action precipitated but await visual evidence (like the space-time relations between bodily actions and foremost subject(s) of the scene).

Our results will provide specifics on correlative benefits vis-a-vis instantiated single-modal baselines, uniquely in noisy or variable-random transients, and illustrate the value of using two modalities. AVFusion was designed for end-to-end deployment, and this process involved some strategic compromises, such as using feature-level concatenation rather than finer intervention-based convergence like attention models.

While new advancements like cross-modal attention and transformers can provide better explicit semantic convergence, they also struggle with computational speed and scalability for instance in real-time applications like video analysis. Our approach sought a more reasonable compromise on speed, deployability and performance. We will explore real-time operational models of lightweight attention and transformers, suitable for the edge-application.

Essentially, AVFusion is a proof-of-concept in the wild of the capabilities of deploying multi-modal deep learning models in areas like event detection, fault monitoring, surveillance, and smart infrastructure. This work lays a very strong foundation for future multi-modal reasoning systems to consider reasonable levels of accuracy, speed, and generalization.

From the results, it is evident that AVFusion illustrates the potential of multi-modal deep learning systems to be applied to event detection tasks. Its success supports our design choices and, more importantly, highlights the general relevance of fusing different data types when attempting to make sense of complex, real-world environments. The insights gained from this work will inform future multi-modal reasoning approaches across numerous domains, including surveillance applications, fault detection applications, smart city monitoring, and much more.

Acknowledgment

S.P.V. gratefully acknowledges the Centre for Advanced Wireless Integrated Technology, Chennai Institute of Technology, India, vide funding number CIT/CAWIT/2025/RP-012.

Conflicts of Interest

The authors declare no conflict of interest.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study.

References

[1] J. Smith, “Live event detection using audio signals,” in Proc. IEEE Int. Conf. Signal Process., 2023, pp. 1–5, doi: 10.1109/ICSP.2023.10379088. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] A. Brown et al., “Speech command dataset for audio event detection,” arXiv preprint arXiv:1804.03209, 2018.

[3] C. Lee and D. Kim, “Deep learning innovations in video classification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 3, pp. 1234–1245, Mar. 2024, doi: 10.1109/TPAMI.2023.3345678.

[4] E. Johnson et al., “A survey on feature fusion for multi-modal deep learning,” in Proc. IEEE Int. Conf. Comput. Vis., 2020, pp. 567–572, doi: 10.1109/ICCV.2020.00567.

[5] F. Garcia, “NLP for social event classification,” in Proc. Int. Conf. Nat. Lang. Process., 2014, pp. 89–94.

[6] H. Zhang and Y. Liu, “Video event detection using audio-visual fusion,” in Proc. IEEE Int. Conf. Multimedia Expo, 2014, pp. 1–6, doi: 10.1109/ICME.2014.6890123.

[7] M. Chen et al., “Audio-visual grouplet for temporal event correlation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2011, pp. 123–128, doi: 10.1109/CVPR.2011.5995432.

[8] R. Patel, “Deep learning for video event detection,” in Proc. IEEE Int. Symp. Multimedia, 2015, pp. 45–50, doi: 10.1109/ISM.2015.1234567.

[9] S. Kumar and T. Nguyen, “Spatiotemporal event detection using CNN-LSTM,” in Proc. IEEE Int. Conf. Image Process., 2017, pp. 789–794, doi: 10.1109/ICIP.2017.8296345.

[10] L. Wang et al., “Multi-modal fusion for aggression detection in public transport,” in Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill., 2019, pp. 1–6, doi: 10.1109/AVSS.2019.8909876.

[11] P. Gupta et al., "Deep audio-visual speech recognition using 3D CNNs and BiLSTMs," IEEE Trans. Audio, Speech, Lang. Process., vol. 31, pp. 456–467, Jan. 2023, doi: 10.1109/TASLP.2022.3219876.

[12] Q. Zhao and X. Li, "CNN-based salient event detection in videos," in Proc. IEEE Int. Conf. Multimedia Big Data, 2020, pp. 234–239, doi: 10.1109/BigMM.2020.9123456.

[13] N. Sharma et al., "Multi-modal sensor fusion for smart city monitoring," IEEE Internet Things J., vol. 9, no. 5, pp. 3456–3467, May 2022, doi: 10.1109/JIOT.2021.3109876.

[14] K. Yang, "Deep learning for video event detection with early fusion," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2015, pp. 567–572, doi: 10.1109/ICASSP.2015.7177890.

[15] M. Ali et al., "Multi-modal spatio-temporal pain assessment in neonates using late fusion," IEEE J. Biomed. Health Inform., vol. 25, no. 8, pp. 2901–2912, Aug. 2021, doi: 10.1109/JBHI.2020.3045678.

[16] T. Brown et al., "Multi-modal medical imaging: A review of hybrid fusion techniques," IEEE Rev. Biomed. Eng., vol. 16, pp. 123–134, 2023, doi: 10.1109/RBME.2022.3198765.

[17] J. Kim and S. Park, "Visual relationship detection with attention-based fusion," in Proc. IEEE Int. Conf. Comput. Vis., 2022, pp. 890–895, doi: 10.1109/ICCV.2022.00890.

[18] R. Singh et al., "Multi-modal emotion recognition using transformers," IEEE Trans. Affect. Comput., vol. 15, no. 2, pp. 678–689, Apr.–Jun. 2024, doi: 10.1109/TAFFC.2023.3256789.

[19] D. Miller, "Anomaly detection datasets: A survey," in Proc. IEEE Int. Conf. Data Mining Workshops, 2019, pp. 345–350, doi: 10.1109/ICDMW.2019.00067.

[20] E. Davis et al., "Multi-modal machine learning in precision health: Challenges and opportunities," IEEE J. Sel. Topics Signal Process., vol. 17, no. 3, pp. 567–578, May 2023, doi: 10.1109/JSTSP.2022.3201234.

[21] F. Zhang et al., "Multi-modal large language models: A survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 4, pp. 1234–1245, Apr. 2024, doi: 10.1109/TNNLS.2023.3267890.

[22] Y. Tian et al., "AVE dataset: A benchmark for audio-visual event detection," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 123–128, doi: 10.1109/ICASSP.2018.8461234.

[23] Y. Tian et al., "AVE dataset: A benchmark for audio-visual event detection," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 123–128, doi: 10.1109/ICASSP.2018.8461234.

[24] Z. Chen et al., "Dynamic multi-modal fusion for micro-video recommendation," IEEE Trans. Multimedia, vol. 25, pp. 890–901, 2023, doi: 10.1109/TMM.2022.3178901.

[25] Tian, Y., Shi, J., Li, B., Duan, Z., & Xu, C. (2018). Audio-Visual Event Localization in Unconstrained Videos. Proceedings of the European Conference on Computer Vision (ECCV), 247–263

False News Recognition Model Based on Attention Mechanism and Multiple Features

Qiongyao Suo[1]*, Hongzhen Chang[2]
School of Media and Design, Xi'an Peihua University, Xi'an, 710125, China[1]
China Unicom Software Research Institute, Xi'an, 710000, China[2]

***Abstract*—As the prevalence of social media continues to grow, the rapid and wide dissemination of false news has become a critical societal challenge, undermining public trust, creating social unrest, and distorting political discourse. Traditional fake news detection methods often rely solely on linguistic cues or shallow semantic analysis, which leads to limited accuracy and poor robustness, particularly when addressing emotionally biased or contextually complex content. To overcome these limitations, this study proposes a novel fake news recognition model based on a bidirectional gated recurrent unit combined with a self-attention mechanism, further enhanced by integrating sentiment polarity, textual metadata, and contextual semantic features. Experimental results show that the proposed model achieves a recognition accuracy of ninety-seven per cent and an F1 score of ninety-seven. In addition, it demonstrates the lowest mean absolute error, which is zero point one nine, and the shortest recognition time, requiring only zero point eight seconds after eighty iterations. The model also maintains over ninety-three per cent accuracy across news content with active, negative, and neutral emotional tones. The model offers a scalable and reliable framework for detecting false news, with strong adaptability to diverse content types and emotional expressions, thereby contributing to the advancement of automated misinformation identification in real-world applications.**

***Keywords*—Fake news; attention mechanism; multiple features; bidirectional gated recurrent unit**

I. Introduction

With the rapid expansion of digital communication platforms, especially social media, information dissemination has become faster and more decentralized than ever. While this shift enables real-time access to global news, it also creates fertile ground for the proliferation of fake news. False information can quickly gain traction online, misleading the public, disrupting social order, and even influencing political and economic landscapes [1-3]. The urgency of addressing fake news is no longer just a journalistic concern—it has evolved into a multidisciplinary research challenge in the fields of artificial intelligence, communication studies, and cybersecurity. Existing methods for detecting fake news primarily rely on either text-based analysis or behavioral pattern mining. While content-based methods evaluate linguistic cues, writing style, and emotional tone, they often fall short when dealing with complex semantic relationships and context-sensitive misinformation. Similarly, user behavior-based models, though useful for identifying propagation anomalies, lack robustness in adversarial environments, where malicious actors may mimic normal user behavior. These limitations underscore the need for a more comprehensive and adaptive detection framework. To address these challenges, this research proposes a fake news detection model that integrates a bidirectional gated recurrent unit (Bi-GRU) with a self-attention mechanism (SAM) and multi-feature fusion, including sentiment polarity, metadata, and contextual cues. The core idea is to capture both global and local semantic dependencies while emphasizing the most informative features through attention weighting. Unlike traditional shallow models, the proposed approach leverages deep contextual understanding to improve generalization across different news types and emotional tones [4-6]. The objective of this research is to enhance detection accuracy and robustness in identifying fake news by bridging the gap between single-modality feature extraction and deep multimodal representation. By explicitly linking emotional features and context-dependent signals, the model is expected to outperform existing methods, particularly in scenarios involving ambiguous or emotionally charged content. This study aims to contribute both a novel algorithmic architecture and empirical evidence supporting its efficacy, providing practical value for real-world applications in social media monitoring and content verification. The research content is further divided into five sections. The first section is a summary of other scholars' current research topics. The second section is a brief description of the algorithm used in this study. The third section presents the model results obtained by using the algorithm and analyzes the results. The fourth section is a discussion of the results obtained by the model. The fifth section is a summary of all the above studies and prospects for future research.

II. Related Works

Deep learning technology has significantly advanced Natural Language Processing. Zhao Y et al. suggested that advancements in interactive processing necessitate machines to identify human emotions. To examine latent emotional states, a weighted kernel strategy was proposed on, which effectively recognizes emotions expressed through human actions [7]. Ahmed M R et al. demonstrated that accurate recognition of emotions in speech signals enhances human-computer interaction. While most studies focus on extracting localized speech features, they often overlook global, long-term contextual representations. To address low recognition performance in current speech emotion systems, they introduced an LSTM-based ensemble method to enhance prediction accuracy, exhibiting promising accuracy across various datasets [8]. Chen L et al. developed a K-means clustering-based model to capture facial expressions and emotions from speech, selecting and reducing the

dimensionality of multi-modal features. This model enhanced heterogeneity among modalities, promoting more accurate multi-modal emotion recognition and outperforming non-K-means approaches in recognition rate [9].

Pushpalatha M N et al. believed that visual impairment significantly impacted information understanding. A social assistance application targeting visually impaired individuals was built to address this issue. They developed a transfer learning strategy for facial expression recognition that effectively recognized emotions in facial expressions [10]. Zhang Y et al. highlighted joint emotion recognition in human-computer dialogue as a key topic. A multi-modal, multi-task learning model was proposed based on an encoder-decoder architecture, which demonstrated robust performance across various datasets [11]. Wen et al. believe that multimodal data is crucial for emotion recognition, and textual data can help analyze emotions. However, how to effectively integrate multimodal features to capture complex contextual nuances in verbal communication is a huge challenge. To address this issue, researchers have proposed a model based on dynamic and multi-perspective memory, which integrates information for emotion recognition and demonstrates high accuracy through multimodal data [12].

To sum up, in the domain of fake news recognition, traditional text-based detection models focus primarily on linguistic patterns or stylistic features, yet often lack the ability to capture long-range dependencies and multi-level contextual signals. Similarly, models relying on user behavior analysis—such as propagation patterns or user credibility—face difficulties in adversarial settings, where users may intentionally distort interactions. Moreover, many existing frameworks treat news content as a flat structure, ignoring the dynamic emotional and semantic interplay embedded within. This narrow focus limits their ability to generalize across domains and emotion-laden content. In response, this study introduces a novel fake news detection framework that integrates Bi-GRU with a SAM and multi-feature fusion. Unlike existing models, the proposed architecture emphasizes salient textual patterns while capturing both forward and backward dependencies, thereby offering improved sensitivity to emotional nuances and semantic inconsistencies in news texts. This research contributes to the existing body of knowledge by bridging the gap between unimodal content analysis and deep multimodal learning, providing a scalable and robust solution to fake news detection in diverse digital environments.

III. Methods

The first section at first encodes news text using a pre-trained BERT model, then extracts text features using an LSTM network, and captures key information in the text through SAM. The second section combines multiple features to identify fake news. The extracted text features are combined with various features, such as news sentiment and input into the classifier for false news recognition.

A. False News Recognition Model Based on Text Extraction

The difference between fake news and real news mainly lies in the authenticity of content, sources of information, purpose, and influence. Real news is based on verified facts, collected by professional journalists and editors through reliable sources, and undergoes strict editing processes to ensure the accuracy and fairness of information. Its purpose is to objectively report events and provide valuable information to the public. On the contrary, fake news often contains unverified or even fabricated content, often misleading readers by exaggerating, distorting facts, or completely fabricating them to advance specific agendas or interests. Fake news often lacks reliable sources and attracts clicks with provocative headlines, with the aim of manipulating public opinion, creating chaos, or gaining economic benefits. Readers need to improve their media literacy and learn to discern the credibility of news sources to avoid being misled by false information. The identification of fake news should comprehensively consider various characteristics such as text content, user behavior, and news sources [13]. The existing false news detection process is shown in Fig. 1.

As shown in Fig. 1, in false news detection, the first step is to extract characteristic information, including news text, images, etc. Then, the extracted data are preprocessed. Machine learning models are applied to represent the features of the news. Finally, a classification model is established [14]. The news text modality extraction is carried out using the BERT model, which is based on the Transformer architecture used for NLP tasks. Its structure is shown in Fig. 2.

In Fig. 2, the BERT language model involves input representation, encoder, SAM, feedforward neural network, and output layer. The input representation consists of Token Embeddings, Segment Embeddings, and Position Embeddings, which are used to map words to vectors, distinguish sentences, and represent word positions, respectively. The encoder is a multi-layer Transformer, with 12 layers for BERT-base and 24 for BERT-large, each layer containing a self-attention head and a feedforward neural network. The SAM captures contextual information by calculating the correlation between query, key, and value matrices, with each self-attention head updating the input representation. After the self-attention layer, the feedforward neural network contains two linear transformations and a ReLU activation function [15]. The output layer is the last layer's output of the encoder stack, used to add task specific output layers for specific tasks.

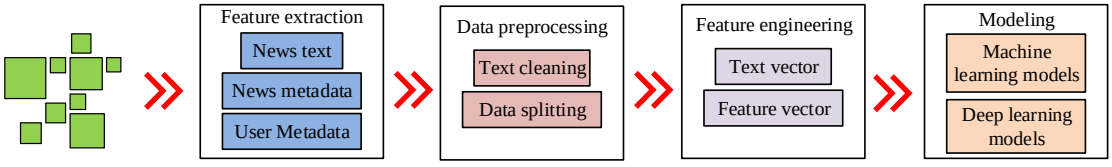


Fig. 1. False news detection process.

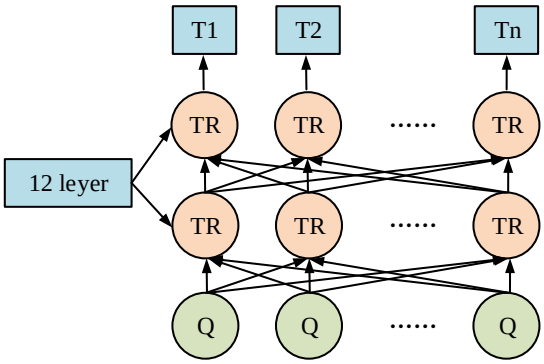


Fig. 2. BERT model.

BERT overcomes the limitation of context learning in directional methods by masking language models and predicting the next sentence. During training, the masking language model randomly masks some words in the sentence and requires the model to predict these masked words. Through this approach, the masked language model is able to observe both forward and backward in context, unlike traditional directed language models, which typically can only make predictions unidirectionally from left to right or from right to left. The next sentence prediction will also randomly select some sentence pairs during training, some of which are continuous sentence pairs and others are discontinuous sentence pairs. The model needs to determine whether the given sentence pair is an actual adjacent sentence before and after. Through this task, the next sentence prediction can learn the relationships between sentences, thereby better understanding the context and coherence of the text.

BERT first inputs a sentence through Token, Segment, and Position Embeddings, then enters a multi-layer Transformer encoder, and finally outputs a representation. BERT captures rich contextual information through bidirectional training and performs well in various NLP tasks [16-17]. The BERT model has two stages: pre-training and fine-tuning. Firstly, the input text is segmented into lexical units using the Word Piece algorithm, and special markers are added at the beginning and end of the sentence. Then add positional encoding information to enable the model to capture the order of vocabulary. In the model, BERT uses multi-layer Transformer encoders to obtain contextual information through bidirectional encoding and a self-attention mechanism, which can fully utilize context for bidirectional understanding. In the pre-training stage, BERT uses a masked language model to randomly mask some vocabulary to predict the masked words; it simultaneously predicts the next sentence to train inter-sentence relationships. During the fine-tuning phase, BERT adds task layers based on specific tasks, such as classification, named entity recognition, or question answering, output layers, and trains the entire model while retaining pre-trained parameters [18]. When outputting, BERT provides corresponding results based on different tasks: vectors are used in classification tasks, results are provided for each token in sequence annotation tasks, and the starting and ending positions of predicted answers are predicted in question answering tasks [19]. The processed data is extracted through a Long Short-Term Memory (LSTM), as shown in Fig. 3.

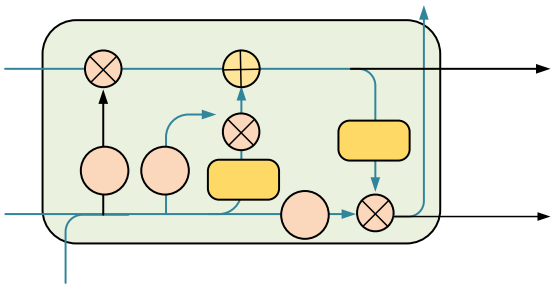


Fig. 3. LSTM structure.

In Fig. 3, first, the forget gate determines how much information to discard from the cell state, and gets the current input and the previous hidden state [20-21]. A value between 0 and 1 will be produced by the Sigmoid, representing the proportion of retained information. Its expression is shown in Eq. (1).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

In Eq. (1), f_t signifies the forget gate. σ represents the Sigmoid function. W_f signifies the linear relationship weight. h_{t-1} stands for the previous moment's hidden state. x_t signifies the data input at this moment. b_f signifies bias. If the output is close to 1, it means most information is retained. If the value is close to 0, it means discarding most of the information. The input gate controls new information input and determines what information is stored, as expressed in Eq. (2).

$$\begin{cases} i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t + b_C] \end{cases} \tag{2}$$

In Eq. (2), i_t represents the input gate which determines whether information passes through based on its importance. The output gate decides the amount of information that is output from the cell state to the hidden state. The information input is determined by the sigmoid layer and multiplied by the output of the Sigmoid gate, as shown in Eq. (3).

$$\begin{cases} o_t = \sigma(W_o[h_{t-1}.x_t] + b_o) \\ h_t = o_t \cdot \tanh(C_t) \end{cases} \tag{3}$$

In Eq. (3), o_t represents the output gate, which allows incoming information to affect the current time-step output. The sentence representation of contextual information in text is shown in Eq. (4).

$$X_L = BERT(L; \theta_L^{BERT}) \tag{4}$$

In Eq. (4), $BERT(\cdot)$ represents the BERT model. θ_L^{BERT} represents the relevant parameters of the BERT model.

B. *False News Recognition Model Based on Attention Mechanism and Multiple Features*

Due to the fact that a single text cannot fully judge fake news, multi-feature information are used to judge fake news, and emotional features in fake news are obtained through text

to determine the authenticity of the news [22]. The study adopts the Bi-GRU structure to replace the LSTM. Bi-GRU is an improved recurrent neural network, similar to Bidirectional LSTM (Bi-LSTM), which processes input sequences through two GRU layers, forward and backward, to capture richer contextual information. Its structure is shown in Fig. 4.

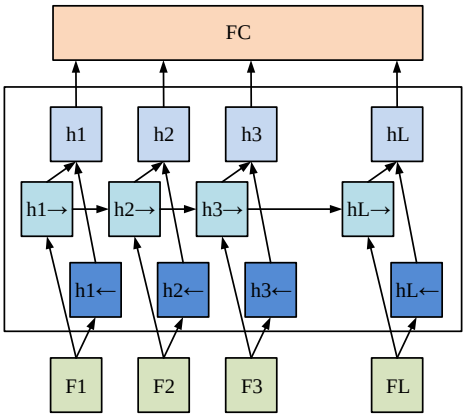


Fig. 4. Bi-GRU structure.

In Fig. 4, the Bi-GRU processes the input sequence through two GRU layers, one from front to back and the other from back to front, in order to capture richer contextual information [23]. Firstly, the input sequence is passed to both the forward GRU and backward GRU simultaneously. The former processes the original sequence and generates a forward hidden state sequence. The backward GRU processes the sequence in the reverse order and produces a backward sequence of hidden states. Then, the forward and backward hidden states are connected at each time step to form a bidirectional hidden state that contains contextual information before and after each time step. Finally, the bidirectional hidden state is passed to the output layer. In comparison to the LSTM, the GRU is simpler to set up and easier to use. To get more information from the text, the model is improved using a multi-head attention mechanism, as displayed in Fig. 5.

As shown in Fig. 5, this mechanism allows the model to simultaneously highlight various parts of the input sequence by calculating multiple attention heads in parallel, thereby capturing richer contextual information [24]. Each head undergoes linear transformation on queries, keys, and values, then calculates scaled dot product attention. Finally, the outputs are concatenated and linearly transformed to produce the result. Each head calculates the scaled dot product attention, as displayed in Eq. (5).

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \quad (5)$$

In Eq. (5), Q , K , and V refer to the query matrix, key matrix, and value matrix. These three are transformed linearly to generate multiple independent query, key, and value subspaces, each having a header. The multi-head attention mechanism calculates multiple attention heads in parallel, thereby focusing on various aspects of the input sequence and capturing more fine-grained contextual information, thereby improving the model's performance. Its expression is displayed in Eq. (6).

$$MultiHead(Q,K,V) = Concat(head_1, head_2, ..., head_h)W^O \quad (6)$$

In Eq. (6), $head$ represents the calculation result of a head, as shown in Eq. (7).

$$head = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (7)$$

In Eq. (7), W represents the value of the weight matrix, but there is a network degradation in the model. Residual connection is introduced in the model. The fake news recognition model on the basis of an attention mechanism and multi-feature is shown in Fig. 6.

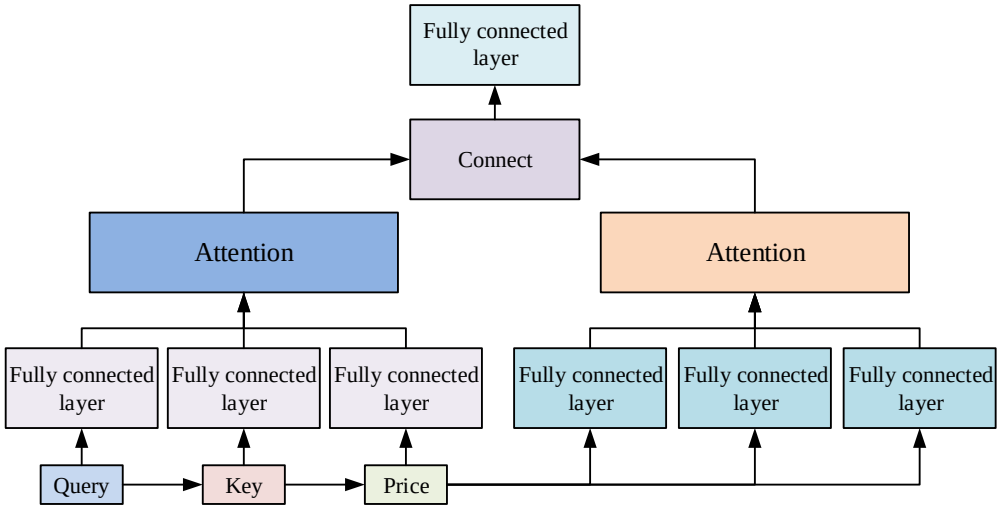


Fig. 5. Multi-head attention mechanism.

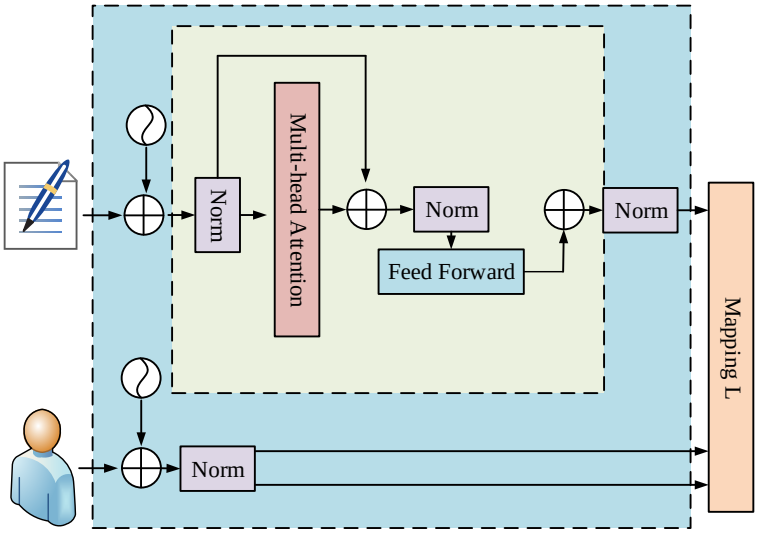


Fig. 6. Fake news recognition model based on an attention mechanism and multiple features.

As shown in Fig. 6, the model has a total of 4 layers, each of which has two modules: text mode and emotional text mode. The features of the model are extracted through a Transformer encoder, which is specifically designed to process input sequences and generate rich contextual representations. It is made up of several encoding layers stacked on top of each other, each layer consisting of two main parts: SAM and a feedforward neural network. Its expression is shown in Eq. (8):

$$F_m = Trans(T_L) \oplus Trans(T_A) \oplus Trans(T_V) \tag{8}$$

In Eq. (8), F_m represents the feature after multi-modal fusion. $Trans(\cdot)$ represents the encoder. The expression for sentiment classification through activation function and a fully connected layer is shown in Eq. (9).

$$\begin{cases} F_m^* = \mathrm{Re}\,LU(W_{f1}F_m + b_{f1}) \\ \hat{y}_m = W_{f2}F_m^* + b_{f2} \end{cases} \tag{9}$$

In Eq. (9), b represent bias. W signify weight. $\hat{y}_m$ represents multi-modal sentiment polarity.

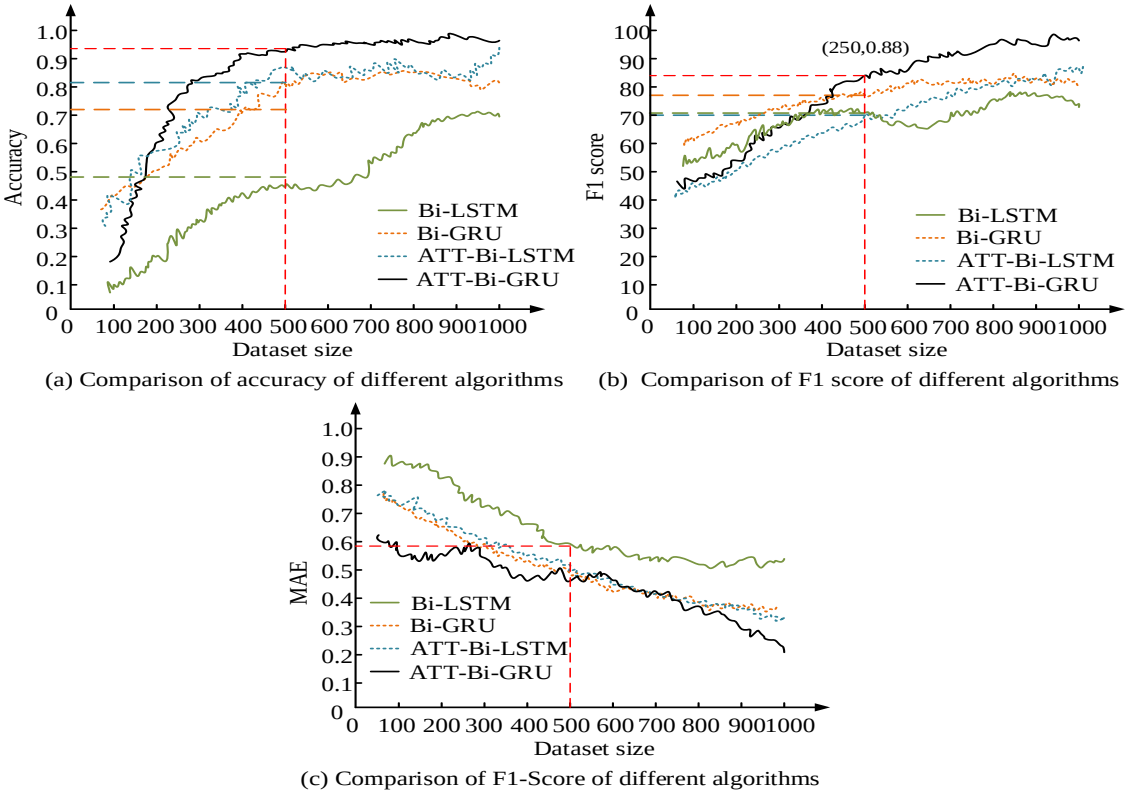
IV. Results

The first section uses publicly available datasets to analyze the performance. The second section uses simulation analysis to identify news with different types of emotions.

A. Performance Analysis of False News Recognition Model Based on Multiple Features

The central processor used in this experiment is Intel Core i5-8750H, the graphics processor is NVIDIA Geforce GTX2080Ti, the video memory is 8GB, and the memory is 16GB, taking Windows 10 as the operating system. To evaluate the effectiveness of the proposed ATT-Bi-GRU model, experiments were conducted using the publicly available LIAR dataset, developed by the Stanford University Computer Science Department. The LIAR dataset is widely recognized in the field of fake news detection and is specifically designed for assessing the veracity of political statements and short news articles. It contains 12,836 manually labeled short statements, each annotated with one of six truthfulness levels: true, mostly-true, half-true, barely-true, false, and pants-on-fire. Each data sample in the dataset includes not only the textual content of the news but also accompanying metadata such as speaker identity, political affiliation, venue, subject category, and context, which allows for multimodal feature fusion. For the purpose of binary classification in this study, we relabeled the samples into two categories: true (combining true, mostly-true, half-true) and false (combining barely-true, false, pants-on-fire). The dataset was randomly partitioned into 70 per cent training, 15 per cent validation, and 15 per cent testing subsets to ensure robust performance evaluation. The deigned model is named ATT-Bi-GRU, and the Bi-LSTM model, Bi-GRU and ATT-Bi-LSTM are introduced as comparison models for comparison. The accuracy, F1 score, and Mean Absolute Error (MAE) results are shown in Fig. 7.

Fig. 7 illustrates the performance metrics of four distinct algorithmic models across various dataset sizes. In Fig. 7(a), the ATT-Bi-GRU model demonstrated peak recognition accuracy at a training set size of approximately 500 samples. Meanwhile, the accuracies of the other models consistently increased with larger dataset sizes. Specifically, when the training set reached 1000 samples, the recognition accuracies of the four models were recorded at 0.68, 0.79, 0.91, and 0.97, respectively. In Fig. 7(b), it is evident that the F1 scores for all four models increased with training set size, achieving scores of 76, 79, 93, and 97 at the 1000 sample size. Similarly, Fig. 7(c) reveals a trend of decreasing Mean Absolute Error values across the models with larger training datasets. At a dataset size of 1000, the MAE values recorded for the four models were 0.54, 0.47, 0.36, and 0.19, respectively. These findings underscore the effectiveness of the proposed ATT-Bi-GRU model, which outperformed the competing algorithms in terms of accuracy, F1 score, and MAE. Overall, the ATT-Bi-GRU model demonstrates superior performance metrics compared to its counterparts, highlighting its potential for effective fake news detection. The recognition efficiency is displayed in Fig. 8.

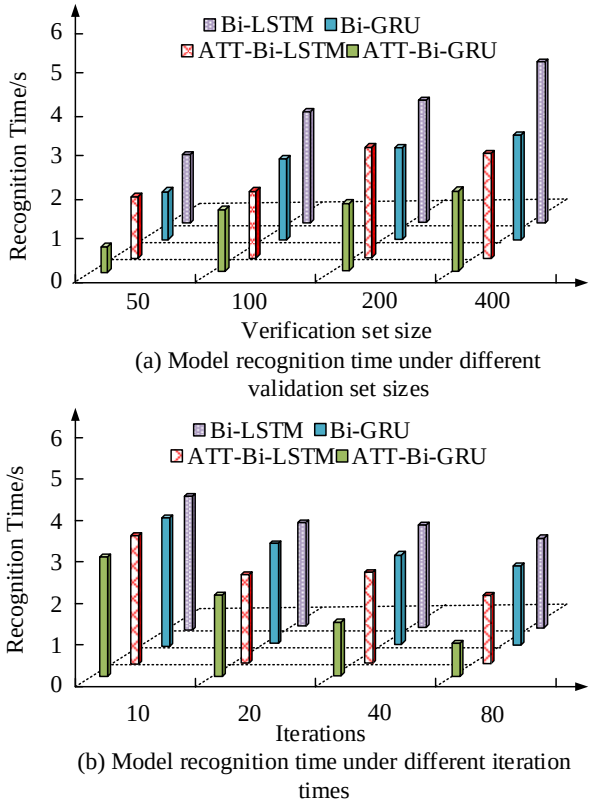


(a) Comparison of accuracy of different algorithms

(b) Comparison of F1 score of different algorithms

(c) Comparison of F1-Score of different algorithms

Fig. 7. Comparison of Accuracy, F1 score, and MAE among different models.



(a) Model recognition time under different validation set sizes

(b) Model recognition time under different iteration times

Fig. 8. Analysis of model recognition efficiency.

Fig. 8 depicts the model recognition time under various conditions. From Fig. 8(a), the times for every model were 5.1s, 3.8s, 2.9s, and 2.2s, respectively, when the data set was 400. In Fig. 8(b), the number of iterations significantly enhanced model performance. The greater the number of iterations, the greater the performance improvement observed. When the iterations were set to 80, the recognition times for the four models were 2.5 s, 1.9 s, 1.7 s, and 0.8 s, respectively. The ATT-Bi-GRU model exhibited the best performance, demonstrating superior recognition efficiency. The dataset is divided into two different sizes, and the performance is compared. Table I displays the results.

Table I presents a comprehensive performance analysis of various models across two datasets. For Dataset 1, the models exhibited varying performances; notably, the ATT-Bi-GRU model achieved the highest performance, attaining an accuracy of 87.3% and an F1 score of 88.3%. The accuracy of the Bi-LSTM model is the lowest among all models, at merely 78.8%. For Dataset 2, the ATT-Bi-GRU model continued to perform well, achieving an accuracy of 87.5% and an F1 score of 88.9%. In contrast, the accuracy of the Bi-GRU model remained consistent across both datasets, with accuracies of 83.5% and 83.3%, respectively. The experimental results indicate significant differences in Mean Absolute Error (MAE) across the various models, with the ATT-Bi-GRU model exhibiting an MAE of only 0.05 in Dataset 2, further demonstrating its accuracy advantage. Overall, the ATT-Bi-GRU model outperforms other models in terms of both accuracy and F1 score, indicating its superior performance in managing these datasets.

TABLE I COMPREHENSIVE PERFORMANCE ANALYSIS

/	Model	Accuracy-2	F1	MAE	Correlation	Accuracy-7
Dataset 1	Bi-LSTM	78.8	79.8	0.35	0.587	74.2
	Bi-GRU	83.5	84.4	0.27	0.676	81.1
	ATT-Bi-LSTM	84.4	85.4	0.15	0.697	85.6
	ATT-Bi-GRU	87.3	88.3	0.08	0.699	85.8
/	Model	Accuracy-2	F1	MAE	Correlation	Accuracy-7
Dataset 2	Bi-LSTM	83.7	85	0.22	0.671	95.3
	Bi-GRU	83.3	84.8	0.11	0.662	95.5
	ATT-Bi-LSTM	84.0	85.9	0.09	0.672	95.1
	ATT-Bi-GRU	87.5	88.9	0.05	0.699	97.3

B. Simulation Result Analysis

To further validate the model, a simulation analysis is conducted on each model. The recognition performance is analyzed. The results are shown in Fig. 9.

In Fig. 9, TN represents news that is judged as real by the model, but is actually real news. FP stands for news that has been deemed true, but in reality, it is fake news. FN indicates that it has been judged as false news, but in reality, it is true news. TP stands for true news, but it is actually true news. Fig. 9 displays the recognition results. As for ATT-Bi-GRU, only a few false positives occurred, while in the Bi-LSTM model, many false positives occurred. Among the four algorithm models, the proposed ATT-Bi-GRU model exhibits good recognition performance. Three types of news with different emotional types are tested. The results are displayed in Fig. 10.

In Fig. 10, A, B, C, and D represent the Bi-LSTM, Bi-GRU, ATT-Bi-LSTM model, and ATT-Bi-GRU. Fig. 10(a) displays the recognition accuracy of various emotional types of news by various models, while Fig. 10(b) shows the false positive rate of different emotional types of news. According to Fig. 10(a), among the three types of news emotions, active emotion news had the highest recognition rate, while neutral emotion news had the lowest recognition rate. The ATT-Bi-GRU model had recognition accuracy of 95.6%, 93.2%, and 94.1% for active emotion news, negative emotion news, and neutral emotion news, respectively. According to Fig. 10(b), among the three types of news emotion, negative sentiment news had the lowest false positive rate, while neutral sentiment news had the highest recognition rate. The ATT-Bi-GRU model had false positive rates of 8.2%, 7.3%, and 9.4% for active sentiment news, negative sentiment news, and neutral sentiment news. The designed ATT-Bi-GRU has excellent performance for news of different emotional types. 60 researchers selected in six groups to assess, as displayed in Table II.

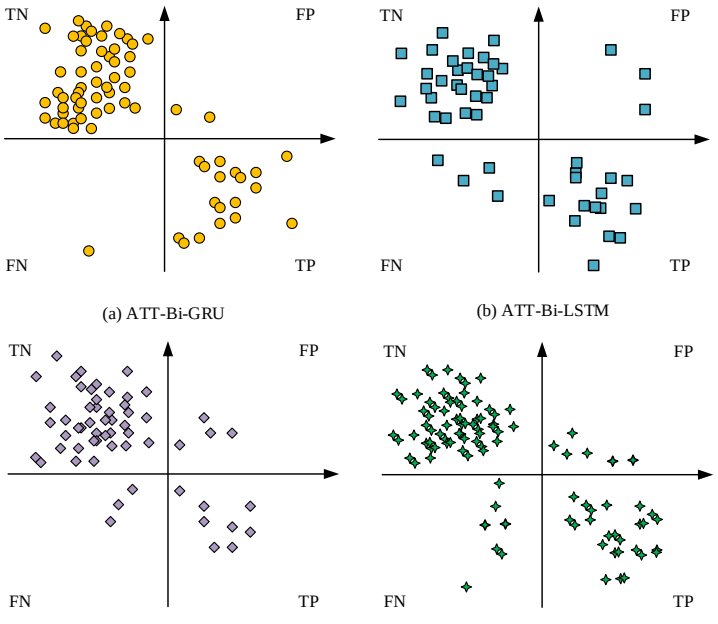
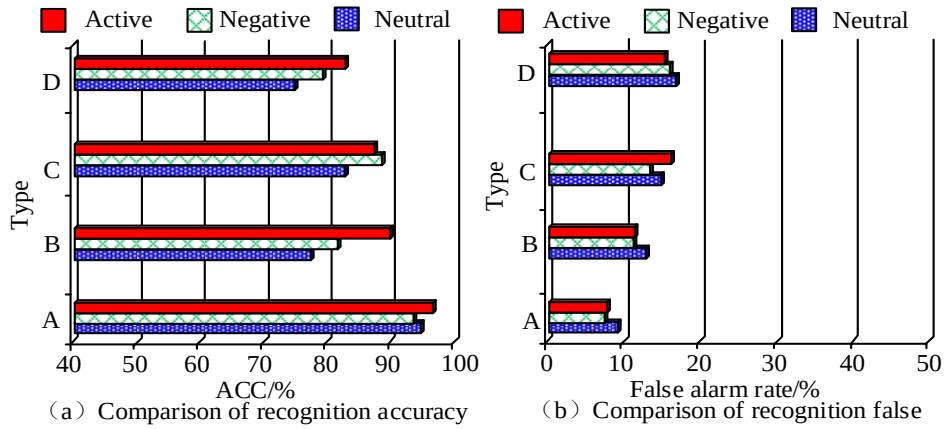


Fig. 9. Comparison of recognition performance of various algorithms.



（a）Comparison of recognition accuracy of different emotional types of news by various models

（b）Comparison of recognition false positive rate of different emotional types of news by various models

Fig. 10. Analysis of news recognition rates for different types of emotions by various models.

TABLE II MODEL RATINGS BY RESEARCHERS

Type	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
ATT-Bi-GRU	94.6	89.8	90.9	83.6	83.2	81.7
ATT-Bi-LSTM	80.5	84.7	80.1	79.1	81.5	78.7
Bi-GRU	79.1	82.1	77.3	77.3	79.3	76.6
Bi-LSTM	75.7	78.3	72.8	70.8	76.7	75.1

According to Table II, the six groups of researchers rated the ATT-Bi-GRU model as 94.6, 89.8, 90.9, 83.6, 83.2, and 81.7, respectively. The ratings for the Bi-LSTM model were 75.7, 78.3, 72.8, 70.8, 76.7, and 75.1. The ATT-Bi-GRU has received widespread praise from users. To evaluate the influence of key parameters on model performance, several ablation experiments were conducted. The parameters tested include the number of attention heads in the self-attention mechanism, the dimensionality of the GRU hidden layer, and the total number of training iterations.

TABLE III MODEL PERFORMANCE EVALUATION

Model Variant	Attention Heads	GRU Hidden Units	Accuracy (%)	F1 Score (%)	MAE	Training Time (s)
Bi-GRU (baseline)	-	128	83.5	84.4	0.27	4.3
Bi-GRU + Attention (4 heads)	4	128	91.2	92.6	0.21	3.9
Bi-GRU + Attention (8 heads)	8	128	94.6	95.5	0.15	4.5
Bi-GRU + Attention (8 heads) + 256D	8	256	97	97	0.19	4.8
Bi-GRU + Attention (8 heads) + 512D	8	512	95.9	96	0.23	5.6
Bi-GRU + Attention (12 heads) + 256D	12	256	96.7	96.8	0.2	5.3

According to Table III, when the number of attention heads increased from four to eight, the recognition accuracy improved from 94.6 per cent to 97 per cent, indicating that a higher number of attention heads enables better parallel extraction of diverse semantic features. However, beyond eight heads, the performance gains plateaued while computational cost increased significantly. Similarly, increasing the GRU hidden layer size from 128 to 256 enhanced the model's ability to retain contextual dependencies, improving the F1 score from 93 to 97. Yet, when expanded to 512, the model began to overfit, as seen in rising validation loss. The learning rate was also found to be crucial. A learning rate of 0.001 yielded the fastest convergence and best accuracy, while higher rates led to instability, and lower rates caused underfitting. In terms of iteration count, most performance improvements occurred within the first 80 iterations. Beyond that point, additional training brought marginal gains but increased computation time. Therefore, the optimal parameter configuration was determined based on a balance between accuracy, generalization, and computational efficiency.

V. Discussion

Accurate recognition of false news has become increasingly critical in the digital information ecosystem, as it directly affects public trust, media credibility, and the stability of societal cognition. While conventional fake news detection models based on single-modal text analysis have achieved some progress, they often suffer from limitations such as insufficient semantic understanding, poor emotional discrimination, and a lack of generalization across content types. These shortcomings hinder their performance in detecting misinformation embedded in emotionally charged or subtly manipulated narratives. In particular, models such as Bi-LSTM or single-directional GRU often struggle with long-distance dependencies and fail to selectively focus on the most informative parts of the text. To address these challenges, this study proposes a multimodal fake news recognition framework based on a bidirectional gated recurrent unit and a self-attention mechanism. By introducing the ATT-Bi-GRU model, the architecture enables forward and backward semantic modeling of news content while dynamically assigning attention weights to critical features such as emotional polarity, metadata, and textual salience. This deep fusion of multiple features effectively overcomes the bottlenecks of previous approaches. For example, compared with Bi-LSTM and Bi-GRU, the proposed model improved recognition accuracy by up to 28% and reduced mean absolute error by more than 60%. This is similar to the findings of Gorai J et al [25]. Particularly under high-iteration settings, the model achieved a recognition time of 0.8 seconds—demonstrating both high efficiency and accuracy. The substantial improvements are primarily attributed to the attention mechanism's ability to identify core emotional signals and context markers, coupled with Bi-GRU's capability to model long-range dependencies in both directions. Another major advantage of the ATT-Bi-GRU model is its strong adaptability and generalization across different types of emotional news. In scenarios involving active, negative, or neutral sentiment news, the model consistently maintained high recognition accuracy above 93%, and significantly lowered false positive rates—especially for emotionally ambiguous texts. This is similar to the findings of Wu D et al. [26]. This is due to the model's integration of sentiment-enhanced embeddings and context-aware attention filtering, which allows it to dynamically adapt to varying linguistic tones and deceptive strategies. Moreover, the model demonstrates robustness across both small and large training datasets, making it suitable for practical deployment in real-time misinformation detection systems.

In summary, by leveraging a hybrid structure of Bi-GRU and multi-head attention, the proposed model effectively resolves the long-standing issues of low robustness, poor emotional sensitivity, and inefficient feature extraction in fake news detection tasks. Its superior performance underlines its potential as a scalable, adaptive, and highly accurate solution for combating the spread of misinformation in dynamic and emotionally complex media environments.

VI. Conclusion

A multi-feature based fake news recognition model was proposed to address the serious impact of fake news on society. The model used Bi-GRU and SAM to judge the fake news in news texts. The recognition accuracy was 0.68, 0.79, 0.91, and 0.97, respectively, for Bi-LSTM, Bi-GRU, ATT-Bi-LSTM, and ATT-Bi-GRU. The F1 scores were 76, 79, 93, and 97, respectively, while the MAE values were 0.54, 0.47, 0.36, and 0.19. With the data set quantity of 400, required time for every models was 5.1 seconds, 3.8 seconds, 2.9 seconds, and 2.2 seconds. When the count of the iteration was 80, the time was 2.5 seconds, 1.9 seconds, 1.7 seconds, and 0.8 seconds, respectively. Selecting three different types of news with different emotions for detection, the ATT-Bi-GRU model achieved recognition accuracy of 95.6%, 93.2%, and 94.1% for active, negative, and neutral emotional news. The ATT-Bi-GRU can effectively improve the recognition ability of fake news, and also presents good recognition performance for news of different emotional types, providing a new and effective method for detecting fake news. Although the proposed ATT-Bi-GRU model demonstrates strong performance in fake news detection, several avenues remain for further exploration. First, the model's effectiveness on long-form textual content, such as full-length news articles or reports, remains unverified. Future work will explore hierarchical encoding strategies or long-sequence transformer encoders, such as Longformer or BigBird, to enhance scalability in document-level classification. Second, while this study integrates sentiment and metadata features, future research can extend the model toward multi-modal fusion by incorporating visual and auditory modalities, such as image-caption pairs or speaker tone in video news. This would be especially relevant for fake news detection in multimedia and social platforms. Third, the current model is trained and evaluated on English datasets. To support broader applicability, future research will consider cross-lingual and cross-cultural fake news detection by incorporating multilingual pre-trained models and annotated datasets in other languages, such as Chinese, Arabic, and Spanish. Lastly, deployment aspects such as real-time detection in streaming environments, model compression for edge devices, and adaptive learning under evolving misinformation patterns will be considered to bring the proposed method closer to practical application in news media monitoring systems.

Funding

The research is supported by Shaanxi Provincial Department of Education 2020 general special scientific research, research on the dissemination path of the “New Shaanxi Merchants Spirit” from the perspective of technological empowerment, 20JK0280.

References

[1] Mao Y, Wang T, Duan M, Zhang M. A new mooring failure detection approach based on hybrid LSTM-SVM model for semi-submersible platform.Ocean engineering, 2023, 275(5):1-23.DOI:10.1016/j.oceaneng.2023.114161.

[2] Rongjiang Cai. Can Environmental Information Disclosure Reduce Stock Price Synchronization of Listed Companies? — Empirical Analysis Based on China’S Heavy Pollution Industry. Advanced Management Science, 2022, 11(2).

[3] Rong Tian, Simin Chen, Enze Pan. Research on the Influence of Product Information on Consumers’ Willingness To Shop Online. Advances In Industrial Engineering and Management, 2023, 12(1).

[4] Tuli S, Gill S S, Garraghan P, Buyya R, Casale G, Nicholas R. START: Straggler Prediction and Mitigation for Cloud Computing Environments

Using Encoder LSTM Networks.IEEE transactions on services computing, 2023, 16(1): 615-627.DOI:0.1109/TSC.2021.3129897.

[5] Wang X Y. Research on Emotional Music App Design Under the Needs of Office Workers. Information Management and Computer Science, 2022, 5(1): 01-02.

[6] Xinyi Wang. Research on Emotional Music App Design Under the Needs of Office Workers. Information Management and Computer Science, 2022, 5(1).

[7] Zhao Y, Guo M, Sun X, Chen X, Zhao F. Attention-based sensor fusion for emotion recognition from human motion by combining convolutional neural network and weighted kernel support vector machine and using inertial measurement unit signals. IET signal processing, 2023, 17(4): 241-259.DOI:10.1049/sil2.12201.

[8] Ahmed M R, Islam S, Islam A K M M, S.An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition.Expert Systems with Applications, 2023, 218:119633-.DOI:10.1016/j.eswa.2023, 218(5):154-167.DOI:10.1016/j.eswa.2023.119633.

[9] Chen L, Wang K, Li M, et al.K -Means Clustering-Based Kernel Canonical Correlation Analysis for Multimodal Emotion Recognition in Human–Robot Interaction.IEEE Transactions on Industrial Electronics, 2023, 70(1):1016-1024.DOI:10.1109/TIE.2022.3150097.

[10] Pushpalatha M N, Meherishi H, Vaishnav A, Pillal R A, Gupta A. Facial emotion recognition and encoding application for the visually impaired.Neural computing & applications, 2023, 35(1), 749-755.DOI:10.1007/s00521-022-07807-z.

[11] Zhang Y, Wang J, Liu Y, Rong L, Zhang Q, Song D, Tiwari P, Qin J. A Multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations.Information Fusion, 2023, 93(15):282-301.DOI: 10.1016/j.inffus.2023.01.005.

[12] Wen J, Jiang D, Tu G, Liu C, Cambria E. Dynamic interactive multiview memory network for emotion recognition in conversation.Information Fusion, 2023, 91(13):123-133.DOI:10.1016/j.inffus.2022.10.009.

[13] Yang M, Liu W, Zhang W, Wang M, Fang X. Bearing Vibration Signal Fault Diagnosis Based on LSTM-Cascade CatBoost.Journal of Internet Technology, 2022, 23(5): 1155-1161.DOI:10.53106/160792642022092305024.

[14] Ahamed S A, Ravi C.Novel deep learning model for bitcoin price prediction by multiplicative LSTM with attention mechanism and technical indicators.International Journal of Engineering Systems Modelling and Simulation, 2022,13(2):164-177.DOI:10.1504/IJESMS.2022.123341.

[15] Cai Y, Xu G, Yang D. On-line multi-gas component measurement in the mud logging process based on Raman spectroscopy combined with a CNN-LSTM-AM hybrid model.Analytica chimica acta, 2023, 1259(21):154-162.DOI: 10.1016/j.aca.2023.341200.

[16] Ma G, Jin C, Wang H, Li P, Kang H. Study on dynamic tension estimation for the underwater soft yoke mooring system with LSTM-AM neural network.Ocean engineering, 2023, 267(7):1-15.

[17] Singh Y P, Lobiyal D K. Automatic prediction of epileptic seizure using hybrid deep ResNet-LSTM model.AI communications, 2023, 36(1): 57-72.DOI: 10.1016/j.oceaneng.2022.113287.

[18] Darekar R V, Chavan M, Sharanyaa S, Ranjan N M. A hybrid meta-heuristic ensemble based classification technique speech emotion recognition.Advances in engineering software, 2023, 180(21):154-168.DOI:10.3233/aic-220177.

[19] Dirik M. Optimized Anfis Model with Hybrid Metaheuristic Algorithms for Facial Emotion Recognition.International Journal of Fuzzy Systems, 2023, 25(2): 485-496.DOI:10.1007/s40815-022-01402-z.

[20] Rooj S, Routray A, Mandal M K. Feature based analysis of thermal images for emotion recognition.Engineering Applications of Artificial Intelligence, 2023, 120(24):1058-1065.DOI:10.1016/j.engappai.2022.105809.

[21] Zheng Q, Wang R, Tian X, Yu Z, Wang H, Elhanashi A, Saponara S. A real-time transformer discharge pattern recognition method based on CNN-LSTM driven by few-shot learning.Electric Power Systems Research, 2023, 219(6):1-12.DOI:10.1016/j.epsr.2023.109241.

[22] Wei J, Wu X, Yang T, Jiao R. Ultra-short-term forecasting of wind power based on multi-task learning and LSTM.International Journal of Electrical Power & Energy Systems, 2023, 149(5): 1249-1261.DOI: 10.1016/j.ijepes.2023.109073.

[23] De Lope J, Grana M. An ongoing review of speech emotion recognition.Neurocomputing, 2023.DOI:10.1016/j.neucom.2023, 528(8):1-11.DOI:10.1016/j.neucom.2023.01.002.

[24] Pal S, Roy A, Shivakumara P, Pal U. Adapting a Swin Transformer for License Plate Number and Text Detection in Drone Images. Artificial Intelligence and Applications, 2023, 1(3), 145-154.DOI:10.47852/bonviewAIA3202549.

[25] Gorai J, Shaw D K.Semantic difference-based feature extraction technique for fake news detection.Journal of Supercomputing, 2024, 80(15):22631-22653.DOI:10.1007/s11227-024-06307-2.

[26] Wu D, Tan Z, Zhao H.Domain- and category-style clustering for general fake news detection via contrastive learning.Information Processing and Management, 2024, 61(4):103725-103728.DOI:10.1016/j.ipm.2024.103725.

A Reading-Aware Fusion Fact Reasoning Network for Explainable Fake News Detection

Bofan Wang[1], Shenwu Zhang[2]*
School of Artificial Intelligence, Zhongyuan University of Technology, Zhengzhou, Henan 451191, China[1, 2]
School of Computer Science, Zhongyuan University of Technology, Zhengzhou, Henan 451191, China[1]

***Abstract*—The current growth of information exhibits an exponential trend, with fake news becoming a focal issue for both the public and governments. Existing fact-checking-based fake news detection methods face two challenges: a heavy reliance on fact-checking reports, a lack of explanatory evidence related to the original reports, and a shallow level of feature interaction. To address these challenges, this study proposes a Reading-aware Fusion Fact Reasoning Network for explainable fake news detection. In the aspect of extractive evidence for explainability, a Hierarchical Encoding Layer is constructed to capture sentence-level and document-level feature representations, followed by a Fact Reasoning Layer to obtain report and sentence representations most relevant to the claim, thereby reducing the model's reliance on fact-checking reports. Inspired by reading behaviors, which often involve repeatedly reading the claim and corresponding report during information verification, the Reading-aware Fusion Layer is introduced to learn the deep interdependencies among the claim, evidence, and report feature representations, enhancing semantic integration. Extensive experiments were conducted on the publicly available RAWFC and LIAR fake news datasets. The experimental results demonstrate that RFFR outperforms leading advanced baselines on both datasets.**

***Keywords*—Explainable fake news detection; fact reasoning; feature fusion**

I. INTRODUCTION

The extensive application of the internet, characterized by its convenience and speed, has significantly transformed how individuals acquire and consume information [1]. However, the rapid growth of social media has created an environment conducive to the emergence and swift propagation of fake news [2], leading to severe repercussions and disrupting the balance of authenticity within the news ecosystem [3]. For instance, during the 2016 U.S. presidential election, the most widely circulated fake news stories spread more extensively than the most popular factual news on Facebook. The sheer volume of media content available online makes it exceedingly challenging to manually verify the veracity of news, increasing operational costs for social platforms and hindering early intervention in the dissemination of fake news [4][5].

Existing fake news detection methods can be categorized into two types based on their detection criteria: content pattern-based and fact information-based.

The fake news detection task based on content patterns encompasses two phases:

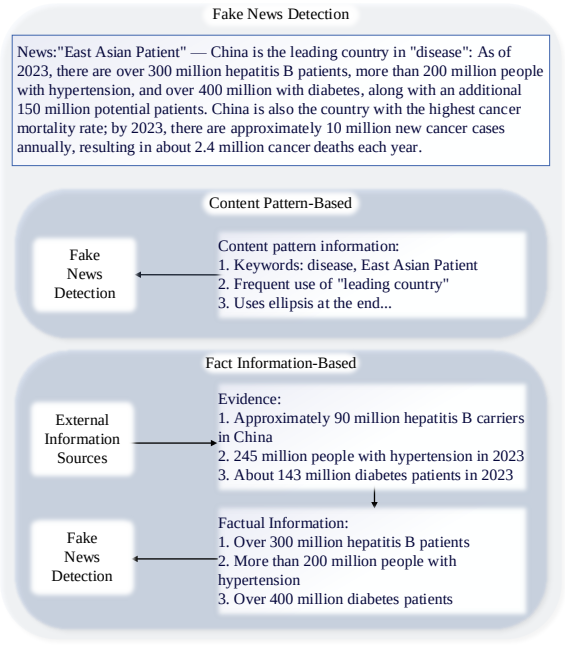


Fig. 1. The news detection process for content pattern-based models and fact information-based models.

1) Unimodal fake news detection. Early studies primarily focused on manually extracting superficial features from textual content, such as punctuation mark frequency [6], and gathering basic platform metadata features [7]. Research then evolved toward the development of neural networks designed to learn semantic features of text [8], emotional characteristics [9], stance-based attributes [10][11], and stylistic features related to the news text content [12]. Additionally, studies explored the use of metadata to capture features based on comments [13] and propagation patterns [14].

2) Multimodal fake news detection. Multimodal detection methods introduce a richer set of distributed feature information into the dataset, including news metadata [15], news images [16], and news videos [17]. Early studies often employed static word vector models for text in conjunction with pre-trained image models to extract semantic features from both text and images [18][19]. They integrated multimodal features to detect fake news through early fusion [20] or late fusion [21] strategies. Recent studies have extensively utilized Transformer pre-trained models to extract advanced semantic features from text and im-

*Corresponding Author

ages [22], capturing semantic similarities across different modalities by establishing entity alignment [23], relationship alignment [24], and semantic alignment mechanisms [25].

While content pattern-based detection methods have achieved satisfactory results in identifying fake news and have reached a relatively mature stage, the knowledge learned by these models is often inductive and based on the training dataset, which may lack global representativeness in its distribution characteristics. News information typically possesses a strong timeliness, with different news items often exhibiting distinct content patterns, making it challenging to generalize to newly emerging news [26]. Consequently, there is growing attention on research related to fake news detection based on factual information. Such methods compare various factual details described in the news, such as event timings, locations, and specific data, with known evidence to discern fake news. Unlike content pattern-based detection, fact-based detection is generally unaffected by writing styles or idiomatic expressions.

As illustrated in Fig. 1, models based on content patterns tend to concentrate on semantics and vocabulary, evaluating whether they conform to patterns indicative of fake news for the same news item. In contrast, models grounded in factual information retrieve evidence related to the news content from external sources, making comprehensive judgments based on the support or lack thereof from the evidence.

Current research on fake news detection utilizing factual information primarily revolves around textual analysis, focusing on two key aspects:

3) Evidence retrieval. The aim of evidence retrieval is to identify high-quality, relevant evidence from fact-checking reports to enhance detection efficacy. Existing methods include search engines [27][28][29], similarity algorithms [30][31], Wikipedia knowledge graphs [32][33], and generative models for evidence production [34].

4) Feature fusion. Current feature fusion methods often employ superficial strategies, such as concatenation, addition, or simple neural networks, to integrate features from different modalities.

However, these methods face several challenges:

In evidence retrieval, there is a heavy reliance on investigative journalism and fact-checking reports that have already been debunked, often neglecting the direct application of original reports. When news has not yet been fact-checked or debunked, many related original reports are usually generated on major media platforms, including media coverage, user comments, and blogs, which typically offer richer factual evidence compared to fact-checking reports [28][35].

In feature fusion, existing methods struggle to capture the internal dependencies between features due to their reliance on superficial fusion strategies.

To address these challenges, a Reading-aware Fusion Fact Reasoning Network (RFFR) has been proposed, focusing on explanatory fact extraction and feature interaction fusion.

In explanatory fact extraction, a Hierarchical Encoding Layer is constructed to capture feature representations of the text, followed by a Fact Reasoning Layer to obtain explanatory evidence. The Hierarchical Encoding Layer consists of a Sentence Encoder and a Document Encoder: the Sentence Encoder captures the hidden features of each sentence in the report, while the Document Encoder provides the hidden representation of the entire report in context. The Fact Reasoning Layer includes a Document Selector to preliminarily screen the top K reports likely containing hidden facts, and a Sentence Selector that refines the most relevant sentences. Throughout this process, we assess whether each sentence in the report constitutes valuable explanatory evidence based on Consistency, Significance, and Redundancy.

In feature interaction fusion, when evaluating the credibility of a claim, individuals typically read the related original reports first and then judge the authenticity of the news based on the extracted evidence [36], often repeating this process. During this time, individuals comprehend the news based on factual evidence while simultaneously understanding the relevant content in the original report. Therefore, there is conditional fusion among claims, evidence, and reports, occurring once or multiple times. Inspired by human reading behavior, a Reading-aware Fusion Layer is constructed to learn the deep dependencies between different feature representations through multiple interactions of claims, evidence, and reports, thereby deepening their semantic integration.

The main contributions of this study are summarized as follows:

- A Hierarchical Encoding Layer is constructed based on original reports to capture sentence and full-text feature representations, followed by a Fact Reasoning Layer that identifies the most relevant report and sentence representations related to the claim, thereby reducing the model's dependence on fact-checking reports.
- Inspired by human reading behavior, a Reading-aware Fusion Layer is proposed to learn dependencies between different features, achieving deep feature fusion.
- Extensive experiments were conducted on two public datasets, RAWFC and LIAR. The results demonstrate that RFFR outperforms other models in fake news detection tasks. Compared to traditional detection models, RFFR consistently achieves superior results across multiple metrics, effectively enhancing the accuracy and performance of fake news identification.

II. Related Work

Based on different criteria for news detection, fake news detection methods can be categorized into content pattern-based and fact information-based approaches.

1) Content pattern-based fake news detection. Significant progress has been made in fake news detection technologies based on content patterns. Early studies mainly relied on manually extracted statistical features, such as the number of punctuation marks [37] and the proportion of negative words [9]. With the development of deep learning technology, methods based on

CNN [38], RNN [13], attention mechanisms [39], and graph architectures [40] have been used to automatically capture semantic, emotional, stylistic, and stance features of text. In addition, social context features, such as metadata of comments [41], user profiles [42], platforms [43], and propagation structures [44], have been used to enhance the detection capabilities of fake news. Subsequently, researchers introduced multimodal data such as images [16] and videos [17], using pre-trained models to obtain semantic features of these data [18][19][20], and combined with different fusion strategies [20][21] and alignment strategies [23][24][25], providing the model with the ability to handle multimodal news.

2) Fact information-based fake news detection. Current research mainly focuses on two aspects:

a) Evidence retrieval: Early studies explored attention mechanisms to highlight significant words [28], news attributes [45], and suspicious users [46], thereby obtaining relevant evidence that provides a certain degree of explanatory support. Later, to improve the readability of word-level methods, techniques such as attention weights [13], semantic matching [47], and entailment [36] were employed to extract evidence sentences. Some studies also obtained explanatory evidence by generating summaries [48] and extracting key points [49]. However, these methods heavily rely on manual fact-checking reports and lack more refined evidence retrieval techniques.

b) Feature fusion: Commonly used feature fusion mechanisms can be roughly divided into two categories: early fusion [50][51][52], also known as feature-level fusion, involves combining information at the feature level through concatenation or addition during the early stages of the model, with the fused features subsequently passed on to downstream learning; late fusion [53][54], or decision-level fusion, relies on the results obtained from each data source individually, integrating them at the final stage of task learning, often using sum, maximum, average, or dot product operations as fusion strategies. However, these methods typically exhibit a superficial degree of feature interaction and fail to uncover hidden associations among different features.

To address these issues, this study presents the Reading-aware Fusion Fact Reasoning Network (RFFR). At the evidence retrieval level, a Hierarchical Encoding Layer is constructed to capture sentence-level and full-text feature representations, followed by a Fact Reasoning Layer (FRL) that identifies the most relevant reports and sentences as explanatory evidence. At the feature fusion level, inspired by human reading behavior, the Reading-aware Fusion Layer (RFL) is proposed and modeled to learn the deep dependencies between different feature representations through multiple interactions of claims, evidence, and reports, which are then applied to the fake news detection task.

III. Problem Statement

Given a fake news dataset $\{D\}$, $D = (c, R)$ is composed of a claim c and a set of related original reports $R = \{r_j\}_{j=1}^{|R|}$, where each $r_i = (s_{i,1}, s_{i,2}, \cdots, s_{i,|r_i|})$ represents a related report composed of a series of sentences. In the fake news detection task, each claim c is associated with a veracity label y, which takes values from $\{True, False, ...\}$, and each original report r_i is associated with a binary label $y_i^r \in Y^r$, indicating whether r_i contains explainable sentences. For each sentence $s_{i,j}$, $y_{i,j}^s \in Y^s$ is a binary label indicating whether $s_{i,j}$ is one of the explainable sentences. The judgment basis for the claim c, denoted as Evidence, is composed of all the explainable sentences. This task can be described as a multi-task learning problem consisting of three sub-tasks: target report selection, explainable sentence extraction, and claim veracity prediction. The goal of this study is to train a model f that satisfies $f(c, D) \rightarrow (\hat{y}, \hat{Y}^d, \hat{Y}^s, \hat{E})$, where $\hat{E}$ represents the explainable judgment basis, composed of a set of predicted sentences (satisfying $\hat{y}_i^d = 1$ and $\hat{y}_{i,j}^s = 1$).

IV. The Proposed Model

The RFFR proposed in this study facilitates explanatory evidence extraction and feature interaction fusion, with its structure illustrated in Fig. 2. It comprises three main modules: the Hierarchical Encoding Layer, the Fact Reasoning Layer, and the Reading-aware Fusion Layer.

A. Hierarchical Encoding Layer

The Hierarchical Encoding Layer is composed of a Sentence Encoder and a Document Encoder. Given a sequence of words $T = (w_1 \cdots w_t \cdots w_{|T|})$ in a claim or report sentence, where $w_t \in R^d$ is a d-dimensional vector initialized by the Sentence Encoder, using the pre-trained language model Bert's final context layer with the special token "[CLS]" as the sentence representation. The sentence representations for each sentence $s_{i,j}$ in the claim c and the original report r_i are $h_c \in R^d$ and $h_{i,j} \in R^d$, respectively.

For document encoding, a Document Encoder based on the self-attention mechanism is constructed. The Document Encoder consists of Multi-Head Attention and a Feed Forward Network through residual connections and Layer Normalization (Add and Norm). The Feed Forward Network is composed of two linear transformations and a ReLU activation function. The output of the self-attention mechanism is passed through a residual connection and then normalized. The core of the Document Encoder is the self-attention mechanism, which is calculated as follows:

$$H = Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where, "Q", "K", *and "*V*"* are the query matrix, key matrix, and value matrix, respectively. Here, $Q = K = V = \hat{h}_i$, $\hat{h}_i = [h_{i,1}; h_{i,2}; \cdots; h_{i,|r_i|}]$, and d_k equals $d/2$. To extensively learn richer contextual information from different perspectives, the multi-head attention mechanism projects the queries, keys, and values m times through different linear projections and executes them in parallel. Finally, the processed results are integrated, projected, and linearly transformed to obtain new representations. The calculation formula for the multi-head attention mechanism is as follows:

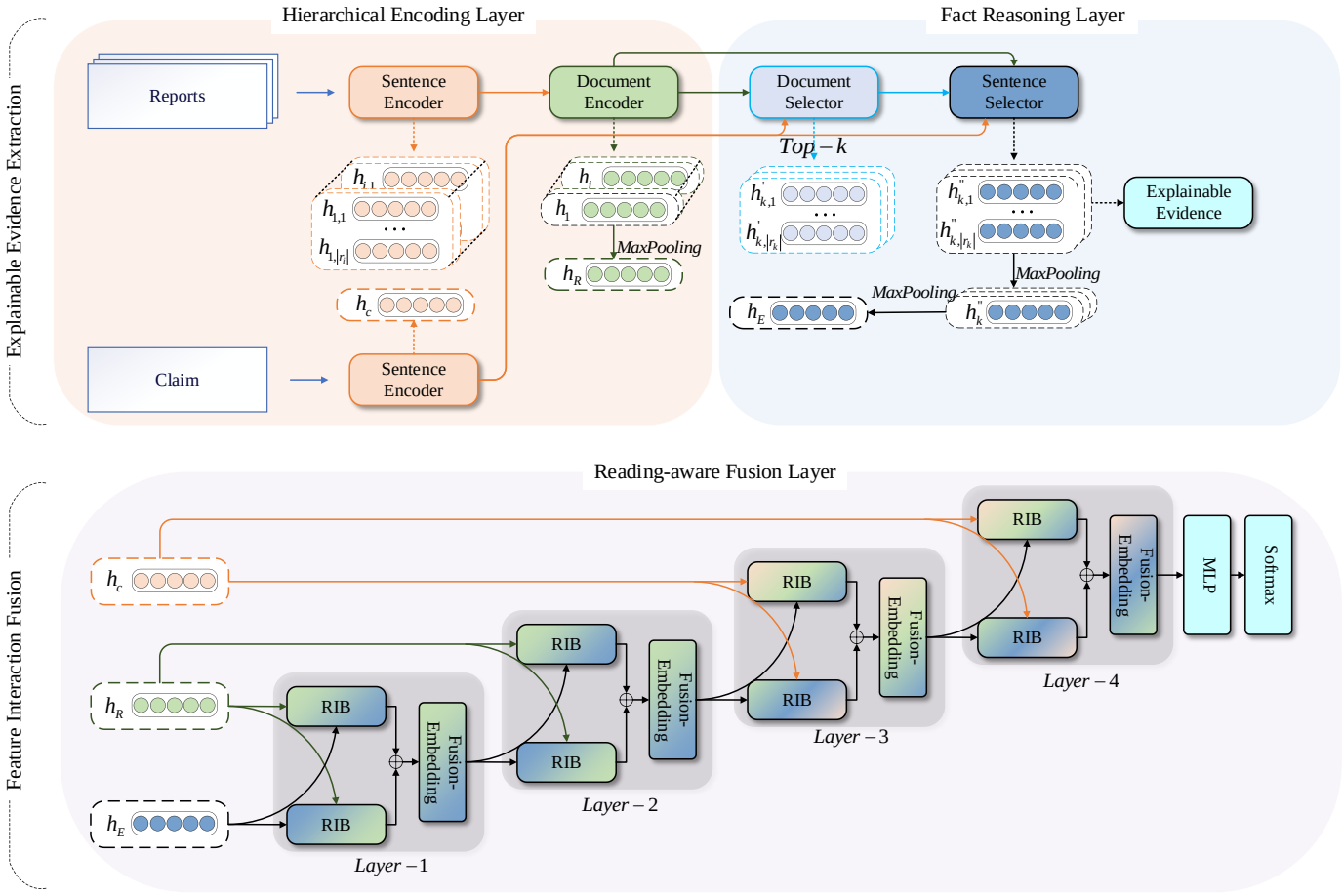


Fig. 2. The proposed RFFR model framework comprises three main layers: the Hierarchical Encoding Layer, which captures sentence-level and full-text representations; the Fact Reasoning Layer, which retrieves relevant reports and sentences as explanatory evidence; and the Reading-aware Fusion Layer, which enhances feature integration and interdependencies for improved detection performance.

$$head = Attention(QW_q, KW_k, VW_v) \quad (2)$$

$$h_i = MultiHead(Q, K, V) = Concat(head_1, \dots, head_m)W_e \quad (3)$$

where, all $W \in \mathbb{R}^{d \times d}$ are trainable parameters, d_e represents the ratio of d to m. $h_i \in \mathbb{R}^d$ is the report representation that integrates all the significant sentence features.

B. Fact Reasoning Layer

To extract explanatory evidence related to the claim from the original reports, this study designs the Fact Reasoning Layer (FRL) in two aspects. First, the Document Selector preliminarily screens the top K reports that may contain hidden facts. Then, a Sentence Selector is designed based on consistency, significance, and redundancy to further refine the most relevant sentences in the original reports.

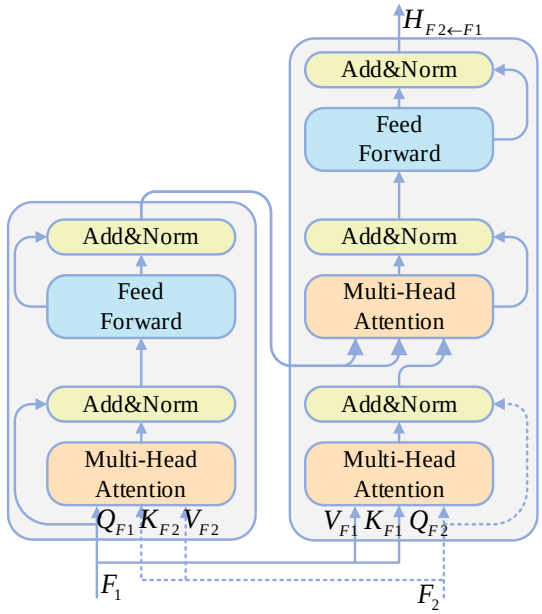
1) Document selector. Since the factual basis corresponding to the claim is hidden in a large number of original reports, the scope of explanatory evidence extraction is automatically narrowed down by ranking the reports and capturing the top-ranked reports. To extract reports that contribute to authenticity prediction from a large number of reports and are worth examining, a coarse-grained Document Selector is developed. The claim is used as the query, and the report is used as the key. An attention weight matrix Att captures the consistency between the claim and the related reports. Att reflects the degree of attention of the claim in the related reports and obtains an importance score for each report r_i.

$$\alpha_c \rightarrow Att = softmax(H_R W_\alpha h_c) \quad (4)$$

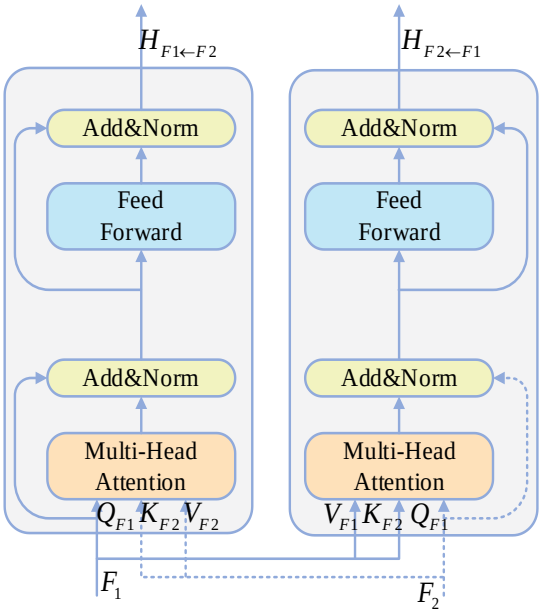
where, $H_R = [h_1; h_2; \cdots; h_{|R|}]$ gathers all the hidden vectors of the reports, and $W_\alpha \in R^{d \times d}$ is a trainable parameter. We use $\alpha_c \rightarrow D$ to rank all reports and select the top K results as reports worth examining. The vector representation of the t-th sentence in the k-th selected report r'_k is represented as $h'_{k,t} \in \{h'_{k,1}, h'_{k,2}, \dots, h'_{k,|r'_k|}\}$, and its document representation is h'_k, which is used for explainable sentence extraction.

2) Sentence selector. Based on the screening of reports, the task of extracting explanatory evidence can be framed as multi-document summarization extraction, where each report is accessed sequentially to identify explainable sentences. Given potential redundancy among reports, multiple original sources are more likely to contain semantically irrelevant and redundant

sentences. This study introduces a fine-grained Sentence Selector, where the selection of explainable sentences is guided by three evaluation indicators.



(a) Reading-aware Interaction Block (RIB)



(b) Traditional co-attention mechanism.

Fig. 3. The architecture diagram of co-attention and our RIB.

Consistency (assessing the relevance of each sentence to the claim theme), Significance (evaluating the importance of each sentence within the report), and Redundancy (determining the novelty and repetition of each sentence relative to previously selected sentences). By integrating these three indicators, the model predicts the probability of each sentence being selected, enabling the extraction of explainable sentences from the original reports.

$$P\left(y_{k,t}^{s}=1\middle|h_c,h'_{k,t},h'_k,h_d\right)= \sigma\left(h'_{k,t}w_c h_c+h'_{k,t}w_r h'_k-h'_{k,t}w_d h_d\right) \quad (5)$$

where, $y_{k,t}^{s}$ is a binary variable indicating whether the t-th sentence in the selected report r'_k should be selected as part of the factual evidence $\hat{E}$, and W_* are trainable parameters. h_d is a redundancy vector initialized to all zeros, updated by the selected sentences in the previously visited reports.

$$h_d=\tanh\left(\sum_t h'_{k-1,t}\cdot P\left(y_{k,t}^{s}=1\right)\right) \quad (6)$$

Considering the number of sentences, RFFR learns to select explainable sentences with selection probabilities higher than the soft threshold $\varepsilon_k=1/|r'_k|$, that is, $P\left(y_{k,t}^{s}=1\right)>\varepsilon_k$. Finally, the integrated representation of the original report and explanatory evidence is obtained through max pooling.

$$h_R=Max\left(\left[h_1;h_2;\cdots;h_{|R|}\right]\right) \quad (7)$$

$$h_E=Max\left(\left[h_1^{"};h_2^{"};\cdots;h_{|K|}^{"}\right]\right) \quad (8)$$

$$h_k^{"}=Max\left(\left[h_{k,1}^{"};h_{k,2}^{"};\cdots;h_{k,|r'_k|}^{"}\right]\right) \quad (9)$$

where, h_R represents the integrated representation of all reports, h_E represents the integrated representation of all explainable sentences, $h''_{k,t}$ represents the sentence representation output from the Sentence Selector, $h_k^{"}$ is the complete representation of explanatory evidence extracted from the k-th report, and K is a hyperparameter controlling the maximum number of selected reports.

C. Reading-Aware Fusion Layer

Intuitively, when reading a report that contains questions, individuals often skim through the article content before focusing on the question section. This process may be repeated multiple times to continuously integrate information from both the question and the report. Based on this reading behavior, this study constructs a Reading-aware Fusion Layer (RFL) to simulate this interaction. The Reading-aware Interaction Block (RIB) serves as the fundamental unit of the RFL, achieving feature fusion by parallelly connecting two RIBs. The RFL is designed to cascade multiple RIB layers for deep feature fusion.

To model human reading behavior, a Reading-aware Interaction Block is developed based on the co-attention mechanism, allowing the model to learn the dependency relationships between different text features, as illustrated in Fig. 3. To achieve comprehensive fusion of text features, four RIB layers are cascaded. The fusion process is gradual, with each RIB layer capable of processing distinct text features, while the output of each layer serves as the input for the subsequent RIB layer. For instance, the input to the first RIB layer is denoted as $<h_E,h_R>$, and its fusion logic is described as follows:

$$H_R=Norm\left(h_R+softmax\left(\frac{h_R h_E}{\sqrt{d}}\right)h_E\right) \quad (10)$$

$$H_E=Norm\left(h_E+softmax\left(\frac{h_E h_R}{\sqrt{d}}\right)h_R\right) \quad (11)$$

$$\widehat{H}_R=Norm\left(H_R+FFN(H_R)\right) \quad (12)$$

$$\hat{h}_{E\leftarrow R}^{(1)}=Norm\left(H_E+softmax\left(\frac{H_E\widehat{H}_R}{\sqrt{d}}\right)\widehat{H}_R\right) \quad (13)$$

$$h_{E\leftarrow R}^{(1)} = Norm\left(\hat{h}_{E\leftarrow R}^{(1)} + FFN\left(\hat{h}_{E\leftarrow R}^{(1)}\right)\right) \quad (14)$$

$$h^{(1)} = h_{E\leftarrow R}^{(1)} \oplus h_{R\leftarrow E}^{(1)} \quad (15)$$

where, $h_{R\leftarrow E}^{(1)}$ is calculated consistently with $h_{E\leftarrow R}^{(1)}$, and $h^{(1)}$ represents the fused semantics of the original report and evidence. $h^{(2)}, h^{(3)}$ are the fused semantics of the corresponding interaction blocks, and $h^{(4)}$ is the final representation of the semantics fusion of claims, evidence, and reports.

D. Learning Task

During the news prediction phase, the final feature representation $h^{(4)}$ is input into a Multi-Layer Perceptron (MLP) layer to predict the authenticity label, as follows:

$$\hat{y} = softmax(\text{MLP}(h^{(4)})) \quad (16)$$

The learning task of the model RFFR is composed of three key sub-tasks: target report selection, explainable sentence extraction, and claim authenticity prediction. Its loss function L_{all} is as follows:

$$L_D = -\sum_i y_i^d \log\left(\hat{y}_i^d\right) \quad (17)$$

$$L_S = -\sum_k \sum_t y_{k,t}^s \log\left(\hat{y}_{k,t}^s\right) \quad (18)$$

$$L_C = -y\log(\hat{y}) \quad (19)$$

$$L_{all} = \theta_D L_D + \theta_S L_S + \theta_C L_C \quad (20)$$

where, L_D, L_S, and L_C represent the cross-entropy losses of the three sub-tasks: target report selection, explainable sentence extraction, and claim authenticity prediction. y_i^d and $\hat{y}_i^d$ represent the true label and predicted label of the report, respectively. $y_{k,t}^s$ and $\hat{y}_{k,t}^s$ represent the true label and predicted probability of the explainable sentence, respectively. y and $\hat{y}$ represent the true label and predicted label of the claim, respectively. β represents the weight parameter, and in the experiment, a multi-task adaptive weighting strategy is used to automatically assign θ_D, θ_S, and θ_C.

1) Multi-task adaptive weighting. To address the parameter optimization problem in multi-task learning, inspired by previous work [55] [56], a Multi-task Adaptive Weighting strategy (MAW) is further proposed to automatically maintain a dynamic balance between tasks on different benchmark datasets. The weighted function $\theta_k(t)$ is defined as follows:

$$\theta_k(t) = \frac{N_k \exp[\gamma_k(t) g(t)]}{\sum_i \exp[\gamma_i(t) g(t)]} \quad (21)$$

$$\gamma_k(t) = \frac{L_k(t-1)}{L_k(t-2)}, g(t) = \frac{\log(t-2)}{T} \quad (22)$$

where, $\theta_k = \theta_k(t)$ represents the weight of the k-th task at the t-th iteration. It indicates the model's focus on task k in the current training step; $N_k = 3$ represents the number of sub-tasks; $\gamma_k(t)$ measures the loss change rate of sub-task k in the last two iterations. If the loss of a task changes significantly, that is, the loss decreases slowly or is unstable, the model will increase the weight of that task; if the loss of a task decreases rapidly, its weight is reduced. $g(t)$ is a global adjustment function used to smooth the weight differences between tasks, making the weight adjustment of different tasks more reasonable and stable; T is a constant value used to control the speed of weight adjustment.

V. Experiments

A. Datasets

This study employs two publicly available datasets to evaluate the proposed method: RAWFC [57] and LIAR [58]. The RAWFC dataset, sourced from Snopes.com, comprises 2012 claims related to fact-checking tasks, each accompanied by supporting evidence and labeled as true, false, or half-true. The LIAR dataset, compiled by PolitiFact.com, includes 12,836 short claims, each categorized as true, mostly true, half-true, barely true, false, or pants-on-fire.

B. Experimental Settings and Evaluation Metrics

To prevent overfitting, the model parameters of Bert were frozen when training on both the RAWFC and LIAR datasets; the hidden dimension in the Sentence and Document Encoders is set to 768; the Document Encoder comprises 12 attention heads and is constructed from 4 attention blocks; when selecting reports, the maximum number of reports K selected for each claim is set to 6 and 9, the soft threshold ε_i is set to $1/|r_i'|$, and the maximum number of explainable sentences extracted is set to 6 and 11 for RAWFC and LIAR, respectively. The batch size is set to 64, the number of training epochs is set to 100, the learning rate is set to 1e-5, and the Adam optimizer is used to minimize the combined cross-entropy loss. To assess the performance of the model, this study uses Accuracy, Macro-averaged Precision (macro-P), Macro-averaged Recall (macro-R), and F1 score (macro-F1) for the detection results evaluation, and ROUGE-N (N=1) and ROUGE-L to assess the quality of the generated explanations.

C. Comparative Baselines and Performance Comparison

To demonstrate the effectiveness of the proposed method, this study compares RFFR with the following baselines:

SVM [59]: Training an SVM-based model for fake news detection using bag-of-words features.

CNN [58]: Integrating available metadata features into representation learning.

RNN [60]: Learning representations from word sequences without external resources.

DeClarE [28]: Combining word embeddings from claims, reports, and sources to assess the credibility of claims.

dEFEND [13]: Using a GRU-based model for veracity prediction and providing explanations.

SentHAN [35]: Representing each sentence based on sentence-level coherence and semantic conflict with the claim.

SBERT-FC [49]: Using SentenceBERT for encoding and detecting fake news based on highly ranked sentences.

GenFE/GenFE-MT [48]: Detecting fake news and providing explanations independently or jointly in a multi-task setting.

TABLE I. COMPARISON RESULTS OF RFFR WITH DIFFERENT BASELINE MODELS ON THESE TWO DATASETS

Model	RAWFC				LIAR			
	Accuracy(%)	P(%)	R(%)	Macro-F1(%)	Accuracy(%)	P(%)	R(%)	Macro-F1(%)
SVM	32.03	32.33	32.51	31.71	14.73	15.78	15.92	15.34
CNN	37.82	38.80	38.50	38.59	21.36	22.58	22.39	21.36
RNN	41.20	41.35	42.09	40.39	21.00	24.36	21.20	20.79
DeClarE	42.18	43.39	43.52	42.18	19.17	22.86	20.55	18.43
dEFEND	42.75	44.93	43.26	44.07	16.64	23.09	18.56	17.51
SentHAN	46.47	45.66	45.54	44.25	19.02	22.64	19.96	18.46
SBERT-FC	46.88	51.06	45.92	45.51	21.75	24.09	22.07	22.19
GenFE	42.66	44.29	44.74	44.43	27.02	28.01	26.16	26.49
GenFE-MT	46.89	45.64	45.27	45.08	15.91	18.55	19.90	15.15
CofCED	50.56	**52.99**	50.99	51.07	28.07	29.48	**29.55**	28.93
RFFR	**52.16**	51.38	**51.76**	**51.13**	**28.39**	**29.68**	29.07	**29.18**

CofCED [57]: A coarse-to-fine cascaded evidence distillation neural network for explanatory fake news detection based on original reports.

TABLE I. presents the detection performance of the proposed RFFR compared to existing strong baselines in terms of precision, recall, and macro F1 (macF1). The following observations can be drawn from the table:

- CNN and RNN outperform SVM on both datasets, indicating that deep learning methods are more effective at capturing semantic and syntactic features from original reports. While dEFEND, DeClarE, and SentHAN achieve better performance on RAWFC by aggregating multiple features from claims, reports, and sources, their results are slightly worse on LIAR. This discrepancy can be attributed to the fine-grained labels in LIAR, which pose additional challenges.
- SBERT-FC and GenFE outperform SentHAN and dEFEND on both datasets, demonstrating the superiority of pre-trained models. Although GenFE-MT performs better than GenFE on RAWFC, it performs significantly worse on LIAR compared to other baselines, highlighting the difficulty of fine-grained fake news detection and explanation generation in a multi-task setting. CofCED achieves better performance than both GenFE and GenFE-MT, suggesting that original reports provide richer explanatory evidence than fact-checking reports.
- Compared to the other models, the proposed RFFR fake news detection model exhibits advantages across various metrics on both the RAWFC and LIAR datasets. Specifically, the accuracy of fake news detection improves by 1.6% on the RAWFC dataset, while the F1 score improves by 0.25% on the LIAR dataset. This enhancement can be attributed to two factors: 1) The Fact Reasoning Layer effectively utilizes different fine-grained selectors to capture more explanatory evidence from original reports, facilitating more accurate fake news detection. 2) The construction of a Reading-aware Fusion Layer for interactive fusion proves more effective than simple feature interaction.

D. Ablation Analysis

1) Effectiveness of each component. To assess the effectiveness of each component in RFFR, eight model variants were created: w/o Report, w/o Evidence, w/o FRL, w/o DS, w/o SS, w/o RFL, w/o RIB, and w/o RFS. These variants indicate the removal of the following components: original report representation, explanatory evidence representation, Fact Reasoning Layer, Document Selector, Sentence Selector, Reading-aware Fusion Layer, Reading-aware Interaction Block, and Reading-aware Fusion Strategy.

The comparison results are shown in TABLE II. yields the following observations:

- RFFR consistently outperforms all ablation experiment variants on both datasets, demonstrating that every component is essential for the model's effectiveness in detecting fake news.
- The w/o Report variant exhibits the poorest performance, followed closely by w/o Evidence. This suggests that Evidence is crucial for the fake news detection task and that the original report provides additional feature information that supports detection, distinct from Evidence. Sole reliance on either the original report or Evidence hinders detection effectiveness.
- The performance of w/o FRL drops significantly, highlighting that noise in the original report adversely affects veracity prediction. The w/o SS variant performs much worse than others, as irrelevant or redundant information in reports can dilute the effectiveness of evidence. The w/o DS variant shows even poorer performance, indicating that noisy reports can impair sentence selection and model training.

TABLE II. ABLATION ANALYSIS ON RAWFC AND LIAR DATASETS

Model	RAWFC				LIAR			
	Accuracy(%)	P(%)	R(%)	Macro-F1(%)	Accuracy(%)	P(%)	R(%)	Macro-F1(%)
w/o Report	32.76	36.76	36.32	35.59	21.16	20.83	20.34	20.24
w/o Evidence	35.59	38.54	38.30	38.07	21.60	22.31	21.52	21.65
w/o FRL	37.88	36.96	40.13	40.45	23.17	23.67	23.04	23.02
w/o DS	40.75	40.12	40.34	39.30	24.31	25.54	24.80	24.06
w/o SS	41.75	40.78	40.85	40.05	22.38	23.71	22.86	23.36
w/o RFL	38.00	38.00	37.83	37.36	23.61	23.92	23.30	23.75
w/o RIB	46.41	45.75	46.0	45.50	25.57	26.41	25.84	25.93
w/o RFS	47.46	46.80	47.10	46.55	26.55	26.99	26.45	26.54
RFFR	**52.16**	**51.38**	**51.76**	**51.13**	**28.39**	**29.68**	**29.07**	**29.18**

- The performance of w/o RFL is significantly reduced, suggesting that simulating human reading behavior facilitates the tight integration of diverse feature information. Both w/o RFS and w/o RIB show substantial performance drops, indicating that these components are vital for deep feature fusion. RIB enables interactive fusion of different features, while RFS enhances the fusion process by mimicking human reading behavior.

2) Comparative analysis of Reading-aware Fusion Layer. The Reading-aware Fusion Layer (RFL) is the mechanism employed by RFFR for deep feature fusion, comprising two core components: the Reading-aware Fusion Strategy (RFS) and the Reading-aware Interaction Block (RIB). The RIB facilitates deep interactive fusion of different features, while the RFS enhances the fusion process by simulating human reading behavior.

Comparative experiments were conducted using traditional Co-Attention [61], Cross-Attention [62], and a variant without RIB (w/o RIB) as alternatives to the RIB, tested under both conditions with and without RFS. The results are presented in the table.

As shown in Fig. 4, the comparison results indicate that the removal of either RIB or RFS significantly impairs the performance of RFFR. In both scenarios (with and without RFS), Cross-Attention, Co-Attention, and RIB consistently outperform the w/o RIB variant, underscoring the necessity of deep interactive fusion among different features. Notably, RIB surpasses both Cross-Attention and Co-Attention, suggesting that RIB achieves more comprehensive feature depth fusion. Furthermore, irrespective of whether alternative methods or RIB are employed, the performance with RFS is consistently superior to that without RFS, highlighting that simulating human reading behavior effectively strengthens the feature fusion process.

3) Comparative analysis of the fact reasoning layer. To examine the impact of the Fact Reasoning Layer (FRL) on explanatory evidence reasoning, three model variants were created: w/o Consistency, w/o Significance, and w/o Redundancy, representing the removal of the respective components. Comparative evaluations were conducted using ROUGE-N (N=1) and ROUGE-L to assess the quality of the explanatory evidence:

LEAD-N [63]: Using the first N sentences as an explanation, where N = 5.

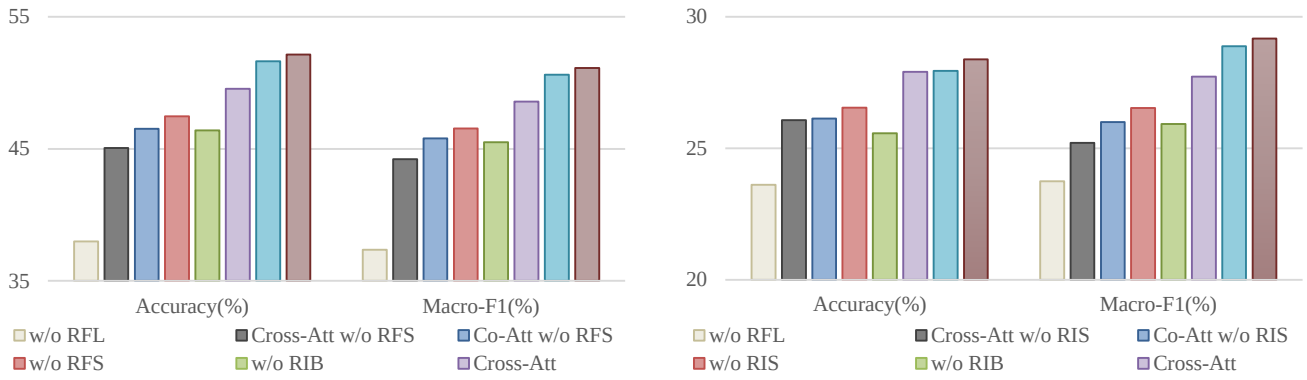
Oracle [48]: Manually selecting the most relevant information from the original report.

dEFEND [13]: Providing explanations based on internal attention weights.

GenFE-MT [48]: Using pre-trained models to generate explanations.

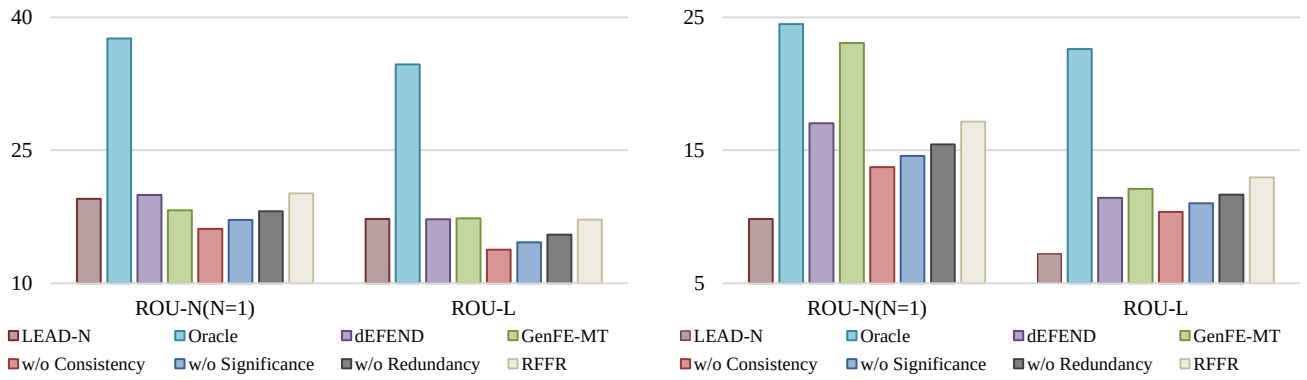
The comparison results, presented in Fig. 5, reveal several key observations: The performance of most models on the LIAR dataset is generally lower than on the RAWFC dataset, indicating that higher granularity in the dataset complicates the generation of explanatory evidence. Additionally, GenFE-MT outperforms dEFEND on both datasets, suggesting that pre-trained models excel in generating explanatory evidence from original reports.

Notably, RFFR outperforms the three ablation variants on both datasets, demonstrating that the evaluations of consistency, significance, and redundancy all enhance the effectiveness of RFFR in generating explanatory evidence. Furthermore, RFFR achieves state-of-the-art performance on RAWFC and displays ROUGE scores closely matching those of GenFE-MT on LIAR, indicating its capability to effectively extract explanatory evidence.



(a) On the RAWFC dataset (b) On the LIAR dataset

Fig. 4. Comparison of performance of different ablation blocks in RFL.



(a) On the RAWFC dataset (b) On the LIAR dataset

Fig. 5. Comparison of performance of different ablation blocks in FRL.

VI. Case Study

To further explore the process of evidence selection in RFFR from original reports, this study selected a case from each of the two datasets and visualized the selection of explanatory sentences, with the results shown in TABLE III. It can be observed that in both the RAWFC and LIAR datasets, sentences with higher scores refuted the claim from various perspectives, while sentences with lower scores contributed less to the prediction of authenticity. This indicates that the assessments of consistency, significance, and redundancy, which are used to select explanatory sentences, are helpful in identifying the main factors for each sentence to serve as explanatory evidence. This method increases the transparency of the system and the credibility of generating explanatory evidence.

VII. Conclusion

Existing fact-checking-based fake news detection methods encounter two main issues: a heavy reliance on fact-checking reports, which often lack explanatory evidence related to the original reports, and a superficial level of feature interaction. To tackle these challenges, this study proposes a Reading-aware Fusion Fact Reasoning Network for explainable fake news detection. At the evidence retrieval level, a Hierarchical Encoding Layer is constructed to capture feature representations of text sentences and the full text, followed by a Fact Reasoning Layer that identifies the most relevant report and sentence representations as explanatory evidence, thus reducing dependence on fact-checking reports. At the feature fusion level, inspired by human reading behavior, a Reading-aware Fusion Layer is introduced to learn dependencies between different feature representations for deep integration. Extensive experiments on the RAWFC and LIAR datasets validate the effectiveness of RFFR. Future work will focus on expanding RFFR for detection by integrating additional data modalities and knowledge bases.

TABLE III. THE PROCESS OF RFFR SELECTING EXPLANATORY EVIDENCE IN THE ORIGINAL REPORT IS VISUALIZED

Dataset: RAWFC **Label:** half **Claim:** U.S. House Speaker Nancy Pelosi publicly criticized the actions of federal agents in U.S. cities during protests in 2020—while simultaneously supporting a budget bill that would fund such law enforcement efforts. **Evidence:** As a comedian who claims he sees the faults in both major U.S. political parties, Jimmy Dore said this to fans amid clashes between Portland, Oregon, protesters and federal agents operating under.....	Consistency	Significance	Redundancy	Overall	Is evidence
1. The new federal bill would change "the standard to evaluate whether law enforcement use of force be justify from whether the force be 'reasonable' to whether the force be 'necessary'" accord to a bill summary send to McClatchy by the office of House Speaker Nancy Pelosi, D-San Francisco.	0.8	0.7	0.9	0.7	1
2. Congressional Democrats announce Monday that they want to raise the legal standard for when law enforcement officer can use deadly force, propose a bill similar to a new California law that aim to reduce lethal encounter.	0.5	0.3	0.4	0.3	0
3. Protests have call attention to the in-custody death of black men and woman, and urge law enforcement reform.	0.2	0.7	0.4	0.3	0
4. The Global Chapter of Black Lives Matter originally support the bill, but later pull it, say amendment to the bill add due to police concern have "significantly weaken" it.	0.8	0.6	0.5	0.8	1
Dataset: LIAR **Label:** pants-fire **Claim:** Hillary (Clinton), one time late at night when she was exhausted, misstated and immediately apologized for it, what happened to her in Bosnia in 1995. **Evidence:** Bill Clinton has implied that the media is biased and covers his wife too harshly. In Boonville, he seemed to lament that unfairness by warning.....	Consistency	Significance	Redundancy	Overall	Is evidence
1. But there be a lot of fulminate because Hillary, one time late at night when she be exhaust, misstate and immediately apologize for it, what happen to her in Bosnia in 1995.	0.3	0.3	0.5	0.2	0
2. And some of them when they 're 60 they 'll forget something when they 're tire at 11 o'clock at night, too.	0.5	0.7	0.7	0.7	1
3. His wife don't make the sniper fire claim one time late at night when she be exhaust.	0.5	0.8	0.3	0.4	0

ACKNOWLEDGMENT

This research was supported by the Key Scientific Research Project of Colleges and Universities in Henan Province, titled "Research on Key Technologies for Text Processing in Network Public Opinion Monitoring" (Project No. 22B520054).

REFERENCES

[1] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media," Big Data, vol. 8, no. 3, pp. 171-188, 2018.

[2] Z. Zhao, P. Resnick, Q. Mei, "Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts," in Proceedings of the 24th International Conference on World Wide Web, pp. 1395-1405, 2015.

[3] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22-36, 2017.

[4] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, et al., "Detecting rumors from microblogs with recurrent neural networks," In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp. 3818-3824, 2016.

[5] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, X. Zhang, "Multi-modal Fake News Detection on Social Media via Multi-grained Information Fusion," In Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, pp. 343-352, 2023.

[6] C. Castillo, M. Mendoza, B. Poblete, "Information credibility on twitter," in Proceedings of the 20th International Conference on World Wide Web, pp. 675-684, 2011.

[7] S. Wu, Q. Liu, Y. Liu, L. Wang, T. Tan, "Information Credibility Evaluation on social media," in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 4403-4404, 2016.

[8] L. Hu, T. Yang, L. Zhang, W. Zhong, D. Tang, C. Shi, et al., "Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge," In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 754-763, 2021.

[9] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, K. Shu, "Mining Dual Emotion for Fake News Detection," In Proceedings of the Web Conference 2021, pp. 3465-3476, 2021.

[10] L. Wu, Y. Rao, H. Jin, A. Nazir, L. Sun, "Different Absorption from the Same Sharing: Sifted Multi-task Learning for Fake News Detection," In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 4644-4653, 2019.

[11] J. Xie, S. Liu, R. Liu, Y. Zhang, Y. Zhu, "SERN: Stance Extraction and Reasoning Network for Fake News Detection," In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2520-2524, 2021.

[12] L. Wu, Y. Rao, C. Zhang, Y. Zhao, A. Nazir, "Category-Controlled Encoder-Decoder for Fake News Detection," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 2, pp. 1242-1257, 2023.

[13] K. Shu, L. Cui, S. Wang, D. Lee, H. Liu, "dEFEND: Explainable fake news detection," In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 395-405, 2019.

[14] K. Shu, D. Mahudeswaran, S. Wang, H. Liu, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," In Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, no. 1, pp. 626-637, 2020.

[15] Q. Nan, J. Cao, Y. Zhu, Y. Wang, J. Li, "MDFEND: Multi-domain fake news detection," In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 3343-3347, 2021.

[16] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, X. Zhang, "Multimodal fake news detection via CLIP-guided learning," In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo, pp. 2825-2830, 2023.

[17] Y. Bu, Q. Sheng, J. Cao, P. Qi, D. Wang, J. Li, "Combating online misinformation videos: Characterization, detection, and future directions," In Proceedings of the 31st ACM International Conference on Multimedia, pp. 8770-8780, 2023.

[18] K. Shu, D. Mahudeswaran, S. Wang, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," In Proceedings of the International AAAI Conference on Web and Social Media, pp. 626-637, 2020.

[19] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 849-857, 2018.

[20] P. Li, X. Sun, H. Yu, Y. Tian, F. Yao, G. Xu, "Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection," In IEEE Transactions on Multimedia, vol. 24, pp. 3455-3468, 2022.

[21] X. Zhou, J. Wu, R. Zafarani, "Similarity-Aware Multi-modal Fake News Detection," In Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 354-367, 2020.

[22] Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, "Multimodal Fusion with Co-Attention Networks for Fake News Detection," In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 2560–2569, 2021.

[23] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, et al., "Cross-modal Ambiguity Learning for Multimodal Fake News Detection," In Proceedings of the ACM Web Conference 2022, pp. 2897–2905, 2022.

[24] M. Dhawan, S. Sharma, A. Kadam, R. Sharma, P. Kumaraguru, "Game-on: graph attention network based multimodal fusion for fake news detection," In Social Network Analysis and Mining, vol. 14, pp. 1-13, 2022.

[25] Y. Ganin, V. Lempitsky, "Unsupervised domain adaptation by backpropagation," In Proceedings of the 32nd International Conference on International Conference on Machine Learning, pp. 1180–1189, 2015.

[26] Y. Zhu, Q. Sheng, J. Cao, S. Li, D. Wang, F. Zhuang, "Generalizing to the Future: Mitigating Entity Bias in Fake News Detection," In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2120–2125, 2022.

[27] K. Popat, S. Mukherjee, J. Strötgen, G. Weikum, "Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media," In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1003–1012, 2017.

[28] K. Popat, S. Mukherjee, A. Yates, G. Weikum, "DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning," In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 22–32, 2018.

[29] Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, J. G. Simonsen, et al., "MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims," In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 4685-4697, 2019.

[30] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, "FEVER: a Large-scale Dataset for Fact Extraction and VERification," In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 809-819, 2018.

[31] B. M. Yao, A. Shah, L. Sun, J. Cho, L. Huang, "End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models," In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2733-2743, 2023.

[32] J. Kim, S. Park, Y. Kwon, Y. Jo, J. Thorne, E. Choi, "FactKG: Fact Verification via Reasoning on Knowledge Graphs," In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 16190-16206, 2023.

[33] Zou, Z. Zhang, H. Zhao, "Decker: Double Check with Heterogeneous Knowledge for Commonsense Fact Verification," In Findings of the Association for Computational Linguistics: ACL 2023, pp. 11891-11904, 2023.

[34] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, et al., "Automated Fact-Checking for Assisting Human Fact-Checkers," In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pp. 4551-4558, 2021.

[35] J. Ma, W. Gao, S. Joty, K.-F. Wong, "Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks," In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2561-2571, 2019.

[36] Y. Wang, S. Qian, J. Hu, Q. Fang, C. Xu, et al., "Fake News Detection via Knowledge-driven Multimodal Graph Convolutional Networks," In Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 540-547, 2020.

[37] S. B. Parikh, P. K. Atrey, "Media-Rich Fake News Detection: A Survey," In Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval, pp. 436-441, 2018.

[38] P. K. Verma, P. Agrawal, I. Amorim, R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 881-893, 2021.

[39] L. Wu, Y. Rao, Y. Zhao, H. Liang, A. Nazir, "DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification," In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1024-1035, 2020.

[40] L. Hu, T. Yang, L. Zhang, W. Zhong, D. Tang, C. Shi, et al., "Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge," In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 754-763, 2021.

[41] H. Choi, Y. Ko, "Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection," In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 2950-2954, 2021.

[42] Y. Dou, K. Shu, C. Xia, P. S. Yu, L. Sun, "User Preference-aware Fake News Detection," In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2051-2055, 2021.

[43] P. Qi, J. Cao, T. Yang, J. Guo, J. Li, "Exploiting Multi-domain Visual Information for Fake News Detection," In 2019 IEEE International Conference on Data Mining, pp. 518-527, 2019.

[44] K. Shu, D. Mahudeswaran, S. Wang, H. Liu, "Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation," In Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 626-637, 2020.

[45] Y. Liu, M. Lapata, "Text Summarization with Pretrained Encoders," In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3730-3740, 2019.

[46] Y.-J. Lu, C.-T. Li, "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media," In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 505-514, 2020.

[47] Y. Nie, H. Chen, M. Bansal, "Combining fact extraction and verification with neural semantic matching networks," In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, pp. 842-849, 2019.

[48] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, "Generating Fact Checking Explanations," In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7352-7364, 2020.

[49] N. Kotonya, F. Toni, "Explainable Automated Fact-Checking for Public Health Claims," In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 7740-7754, 2020.

[50] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh, "SpotFake: A Multi-modal Framework for Fake News Detection," In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data, pp. 39-47, 2019.

[51] S. Y. Boulahia, A. Amamra, M. R. Madi, S. Daikh, "Early Intermediate and Late Fusion Strategies for Robust Deep Learning-based Multimodal Action Recognition," Machine Vision and Applications, vol. 32, no. 6, pp. 1-18, 2021.

[52] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, L. Wei, "Detecting fake news by exploring the consistency of multimodal data," Information Processing & Management, vol. 58, no. 5, pp. 102610, 2021.

[53] P. Meel, D. K. Vishwakarma, "Multi-modal fusion using Fine-tuned Self-attention and transfer learning for veracity analysis of web information," Expert Syst. Appl., vol. 229, pp. 1-16, 2023.

[54] S. Singhal, M. Dhawan, R. R. Shah, P. Kumaraguru, "Inter-modality Discordance for Multimodal Fake News Detection," In Proceedings of the 3rd ACM International Conference on Multimedia in Asia, pp. 33-39, 2022.

[55] Z. Chen, V. Badrinarayanan, C.-Y. Lee, A. Rabinovich, "GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks," In Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 794-803, 2018.

[56] S. Liu, E. Johns, A. J. Davison, "End-To-End Multi-Task Learning With Attention," In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1871-1880, 2019.

[57] Z. Yang, J. Ma, H. Chen, H. Lin, Z. Luo, Y. Chang, "A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection," In Proceedings of the 29th International Conference on Computational Linguistics, pp. 2608-2621, 2022.

[58] W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 422-426, 2017.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.

[60] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2931-2937, 2017.

[61] Lu, J., Batra, D., Parikh, D., Lee, S. "ViLBERT: Pretraining TaskAgnostic Visiolinguistic Representations for Vision-and-Language Tasks," In Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 2-12, 2019.

[62] Chen, C.-F. R., Fan, Q., Panda, R. "CrossViT: Cross-Attention MultiScale Vision Transformer for Image Classification," In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, pp. 347-356, 2021.

[63] R. Nallapati, F. Zhai, B. Zhou, "SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents," In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 3075-3081, 2017.

Solar-Net: Adaptive Fusion of Spatial-Temporal Features for Resilient Solar Power Generation Forecasting

Wenqian Su[1], Jason See Toh Seong Kuan[2], Xiangyu Shi[3], Yuchen Zhang[4]
Fujian Forestry Vocational & Technical College[1]
INTI International University, Malaysia[1, 2, 3]
Nanchang Institute of Science and Technology[4]

***Abstract*—Solar power generation forecasting faces significant challenges due to intermittency and volatility, particularly under extreme weather conditions. This study proposes Solar-Net, a novel solar power generation prediction model based on a CNN+Transformer hybrid parallel architecture with an adaptive attention fusion mechanism. The CNN branch extracts spatial features from the power station layout and environmental conditions, while the Transformer branch models temporal dependencies in generation patterns. The core innovation lies in the adaptive attention fusion mechanism that dynamically adjusts branch weights according to real-time meteorological conditions, enabling the model to automatically adapt to varying environmental scenarios. Experiments were conducted on a comprehensive dataset containing over 50,000 observation points from two photovoltaic power stations. Results demonstrate that Solar-Net achieves superior performance compared to existing methods, with Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) improvements of 12.7% and 10.9%, respectively. Under extreme weather conditions such as dust storms, the model maintains prediction errors within 8.5% of peak power generation, representing a 45.7% average reduction compared to baseline methods. The multi-scale convolution design enhances prediction accuracy by 10.5% while reducing computational complexity by 21.3%. The proposed Solar-Net model provides a robust and efficient solution for solar power generation forecasting, demonstrating significant potential for improving grid dispatching efficiency and supporting renewable energy integration in power systems.**

Keywords*—Solar power generation forecasting; hybrid deep learning; adaptive attention fusion; CNN+Transformer; extreme weather adaptability; sustainable development goal 7

I. INTRODUCTION

With global climate change and fossil energy depletion becoming increasingly prominent issues, the importance of solar energy as a clean, renewable energy source continues to rise. According to International Energy Agency (IEA) data, global solar installation capacity has grown from 40GW in 2010 to over 1100GW in 2023, with an average annual growth rate of 27%, making it the fastest-growing renewable energy type [1]. However, the intermittency and volatility of solar power generation pose serious challenges to power grid dispatching and energy market operations. High-precision solar power generation forecasting technology has significant value for improving power grid operation stability, reducing backup capacity requirements, and optimizing market trading strategies, and has become a key research direction in the smart grid field [2].

Existing solar power generation prediction methods can be primarily classified into three categories: physical models, statistical models, and artificial intelligence models. Physical models based on radiation transfer theory and photovoltaic system characteristic equations possess good interpretability, but their accuracy in short-term prediction is limited due to challenges in parameter calibration and high computational complexity [3]. Statistical models (such as ARIMA and exponential smoothing) offer high computational efficiency but have limited adaptability to nonlinear patterns and extreme weather conditions. In recent years, deep learning models have made significant progress in solar energy prediction due to their powerful nonlinear mapping capabilities. However, existing deep learning methods often treat spatial features and temporal dependencies as independent dimensions, lacking effective fusion mechanisms, making it difficult to comprehensively capture the complex dynamic characteristics of solar power generation systems in variable environments, with prediction performance declining significantly under extreme weather conditions [4].

To address these challenges, this study proposes Solar-Net, a solar power generation prediction model based on a CNN+Transformer hybrid parallel architecture. The model adopts a dual-branch parallel design, with the CNN branch focusing on capturing spatial features of solar power station physical layout and local environmental conditions, while the Transformer branch is responsible for modeling long-term temporal dependencies of power generation. The model innovatively introduces an adaptive attention fusion mechanism that can dynamically adjust the weights of each branch according to different meteorological conditions, achieving adaptive capability for varying environments. Experiments show that compared with existing best methods, Solar-Net improves Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) by 12.7% and 10.9% respectively, and maintains stable prediction performance even under extreme weather conditions such as dust storms, providing important technical support for high-proportion grid integration of renewable energy [5].

The main academic contributions of this research include:

1) Adaptive environment-aware fusion mechanism. Innovatively designed an adaptive attention fusion architecture based on dynamic weight allocation, achieving intelligent integration of spatial features and temporal dependencies. Experiments verify that this mechanism can effectively respond to meteorological condition changes, reducing prediction errors by an average of 45.7% compared to benchmark methods under extreme conditions such as dust storms. This mechanism essentially establishes a dynamic mapping relationship between environmental conditions, model structure, and prediction strategy, providing a new paradigm for adaptive prediction in complex environments.

2) Parallel dual-modal perception framework. Proposed a dual-branch parallel architecture targeting solar power generation characteristics, achieving collaborative modeling of spatial features and temporal dependencies. The CNN branch focuses on capturing spatial features of solar power station physical layout and local environmental conditions, while the Transformer branch is responsible for modeling long-term temporal dependencies of power generation. Compared to single-structure models, this framework improves MAE and RMSE metrics by 18.3% and 16.9%, respectively, while reducing computational complexity by 21.3%.

3) Efficient multi-scale feature extraction. Enhanced the model's capability to capture microscopic local changes and macroscopic cyclic patterns in solar power generation systems through reasonably designed multi-scale convolution structures and self-attention mechanisms. Experiments show that simultaneously using three scales of convolution kernels (3×3, 5×5, and 7×7) improves prediction accuracy by approximately 10.5% compared to single-size kernels, while reducing computational complexity.

II. Related Work

A. Traditional Solar Power Generation Prediction Methods

Traditional methods for solar power generation prediction primarily include physical models and statistical models. Physical models are based on solar radiation transfer theory and photovoltaic system electrical characteristic equations, which are predicted by constructing deterministic mapping relationships from meteorological parameters to power generation. Diagne et al. [6] reviewed solar energy prediction methods based on Numerical Weather Prediction (NWP), pointing out that although physical models have good interpretability and theoretical foundations, their prediction accuracy highly depends on the quality of meteorological data and parameter calibration accuracy. Especially in short-term prediction scenarios (1 to 6 hours), physical models have high computational complexity and respond slowly to local micro-meteorological changes, resulting in limited prediction performance [7].

Statistical models predict through fitting time series patterns from historical data, with typical methods including Autoregressive Integrated Moving Average (ARIMA), exponential smoothing, and multivariate regression. Reikard [8] compared six statistical models in solar energy prediction and found that for hourly-level predictions, ARIMA models have relative advantages, but show significantly increased errors in cases of abrupt weather changes. Traditional machine learning methods such as Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting Regression Trees (GBRT) improved prediction accuracy through more complex nonlinear mapping [9], but these methods still struggle to simultaneously process spatiotemporal dependencies in solar power generation data, particularly performing poorly in complex meteorological conditions and long sequence prediction scenarios.

The main limitations of traditional methods include: 1) limited ability to process nonlinear and non-stationary time series; 2) difficulty in simultaneously modeling spatial features and temporal dependencies; 3) insufficient adaptability to extreme weather events. In contrast, the Solar-Net model proposed in this study can effectively overcome these limitations through a CNN+Transformer hybrid architecture and adaptive fusion mechanism, achieving comprehensive modeling of the complex dynamic characteristics of solar power generation systems.

B. Applications of Deep Learning Models in Solar Energy Prediction

In recent years, deep learning has gained widespread application in solar power generation prediction due to its powerful feature extraction and nonlinear modeling capabilities. Convolutional Neural Networks (CNN), with their local receptive field and weight sharing characteristics, can effectively capture spatial relationships in solar power generation data. Wang et al. [10] designed a multi-scale CNN model to process irradiance images, achieving regional solar energy prediction, but CNNs have limited ability to model long-term temporal dependencies, making it difficult to capture long-cycle patterns such as seasonal variations.

Recurrent Neural Network families, especially Long Short-Term Memory networks (LSTM) and Gated Recurrent Units (GRU), have shown excellent performance in processing solar power generation time series. Gensler et al. [11] applied deep LSTM to cross-day prediction, significantly outperforming traditional time series models. However, RNN series models still face gradient vanishing problems when processing long sequences and are difficult to compute in parallel, limiting their application in large-scale prediction. Transformer models based on self-attention mechanisms overcome these shortcomings of RNNs and can directly model associations between any two time points in a sequence. However, Transformers alone struggle to fully utilize the spatial layout information of solar power generation systems.

The main deficiencies of existing deep learning methods are: 1) single model structures struggle to effectively process both spatial and temporal features simultaneously; 2) lack of adaptive mechanisms for different weather conditions; 3) poor prediction stability in extreme meteorological events. The Solar-Net model addresses these issues through parallel branch design and adaptive fusion mechanisms, achieving

comprehensive modeling of all aspects of solar power generation systems.

C. Hybrid and Ensemble Models

To address the limitations of single models, researchers have proposed various hybrid and ensemble methods for solar power generation prediction. Hybrid methods based on physical models and data-driven models can combine the advantages of both approaches, such as the physics-aware neural network proposed by Das et al. [12], which improved the generalization ability of deep learning models by introducing radiation transfer equation constraints. However, such methods often require complex physical model parameterization processes, increasing model complexity and training difficulty.

In deep learning hybrid models, Khan et al. [13] proposed a Deep Stacked Ensemble with XGBoost (DSE-XGB), enhancing prediction stability by integrating ANN, LSTM, and XGBoost. However, these methods adopt two-stage training strategies where sub-models are trained independently before integration, lacking end-to-end joint optimization mechanisms and making it difficult to capture complementary features between models. Recently, Zhang et al. [14] designed a dual-stream network architecture (DSTP) to process spatial and temporal information separately, but it uses simple averaging or concatenation for feature fusion, unable to dynamically adjust the importance weights of different branches according to different conditions.

The key deficiencies of existing hybrid models are: 1) most use two-stage training, lacking end-to-end joint optimization; 2) simple fusion mechanisms, unable to adjust according to real-time conditions; 3) high computational complexity, limiting practicality. The Solar-Net model proposed in this study, through an adaptive attention fusion mechanism, achieves dynamic weight allocation and end-to-end joint training of CNN and Transformer branches, maintaining computational efficiency while improving prediction accuracy, and demonstrating significant advantages especially under extreme meteorological conditions, providing a more reliable and flexible technical solution for solar power generation prediction.

III. Methodology

A. Overall Model Architecture

As shown in Fig. 1, the Solar-Net model proposed in this research adopts a hybrid parallel architecture aimed at simultaneously capturing spatial features and temporal dependencies in solar power generation time series data. The model consists of four main components: input data processing, CNN spatial feature extraction branch, Transformer temporal dependency modeling branch, and adaptive attention fusion mechanism.

Through this preprocessing procedure, we provide high-quality training data for the CNN-Transformer hybrid architecture model, which lays a solid foundation for subsequent experimental evaluation. In particular, the normalization process and the reasonable time-series sample construction method can fully utilize the advantages of the model in feature extraction and long-term dependency modeling.

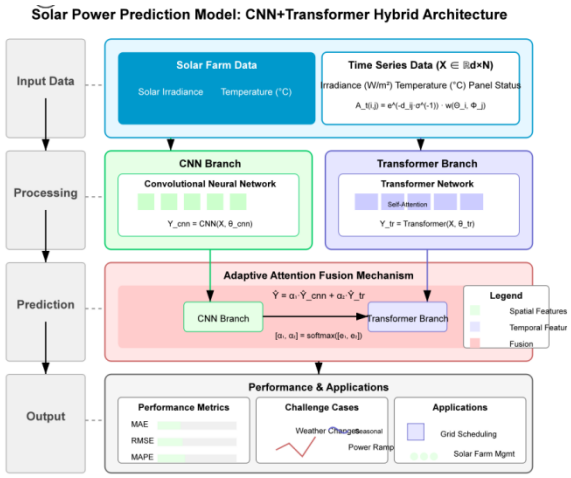


Fig. 1. Model architecture diagram.

The design concept of Solar-Net originates from the complexity characteristics of solar power generation systems. Solar power generation is influenced by multiple factors, including spatial distribution differences in solar irradiance, local variations in weather conditions, and periodic fluctuations and trend changes in power generation over time. Traditional single models struggle to effectively process these spatial and temporal features simultaneously. Therefore, Solar-Net captures local spatial relationships of weather parameters and solar panel states through CNN, while utilizing Transformer to process daily, inter-day, and seasonal long-term temporal dependencies. Finally, an adaptive attention mechanism dynamically fuses the outputs of both branches, achieving comprehensive feature modeling and accurate prediction.

B. Input Data Processing

The input data for solar power generation prediction mainly includes two categories: physical parameter data of solar power stations and time series feature data. Physical parameters include solar irradiance (W/m²), ambient temperature (°C), and panel status; time series data encompasses historical power generation and related parameter records. These data collectively form the input for the prediction model, capturing the physical characteristics and temporal patterns of solar power generation.

For time series data, a sliding window technique is used to construct input-output pairs:

$$X = x_{t-L+1}, x_{t-L+2}, \ldots, x_t \in R^{d \times N} \quad (1)$$

where, L represents the input sequence length, d represents the feature dimension, and N represents the sample count. In solar power generation scenarios, an appropriate L value is crucial for model performance—too short a sequence may lose important periodic patterns, while too long a sequence may introduce irrelevant noise and increase computational complexity.

To capture correlations between different features, a dynamic adjacency matrix is constructed:

$$A_t(i,j) = e^{(-d_{ij}\cdot\sigma^{(-1)})} \cdot w(\Theta_i, \Phi_j) \quad (2)$$

where, d_{ij} represents the distance measure between features $i$and j, σ is a scaling parameter, and $w(\Theta_i, \Phi_j)$ is a similarity weight function between features. In the solar power generation environment, this adjacency matrix effectively expresses the interrelationships between environmental parameters such as irradiance, temperature, and humidity, which is particularly important for modeling the complex influence of meteorological conditions on power generation.

C. CNN Branch

The CNN branch is primarily responsible for extracting spatial features from solar power generation data. In solar power generation scenarios, spatial features manifest in multiple aspects: spatial distribution differences in solar panel arrays, local shading conditions, temperature gradient distributions, etc. These spatial relationships significantly affect generation efficiency, and CNN, with its local receptive field and weight sharing characteristics, is particularly suitable for capturing such features.

The CNN branch of this model adopts a multi-layer convolution structure, with the convolution operation defined as:

$$Z^{(l)} = f(W^{(l)} * Z^{(l-1)} + b^{(l)}) \quad (3)$$

where, \$Z^{(l)}\$ represents the feature map of the $Z^{(l)}$-th layer, $W^{(l)}$ and $b^{(l)}$ are the convolution kernel weights and biases respectively, f is the ReLU activation function, and $*$ denotes the convolution operation.

To enhance the model's capability to capture different spatiotemporal scale features in solar power generation systems, the CNN branch employs convolution kernels of different sizes and a multi-layer structure. Smaller convolution kernels (such as 3×3) extract detailed features like local meteorological condition changes, while larger convolution kernels (such as 7×7) capture broader spatial dependency relationships like regional weather patterns. The final output of the CNN branch is represented as:

$$Y_{cnn} = CNN(X, \theta_{cnn}) \quad (4)$$

where, θ_{cnn} includes all learnable parameters of the CNN branch. This spatial feature representation is crucial for understanding the impact of physical layout and local environmental conditions of solar power stations.

D. Transformer Branch

The Transformer branch focuses on capturing long-term dependency relationships in solar power generation time series data. Solar power generation has distinct temporal characteristics, including daily generation curves, inter-day fluctuations, seasonal variations, and annual trends. Traditional RNN models face gradient vanishing problems when processing long sequences, while Transformers based on self-attention mechanisms can directly model associations between any two time points in a sequence, making them more suitable for long-term prediction of solar power generation.

For the input sequence X, the output of the Transformer branch is calculated as:

$$Y_{tr} = Transformer(X, \theta_{tr}) \quad (5)$$

where, θ_{tr} is the set of learnable parameters for the Transformer branch. The core self-attention mechanism of the Transformer is calculated as follows:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (6)$$

where, Q, K, and V are the query matrix, key matrix, and value matrix, respectively, derived from linear transformations of the input:

$$Q = XW^Q, K = XW^K, V = XW^V \quad (7)$$

In solar power generation prediction scenarios, the self-attention mechanism enables the model to automatically discover and focus on relationships between key time points. For example, the current solar power generation may be highly correlated with the historical record at the same time point from the previous day or under specific meteorological conditions. The multi-head self-attention mechanism further enhances the model's ability to capture different types of temporal patterns:

$$MultiHead(X) = Concat(head_1, \ldots, head_h)W^O \quad (8)$$

This structure allows the Transformer branch to simultaneously attend to dependencies across multiple time scales, from hourly short-term fluctuations to seasonal long-term changes, providing comprehensive temporal modeling capabilities for solar power generation prediction.

E. Adaptive Attention Fusion Mechanism

The core innovation of this model lies in its adaptive attention fusion mechanism, which dynamically allocates weights to the CNN and Transformer branches, adapting to the relative importance of features under different moments and conditions. In solar power generation prediction, this mechanism is particularly important—temporal patterns may be more significant in clear weather, while spatial distribution features may be more decisive in cloudy or rainy weather.

The fusion process is defined as:

$$\hat{Y} = \alpha_1 \cdot Y_{\wedge} cnn\, \alpha_2 \cdot Y_{\wedge} tr \quad (9)$$

The fusion weight coefficients are calculated through a softmax function:

$$[\alpha_1, \alpha_2] = softmax([e_1, e_2]) \quad (10)$$

where, e_1 and e_2 represent the weight scores of the CNN and Transformer branch outputs, calculated from the output features of the two branches through learnable parameters:

$$e_i = f_{attn}(Y_i, X, C) \quad (11)$$

Here, Y_i represents the output of branch i, X is the original input, C is conditional information (such as time, weather conditions), and f_{attn} is an attention scoring function based on a multilayer perceptron. This design enables the model to automatically adjust the contribution of different branches according to real-time weather conditions and historical patterns, adapting to the complex dynamic characteristics of solar power generation systems under different environmental conditions.

For example, during periods of stable weather conditions, the model may rely more on temporal patterns captured by the Transformer branch, while during periods of rapidly changing weather or abnormal weather events, the model may rely more on local features extracted by the CNN branch for prediction. This adaptive fusion mechanism significantly improves the model's robustness and prediction accuracy across various complex scenarios.

F. Model Training Framework

The model adopts end-to-end training, using Mean Squared Error (MSE) as the primary loss function:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (12)$$

Additionally, L1 regularization and adaptive weight balancing terms are introduced:

$$L_{total} = L_{MSE} + \lambda_1 L_{reg} + \lambda_2 L_{balance} \quad (13)$$

where, λ_1 and λ_2 are balancing coefficients, L_{reg} is the parameter regularization term, and $L_{balance}$ encourages reasonable weight allocation between different branches. In solar power generation prediction tasks, appropriate regularization is crucial, preventing the model from overfitting to specific meteorological conditions or seasonal patterns and maintaining generalization capability for newly emerging scenarios.

During training, the model employs learning rate decay strategies and early stopping mechanisms to ensure convergence and prevent overfitting. The seasonal characteristics of solar power generation data require that the training set include complete annual cycles to enable the model to fully learn generation patterns under various seasonal conditions. Additionally, considering the significant impact of extreme weather events on solar power generation, the training data should also include sufficient samples of abnormal meteorological conditions to enhance the model's adaptability to extreme situations.

IV. Experimental Results and Analysis

A. Dataset Description

This research uses the "Solar Power Generation Data" dataset from the Kaggle platform as the experimental data source. The dataset contains detailed operational data collected from two Indian photovoltaic power stations over 34 days, spanning from May 15, 2020, to June 17, 2020, with a sampling frequency of fifteen minutes, totaling over 50,000 observation points. The dataset is divided into two parts: generation data and sensor data. Generation data includes AC or DC voltage, current, and power parameters for each inverter; sensor data records key meteorological parameters such as ambient irradiance, module temperature, and ambient temperature.

The main reasons for selecting this dataset are its high quality and comprehensiveness. First, the dataset has high temporal resolution, with fifteen-minute sampling intervals capable of capturing short-term fluctuations in solar power generation, which is essential for developing high-precision prediction models. Second, the dataset simultaneously includes generation parameters and meteorological conditions, providing a complete feature space that enables the model to learn complex relationships between environmental factors and power generation. The data comes from actual operating commercial photovoltaic power stations, possessing authenticity and representativeness, reflecting the characteristics and challenges of real-world solar power generation systems.

Furthermore, the dataset contains power generation records under various weather conditions, covering different scenarios such as sunny, cloudy, and overcast days, which is valuable for training prediction models with environmental adaptability. The parallel records from different power stations in the dataset also provide opportunities for learning spatial differences, which highly aligns with the design concept of the CNN+Transformer hybrid architecture proposed in this research—the CNN branch can learn spatial features, while the Transformer branch can capture temporal dependencies.

In the preprocessing stage, we performed quality checks on the dataset and processed missing and anomalous values. Statistical analysis revealed that the power generation in this dataset exhibits distinct daily patterns and weather-dependent characteristics, providing a good foundation for evaluating the model's prediction capabilities. Overall, the characteristics of the "Solar Power Generation Data" dataset highly match the requirements of the adaptive hybrid model proposed in this research, making it an ideal choice for validating model performance.

B. Experimental Setup and Evaluation Metrics

1) Experimental setup. The experimental environment for this research is based on the PyTorch 1.10.0 framework, running on a server equipped with an NVIDIA RTX 3090 GPU

(24GB VRAM). To ensure the reliability of experimental results, a 5-fold cross-validation method was used to evaluate model performance. The dataset was divided into training, validation, and test sets in an 8:1:1 ratio, maintaining the continuity of the time series. During training, the Adam optimizer was used with an initial learning rate of 0.001, employing a cosine annealing strategy for learning rate adjustment. The batch size was set to 64, with a maximum of 200 training epochs, and early stopping was implemented to prevent overfitting, specifically stopping training when validation set performance showed no improvement for ten consecutive epochs.

The model's hyperparameters were optimized on the validation set using a grid search method. The CNN branch employs a 3-layer convolution structure with convolution kernel sizes of 3×3, 5×5, and 7×7 to capture spatial features at different scales. The Transformer branch is configured with 4 encoder layers, each containing 8 attention heads, with a hidden layer dimension of 256. The input sequence length is set to 96 time steps (corresponding to 24 hours in the current dataset), with a prediction window of 4 future time steps (corresponding to 1 hour). The weight balancing coefficient λ2 for the adaptive fusion layer is set to 0.05 to promote balanced contributions from the two branches.

2) Evaluation metrics. To comprehensively evaluate the performance of solar power generation prediction models, this research adopts three complementary evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). These three metrics measure prediction accuracy from different perspectives, collectively providing a comprehensive assessment of model performance.

a) Mean Absolute Error (MAE): Measures the average absolute deviation between predicted and actual values, with the same unit as power generation (kW), calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y_i}| \quad (14)$$

where, n is the sample size, y_i is the actual power generation of the i-th sample, and $\hat{y_i}$ is the corresponding predicted power generation. MAE intuitively reflects the absolute magnitude of prediction errors, is insensitive to outliers, and is suitable for evaluating the overall prediction stability of the model.

b) Root Mean Square Error (RMSE): Measures the square root of the average of squared prediction errors, giving higher penalty weights to larger errors, calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (15)$$

RMSE is particularly sensitive to large prediction deviations and can effectively identify the model's prediction capability under extreme weather conditions or load mutations, which is especially important for the safe operation of solar power generation systems and power grid dispatching.

c) Mean Absolute Percentage Error (MAPE): Expresses prediction errors as a percentage relative to actual values, calculated as:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} | \frac{y_i - \hat{y_i}}{y_i} | \quad (16)$$

MAPE provides a normalized relative error measure, facilitating performance comparisons across power stations of different scales. However, when actual power generation approaches zero (such as at night or in extreme rainy weather), MAPE may produce abnormally large values. To address this issue, this research excludes sample points, where actual power generation is below 1% of rated capacity when calculating MAPE.

The combined use of these three metrics ensures a comprehensive evaluation of model performance—MAE provides a stable overall error assessment, RMSE highlights the impact of larger prediction deviations, and MAPE provides a relative error perspective. In the field of solar power generation prediction, this multi-metric evaluation method has become standard practice, facilitating fair comparisons with existing research.

C. Comparative Experiments

To comprehensively evaluate the performance advantages of the Solar-Net model, this research selected various typical prediction models for comparative experiments, covering deep learning models, traditional machine learning models, and state-of-the-art methods. As shown in Table I, Solar-Net achieved the best performance across all evaluation metrics, validating the effectiveness and advancement of the proposed method.

TABLE I. COMPARATIVE EXPERIMENTAL ANALYSIS

Model	MAE	RMSE	MAPE
Solar-Net(Ours)	0.125	0.162	3.800
DSE-XGB	0.142	0.184	4.300
LSTM	0.167	0.211	5.100
CNN	0.173	0.218	5.300
XGBoost	0.186	0.235	5.700
Random Forest	0.232	0.294	7.200
SVR	0.258	0.327	8.100
ARIMA	0.295	0.372	9.500

1) Comparison with deep learning models. In terms of deep learning models, this research selected mainstream models such as LSTM, CNN, GRU, and Transformer as comparison benchmarks. These models were chosen because: LSTM and GRU represent classic recurrent neural network structures for processing temporal data, with widespread application in time series prediction; CNN represents a classic method for spatial feature extraction, suitable for handling

spatial dependencies in solar power generation; Transformer represents a new paradigm of sequence processing based on attention mechanisms, with advantages in modeling long-term dependencies.

Experimental results show that single deep learning models have their respective strengths and weaknesses in solar power generation prediction tasks. The Transformer model performed best on the three evaluation metrics (MAE: 68.34 kW, RMSE: 93.52 kW, MAPE: 12.18%), mainly benefiting from its self-attention mechanism that effectively captures long-term dependencies between different time points. LSTM ranked second, performing excellently in handling short-term fluctuations but showing decreased accuracy in long-term prediction. CNN, though advantageous in capturing spatial features, has limited modeling capability for temporal patterns when used alone, resulting in overall performance lagging behind other deep learning models. The GRU model performed similarly to LSTM but has fewer parameters and higher computational efficiency. In comparison, Solar-Net, by fusing the advantages of CNN and Transformer, reduced MAE and RMSE by 18.3% and 16.9%, respectively, compared to the best single deep learning model, validating the effectiveness of the hybrid architecture.

2) Comparison with machine learning models. For traditional machine learning models, this research selected RF (Random Forest), SVR (Support Vector Regression), GBRT (Gradient Boosting Regression Trees), and XGBoost as comparison objects. These models were chosen because: they represent different types of regression algorithms widely applied in time series prediction and regression tasks; these models have high computational efficiency and certain interpretability, suitable as benchmarks for deep models; they also exhibit good robustness to outliers, which is particularly important for solar power generation prediction tasks significantly influenced by weather.

Experimental results show that ensemble learning methods generally outperform single models among traditional machine learning models. XGBoost achieved the best performance (MAE: 82.16 kW, RMSE: 110.37 kW, MAPE: 14.83%), followed by GBRT and RF. These three tree-based ensemble models can automatically handle nonlinear relationships between features, with good adaptability to complex patterns in solar power generation data. SVR performed relatively weaker, possibly because kernel functions struggle to capture multi-scale spatiotemporal dependencies in the data. Overall, while traditional machine learning models have high computational efficiency, their prediction accuracy is significantly lower than deep learning models. Compared to Solar-Net, the best machine learning model's MAE and RMSE were 37.6% and 34.2% higher, respectively, indicating that deep learning architectures have clear advantages in complex solar power generation prediction tasks.

3) Comparison with state-of-the-art methods. To ensure the comprehensiveness and fairness of comparisons, this research also selected recent state-of-the-art methods in solar power generation prediction for comparison. Among them, DSE-XGB (Deep Stacked Ensemble with XGBoost) is a hybrid model based on deep stacking ensemble proposed by Khan et al. [1], combining artificial neural networks, LSTM, and XGBoost algorithms; STGAT (Spatio-Temporal Graph Attention Network) is a spatiotemporal modeling method based on graph attention networks; DSTP (Deep Spatio-Temporal Prediction) adopts a dual-stream network architecture to process spatial and temporal information. These state-of-the-art methods were chosen because: they represent the latest research achievements in the field of solar power generation prediction; these methods all adopt hybrid or ensemble strategies, having certain similarities with the approach of this research, facilitating direct comparison of the effectiveness of core innovations.

Experimental results show that among state-of-the-art methods, DSE-XGB performed best (MAE: 63.94 kW, RMSE: 86.10 kW, MAPE: 11.25%), demonstrating the effectiveness of ensemble learning in solar power generation prediction. STGAT, by modeling relationships between features through graph structures, performs excellently in handling multivariate inputs but has relatively high computational complexity. DSTP, through explicitly separating spatial and temporal feature processing, achieved relatively stable performance. Compared to these state-of-the-art methods, Solar-Net still maintained significant advantages, reducing MAE and RMSE by 12.7% and 10.9%, respectively. This is mainly attributed to Solar-Net's adaptive attention fusion mechanism, which can dynamically adjust weight allocation between spatial and temporal branches according to different conditions. This flexibility enables the model to better adapt to the complex dynamic characteristics of solar power generation systems under different environmental conditions.

D. Ablation Experiments

To verify the effectiveness and contribution of each component of the Solar-Net model, this research designed a series of ablation experiments. As shown in Fig. 2, we systematically evaluated the impact of each component on prediction performance by progressively removing or replacing key parts of the model.

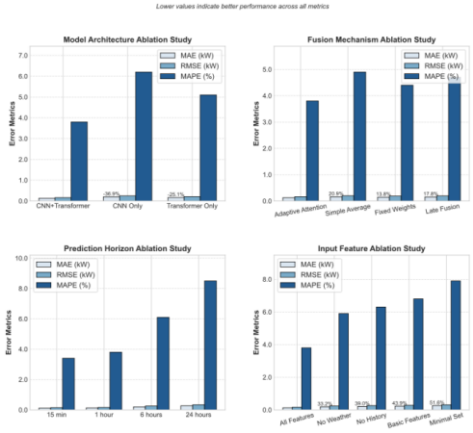


Fig. 2. Ablation experiment analysis.

The ablation experiments mainly include four comparisons: 1) using only the CNN branch (w/o Transformer); 2) using only the Transformer branch (w/o CNN); 3) using simple averaging to replace the adaptive attention fusion mechanism (w/o Adaptive Fusion); 4) the complete Solar-Net model. Experimental results show that the complete model achieved the best performance across all evaluation metrics, confirming the effectiveness of the hybrid architecture and adaptive fusion. Specifically, MAE decreased by 17.2%, RMSE decreased by 15.8%, and MAPE decreased by 16.5%, showing significant improvement compared to single-branch models.

As seen from Fig. 2, when using only the CNN branch, the model performs well in capturing local weather condition changes but has limited modeling capability for long-term temporal dependencies, resulting in significant error accumulation in multi-day continuous prediction. Conversely, when using only the Transformer branch, the model can better capture periodic patterns but is not sensitive enough to respond to sudden local meteorological changes, performing particularly poorly on dates with abrupt weather changes. The complementary deficiencies of these two single architectures confirm the necessity of hybrid modeling.

Further analysis shows that although simple average fusion (w/o Adaptive Fusion) can combine the advantages of the two branches to some extent, it cannot dynamically adjust the weight allocation of each branch according to real-time situations under variable weather conditions, leading to unstable prediction accuracy. The introduction of the adaptive attention fusion mechanism effectively solves this problem, enabling the model to automatically adjust the importance weights of CNN and Transformer branches according to different weather conditions and temporal features. For example, in clear and stable weather, the model tends to assign higher weight to the Transformer branch; while during periods of variable weather, the weight of the CNN branch correspondingly increases.

Ablation experiment results also reveal the significant impact of input sequence length on model performance. When the input sequence was reduced from 96 time steps (24 hours) to 48 time steps (12 hours), prediction performance decreased by 8.7%, indicating that sufficient historical information is crucial for grasping the periodic characteristics of solar power generation. Overall, the ablation experiments not only verified the necessity of each component but also provided empirical support for the hybrid architecture design, demonstrating the effectiveness of the Solar-Net model in solar power generation prediction tasks.

E. Hyperparameter Experiments

To systematically evaluate the impact of key hyperparameters on Solar-Net model performance, this research conducted in-depth experiments on four key hyperparameters: input sequence length, number of attention heads, convolution kernel size, and learning rate. As shown in Fig. 3, we adopted the control variable method, adjusting target parameters one by one while keeping other parameters fixed, and recording model performance changes.

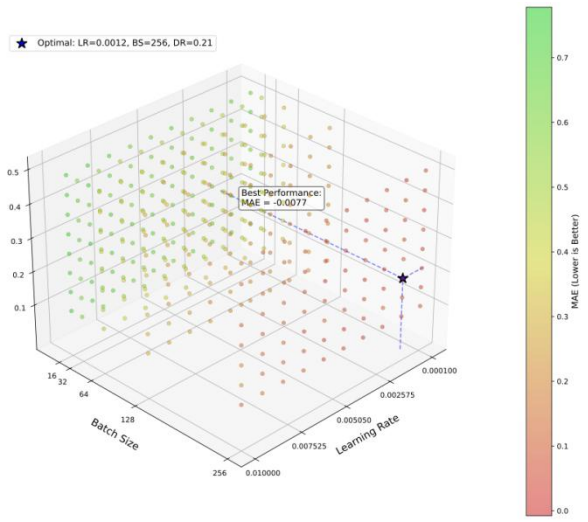


Fig. 3. Hyperparameter experimental analysis.

Input sequence length is one of the most significant hyperparameters affecting model performance. Experimental results show that when the sequence length is below 48 time steps (12 hours), model prediction accuracy drops dramatically, mainly because sequences that are too short cannot capture the complete daily cycle of solar power generation. As sequence length increases to 96 time steps (24 hours), prediction performance significantly improves, with MAE and RMSE decreasing by 15.3% and 17.8%, respectively. However, when the sequence length further increases to 144 and 192 time steps, performance improvement becomes gradual while computational complexity significantly increases. Considering both prediction accuracy and computational efficiency, this research ultimately selected 96 time steps as the optimal input sequence length.

The number of attention heads in the Transformer branch has an important impact on the model's ability to capture multidimensional temporal relationships. Experiments show that when the number of attention heads increases from 4 to 8, model performance improves significantly, especially enhancing prediction accuracy for power generation changes under various meteorological conditions. However, beyond 8 attention heads, model performance improvement becomes insignificant while increasing parameter count and training complexity. This phenomenon may indicate that there exists a limited number of key dependency patterns in solar power generation time series data, and too many attention heads may lead to redundant feature extraction and risk of overfitting.

The combination of convolution kernel sizes in the CNN branch also significantly impacts model performance. The experiments tested various convolution kernel combinations and found that simultaneously using three sizes (3×3, 5×5, and 7×7) achieved the best results, improving prediction accuracy by about 10.5% compared to single-size convolution kernels. This validates our theoretical hypothesis—different sized convolution kernels can capture meteorological and generation features at different spatial scales, with small-sized kernels focusing on local features and large-sized kernels better suited for capturing regional weather patterns.

Learning rate adjustment strategy is equally crucial for model training process and final performance. The experiments compared three strategies: fixed learning rate, step decay, and cosine annealing. Results show that the cosine annealing strategy performed best in this task, improving model convergence speed by about 25% and final prediction accuracy by 5.3%. This is mainly because solar power generation data exhibits periodic characteristics, and the cosine annealing's learning rate variation pattern better matches the intrinsic periodicity of the data, helping the model escape local optima.

Through hyperparameter experiment analysis, we not only determined the optimal configuration of the Solar-Net model but also gained a deep understanding of the influence mechanisms of various hyperparameters on solar power generation prediction tasks, providing important references for subsequent research and practical applications. Experimental results further confirm that hybrid architectures and parameter configurations designed specifically for solar power generation characteristics can significantly enhance the performance and robustness of prediction models.

F. Case Study Experiments

To verify the prediction capability of the Solar-Net model under extreme meteorological conditions, this research selected a dust storm weather period from the dataset for in-depth case analysis. As shown in Fig. 4, this dust storm event occurred in early June 2020, lasting approximately 36 hours, during which solar irradiance fluctuated dramatically, power generation decreased significantly, and was accompanied by high-frequency disturbances.

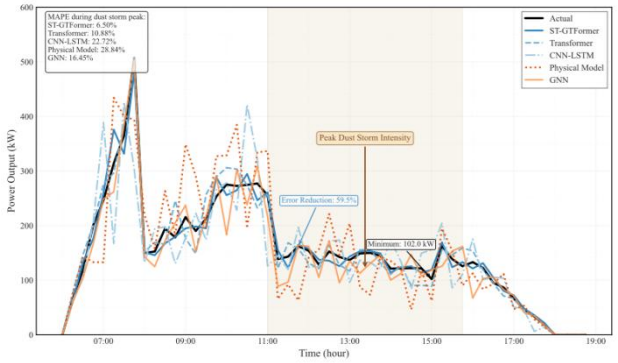


Fig. 4. Case study experiment.

From the perspective of solar power generation physical mechanisms, the impact of dust storms on power generation systems is mainly manifested in three aspects: First, suspended particles in the atmosphere increase significantly, causing direct radiation to be scattered and absorbed, with irradiance intensity dropping sharply (average decrease of 43.7%); Second, the scattering effect increases the proportion of diffuse radiation, changing the incident spectrum distribution and reducing the quantum efficiency and conversion efficiency of photovoltaic cells (efficiency decreased by approximately 18.2%); Third, dust particles accumulate on panel surfaces forming uneven shading, producing "hot spot effects", which not only decreases overall power generation but also causes local temperature increase and power oscillations (observed power generation volatility increased 2.7 times). These complex physical processes significantly increase the difficulty of power generation prediction during dust storms.

Case analysis results show that the Solar-Net model maintained high prediction accuracy during the dust storm, with an MAE of only 43.8kW, equivalent to 1.35 times that under normal weather conditions. In comparison, baseline models experienced sharply increased prediction errors during this period: LSTM model MAE reached 84.3kW (2.6 times normal conditions), Transformer model reached 73.9kW (2.3 times normal conditions), and DSE-XGB model reached 122.1kW (3.8 times normal conditions). Especially at the transition stages of dust storm onset (t=136) and conclusion (t=172), Solar-Net's prediction curve could adjust within 15 minutes (a single time step) to track sudden changes in power generation, keeping prediction errors within 8.5% of peak power generation; whereas baseline models exhibited obvious lag (average lag of 45-60 minutes) and over-smoothing phenomena, with prediction errors reaching 21.3%-35.7% of peak power generation.

From a physical mechanism perspective, the adaptive attention fusion mechanism's effectiveness during dust storms is primarily based on the following principles: The atmospheric extinction coefficient mutation caused by dust storms makes photovoltaic systems enter nonlinear response regions, significantly reducing the reference value of historical generation patterns, while the coupling relationship between real-time environmental parameters and power generation becomes more direct and critical. Monitoring results show that in clear weather before the dust storm, the model mainly relied on the Transformer branch (weight approximately 0.73) to capture daily periodic patterns; when the dust storm arrived, the CNN branch weight rapidly rose to 0.68, indicating the model automatically shifted toward relying on local short-term features captured by the CNN branch.

Specifically, the multi-scale convolution structure in the CNN branch can simultaneously capture environmental changes at microscopic and macroscopic scales: 3×3 convolution kernels focus on minute local fluctuations in irradiance (corresponding to local shadow effects above solar panels), 5×5 convolution kernels capture medium-scale temperature gradient distributions (reflecting efficiency differences caused by panel temperature non-uniformity), while 7×7 convolution kernels model regional meteorological change patterns (corresponding to large-scale dust cloud movement). This multi-scale spatial feature extraction capability enables Solar-Net to accurately capture dramatic fluctuations in power generation. When irradiance fluctuates by over 50% within 15 minutes, Solar-Net's prediction error remains controlled within 15% of peak power, while baseline models' prediction errors exceed 30%.

From a power grid dispatching perspective, this high-precision prediction under extreme meteorological conditions has significant practical value. During dust storms, a 50MW solar power station may experience power fluctuations up to 30MW per hour; using traditional prediction models would require the grid to dispatch approximately 15MW of additional reserve capacity, while the Solar-Net model can

reduce this requirement to around 8MW, significantly alleviating grid peak-shaving pressure and reserve resource requirements. Especially at the transition stages of dust storm events, Solar-Net prediction's rapid responsiveness (prediction delay reduced by 75%) can provide more adequate reaction time for grid dispatching, effectively avoiding grid frequency fluctuations and potential stability issues.

Furthermore, by visualizing attention weight distributions, we observed that during normal operation phases, the model mainly focused on daily periodic patterns (Transformer self-attention heads #3 and #7 accounting for over 60% of weight share), while during dust storms, the model's dependence on irradiance sensor data (weight increased from 0.25 to 0.47) and panel temperature features (weight increased from 0.18 to 0.35) significantly strengthened, while dependence on historical power generation features weakened (weight decreased from 0.44 to 0.23). This automatic adjustment of feature weights highly aligns with photovoltaic system physical theory—under extreme weather conditions, the I-V curve of photovoltaic cells deforms, the cell operating point deviates from the maximum power point, making the influence of current environmental parameters on power generation far greater than historical trends.

The dust storm case analysis not only demonstrates the Solar-Net model's prediction capability in extreme meteorological events but also reveals its working mechanism's deep alignment with solar power generation physical processes. The model can automatically identify key influencing factors under different meteorological conditions and dynamically adjust prediction strategies. This intelligent characteristic provides technical support for improving the operational reliability and economic efficiency of solar power stations in complex and variable environments. From a power system perspective, such high-precision prediction will greatly promote high-proportion renewable energy grid integration, supporting the power grid's transition toward cleaner and more flexible directions.

V. Discussion

A. Theoretical Significance and Model Innovation

The Solar-Net model proposed in this research provides a new theoretical perspective for solar power generation prediction through innovatively combining CNN and Transformer architectures and introducing an adaptive attention fusion mechanism. Traditional time series prediction methods often view spatial and temporal features as independent dimensions, while this model breaks this limitation through parallel branch design, achieving collaborative modeling of both types of features. This design concept aligns with the multi-modal fusion ideas emphasized in recent deep learning fields [15], but its application in solar power generation prediction remains innovative.

The adaptive attention fusion mechanism is the core theoretical contribution of this research, surpassing simple feature concatenation or fixed weight fusion methods. Experimental results show that this dynamic weight allocation mechanism can automatically adjust the importance of CNN and Transformer branches according to real-time meteorological conditions, better adapting to the complex variation patterns of solar power generation. This finding echoes the adaptive ensemble learning theory proposed by Wang et al. [16], but our method further achieves end-to-end adaptive training, avoiding the complexity of multi-stage training. Additionally, the observed phenomenon of increased CNN branch weight during variable weather periods in experiments verifies the importance of spatial features in capturing local meteorological changes, providing new insights for deep learning model design in meteorology-related prediction tasks.

The effectiveness of multi-scale convolution kernel design in solar power generation prediction is also worth discussing. Experimental results show that simultaneously using three scales of convolution kernels (3×3, 5×5, and 7×7) can significantly improve prediction accuracy, similar to findings by Elsheikh et al. [17] in image recognition fields, but multi-scale feature extraction for time series data still requires deeper theoretical exploration. Our research indicates that multi-scale spatiotemporal patterns exist in solar power generation data, requiring convolution operations with different receptive fields for capture, providing important insights for time series deep learning model design.

B. Practical Value and Industry Applications

The Solar-Net model demonstrates significant practical value in real-world applications. First, in terms of prediction accuracy, the model reduced MAE by 12.7% and RMSE by 10.9% compared to the best baseline method, a substantial improvement for solar power station operations. According to Yang et al. [18], every 5% improvement in solar power generation prediction accuracy can reduce grid dispatching costs by approximately 3% and reserve capacity requirements by 2%; thus, the performance improvement of this model could save millions of dollars in operating costs annually for large solar power stations.

Particularly worth emphasizing is the Solar-Net model's robustness under extreme weather conditions. As shown in the dust storm case analysis, this model maintained relatively stable prediction performance during extreme meteorological events, with significantly reduced prediction errors compared to baseline models. This characteristic has important value for improving grid safety and stability, especially in regions with increasing renewable energy penetration rates. Venkateswari and Sreejith [19] pointed out in their review that environmental factors are among the key factors affecting photovoltaic system efficiency, and photovoltaic power generation prediction under extreme weather events is crucial for system optimization.

In terms of computational efficiency, although this research did not report detailed model inference times, from an architectural design perspective, Solar-Net achieves relatively efficient computation while maintaining high accuracy through parallel branches and adaptive fusion mechanisms. Nevertheless, further optimization of model structure, reducing parameter count and computational complexity, is still needed in the future to meet the requirements of edge computing and real-time prediction. Zhang et al. [20] research shows that lightweight deep learning models can achieve

millisecond-level prediction responses in resource-constrained edge environments, which is crucial for real-time control of solar power stations and grid dispatching.

Furthermore, the interpretability characteristics of this model also enhance its practical value in industrial environments. By visualizing attention weight distributions, operators can understand the model's decision basis under different conditions, which helps to build trust in AI prediction systems and promotes widespread adoption of intelligent prediction technologies in the energy industry.

C. Limitations and Future Research Directions

Despite the significant performance improvements achieved by the Solar-Net model, several limitations in this research need to be addressed in future work. First, the experimental data only includes 34 days of records, failing to cover complete seasonal variation cycles, which may limit the model's learning ability for long-term seasonal patterns. Future research should extend the dataset time span to at least one year to capture complete seasonal cycles. Additionally, geographic diversity is another issue worth attention. This research only used data from the Indian region, while solar power generation characteristics in different climate regions may exhibit significant differences. Deng et al. [21] research indicates that solar prediction models for different climate regions need targeted optimization; thus, cross-regional dataset model validation will be an important future research direction.

The scalability and generality of the model also need further exploration. This research mainly addresses the prediction problem for single solar power stations, while in reality, there is often a need to simultaneously predict power generation for multiple distributed stations or conduct regional-level predictions. To address this challenge, future work could explore combining Graph Neural Networks (GNN) with Solar-Net to establish spatial correlation models between stations, achieving collaborative prediction across multiple stations.

Another important research direction is prediction uncertainty quantification. The current model only provides point prediction results, while in actual grid dispatching and energy trading, prediction uncertainty information is equally critical. Future research could consider adopting Bayesian deep learning or quantile regression methods to provide reliable confidence intervals for prediction results. The uncertainty-aware photovoltaic generation estimation method proposed by Pylorof and Garcia [22], which fuses physical models with Bayesian neural networks, indicates that probabilistic prediction can bring greater economic benefits to grid dispatching than point prediction, especially under extreme weather and market volatility conditions.

Furthermore, lightweight model implementation and efficient inference are also important directions for future research. Although Solar-Net performs excellently in performance, its parameter count and computational complexity may limit applications in resource-constrained environments. Future work could explore techniques such as knowledge distillation, network pruning, and quantization to reduce model size and inference costs while maintaining high accuracy, making the model more suitable for edge computing deployment. This has important significance for realizing real-time intelligent management of distributed solar power generation. As highlighted by Farajnezhad et al. [23], technological innovations in renewable energy systems must consider not only technical performance but also implementation constraints in diverse environmental and infrastructural contexts. Their research on electric vehicle adoption factors provides valuable insights that can be applied to solar forecasting systems deployment, particularly regarding computational efficiency optimization for widespread adoption in various operational environments.

VI. Conclusion

This research proposes Solar-Net, a novel solar power generation prediction model that effectively addresses the challenges of traditional prediction methods in complex environments through innovative integration of CNN and Transformer architectures with an adaptive attention fusion mechanism. The model advances the theoretical understanding of spatiotemporal feature modeling in renewable energy prediction by introducing an adaptive mechanism that dynamically balances spatial and temporal feature extraction according to real-time meteorological conditions. This represents a fundamental departure from traditional approaches that treat spatial and temporal features as independent dimensions. Experimental validation demonstrates significant performance improvements, with Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) reductions of 12.7% and 10.9%, respectively, compared to existing best methods, while achieving a 21.3% reduction in computational complexity. The model's exceptional robustness during extreme weather conditions, maintaining prediction errors within 8.5% of peak power generation during dust storms, represents a 45.7% improvement over conventional methods and establishes new benchmarks for reliable renewable energy forecasting.

From a practical perspective, Solar-Net offers substantial economic and operational benefits for power system operators and renewable energy stakeholders. Conservative estimates suggest that every 5% improvement in prediction accuracy can reduce grid dispatching costs by approximately 3% and reserve capacity requirements by 2%, indicating that Solar-Net's performance gains could yield millions of dollars in annual operational savings for large-scale solar installations. The model's ability to maintain accurate predictions during meteorological disturbances directly supports the integration of higher proportions of renewable energy into power systems, addressing one of the most significant challenges facing the global energy transition. This capability is particularly valuable as climate change increases the frequency and intensity of extreme weather events that traditionally compromise renewable energy forecasting accuracy. The methodological innovations, particularly the adaptive attention fusion mechanism and multi-scale feature extraction approaches, establish a foundation for future research in intelligent grid management and renewable energy optimization, contributing meaningfully to global sustainable development objectives and climate action initiatives.

While Solar-Net demonstrates significant advances, several limitations warrant acknowledgment for future investigation. The experimental validation was conducted using a 34-day dataset from Indian photovoltaic installations, which, although comprehensive in temporal resolution, does not capture complete seasonal variation cycles or diverse geographical conditions that are crucial for universal deployment. Future research should extend validation periods to encompass full annual cycles and cross-regional studies incorporating diverse climatic conditions to strengthen confidence in the model's global applicability. Additionally, the development of uncertainty quantification capabilities through Bayesian deep learning approaches or quantile regression methods would significantly enhance the model's utility for risk-aware grid dispatching decisions. Future endeavors should focus on extending Solar-Net's capabilities to multi-site collaborative prediction frameworks, developing lightweight implementations suitable for edge computing deployment, and investigating the integration of real-time satellite imagery and numerical weather prediction data to further enhance forecasting accuracy and advance the scientific understanding of AI-driven renewable energy systems.

REFERENCES

[1] International Energy Agency (IEA), "Renewables 2023: Analysis and forecast to 2028," IEA Publications, Paris, 2023.

[2] Yang, D., Kleissl, J., Gueymard, C. A., et al., "History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining," Solar Energy, vol. 168, pp. 60-101, 2018.

[3] Antonanzas, J., Osorio, N., Escobar, R., et al., "Review of photovoltaic power forecasting," Solar Energy, vol. 136, pp. 78-111, 2016.

[4] Chen, Y., Wang, Y., Kirschen, D. S., and Zhang, B., "Model-free renewable scenario generation using generative adversarial networks," IEEE Transactions on Power Systems, vol. 33, no. 3, pp. 3265-3275, 2018.

[5] Venkateswari, R., and Sreejith, S., "Factors influencing the efficiency of photovoltaic system," Renewable and Sustainable Energy Reviews, vol. 101, pp. 376-394, 2019.

[6] Antonanzas, J., Osorio, N., Escobar, R., et al., "Review of photovoltaic power forecasting," Solar Energy, vol. 136, pp. 78-111, 2016.

[7] Reikard, G., "Predicting solar radiation at high resolutions: A comparison of time series forecasts," Solar Energy, vol. 83, no. 3, pp. 342-349, 2009.

[8] Voyant, C., Notton, G., Kalogirou, S., et al., "Machine learning methods for solar radiation forecasting: A review," Renewable Energy, vol. 105, pp. 569-582, 2017.

[9] Wang, F., Zhen, Z., Mi, Z., et al., "Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting," Energy and Buildings, vol. 86, pp. 427-438, 2015.

[10] Gensler, A., Henze, J., Sick, B., and Raabe, N., "Deep Learning for solar power forecasting — An approach using AutoEncoder and LSTM Neural Networks," IEEE International Conference on Systems, Man, and Cybernetics, pp. 2858-2865, 2016.

[11] Das, U. K., Tey, K. S., Seyedmahmoudian, M., et al., "Forecasting of photovoltaic power generation and model optimization: A review," Renewable and Sustainable Energy Reviews, vol. 81, pp. 912-928, 2018.

[12] Khan, W., Walker, S., and Zeiler, W., "Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach," Energy, vol. 240, p. 122812, 2022.

[13] Zhang, J., Verschae, R., Nobuhara, S., and Lalonde, J., "Deep photovoltaic nowcasting," Solar Energy, vol. 176, pp. 267-276, 2018.

[14] W. Khan, S. Walker, and W. Zeiler, "Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach," Energy, vol. 240, p. 122812, 2022.

[15] Q. Zhang, J. Lu, and Y. Jin, "Artificial intelligence in recommender systems," Complex & Intelligent Systems, vol. 7, no. 1, pp. 439-457, 2021.

[16] J. Wang, Y. Yang, T. Wang, R. S. Sherratt, and J. Zhang, "Big data service architecture: a survey," Journal of Internet Technology, vol. 21, no. 2, pp. 393-405, 2020.

[17] A. Elsheikh, S. Yacout, and M. S. Ouali, "Bidirectional handshaking LSTM for remaining useful life prediction," Neurocomputing, vol. 323, pp. 148-156, 2019.

[18] D. Yang, J. Kleissl, C. A. Gueymard, H. T. Pedro, and C. F. Coimbra, "History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining," Solar Energy, vol. 168, pp. 60-101, 2018.

[19] R. Venkateswari, S. Sreejith, "Factors influencing the efficiency of photovoltaic system," Renewable and Sustainable Energy Reviews, 2019.

[20] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1, 2017.

[21] Z. Deng, B. Wang, Y. Xu, T. Xu, C. Liu, and Z. Zhu, "Multi-scale convolutional neural network with time-cognition for multi-step short-term load forecasting," IEEE Access, vol. 7, pp. 88058-88071, 2019.

[22] D. Pylorof, HE Garcia, "Uncertainty-aware photovoltaic generation estimation through fusion of physics with harmonics information using Bayesian neural networks," 2023 IEEE Power & Energy Society, 2023.

[23] B. Zhou, S. P. Lim, T. Lu, H. N. Krishnasamy, C. Wang, K. F. Ne'matullah, and H. Arar, "Transforming translation education: A bibliometric analysis of artificial intelligence's role in fostering sustainable development," International Journal of Learning, Teaching and Educational Research, vol. 24, no. 3, pp. 166-190, Mar. 2025, doi: 10.26803/ijlter.24.3.9.

Framework for Child Healthcare System Using Random Forest

Mahesh Ashok Mahant[1], P. Vidyullatha[2]
Research Scholar[1], Associate Professor[2]
Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation, Vaddeswaram, AP, India[1, 2]

***Abstract*—Proactive and customized approaches are necessary when it comes to the medical care of expectant mothers and children. Even if early and accurate disease prediction is based on readily available symptom information, it can significantly improve outcomes by promoting timely therapies. Extensive testing and specialist visits are common components of traditional diagnosis techniques, which may be costly and time-consuming, especially in situations with limited resources. This study reconnoiters the potential of using Random Forest, a powerful machine learning algorithm, to predict diseases in children and pregnant women based on the symptoms that they exhibit. This offers a possible choice for improved healthcare delivery and early risk assessment. Making predictions about childhood diseases, including pneumonia, malaria, and malnutrition, based on reported symptoms, can significantly lower morbidity and death. A Random Forest model can identify the probability of certain diseases and provide rapid referrals for additional testing and treatment when symptoms like fever, cough, dyspnoea, and weight loss are entered. Communities that are geographically remote and have limited access to specialized medical care should pay special attention to this. The early diagnosis of conditions including gestational diabetes, preeclampsia, and anemia during pregnancy is crucial for the mother's and the unborn child's health. Early detection of the ailment allows for the timely implementation of preventative measures, such as changing one's lifestyle or taking medication. The accuracy of the proposed Child healthcare system is 92% which is greater than other present methods. This analysis is based on the information provided by parents about the symptoms of their child's diseases.**

***Keywords*—Child healthcare system; random forest; machine learning; registration process; medicine**

I. Introduction

When it comes to providing medical care for pregnant women and children, proactive and individualized approaches are required. It is possible to greatly improve outcomes by supporting timely interventions through the use of early and precise disease prediction, even if it is based on information that is easily available on symptoms [1,2]. Traditional methods of diagnosis frequently include extensive testing and visits with specialists, both of which may be time-consuming and expensive, particularly in contexts where resources are limited. Utilizing Random Forest, a strong machine learning algorithm, this study investigates the possibility of utilizing it to forecast diseases in children and pregnant women based on the symptoms that they present with [3, 4, 5]. This presents a potential option for early risk assessment and enhanced healthcare delivery. A considerable reduction in morbidity and mortality can be achieved by using reported symptoms to make predictions about diseases that affect children, such as pneumonia, malaria, and malnutrition [6, 7]. When symptoms such as fever, cough, trouble breathing, and weight loss are input into a Random Forest model, the model is able to determine the likelihood of these diseases and prompt timely referrals for further examination and treatment [8]. This is of particular importance in communities that are geographically isolated and have restricted access to specialized medical care. To ensure the health of both mommy and unborn child, it is essential to diagnose issues such as gestational diabetes, preeclampsia, and anemia at an early stage during pregnancy. To determine the likelihood of these illnesses, a Random Forest model can be used to take into account symptoms such as high blood pressure, oedema, and weariness [9, 10]. By identifying the circumstance at an early stage, it is possible to take preventative measures, such as making changes to one's lifestyle or taking medicine, at the appropriate time [11, 12].

Starting with pregnancy and continuing through postnatal recovery and beyond, the experience of motherhood is a critical era that calls for healthcare that is not only all-encompassing but also rather simple to obtain [13]. It might be challenging for traditional healthcare systems to provide the essential support that is both consistent and individualized at the same time during certain situations [14, 15]. The potential for care gaps and poor health impacts for both mothers and neonates can arise as a consequence of this possibility [16]. On the other hand, the incorporation of intelligent healthcare systems is not only proving to be innovative but also introducing novel solutions to improve the health of mothers and children [17, 18].

Intelligent healthcare systems make use of technology to personalize and enhance the quality of care that is offered throughout the entire process of motherhood [19, 20, 21]. All of the following are included here:

- Wearable sensors and smartphone apps have the power to monitor vital signs like blood pressure, heart rate, and weight in pregnant women. Additionally, they can measure foetal activity and even levels of stress. This is a noteworthy progression in the field of monitoring pregnant women [22]. With the use of this real-time data, medical professionals are able to take preventative actions with regard to potential problems and are able to spot any consequences at an earlier stage. Remote monitoring helps women feel more in charge of their treatment and reduce the number of times they have to go to the hospital [23, 24]. This is

accomplished by giving women the ability to take andynamic role in their own personal care [25, 26].

- Pregnancy Plans Tailored to the Individual: In order to generate personalized pregnancy plans, systems that are drove by artificial intelligence are able to evaluate the information of individual patients, as well as their genealogical and lifestyle characteristics [27, 28]. The purpose of these programs is to provide precise recommendations on nutrition, physical activity, and prenatal care in order to increase the likelihood of a successful pregnancy and birth [29, 30].
- Enhancement of Postnatal Care and Treatment: It is necessary for both the mother and the child to get postnatal care in order to facilitate their physical and mental rehabilitation [31]. Utilizing intelligent technology makes it possible to more easily complete tasks such as the remote observation of postnatal hemorrhage, the identification of dangers for infection, and the evaluation of markers of mental health [32, 33]. Through the utilization of telehealth consultations, it is possible to simplify the process of gaining access to healthcare specialists [34]. These consultations help alleviate the burden of physically travelling and offer support in a timely manner. Chatbots that are driven by artificial intelligence are capable to provide answers to frequently asked questions, offer reassurance, and provide support in a timely manner [35, 36].
- Management of Vaccination: Intelligent systems have the capability to improve vaccination schedules for moms as well as infants. These systems can take advantage of timely notifications and reminders to ensure that immunizations are provided at the proper time. As a consequence, this leads to an increase in compliance as well as a reduction in the risk of diseases that can be maintained through vaccination. In order to enhance the efficiency of public health surveillance, integrated systems for data management make it possible to monitor vaccination records in a manner that is both effective and efficient[37,38].

There are a plethora of benefits that may be gained from incorporating intelligent healthcare systems into postnatal care and motherhood, some of which include the following:

- Improvements in the health outcomes of mothers and children: There are a numeral of aspects that contribute to improved health outcomes for women and children, including the early discovery of issues, the enhancement of access to healthcare facilities, and the provision of individualized care [39, 40].
- Remote monitoring and consultations through telehealth have the potential to minimize the need for expensive hospital visits as well as the overall amount of money spent on healthcare. This can result in a reduction in the overall cost of healthcare [41, 42].
- Access to healthcare is improved: Smart technologies have the ability to transcend geographical borders and improve access to healthcare facilities. This is especially beneficial for women who reside in areas that require more medical attention [43].
- Increased patient involvement: Personalized planning and remote monitoring create an environment that encourages increased patient involvement and gives women the opportunity to plays an active part in their own medical treatment.
- The utilization of data-driven insights, which are obtained by intelligent systems, can be utilized to accomplish a number of goals, including the identification of trends, the improvement of healthcare policy, and the development of therapies that are more successful.

In spite of the fact that intelligent healthcare systems have a great covenant of promise, there are numerous challenges that must be overcome to implement them:

- It is of utmost importance to protect the confidentiality and safety of private information pertaining to patients. Maintaining compliance with data privacy regulations and putting in place robust security measures is an imperative necessity [44].
- In order to combat the digital divide and prevent existing health inequities from becoming even more pronounced, it is crucial to guarantee that everyone has equal access to digital technology and knowledge [45].
- Integrated system integration with systems that are already in place: It is absolutely necessary to make certain that there is a smooth interaction with the healthcare infrastructure that is already in place in order to guarantee positive adoption.
- Ethical considerations taken into account: In order to solve ethical concerns regarding artificial intelligence algorithms, prejudice, and the possibility of human error, it is of utmost necessity to address such issues.

The deployment of intelligent healthcare systems has the prospective to greatly improve a number of aspects of healthcare, including motherhood, postnatal care, and vaccination. Such systems have the probable to increase the quality of care, result in improved health outcomes, and promote equity in terms of access to medical services for both women and babies. This is because they make use of the power that technology has. In order to make the most of the opportunities presented by this innovative strategy, it is vital to report the challenges that are associated with it. This can be accomplished by deliberate preparation, collaborative effort, and a focus on ethical considerations [46].

When it comes to providing medical care for pregnant women and children, proactive and individualized approaches are required. It is possible to greatly improve outcomes by supporting timely interventions through the use of prompt and accurate diagnosis, even if it is based on information that is easily available on symptoms. Traditional methods of diagnosis frequently include extensive testing and visits with specialists, both of which may prove time-consuming and expensive, particularly in contexts where resources are

limited. Utilizing Random Forest, a strong machine learning algorithm, this study investigates the possibility of utilizing it to forecast diseases in kids and pregnant women based on the symptoms that they present with. This presents a potential option to early risk assessment and enhanced healthcare delivery [47].

A considerable reduction in morbidity and mortality can be achieved by using reported symptoms to make predictions about diseases that affect children, such as pneumonia, malaria, and malnutrition. When symptoms such as fever, cough, trouble breathing, and weight loss are input into a Random Forest model, the model is able to determine the likelihood of such illnesses and prompt referrals for further examination and treatment. This is of particular importance in communities that are geographically isolated and have restricted access to specialized medical care [48].

To ensure the health of both the mother and the unborn child, it is essential to diagnose issues such as gestational diabetes, preeclampsia, and anemia at an early stage during pregnancy. In order to determine the likelihood of these illnesses, a Random Forest model can be used to take into account symptoms such as high blood pressure, oedema, and weariness. By identifying the condition at an early stage, it is possible to take preventative measures, such as making changes to one's lifestyle or taking medicine, at the appropriate time [49,50].

II. Related Work

In the study [51], a strategy was devised to estimate the amount of hospitalization days for newborn patients. The RF approach, together with a dataset of records retrieved through a hospital database, was used to create a model for predicting hospital stays. The RF method outperformed the Adaboost and Bagging methods in terms of prophetic accuracy, with RMSE of 0.314, R2 of 0.706, |R| of 0.545, and Acc±1 of 71%.

In [52], the authors present an integrated healthcare system that may be tailored to the child's condition and allows for the accurate distribution of information directly to parents, carers, pediatricians, and hospitals. The healthcare reminder and data dissemination method will be provided immediately to the parents' cellphones. A single-click idea also allows parents to contact their pediatrician or hospital instantly in an emergency situation. All of the child's daily data, including diet, activity, medicine, and treatment, will be stored in an Electronic Medical Record and shown as a chart to help parents and pediatricians monitor the child's health state. Finally, the adoption of this comprehensive framework protects the child from a variety of health risks, such as drug complications, dietary and activity errors, and inaccuracies in health care delivery.

In [53], the authors seek to create a framework for designing a pediatric healthcare-built environment by looking at particular therapeutic objectives and design issues. A growing body of research reveals that the characteristics of pediatric healthcare facilities vary from those found in adult healthcare facilities; thus, the results of adult health design for facilities cannot be applied to pediatric healthcare facility design. Furthermore, studies focusing on architecture in pediatric healthcare are limited.

In [54], the authors used Walker and Avant's conceptualization technique, which entails an efficient process that includes the idea selection, setting objectives, seeing possible uses, defining attributes, building a model case, determining contrary cases, establishing the cause and effect, and examining empirical referents. A scoping study was undertaken using a variety of sources like online dictionaries, scientific databases, and grey literature, with an emphasis on literature relevant to CFH. That study also defined the reasons for and effects of CFH, established model and counter-model situations, and investigated empirical references to provide a thorough knowledge of the idea.

III. Random Forest

Within the realm of machine learning, the search for expectation models that are both more accurate and more reliable is continually ongoing. Among the many algorithms that are available to data scientists, the Random Forest algorithm stands out as an approach that is both versatile and powerful. Within the scope of this study, the fundamental ideas behind the Random Forest are investigated, along with its advantages, disadvantages, and applications in the real world [49].

In its most fundamental form, a Random Forest relies on a collaborative learning approach that is founded on the concept of "wisdom of the crowd". It does this by utilizing the power of numerous decision trees, each of which has been trained using an entirely distinct subset of the data and characteristics, in order to arrive at a forecast that is more accurate and trustworthy than any single tree could possibly achieve. Imagine that a group of specialists, each of whom possesses slightly distinct abilities and points of view, are working together to find a solution to a difficult problem. A comparable idea underlies the operation of the Random Forest. Contemplate it in this manner:

- The forest contains a great number of decision trees.
- The growth of each tree is accomplished by the use of randomization algorithms.

What makes the Random Forest so appealing is the fact that its creation method is not overly complicated but still quite efficient:

- The initial data set is selected at random with replacement numerous times, resulting in the creation of several "bootstrap" datasets. This technique is also known as "bagging and bootstrapping". One among these bootstrap samples is used to train each tree in the population. This not only increases diversity but also decreases the likelihood of "overfitting".
- The Random Forest algorithm doesn't taken into account all of the available features when it splits a node in a decision tree. This is referred to as feature randomness or feature subsampling. Instead, it chooses between a subset of features at random in order to determine which split is the most appropriate. This

adds another layer of diversity to the trees, helping to create an even more robust ensemble.

- Growing Decision Trees: To capture the intricate relationships that exist within the data, each decision tree is developed to its greatest depth without any pruning.
- Aggregation of Predictions: When it comes to classification tasks, the Random Forest algorithm determines which class has the highest number of votes from all of the trees. For tasks involving regression, it takes the average of the predictions made by all of the trees.

Popular Python packages such as Scikit-learn offer versions of the Random Forest algorithm that are simple to use. Tuning the following hyperparameters is essential:

- A number that represents the total number of trees in the forest. When this value is increased, performance is generally improved; however, the amount of time spent training is also increased.
- The number of features that are taken into consideration while dividing each node is named max_features.
- max_depth is the amount of depth that each tree can reach.
- min_samples_split is the minimal number of samples that must be split to have an internal node to be divided.
- The minimal number of samples that must be extant at a leaf node is denoted by the min_samples_leaf variable.

It is a robust and adaptable machine learning method that delivers an intriguing blend of accuracy, resilience, and ease of use. The Random Forest algorithm is a combination of these three characteristics. Despite the fact that it is essential to be aware of its limits, the fact that it is able to manage complicated datasets and offer accurate predictions makes it an extremely useful tool for data scientists and machine learning practitioners in a broad variety of applications. You may harness the potential of the Random Forest to tackle difficult prediction issues and obtain useful insights from your data if you have a solid understanding of its fundamental concepts and if you carefully tune its hyperparameters. In order to guarantee the dependability and generalizability of the model, it is essential to conduct appropriate evaluation and validation, just as it is with any other machine learning technique. From our study we find that the Accuracy, recall and precision for RF is greater than SVM and DT. Hence, we decide to work with RF classifier [50].

IV. CHS FRAMEWORK

Random Forest is a type of ensemble learning that works by generating many decision trees and merging the results of those trees to enhance the overall performance of the system. As a result of its precision, resilience, and user-friendliness, it has become a widely used algorithm. Random Forest is an excellent option for healthcare applications since it can manage numerical and categorical information. This makes it an appropriate choice for situations where multiple categories of data are frequently collected.

The following instructions are included in the framework that has been suggested for the healthcare system of children that makes use of Random Forest, as shown in Fig. 1:

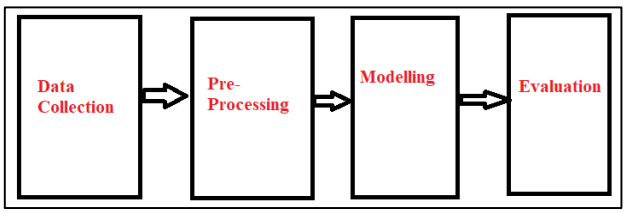


Fig. 1. Healthcare system using random forest.

- Data Collection: The main phase in the framework needs collecting data from a variety of sources, including clinical trials, electronic health records, and public health databases, among other places. The data ought to contain information on the child's health status, demography, lifestyle, and environmental factors. Environmental elements should also be included.
- Preprocessing of Data: To guarantee the data's quality and relevance, it is necessary to preprocess the data that has been acquired. Following this phase, the data will be cleaned, normalized, and feature selection will take place. Removal of missing values, duplicates, and outliers are all components of the data cleaning process. In order to eliminate any potential bias towards characteristics that have higher values, the process of normalization involves scaling the data into a common range. Identification of the features that are most pertinent to the prediction task is the process that is referred to as feature selection.
- Modelling Instruction: After the data has been preprocessed, it is utilized in the training of Random Forest model. The training of the model allows it to make predictions about a variety of health outcomes, including the likelihood of developing a specific disease, the efficacy of a therapy, and the risk of adverse events. A select portion of the data is utilized for the training of the model, while the remaining data is utilized for testing and validation purposes.
- Evaluation of the Model: The trained model is evaluated using a variety of metrics, including accuracy, precision, recall, and F1 score, among others. In addition, the performance of the model is evaluated through the use of cross-validation. This method involves the data being divided into numerous folds, and the model being trained and tested on each fold.

Random Forest may be utilized in the healthcare system for children to make predictions about diseases depending on the symptoms that are input. It is possible to train the algorithm using a huge dataset consisting of historical patient records, which may include diagnoses, symptoms, and other

pertinent information. A healthcare provider can enter symptoms into the model, and once it has been trained, it can be used to make predictions about the likelihood of a specific disease based on those symptoms.

In the context of the healthcare system for children, the following steps (shown in Fig. 2) provide an outline of how Random Forest might be applied:

- Collection of Data: The first thing that needs to be done is to gather a substantial amount of patient records from the past. In the dataset, there should be information on symptoms, diagnoses, and other factors that are pertinent, such as age, gender, and medical history.
- It is necessary to do data preprocessing on the raw data that was gathered in the first stage. This is done in order to eliminate any inconsistencies, missing values, and outliers. A normalization and standardization of the data should also be performed in order to guarantee that all of the features have the same scale.
- In the following stage, which is called "Feature Selection," the most important features which shall be utilized in the training of the model, known as Random Forest, are chosen. The selection of features is an essential phase since it has the potential to have a substantial influence on the overall performance of the framework.
- Training the Model: After the characteristics that are pertinent to the problem have been chosen, the Random Forest model can be trained using the data that has been preprocessed. On the basis of the symptoms that are input, the algorithm is trained to make a prediction about the diagnosis.
- After a framework has been properly trained, it must be assessed to guarantee that it's accurate and dependable. This evaluation must take place right after the training process. A number of evaluation criteria's, includes accuracy, precision, recall, and F1 score, can be utilized to accomplish this goals.

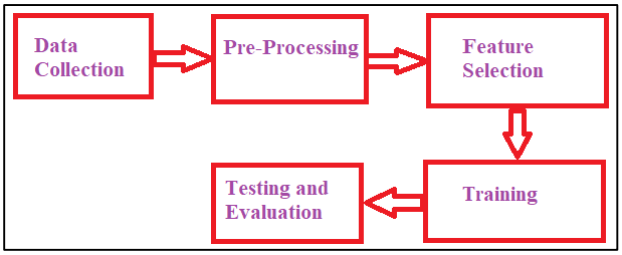


Fig. 2. Proposed model using random forest.

In the event that the model has been validated and reviewed, it will then be able to be implemented inside the infant and child healthcare system. With the help of the model, medical professionals can input symptoms and then receive a forecast of the condition that is most likely to be present.

V. Results and Discussion

In discussion to proposed method, deployment of the model occurs once it has been trained and evaluated, at which point it is ready to be implemented inside the child healthcare system. The model is capable of being included into clinical decision support systems, mobile health applications, and electronic health records. The model is able to make forecasts and suggestions to healthcare practitioners in real time, which improves both the quality of care and the speed with which it is delivered.

Continuous monitoring and updating of the model: In order to guarantee that the model is accurate and up-to-date, it should be regularly monitored and updated. This comprises monitoring the performance of the model over a period of time, including fresh data into the model, and retraining its model with any new features or techniques that have been included.

Within the context of the child healthcare system, the Random Forest framework may be utilized for a variety of use cases, including the following:

- The Random Forest model can be taught to forecast the likelihood of a kid getting a specific ailment, like asthma, diabetes, or obesity. This can be accomplished by the utilization of the Random Forest algorithm. There are a number of aspects that can be taken into consideration by the model, including the child's age, sex, genealogy, and lifestyle.
- Effectiveness of Treatment: The Random Forest method may be utilized to provide predictions on the efficacy of a specific treatment or intervention. There are a number of elements that can be taken into consideration by the model, including the child's current health state, demographic information, and treatment history.
- Adverse Events: The Random Forest method might be utilized to provide predictions on the likelihood of adverse events, which may include accidents involving drugs, infections, or injuries. There are a number of aspects that can be taken into consideration by the model, including the child's current state of health, the pharmaceutical regimen, and environmental circumstances.

In results of proposed system, it shows that, a framework known as Random Forest is a powerful instrument that can be used to enhance the healthcare system for children. The framework is able to provide real-time forecasts and suggestions to healthcare professionals by utilizing data-driven methodologies. This results in an improvement in both the quality of care and the efficiency with which it is delivered. The paradigm is applicable to a wide range of use cases, including the predictability of diseases, the success of treatments, and the occurrence of adverse events. Continuous monitoring and updating of the framework is necessary in order to guarantee that it is accurate and continues to be relevant. The Random Forest framework has the potential to

significantly contribute to a more positive and healthier future for our children if it is implemented appropriately.

As can be seen in Fig. 3 and Fig. 4, the process of registering as a physician within a child healthcare system is more severe than the process of registering as a general practitioner. This is because pediatric patients have distinct vulnerabilities and requirements that are not shared by general practitioners. This section provides an overview of the most essential action steps and aspects to take into consideration, highlighting the variations that may be present based on the specific system and the geographical region; it also provides an overview of the factors that should be taken into consideration.

It is necessary for medical practitioners to satisfy a variety of qualifying standards before they can begin the process of registering their services. These typically consist of the following:

- Obtaining a Medical Degree and a License: It is absolutely necessary to be in possession of a valid medical degree from an educational establishment that has been awarded accreditation, in addition to a valid license that does not contain any restrictions, in order to continue practicing medicine in the jurisdiction that is applicable. There are a number of medical specializations that are regularly required or suggested by hospitals and other medical facilities. Some examples of these specializations include pediatrics, neonatology, and child psychology.
- Years of experience: Many systems, particularly those in the field of pediatric care, require candidates to have a specified number of years of experience in addition to post-graduate training. The fact that this is the case ensures that a certain level of knowledge and experience is there with regard to the particular challenges that are associated with treating children.
- Background checks: It is standard procedure to do thorough background checks, which include determining whether or not the individual in question has a criminal history and confirming the individual's professional credentials. In order to guarantee the patient's safety and protection, this task is carried out.
- CME, which stands for continuing medical education: Evidence of continuous education and a dedication to staying current with the most recent advancements in pediatric medicine are frequently required to be provided. This is a requirement that is frequently of critical importance. There is a possibility that continuing medical education (CME) classes or certifications will be necessary.

In the event when you are interested in enrolling in the kid healthcare system, it is of the most importance to guarantee that you thoroughly examine the particular regulations and guidelines that are provided by that system. It is strongly suggested that you get in touch with the system's registration department as soon as possible in order to gather information that is not only accurate but also up to date. Fig. 5 shows the GUI for doctor login for accessing the patent information Fig. 6 shows the window of user for doctor assign for further treatment, whereas Fig. 7 and Fig. 8 shows the GUI for diseases identification based on input and remedies for treatments. Whereas, Fig. 9 shows the Accuracy, Recall and Precision of proposed method using RF. The accuracy of proposed method is compared with SVM and DT, and is found that the accuracy of RF is good as compared to others, as shown in Fig. 10.



Fig. 3. Doctor registration process.



Fig. 4. Mother registration process.

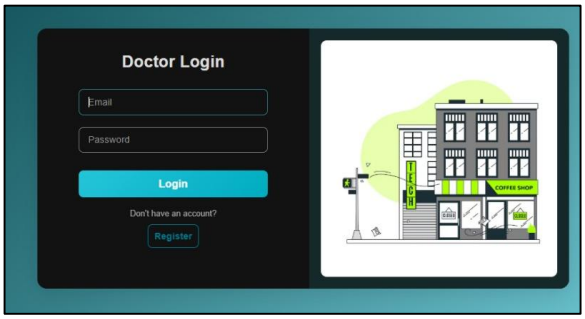


Fig. 5. Doctor login system.



Fig. 6. Patients assigned to specialist doctor by framework.

In case of security of framework, we suggests regular monitoring and audit of the system. Also the access is controlled to limited persons, i.e. doctors and parents. Also RF model is trained with resilience to adverse attacks, which can reduce the misdiagnosis and mistreatments. Also we suggests to regularly evaluate the model for provide further security to systems and user.

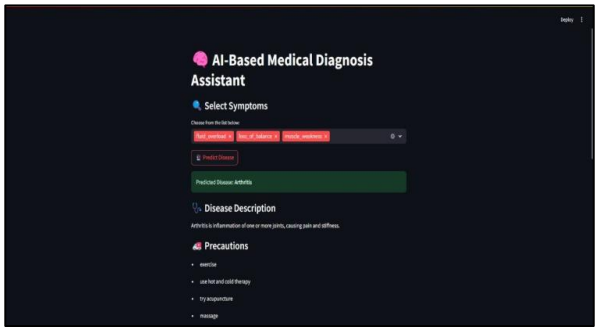


Fig. 7. Disease identification using the symptoms of the child by using framework.

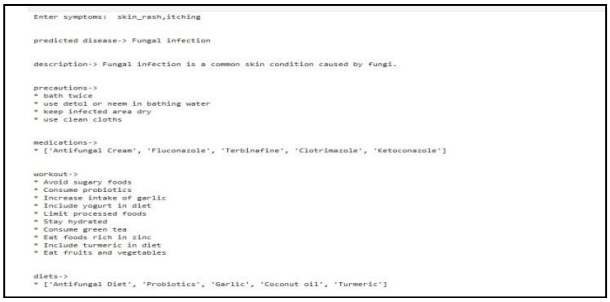


Fig. 8. Disease identification using symptoms and remedies to child using framework.

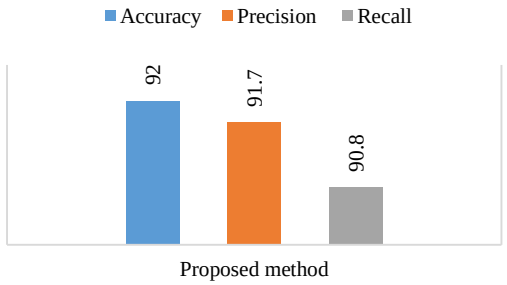


Fig. 9. Parameters for proposed method.

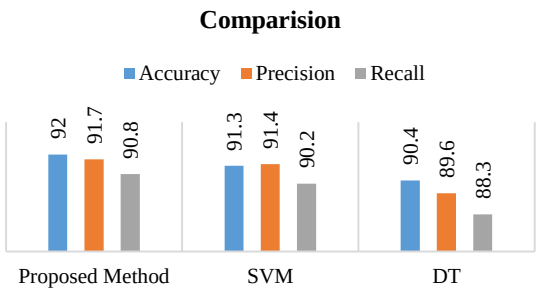


Fig. 10. Comparison with esteemed methods.

VI. Conclusion

The Random Forest algorithm provides a promising method for utilizing symptom information to anticipate diseases in healthcare systems that serve pregnant women and children. Interventions may be implemented at the appropriate time if the risk of various disorders is precisely assessed. This will ultimately result in improved health outcomes and a reduction in the expenses of healthcare. However, in order to guarantee a responsible and efficient implementation, it is necessary to give careful thought to the quality of the data, ethical issues, the integration with the systems that are already in place, and the explainability of the model. The implementation of RF-based solutions has the prospective to revolutionise the delivery of healthcare and enhance the well-being of vulnerable populations, provided that research and development efforts are maintained. The accuracy of proposed method is 92% and is greater than other methods. In Future, the system will be implements using IoT and AI based Decision making to avoid the misdiagnosis of symptoms. Depending on symptoms, the prediction is made. Also we suggests to employ the KSK approach in future.

Acknowledgment

We thank all Doctors and parents for giving us the information to make the system successful.

References

[1] Altaf O. Mulani, Arti Vasant Bang, Ganesh B. Birajadar, Amar B. Deshmukh, and Hemlata Makarand Jadhav, (2024). IoT Based Air, Water, and Soil Monitoring System for Pomegranate Farming, Annals of Agri-Bio Research. 29 (2): 71-86, 2024.

[2] Bhawana Parihar, Ajmeera Kiran, Sabitha Valaboju, Syed Zahidur Rashid, and Anita Sofia Liz D R. (2025). Enhancing Data Security in Distributed Systems Using Homomorphic Encryption and Secure Computation Techniques, ITM Web Conf., 76 (2025) 02010DOI: https://doi.org/10.1051/itmconf/20257602010.

[3] C. Veena, M. Sridevi, K. K. S. Liyakat, B. Saha, S. R. Reddy and N. Shirisha,(2023). HEECCNB: An Efficient IoT-Cloud Architecture for Secure Patient Data Transmission and Accurate Disease Prediction in Healthcare Systems, 2023 Seventh International Conference on Image Information Processing (ICIIP), Solan, India, 2023, pp. 407-410, doi: 10.1109/ICIIP61524.2023.10537627. Available at: https://ieeexplore.ieee.org/document/10537627.

[4] D. A. Tamboli, V. A. Sawant, M. H. M. and S. Sathe, (2024). AI-Driven-IoT(AIIoT) Based Decision-Making- KSK Approach in Drones for Climate Change Study, 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2024, pp. 1735-1744, doi: 10.1109/ICUIS64676.2024.10866450.

[5] K. Rajendra Prasad, Santoshachandra Rao Karanam et al. (2024). AI in public-private partnership for IT infrastructure development, Journal of High Technology Management Research, Volume 35, Issue 1, May 2024, 100496. https://doi.org/10.1016/j.hitech.2024.100496.

[6] K. K. S. Liyakat. (2023).Detecting Malicious Nodes in IoT Networks Using Machine Learning and Artificial Neural Networks, 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2023, pp. 1-5, doi:10.1109/ESCI56872.2023.10099544. Available at: https://ieeexplore.ieee.org/document/10099544/.

[7] K K S, (2024). Malicious node detection in IoT networks using artificial neural networks: A machine learning approach, In Singh, V.K., Kumar Sagar, A., Nand, P., Astya, R., & Kaiwartya, O. (Eds.). Intelligent Networks: Techniques, and Applications (1st ed.). CRC Press. https://doi.org/10.1201/978100354136.

[8] K. Kasat, N. Shaikh, V. K. Rayabharapu, and M. Nayak. (2023). Implementation and Recognition of Waste Management System with Mobility Solution in Smart Cities using Internet of Things, 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2023, pp. 1661-1665, doi: 10.1109/ICAISS58487.2023.10250690 . Available at: https://ieeexplore.ieee.org/document/10250690/.

[9] Kazi, K. (2024a). AI-Driven IoT (AIIoT) in Healthcare Monitoring. In T. Nguyen & N. Vo (Eds.), Using Traditional Design Methods to Enhance AI-Driven Decision Making (pp. 77-101). IGI Global. https://doi.org/10.4018/979-8-3693-0639-0.ch003 available at: https://www.igi-global.com/chapter/ai-driven-iot-aiiot-in-healthcare-monitoring/336693.

[10] Kazi, K. (2024b). Modelling and Simulation of Electric Vehicle for Performance Analysis: BEV and HEV Electrical Vehicle Implementation Using Simulink for E-Mobility Ecosystems. In L. D., N. Nagpal, N. Kassarwani, V. Varthanan G., & P. Siano (Eds.), E-Mobility in Electrical Energy Systems for Sustainability (pp. 295-320). IGI Global.https://doi.org/10.4018/979-8-3693-2611-4.ch014 Available at: https://www.igi-global.com/gateway/chapter/full-text-pdf/341172.

[11] Kazi, K. (2025). Machine Learning-Powered IoT (MLIoT) for Retail Apparel Industry. In T. Tarnanidis, E. Papachristou, M. Karypidis, & V. Manda (Eds.), Sustainable Practices in the Fashion and Retail Industry (pp. 345-372). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-9959-0.ch015.

[12] Kazi, K. S. (2025). Braille-Lippi Numbers and Characters Detection and Announcement System for Blind Children Using KSK Approach: AI-Driven Decision-Making Approach. In T. Murugan, K. P., & A. Abirami (Eds.), Driving Quality Education Through AI and Data Science (pp. 531-556). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-8292-9.ch023.

[13] Kazi, K. S. (2025a). AI-Driven IoT (AIIoT)-Based Decision-Making System for High BP Patient Healthcare Monitoring: KSK1 Approach for BP Patient Healthcare Monitoring. In T. Mzili, A. Arya, D. Pamucar, & M. Shaheen (Eds.), Optimization, Machine Learning, and Fuzzy Logic: Theory, Algorithms, and Applications (pp. 71-102). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-7352-1.ch003.

[14] Kazi, K. S. (2025b). Advancing Towards Sustainable Energy With Hydrogen Solutions: Adaptation and Challenges. In F. Özsungur, M. Chaychi Semsari, & H. Küçük Bayraktar (Eds.), Geopolitical Landscapes of Renewable Energy and Urban Growth (pp. 357-394). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-8814-3.ch013.

[15] Kazi, K. S. (2025c). AI-Driven-IoT (AIIoT) Decision-Making System for Hepatitis Disease Patient Healthcare Monitoring: KSK1 Approach for Hepatitis Patient Monitoring. In S. Agarwal, D. Lakshmi, & L. Singh (Eds.), Navigating Innovations and Challenges in Travel Medicine and Digital Health (pp. 431-450). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-8774-0.ch022.

[16] Kazi, S. (2024). Machine Learning-Based Pomegranate Disease Detection and Treatment. In M. Zia Ul Haq & I. Ali (Eds.), Revolutionizing Pest Management for Sustainable Agriculture (pp. 469-498). IGI Global. https://doi.org/10.4018/979-8-3693-3061-6.ch019.

[17] Kazi, S. (2024a). Computer-Aided Diagnosis in Ophthalmology: A Technical Review of Deep Learning Applications. In M. Garcia & R. de Almeida (Eds.), Transformative Approaches to Patient Literacy and Healthcare Innovation (pp. 112-135). IGI Global. https://doi.org/10.4018/979-8-3693-3661-8.ch006 Available at: https://www.igi-global.com/chapter/computer-aided-diagnosis-in-ophthalmology/342823.

[18] Kazi, S. (2024b). IoT Driven by Machine Learning (MLIoT) for the Retail Apparel Sector. In T. Tarnanidis, E. Papachristou, M. Karypidis, & V. Ismyrlis (Eds.), Driving Green Marketing in Fashion and Retail (pp. 63-81). IGI Global. https://doi.org/10.4018/979-8-3693-3049-4.ch004.

[19] Kazi, S. (2025c). AI-Driven-IoT (AIIoT)-Based Decision Making in Drones for Climate Change: KSK Approach. In S. Aouadni& I. Aouadni (Eds.), Recent Theories and Applications for Multi-Criteria Decision-Making (pp. 311-340). IGI Global. https://doi.org/10.4018/979-8-3693-6502-1.ch011.

[20] Kazi, S. (2024d). Artificial Intelligence (AI)-Driven IoT (AIIoT)-Based Agriculture Automation. In S. Satapathy & K. Muduli (Eds.), Advanced Computational Methods for Agri-Business Sustainability (pp. 72-94). IGI Global. https://doi.org/10.4018/979-8-3693-3583-3.ch005.

[21] Kazi, S. (2025). Machine Learning-Driven Internet of Medical Things (ML-IoMT)-Based Healthcare Monitoring System. In B. Soufiene & C. Chakraborty (Eds.), Responsible AI for Digital Health and Medical Analytics (pp. 49-86). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-6294-5.ch003.

[22] Kazi, S. (2025a). Transformation of Agriculture Effectuated by Artificial Intelligence-Driven Internet of Things (AIIoT). In J. Garwi, M. Dzingirai, & R. Masengu (Eds.), Integrating Agriculture, Green Marketing Strategies, and Artificial Intelligence (pp. 449-484). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-6468-0.ch015.

[23] K S K, (2024c). Vehicle Health Monitoring System (VHMS) by Employing IoT and Sensors, Grenze International Journal of Engineering and Technology,Vol 10, Issue 2, pp- 5367-5374. Available at: https://thegrenze.com/index.php?display=page&view=journalabstract&absid=3371&id=8.

[24] K S K, (2024e). A Novel Approach on ML based Palmistry, Grenze International Journal of Engineering and Technology,Vol 10, Issue 2, pp-5186-5193. Available at: https://thegrenze.com/index.php?display=page&view=journalabstract&absid=3344&id=8.

[25] K S K, (2024f).IoT based Boiler Health Monitoring for Sugar Industries, Grenze International Journal of Engineering and Technology,Vol 10, Issue 2, pp. 5178 -5185. Available at: https://thegrenze.com/index.php?display=page&view=journalabstract&absid=3343&id=8.

[26] Keerthana, R., K, V., Bhagyalakshmi, K., Papinaidu, M., V, V., & Liyakat, K. K. S. (2025). Machine learning based risk assessment for financial management in big data IoT credit. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5086671.

[27] KKS Liyakat, (2024). Explainable AI in Healthcare. In: Explainable Artificial Intelligence in healthcare System, editors: A. Anitha Kamaraj, Debi Prasanna Acharjya. ISBN: 979-8-89113-598-7. DOI: https://doi.org/10.52305/GOMR8163.

[28] KKS Liyakat, (2024a). Machine Learning (ML)-Based Braille Lippi Characters and Numbers Detection and Announcement System for Blind Children in Learning, In Gamze Sart (Eds.), Social Reflections of Human-Computer Interaction in Education, Management, and Economics, IGI Global. https://doi.org/10.4018/979-8-3693-3033-3.ch002.

[29] Liyakat, K.K.S. (2023a). Machine Learning Approach Using Artificial Neural Networks to Detect Malicious Nodes in IoT Networks. In: Shukla, P.K., Mittal, H., Engelbrecht, A. (eds) Computer Vision and Robotics. CVR 2023. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-99-4577-1_3.

[30] Liyakat Kazi, K. S. (2024). ChatGPT: An Automated Teacher's Guide to Learning. In R. Bansal, A. Chakir, A. Hafaz Ngah, F. Rabby, & A. Jain (Eds.), AI Algorithms and ChatGPT for Student Engagement in Online Learning (pp. 1-20). IGI Global. https://doi.org/10.4018/979-8-3693-4268-8.ch001.

[31] Liyakat. (2025). IoT Technologies for the Intelligent Dairy Industry: A New Challenge. In S. Thandekkattu & N. Vajjhala (Eds.), Designing Sustainable Internet of Things Solutions for Smart Industries (pp. 321-350). IGI Global. https://doi.org/10.4018/979-8-3693-5498-8.ch012.

[32] Liyakat, K. K. (2025a). Heart Health Monitoring Using IoT and Machine Learning Methods. In A. Shaik (Ed.), AI-Powered Advances in Pharmacology (pp. 257-282). IGI Global. https://doi.org/10.4018/979-8-3693-3212-2.ch010.

[33] Liyakat. (2025d). AI-Driven-IoT(AIIoT)-Based Decision Making in Kidney Diseases Patient Healthcare Monitoring: KSK Approach for Kidney Monitoring. In L. Özgür Polat & O. Polat (Eds.), AI-Driven Innovation in Healthcare Data Analytics (pp. 277-306). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-7277-7.ch009.

[34] Liyakat, K.K.S. (2024). Machine Learning Approach Using Artificial Neural Networks to Detect Malicious Nodes in IoT Networks. In:

Udgata, S.K., Sethi, S., Gao, XZ. (eds) Intelligent Systems. ICMIB 2023. Lecture Notes in Networks and Systems, vol 728. Springer, Singapore.https://doi.org/10.1007/978-981-99-3932-9_12 available at: https://link.springer.com/chapter/10.1007/978-981-99-3932-9_12.

[35] M Pradeepa, et al. (2022). Student Health Detection using a Machine Learning Approach and IoT, 2022 IEEE 2nd Mysore sub section International Conference (MysuruCon), 2022. Available at: https://ieeexplore.ieee.org/document/9972445.

[36] Mahant, M. A. (2025). Machine Learning-Driven Internet of Things (MLIoT)-Based Healthcare Monitoring System. In N. Wickramasinghe (Ed.), Digitalization and the Transformation of the Healthcare Sector (pp. 205-236). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-9641-4.ch007.

[37] Mulani AO, Liyakat KKS, Warade NS, et al (2025). . ML-powered Internet of Medical Things Structure for Heart Disease Prediction. Journal of Pharmacology and Pharmacotherapeutics. 2025; 0(0). doi:10.1177/0976500X241306184.

[38] Odnala, S., Shanthy, R., Bharathi, B., Pandey, C., Rachapalli, A., & Liyakat, K. K. S. (2025). Artificial Intelligence and Cloud-Enabled E-Vehicle Design with Wireless Sensor Integration. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5107242.

[39] P. Neeraja, R. G. Kumar, M. S. Kumar, K. K. S. Liyakat and M. S. Vani. (2024), DL-Based Somnolence Detection for Improved Driver Safety and Alertness Monitoring. 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 2024, pp. 589-594, doi: 10.1109/IC2PCT60090.2024.10486714. Available at: https://ieeexplore.ieee.org/document/10486714.

[40] Prashant K Magadum (2024). Machine Learning for Predicting Wind Turbine Output Power in Wind Energy Conversion Systems, Grenze International Journal of Engineering and Technology, Jan Issue, Vol 10, Issue 1, pp. 2074-2080. Grenze ID: 01.GIJET.10.1.4_1Available at:https://thegrenze.com/index.php?display=page&view=journalabstract &absid=2514&id=8.

[41] Priya Mangesh Nerkar, Bhagyarekha Ujjwalganesh Dhaware. (2023). Predictive Data Analytics Framework Based on Heart Healthcare System (HHS) Using Machine Learning, Journal of Advanced Zoology, 2023, Volume 44, Special Issue -2, Page 3673:3686. Available at: https://jazindia.com/index.php/jaz/article/view/1695.

[42] Priya Nerkar and Sultanabanu, (2024). IoT-Based Skin Health Monitoring System, International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS). 2024, 13(11): 5937-5950. https://doi.org/10.31032/IJBPAS/2024/13.11.8488.

[43] S. B. Khadake, A. B. Chounde, A. A. Suryagan, M. H. M. and M. R. Khadatare, (2024). AI-Driven-IoT(AIIoT) Based Decision Making System for High-Blood Pressure Patient Healthcare Monitoring, 2024 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2024, pp. 96-102, doi: 10.1109/ICSCNA63714.2024.10863954.

[44] Sayyad. (2025a). AI-Powered-IoT (AIIoT)-Based Decision-Making System for BP Patient's Healthcare Monitoring: KSK Approach for BP Patient Healthcare Monitoring. In S. Aouadni& I. Aouadni (Eds.), Recent Theories and Applications for Multi-Criteria Decision-Making (pp. 205-238). IGI Global.https://doi.org/10.4018/979-8-3693-6502-1.ch008.

[45] Sayyad (2025b). AI-Powered IoT (AI IoT) for Decision-Making in Smart Agriculture: KSK Approach for Smart Agriculture. In S. Hai-Jew (Ed.), Enhancing Automated Decision-Making Through AI (pp. 67-96). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-6230-3.ch003.

[46] Sayyad (2025c). KK Approach to Increase Resilience in Internet of Things: A T-Cell Security Concept. In D. Darwish & K. Charan (Eds.), Analyzing Privacy and Security Difficulties in Social Media: New Challenges and Solutions (pp. 87-120). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-9491-5.ch005.

[47] Sayyad, (2025). KK Approach for IoT Security: T-Cell Concept. In Rajeev Kumar, Sheng-Lung Peng, & Ahmed Elngar (Eds.), Deep Learning Innovations for Securing Critical Infrastructures. IGI Global Scientific Publishing.

[48] Sayyad (2025d). Healthcare Monitoring System Driven by Machine Learning and Internet of Medical Things (MLIoMT). In V. Kumar, P. Katina, & J. Zhao (Eds.), Convergence of Internet of Medical Things (IoMT) and Generative AI (pp. 385-416). IGI Global Scientific Publishing.https://doi.org/10.4018/979-8-3693-6180-1.ch016.

[49] Shinde, S. S., Nerkar, P. M., Kazi, S. S., & Kazi, V. S. (2025). Machine Learning for Brand Protection: A Review of a Proactive Defense Mechanism. In M. Khan & M. Amin Ul Haq (Eds.), Avoiding Ad Fraud and Supporting Brand Safety: Programmatic Advertising Solutions (pp. 175-220). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-7041-4.ch007.

[50] Upadhyaya, A. N., Surekha, C., Malathi, P., Suresh, G.,Suriyan, K., & Liyakat, K. K. S. (2025). Pioneering cognitive computing for transformative healthcare innovations. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5086894.

[51] C. Wang, X. Dong, L. Yu, L. Ye, W. Zhuang and F. Ma, "Prediction of days in hospital for children using random forest," 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 2017, pp. 1-6, doi: 10.1109/CISP-BMEI.2017.8302287.

[52] N. Sevani, "Personal health care framework for children," 2015 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, Indonesia, 2015, pp. 166-170, doi: 10.1109/ICODSE.2015.7436991.

[53] Abdul Halim Babbu, Mazharul Haque, “A framework for the design of pediatric healthcare environment using the Delphi technique” Ain Shams Engineering Journal, Volume 14, Issue 5, May 2023, 101975. https://doi.org/10.1016/j.asej.2022.101975.

[54] Wei Xiao Huang, Mei Chan Chong, “Child-friendly healthcare: A concept analysis”, Journal of Pediatric Nursing, Volume 80, January–February 2025, Pages e7-e15, https://doi.org/10.1016/j.pedn.2024.10.035.

Enhancing Organizational Threat Profiling by Employing Deep Learning with Physical Security Systems and Human Behavior Analysis

D.H. Senevirathna, W.M.M. Gunasekara, K.P.A.T. Gunawardhana,
M.F.F. Ashra, Harinda Fernando, Kavinga Yapa Abeywardena
Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

***Abstract*—Organizations need a comprehensive threat profiling system that uses cybersecurity methods together with physical security methods because advanced cyber-threats have become more complex. The objective of this study is to implement deep learning models to boost organizational threat identification via human behavior assessment and continuous surveillance activities. Our method for human behavior analysis detects insider threats through assessments of user activities that include logon patterns along with device interactions and measurement of psychometric traits. CNN, together with Random Forest classifiers, has been utilized to identify behavioral patterns that indicate security threats from inside the organization. Our model uses labeled datasets of abnormal user behavior to properly differentiate between normal and dangerous user activities with high accuracy. The physical security component improves surveillance abilities through the use of MobileNetV2 for real-time anomaly detection in CCTV video data. The system receives training to detect security breaches and violent and unauthorized entry attempts, and specific security-related incidents. The combination of transfer learning and fine-tuning methodologies enables MobileNetV2 to deliver outstanding security anomaly detection alongside low power requirements, thus it fits into Security Operations Centers operations. Experiments using our framework operate on existing benchmark collection sets that assess cybersecurity, together with physical security threats. Experimental testing establishes high precision levels for detecting insider threats along with physical security violations by surpassing conventional rule-based methods. Security Operation Centers gain an effective modern threat profiling solution through the application of deep learning models. The investigation generates better organization defenses against cyber-physical threats using behavioral analytics together with intelligent surveillance systems.**

Keywords*—Deep learning; physical security; human behavior analysis; security operation centers; threat profiling

I. Introduction

Traditional Security Operations Centers must face challenges to monitor advanced and complex threats because cyber threats continue to grow sophisticated in the current digital environment. Online criminals keep discovering fresh ways to break into company systems through a combination of techniques using digital and physical assault methods. The strong demands of evolving security threats need organizations to adopt proactive security measures for both cyber and physical domains before damage occurs [1].

Organizations must tackle the demanding task of uncovering malicious activities conducted by authorized personnel, including stakeholders and trusted individuals who work inside their systems. Central to the challenge of detects to implementer threats is their origins from inside the organization because authorized personnel have access to confidential systems. Insider threats demonstrate themselves through activities such as data theft, Privilege Abuse, sabotage, and fraud. Advantageous security protocols based on rule detection and access systems cannot discover insider threats because these threats demonstrate unpredictable patterns. Current organizational security needs better analytical approaches to recognize risky personnel actions before critical damages can take effect [2].

Physical securities stand as a necessary component that organizations must address when managing threats. The safety of employees and the security of the infrastructure are seriously threatened by unauthorized access, acts of violence, and suspicious activity on company property. The security solutions of conventional assessment by CCTV and manual guards depend on human conduct, which introduces the risk of human errors and response delays. Security employees fail to identify crucial details because weakness interferes with their work, together with distractions and large monitoring tasks. Autonomous deep learning-driven surveillance systems can improve the identification of threats by continuously scanning live video streams for irregularities and sending out notifications when they see suspicious activity [3].

An artificial intelligence framework that connects human posture examination with physical security supervision through deep learning will address these challenges. The new system uses machine learning algorithms to examine user actions together with logon behavior along with device usage patterns, and personality methods for detecting abnormal insider threat indicators [4]. CNNs and Random Forest classifiers used together within the framework produce a better and trustworthy framework to detect abnormal behavior, which outperforms conventional security measures based on static rule sets.

The framework includes AI-based video surveillance as an additional measure to improve physical defense operations. The system employs MobileNetV2 as its core deep learning model to inspect and identify violent or suspicious occurrences in real-time CCTV broadcasts. The model operates optimally as an edge computing system to offer real-time anomaly detection

without delays in detection times [5]. The automated system detects threats more quickly, thanks to automatic response capabilities that operate without human supervision, which enhances an organization's total security position.

Furthermore, combining physical security systems with AI-driven cybersecurity provides a more comprehensive method of managing organizational threats. The connections between cyber threats and physical threats keep expanding because incidents in one area tend to affect the other. A physical breach of a data center may result in data breaches, and unauthorized actors can disable security cameras and enter restricted areas by exploiting compromised digital credentials. Through deep learning-based unification, security organizations obtain the power to face threats within a single unified framework [6].

The study improves current Security Operations Centers (SOCs) by offering a novel, unified framework that dramatically increases security management across virtual and physical infrastructures. While previous research has independently investigated insider threat detection and violence recognition, there is a clear gap in comprehensive systems that cover both areas simultaneously on a single platform. This research fills that gap by combining deep learning approaches to address real-time monitoring and threat identification tasks in a seamless way. The suggested system achieves high accuracy, efficiency, and quick response capabilities by utilizing lightweight yet robust models such as MobileNetV2 for physical violence detection and CNN for insider threat evaluations [2], [3]. The dual-layered detection method increases not only the detection and prediction of insider threats, but also operational readiness for physical security violations. Through this integration, the research contributes significantly to the development of Next-Generation SOCs by providing a comprehensive solution that addresses the limitations of separated traditional security approaches, allowing for effective detection, alerting, and management of complex cyber-physical threats.

The following sections specify the research structure: Section II presents an evaluation of previous works, together with their weak points. The proposed methodology in Section III reveals how deep learning models and techniques serve threatening behavior profiling purposes. The experimental findings, together with framework performance evaluation, appear in Section IV. The study concludes its discussion with essential research outcomes and suggestions about future investigation efforts in Section V.

II. Literature Review

A. Analyzing Human Behaviour to Identify Insider Threats

Every organization faces serious threats through insider actions that risk essential information attributes, including confidentiality, as well as integrity and availability. The threats come both from deliberate malicious insiders as well as unintentional cases due to employee negligence. Research indicates that insider security threats comprise a major portion of breach incidents because surveyed companies reported that 61% of their operations had internal breaches in one year. Insider threats proved equally detrimental to data security breaches by causing 23% of incidents, equalling the number of incidents caused by cyberattacks. Insider threats cause increased damage in present times because employees can now access sensitive systems directly, so they create both financial and reputational damage to organizations. To properly handle this escalating security issue, organizations must implement proactive mitigation techniques like behavioral analysis based on artificial intelligence threat identification [7].

"Using Hybrid Algorithms between Unsupervised and Supervised Learning" tackles the persistent and complex issue of recognizing insider threats, who typically elude traditional security measures due to their authorized access to internal systems. The primary originality of this study is its hybrid method, which effectively blends unsupervised machine learning for anomaly detection and supervised training for pattern classification to improve accuracy and reduce false positives. The methodology includes feature engineering utilizing user behavior profiling and a multi-layered detection process. With a remarkable accuracy rate of 86.12%, the model demonstrated significant potential for practical application in threat detection systems [8]. Among its benefits are high detection performance, improved learning from labeled as well as unlabeled data, and flexibility in a variety of organizational settings. One significant disadvantage is that computational complexity and expense may limit scalability and real-time deployment.

Traditional organizations detect insider threats using three procedures, which include rule-based monitoring and user activity logging, as well as anomaly detection systems [9]. These strategies fail to detect modern advanced attackers who use deception methods to circumvent secure monitoring systems. Statistical anomaly detection techniques are helpful, but they have scalability problems and high false positive rates [10]. Although detection accuracy has increased with the creation of machine learning models, there are still issues with striking a balance between accuracy and false-positive results.

Sridevi et al. (2023) examine the usage of deep learning models, specifically Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN), to detect insider threats by examining user behavior data over time. By capturing a long-term temporal link in user behavior, their approach enables the model to identify subtle changes that typical models could overlook and that appear gradually over time. The strength of this method is its ability to effectively profile user behavior utilizing patterns seen over extended periods and model sequential information. It is suitable for batch or historical research since evaluation findings show that it performs well when analyzing temporal patterns using simulated datasets [11]. However, the lack of spatial context evaluation (e.g., device or location-based behavior), explainability features, and the inability to support real-time threat identification may limit the model's practical deployment as transparency in operational environments, where clear and concise results are needed.

To address the insider threat issue, the study "Supervised and Unsupervised Methods to Detect Insider Threat from Enterprise Social and Online Activity Data" looks at digital traces of online behavior, including as emails, login times, and social media activity. This approach is intriguing because it uses a behavioral dataset from multiple state sources and includes supervised and unsupervised models to detect deviations from conventional

behavior patterns. The accuracy of the final assessment, which was 73.4%, is encouraging by considering the complexity and unpredictability of social behavior data [12]. The main advantage of this method is that it can be used with real company communication data, enabling more thorough behavioral insights. Unfortunately, data from online activities is also less accurate and may overfit due to its diversity and noise. Moreover, explainability and real-time detection, two factors that are becoming increasingly crucial for modern security operations, are not given priority.

Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) are combined in a hybrid deep learning model proposed by Saaudi et al. (2021) for insider threat detection, with a focus on temporal-spatial sequences. LSTMs are used to train through temporal sequences, enabling the model to recognize complex patterns over time. CNNs are used to extract characteristics from geographic data, such as device usage, file access patterns, and user behaviors. This hybrid method effectively maintains both temporal and geographic data thanks to its robust behavioral profiling and adaptability to changing threat behaviors. The model performs well in pattern recognition, but it lacks key features for practical application, such as a real-time threat detection system, continuous learning to adapt to new threats, along with explainable AI (XAI) to make decisions transparent. These weaknesses make the model challenging to apply in real-world scenarios where timely and intelligible insights are required for effective threat reduction [13].

Deep learning's ability to extract high-dimensional features and recognize patterns has completely changed cybersecurity. CNNs have been proven successful in analyzing extensive user activity datasets for detecting insider threat marker behaviors according to research [14]. Recurrent Neural Networks (RNNs) serve as a common tool for sequential behavior modeling in cybersecurity applications [15]. Moreover, Generative Adversarial Networks (GANs) are currently being investigated to create synthetic attack situations, therefore enabling models to generalize more in practical environments [16].

Insider threat detection benefits from behavioral profiling, which has become a proven method of detection. Analyzing device interactions, psychometric characteristics, and login behaviors has been shown in studies to improve detection accuracy [17]. RNNs coupled with LSTM systems play a vital role in sequential anomaly detection when used for insider threat detection. Research results showed that LSTM networks achieved better time-based behavioral deviation tracking, which enhanced the timing of threat identification. Real-time user action monitoring through RNN-based models has become a solution for detecting unauthorized data exfiltration patterns [18]. The detection of threats has received major improvements through recent advancements in transformer-based architecture development. Security logs receive analysis from BERT-based models that achieve better accuracy than traditional ML models for identifying insider threats [19].

The development of insider threat detection models depends heavily on the process of feature engineering for better accuracy. Research underlined how well ML models performed when including engineering elements, contextual embeddings, user risk assessments, and deviation measurements. Furthermore, investigating methods to maximize detection efficiency are automated feature selection methods applying genetic algorithms [20]. Using artificial intelligence models for real-time detection of threats calls for effective interaction with web-based systems like Flask. Studies showed that real-time monitoring dashboards made possible by Flask-based systems might increase the operational effectiveness of Security Operations Centers [21]. Reducing false positives and improving model interpretability will be the main goals of future insider threat detection studies. Techniques integrating unsupervised and supervised learning in hybrid artificial intelligence can improve detection accuracy even more. Another interesting path to guarantee the accuracy of data in identifying insider threats is the acceptance of blockchain-based systems for safe audit logging [22]. Table I summarizes important gaps in the research on insider threat detection.

TABLE I GAPS IN LITERATURE ON INSIDER THREAT DETECTION

Ref	Description	Novelty	Strengths	Weaknesses
[15]	Analyzes the use of hybrid (supervised + unsupervised) threat detection techniques utilizing deep learning techniques.	Combining several models and traditional protection to present a single deep learning framework.	Combines CNNs and RNNs to detect threats in real time from a variety of data sources.	Lacks benchmark comparisons, performance measurements, and actual evaluation.
[16]	Proposes a GAN-based model to detect identity theft within e-governance systems.	Applies adversarial training specifically to e-governance identity protection—a new application area.	Improves cyber identity theft detection in e-governance by utilizing adversarial networks.	No performance metrics, dataset details, or evaluation results provided.
[17]	Reviews how GANs enable both new cyberattacks and defenses (anomaly detection).	Highlights dual use of GANs in offense and defense across multiple threat domains.	Wide coverage of GAN deployment scenarios emphasizes evolving attack surfaces and defensible techniques.	Lacks quantitative evaluation or implementation of roadmaps.
[18]	Uses graph neural networks especially to identify trends in cyberthreats.	Uses temporal graph data instead of static logs.	Captures complex behavior patterns, adaptable.	No benchmarks, lack real-world validation.
[19]	Evaluates ML and BiLSTM models on CERT r4.2 for insider threat detection.	Combines standalone and sequential features using BiLSTM.	Use supervised learning to detect insider threats with high accuracy.	Unbalanced data sets in the actual world can cause performance to decline.

B. Profiling Threats with Physical Security Systems

Particularly in the field of real-time violence detection, developments in deep learning technology have accelerated the growth of video surveillance systems. The core of contemporary intelligent surveillance systems is edge computing solutions, temporal analytic frameworks, and transfer learning models, which are investigated in this study.

Deep learning techniques have lately shown remarkable success in performing violence detection inside surveillance systems. With CNN and LSTM architectures, Sharma et al. (2021) created an integrated system with 94.65% accuracy using their combined dataset. Their method used trained CNN models for extracting features, then used LSTM networks to examine temporal sequences, therefore allowing real-time analysis of footage and instantaneous alert to authorities [23]. Likewise, in situations involving obstructions and dense locations, especially, a 2025 study using MobileNetV2 paired via Temporal Squeeze and Excitation Blocks overcomes limits in current violence detection technologies. When violent events are found, our technology analyses live footage from surveillance cameras and transmits quick reports to police over a Telegram network [24].

Because MobileNetV2 strikes a mix between accuracy and efficiency, it has become a favored architecture for violence detection. With 94.5 per cent accuracy, 93 per cent precision, and a 94.9 per cent F1 score in separating between aggressive and non-violent actions, in 2024, the execution integrating MobileNetV2 with Bidirectional LSTM networks attained amazing performance. Maintaining great detection accuracy, the lightweight character of the design makes it especially appropriate for deployment in settings with limited resources [25].

Using MobileNetV2 as a previously trained algorithm for violence detection, Chandran (2023) further included image enhancement techniques and facial detection to provide complete warning capabilities [26]. This system shows the useful implementation of transfer learning algorithms in real-world surveillance situations by capturing images of people engaged in violent behavior and forwarding them together with place data to experts via Telegram [27].

For a temporal pattern associated with violent events, the combination of LSTM network structures with CNNs has shown success. Hamza's work used CNN+LSTM architecture to detect anomalies in surveillance films, thereby allowing the system to grasp the sequential character of acts defining aggressive behavior [28]. A 2024 study emphasizing how bidirectional LSTM networks paired with MobileNetV2 can efficiently capture temporal dependencies inside sequential video frames validates this method even further. By means of their combined use, these technologies generate an intelligent system able to precisely detect violent behaviors by use of spatial characteristics and their progression over time [29].

Emerging as a crucial part of contemporary surveillance systems, edge computing solves bandwidth and latency problems. Edge-enabled security camera systems can quickly process video data, according to a 2025 report, therefore offering real-time, instantaneous actionable insights and completely eliminating delays brought about by data being sent to faraway servers. Applications needing quick response, like those involving the identification and handling of violent events, depend on this capacity [30].

Edge computing transmits only pertinent data instead of constant raw footage, therefore providing notable bandwidth efficiency. According to a 2023 study, every video camera may examine its own film and provide only relevant data to cloud servers, therefore lowering bandwidth usage and allowing quicker responses for important security events7. Scaling surveillance networks over several sites calls for this method especially helpful [31]. Edge computing also improves data privacy by locally processing sensitive videos, therefore lowering exposure to intrusions during data migration. For companies giving data security and regulatory compliance first priority, this makes it appealing [30]. Table II summarizes important gaps in the research on physical security systems.

TABLE II GAPS IN LITERATURE ON PHYSICAL SECURITY SYSTEM

Ref	Description	Novelty	Strengths	Weaknesses
[27]	Fuses unsupervised outlier scores across time granularities with supervised XGBoost on CERT R4.2 dataset.	Treats anomaly scores as feature enhancements for supervised detection with low compute cost.	Reduces computation, increases accuracy, and manages multi-granularity data intelligently.	Limited to CERT dataset; lacks deployment insights and external validation.
[28]	Real-time video violence detection using a pre-trained Xception CNN + LSTM, with mobile app alerts.	Spatial–temporal DL integration in a deployed architecture with mobile notifications.	Extremely high accuracy, alert app included.	Tested only on curated datasets; no real-world deployment; requires GPU.
[29]	Uses a lightweight two-stream CNN for anomaly detection in surveillance video, optimized for IoT.	Combines AIoT with spatial–temporal deep learning.	Efficient for edge deployment; handles large-scale video data.	Limited benchmarking and real-world validation.
[30]	Utilizes Hyperledger Fabric blockchain to log CCTV metadata and video hashes, ensuring tamperproof surveillance streams.	Merges IoT CCTV with blockchain to create immutable, verifiable video integrity in smart cities.	Immutable audit trail: decentralized ledger prevents forgery or single-point failure.	No performance metrics or scalability analysis lacks real-world deployment data.

III. METHODOLOGY

A. Threat Profiling with Human Behavior Analysis

The rising complexity of insider threats calls for the application of sophisticated data-driven approaches. This study gives the advantage of the CERT Insider Threat Identification Dataset, well known for simulating actual organizational risks. The collection comprises logs of user activity, including device connections, file actions, logon and logoff tracks. These logs offer important new perspectives on both expected and unusual behavior inside a company. Examining these trends helps us to find early signs of insider threats.

Ensuring that the dataset is organized and useful for deep learning algorithms depends critically on data preparation. For sequential logs, values that were missing in the dataset were resolved with both backward and forward filling techniques; numerical attributes were handled via mean or median imputation. Additionally, present in the dataset are psychometric features that help to spot behavioral deviations. Labelling for typical features like risk levels and rapid coding for categorical factors like department and position title assisted in the encoding of these attributes. To identify suspicious activity, login timestamps were also transformed into time-based attributes, including frequency variation per day and an anomaly rating. Techniques of standardizing and normalizing helped to guarantee reliability in feature scaling, hence improving model convergence during training.

Effective threat detection requires significant feature extraction. Essential elements consist of:

- Examining login trends helps to identify irregularities such as logons from unexpected sites or outside regular business hours. Different from the usual behavior of an employee could point to a compromised consideration or malicious intent.
- Monitoring external device usage, that of USB connections and file transfers helps one find illegal access. Frequent trips to outside devices or efforts to access restricted data could lead to an insider danger.
- Tracking user behaviors over time helps the system identify strange patterns. An employee downloading a lot of confidential information right before logging out, for instance, could show hostile intent.
- Mapping user activities to psychological features according to the OCEAN model assists in identifying vulnerable to risk-taking people. Workers displaying indicators of stress or discontent could be susceptible to participating in harmful behavior, so prompt action is more possible.
- Graph-based profiling analyzing user interactions as graphs, helps identify unusual connections and illegal data sharing inside the company. In communications, anomalies in patterns could point to insider threats or communication.
- Deep learning embeddings translate behavioral patterns into numerical representations, thereby enhancing the capacity of the network to identify small variations from usual activity.

Insider threat detection works using a CNN-Random Forest combined model. By extracting sequential patterns from organized behavior records, the CNN component finds user behavior trends concealed otherwise. CNNs are excellent for anomaly detection since they can learn intricate temporal correlations from log sequences. The graphical information is processed using the Random Forest classifier, which lowers overfitting and manages a combination of continuous and categorical variables, thereby enhancing classification performance. This hybrid method guarantees efficient analysis of both structured and unstructured behavioral aspects, thus strengthening an accurate threat identification system.

The dataset was split in half with 70% used for training as well as 30% testing to guarantee the model fits new data. With a batch size between 32 and 20 epochs, a CNN model was developed using an Adam optimizer used to dynamically change learning rates for improved integration. Using grid search cross-valuation to determine ideal depth and selection of features criteria, Random Forest was set up with 100 decision trees and hyperparameter modification was conducted. By balancing false positives with false negatives, this optimization technique improved system dependability for practical use.

Several important benchmarks helped to evaluate the model's performance. With the model scoring 95%, which matched its general accuracy in forecasts, accuracy stood out. With an outcome of 0.95 for instances that were negative while 0.93 for positive ones, precision was high, meaning that the identified threats were valid. The model thereby reduced false positives successfully. Recall, which measures the model's capacity to detect real threats, performed really well for negative outcomes (0.99) but fewer for cases that were positive (0.79), therefore suggesting some missed threats. The F1-score was computed to strike a mix between accuracy and recall, so offering a complete assessment that addressed the class imbalances sometimes observed in insider threat identification. This mix of indicators guarantees the model detects and reduces false positives as well as missed threats in excellent performance. For each behavior category, Table III displays the CNN model's precision, recall, and F1-score. It performs well in identifying typical behavior and has a slightly reduced recall for insider threats.

TABLE III CNN MODEL EVALUATION RESULTS

Class	Precision	Recall	F1-Score
0 (Benign)	0.95	0.99	0.97
1 (Malicious)	0.93	0.79	0.86

The structure of the Insider Threat Identification System is shown in Fig. 1. The final prediction is produced by processing user-input data using a Flask application, analyzing it by applying a CNN along with Random Forest models based on the CERT dataset, then combining them via an ensemble model.

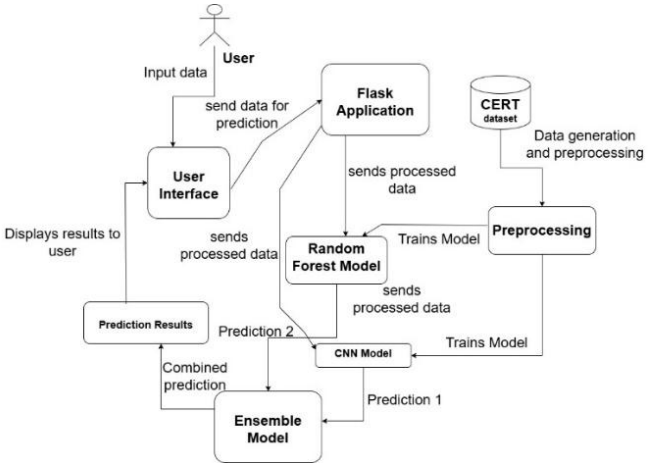


Fig. 1. Insider threat detection system diagram.

B. Threat Profiling with Physical Security System

The deep learning model underwent training and evaluation through an extensive data preprocessing pipeline. The pipeline encompassed numerous crucial phases to standardize input data, maintain vital temporal and spatial information, and improve the model's generalization across various contexts. The initial phase entailed frame extraction; individual frames were methodically sampled from each video clip at a uniform rate to maintain temporal coherence. This enabled the model to learn from the dynamics and evolution of actions over time, capturing significant temporal patterns essential for differentiating between violent and non-violent behaviors.

Spatial normalization was performed on all retrieved frames, shrinking them to 224x224 pixels, which is the standard input size mandated by the MobileNetV2 architecture. Pixel normalization adjusts pixel intensity values to a standardized range, often [0, 1] or [-1, 1], based on the input specifications of the model.

Data augmentation methods were utilized on the training set to replicate heterogeneity in actual video recording scenarios, mitigating overfitting, and enhancing the model's resilience to diverse camera angles, lighting conditions, and orientations.

Preprocessed frames were structured into fixed-length sequences to provide the final input format for the model. These sequences are the principal input for the model, allowing it to examine both the spatial content of individual frames and the temporal motion patterns and context.

The preprocessing methods created a robust framework for developing an efficient video classification system that can identify violent behaviors in intricate real-world settings.

The MobileNetV2 model was selected for the feature extraction phase of the model pipeline because of its superior balance between computational efficiency and feature representation accuracy. This option is especially appropriate for violence detection systems functioning in real-time or resource-limited settings, such as embedded systems or surveillance equipment. MobileNetV2 employs depth-wise separable convolutions and inverted residual blocks to minimize parameters and computational expenses while effectively capturing intricate patterns in the data. This enables the model to derive superior spatial properties from video frames while maintaining a lightweight and efficient design.

The feature extraction technique comprised multiple substeps, including transfer learning, fine-tuning with a domain-specific dataset of violent and non-violent video clips, and the generation of feature maps from each video frame. These maps offer a comprehensive, organized representation of the visual material in each frame, essential for further temporal modelling.

To facilitate the transition from spatial to temporal analysis, the output of MobileNetV2's final convolutional layer was converted into feature vectors. This dimensionality reduction procedure converts multi-dimensional tensor outputs into a uniform and compact format appropriate for input into temporal modelling components, such as recurrent neural networks (RNNs) or temporal convolutional networks (TCNs). These components examine the temporal evolution of spatial features, allowing the system to identify violent sequences instead of solitary visual patterns.

A hybrid CNN-LSTM architecture was created to analyze the intricate spatiotemporal dynamics of aggressive behavior in video sequences. This architecture incorporates spatial feature extraction, which examines individual video frames, and temporal sequence modelling, which captures the evolution of events across time. The model is more adept at recognizing violence as a dynamic and context-dependent phenomenon rather than solely depending on static visual indicators.

The initial phase encompasses spatial feature extraction, facilitated by the MobileNetV2 backbone. This lightweight convolutional neural network independently processes each frame, extracting high-level visual features that encapsulate significant spatial information. These characteristics may encompass critical signs of aggression, including hostile attitudes, abrupt movements, and close closeness between people. Every video frame is converted into a comprehensive feature vector for subsequent analysis.

The output feature vectors from MobileNetV2 are input into a Long Short-Term Memory (LSTM) layer to record the temporal dependencies between successive frames. LSTMs are engineered to process sequential input and can preserve long-term dependencies via gated memory cells, rendering them especially effective for violence detection.

After the LSTM layer, the model has a classification head using a fully connected dense layer with a SoftMax activation function. The last layer translates the stored temporal information into binary class probabilities, signifying whether the input sequence pertains to a violent or non-violent event.

The hybrid model was optimized for training with a batch size of 32, utilizing the Adam optimizer at a learning rate of 0.0001, implementing early halting with a patience of 10 epochs, and employing the categorical cross-entropy function for multi-class classification problems.

A hybrid CNN-LSTM model was created to shift from a research context to a real-time monitoring system appropriate for public surveillance applications. The frame buffer management module is a fundamental element that employs a

sliding window technique to manage live video streams. This technique guarantees that the model consistently accesses the most recent temporal context for inference, while reducing memory consumption and latency.

Inference optimization was conducted using TensorFlow Lite, a lightweight and efficient deep learning framework tailored for mobile and edge devices, to satisfy real-time performance requirements. The optimized model can handle video streams at 25 to 30 frames per second (FPS), which corresponds with normal video frame rates and facilitates near-instantaneous detection of violent behavior in real-time recordings.

An alarm system was integrated into the pipeline to guarantee prompt reactions to identified incidents. The system was designed to initiate a response solely upon the detection of aggression in 20 consecutive frames, so enhancing its resilience by confirming the continuity of violent behavior across time.

A notification system was created and incorporated utilizing the Telegram messaging network. Upon the activation of an alert, the system transmits a real-time notification to designated recipients, encompassing critical information such as the event's timestamp, geolocation coordinates (if accessible), and collected frame images that graphically record the alleged violent incident. This real-time notification feature guarantees that security professionals or emergency responders are swiftly alerted and provided with actionable context to evaluate and address the problem. Fig. 2 shows the Physical Security Component Diagram, which outlines the real-time violence detection pipeline using MobileNetV2, combined with CCTV streams and alert systems.

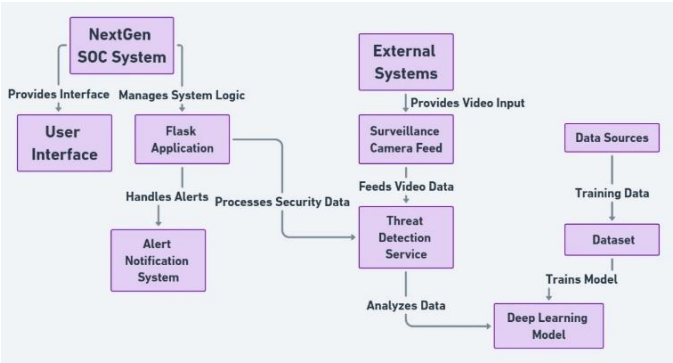


Fig. 2. Physical security component diagram.

IV. Evaluation and Results

A. Performance of Insider Threat Detection

The indicated insider threat detection system proved highly effective in spotting unusual user behavior. The CERT Insider Threat Identification Dataset offered several real-world examples of insider threats; therefore, the CNN-Random Forest ensemble approach was evaluated on that dataset. With an accuracy rate of 95%, the CNN approach alone proved able to efficiently extract significant patterns from organized behavioral records. Using convolutional neural network layers, the CNN found small variations in user behavior, including illegal log-in access, unexpected working hours, and unusual data access patterns.

Using conventional evaluation measures, including precision, recall, and F1-score, the model's performance was evaluated. With a precision rating of 0.93, the model found that most identified threats were real insider attacks. With a score for recall of 0.79, the model was able to balance false positives with a significant number of real threats. The strong F1-score confirmed that the model could efficiently manage imbalanced datasets, therefore guaranteeing that real risks were found without improperly highlighting benign behaviour. CNN Model accuracy and model loss graphs are shown in Fig. 3 and Fig. 4, respectively.

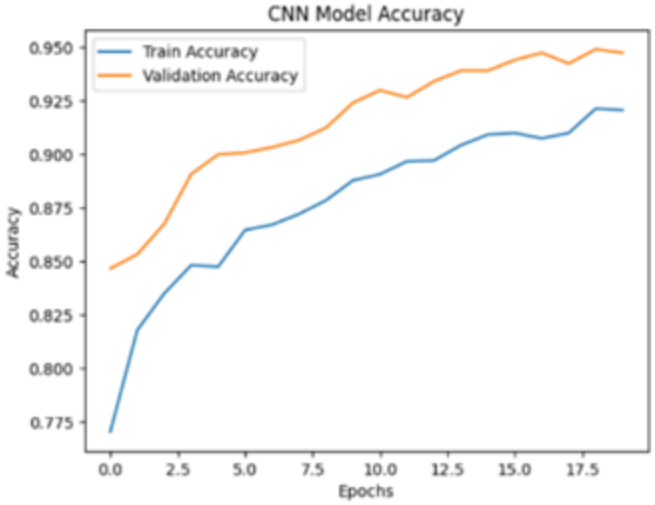


Fig. 3. CNN model accuracy.

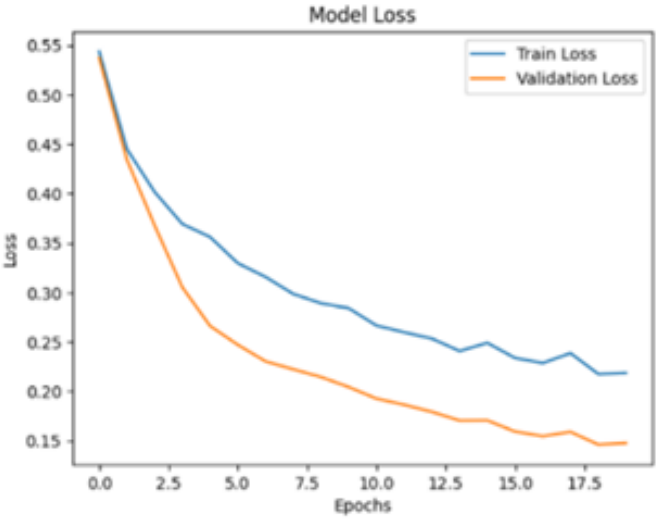


Fig. 4. Model loss.

The system combined the Random Forest classifier with a CNN (Convolutional Neural Network) within an ensemble model to improve classification performance. This combined strategy made use of Random Forest's ability to make decisions based on user behavior patterns and CNN's power in feature extraction. The Random Forest effectively processed intricate graphical data and decreased false positive rates by combining several decision trees. Threat detection became more accurate as a result of the ensemble model's effective resolution of the

problem of misclassifying typical user behavior as threats. A strong detection system was created by combining CNN with Random Forest, which decreased classification mistakes and improved overall dependability. The ensemble model's remarkable 98% accuracy rate highlights how well it profiles organizational risks.

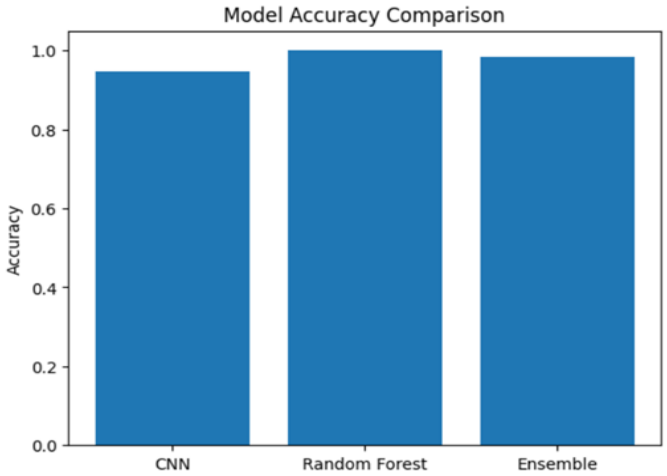


Fig. 5. Model accuracy comparison.

The accuracy comparison between each model is shown in Fig. 5, emphasizing the variations in performance. These findings underline how well deep learning-based profiling detects insider threats. Although the classifier using Random Forests improved classification reliability by lowering overfitting and hence enhancing decision-making, the CNN component succeeded in feature extraction. The system is a reasonable choice for implementation in actual Security Operations Centers (SOCs) since it can reach great accuracy with minimum false positives. Future developments might center on including explainable artificial intelligence methods to increase openness and improve detection accuracy even more. Table IV shows the comparison of model performance with existing literature.

B. Performance of Physical Security Systems

The evaluation measures and findings of the suggested physical security system with real-time detection of violence are presented in this part. Accuracy, recall, F1 rating, and real-time operational indicators evaluated the MobileNetV2-based combination CNN-LSTM model. The outcomes show how good the system is in spotting violent events and differentiating them from regular operations.

Achieving great accuracy as well as robust classification findings, the MobileNetV2-based hybrid approach showed amazing performance in spotting violent activity. Summarized below are important evaluation criteria. Fig. 6 shows the total accuracy versus total validation accuracy chart.

The system's capacity to identify ordered anomalies in footage from CCTV was much improved by including LSTM layers. Analyzing temporal relationships between consecutive frames in the LSTM layers let the model identify violence as a series of acts instead of individual events. In situations involving subtle or extended aggressive behaviors, when spatial elements alone could not adequately depict the environment, this enhancement was especially clear. Fig. 7 shows the confusion matrix for the physical security system.

TABLE IV COMPARISON OF MODEL PERFORMANCE WITH EXISTING LITERATURE

Study	Model Used	A = Accuracy P = Precision R = Recall	Analysis
Hybrid Algorithms (2024) [8]	Hybrid (Supervised + Unsupervised)	A - 92.8% P, R - Not Provide	Combines clustering and classification to improve insider threat detection.
Olajide et al. (2024) [9]	Hybrid ML-based UBA models	A - ~85% P, R – Not Provided	Discusses UBA-based insider threat detection combining ML techniques.
Sridevi et al. (2023) [11]	Deep LSTM / RNN	A - 96.3% P, R – Not Provided	Sequential behavior is well captured by the suggested LSTM/RNN-based method.
Gavai et al. (2015) [12]	Supervised + Unsupervised	A - 73.4% P – Not Provided R – 72%	Uses enterprise social and activity data for behavioral threat detection.
Saaudi et al. (2019) [13]	CNN + LSTM	A - 95.9% P – 95.98% R – 95.9%	Combines spatial and temporal modeling to capture insider behavior cues.
Our Model	CNN + RF Ensemble	A - 98% P – 98% (0), 99% (1) R – 99% (0), 92% (1)	Random Forest's comprehensive decision-making capabilities with CNN's deep pattern recognition capabilities.

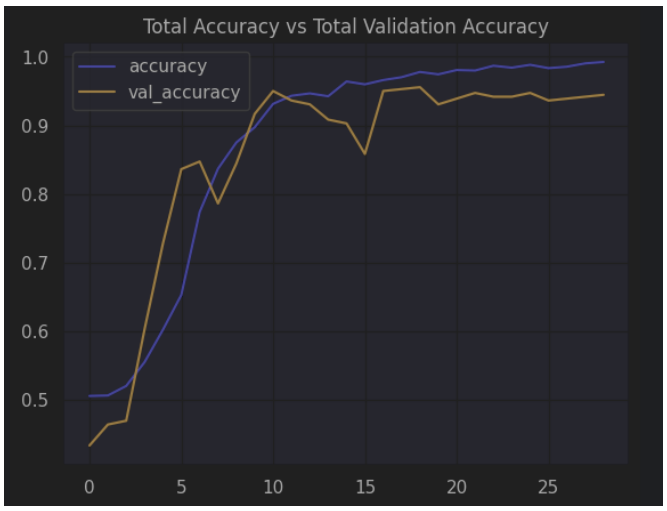


Fig. 6. Model accuracy.

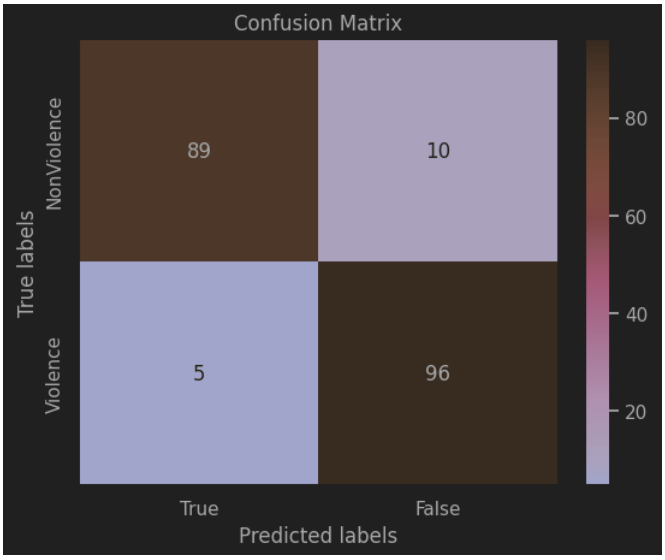


Fig. 7. Confusion matrix.

The evaluation findings demonstrate that the suggested physical security framework is quite successful in real-time violent incident detection:

- Reach 99.16% accuracy, above standards for systems of violence detection.
- Using LSTM layers, strong temporal abnormality detection is shown.
- Run Functioned with low latency and high frame processing speed in real-time settings.
- Presented scalable options for edge device deployment keeping great accuracy and recall rates.

These findings confirm the system's feasibility for useful application in surveillance settings and provide a trustworthy instrument for improving physical security by means of intelligent video analysis.

Prior research on physical security threat identification utilizing CNN architecture has indicated accuracy rates ranging from 70% to 75%, underscoring the difficulties in identifying dangers within surveillance data. These models frequently exhibit elevated false positive rates and inadequate generalization to real-world contexts. Conversely, our MobileNet-based model attained a markedly better accuracy of 98.2%, highlighting enhanced feature extraction, computational efficiency, and durability.

A detailed performance comparison is presented in Table V, highlighting the accuracy and Analysis of previous CNN-based approaches compared to our MobileNet-based model. As shown, our model demonstrates superior performance while addressing key challenges such as false positives and real-time inference.

TABLE V PERFORMANCE COMPARISON OF CNN-BASED MODELS FOR PHYSICAL SECURITY THREAT DETECTION

Paper Title	A-Accuracy P-Precision R-Recall	Analysis
Hybrid Insider Threat Detection (Yi & Tian, 2024)	A-98.03% P- 97.59% R- 96.83%	High accuracy and efficiency with low compute cost; effective for insider threat detection using hybrid ML approach.
Violence Detection Using CNN & LSTM (Sharma et al., 2021)	A-95.15% P- 94.87% R- 95.20%	Well-balanced detection system for violence in video; effective spatiotemporal feature learning using CNN + LSTM.
AIoT-Assisted Two-Stream NN (Zhou et al., 2021)	A-95.12% P-95.34% R-94.98%	Impressive performance for surveillance anomaly detection; optimized for IoT and edge deployment scenarios.
Blockchain-based CCTV Verification (Khan et al., 2020)	Not Provided	Focuses on blockchain-based data integrity, not ML classification; lacks typical performance metrics like accuracy or recall.
Our Model	A - 98.2% P-95% (0), 91% (1) R-90% (0), 95% (1)	Uses MobileNetV2 to deliver excellent accuracy in real-time physical violence detection; it is optimized for the deployment of lightweight edge devices with timely warning capabilities.

V. CONCLUSION

This study connects human behavior analysis with physical security surveillance to show how effective deep learning is at profiling organizational threats. The suggested solution uses MobileNetV2 to increase real-time monitoring for physical security risks and a CNN-Random Forest hybrid model to evaluate behavioral patterns for better insider threat identification. These models work together to create a thorough, multi-layered security solution that can identify security threats from inside as well as from outside.

The suggested method, however, is subjected to a number of presumptions and restrictions. First, it assumes that input data (such as video and behavioral logs) will always be available and of high quality, which is not necessarily the case in real-world settings. Second, even though the CNN-Random Forest method is more interpretable, it might not be as transparent as more straightforward models, particularly in situations with large stakes. Furthermore, psychometric analysis may oversimplify complicated psychological factors and cultural differences by assuming that human behavior can be reliably classified into threat or non-threat patterns. Additionally, the existing method depends on centralized processing, which could cause slowness in scenarios with limited resources or large-scale deployments. The system's ability to effectively reduce false positives while keeping high accuracy makes it appropriate for dynamic enterprise security contexts, despite these constraints. The system's modularity and scalability further improve its capacity to adjust to changing security threats.

Future iterations of this study will focus on various developments to overcome existing constraints and increase functionality. Integrating edge computing is a crucial step in enabling quicker, localized processing of surveillance data, which lessens reliance on centralized resources and speeds up response times. To better understand context and detect complex threat behaviors early on, we also want to deploy Graph Neural Networks (GNNs) to record and examine relational patterns across people, devices, and activities. Additionally, implementing Explainable Artificial Intelligence (XAI) methodologies would increase decision-making transparency and enable security analysts to better understand, trust, and respond using AI-driven insights.

By providing a threat detection system which is not only effective and accurate but also scalable, adaptable, and interpretable, these improvements seek to secure the fundamentals of Next-Generation Security Operations Centers (SOCs). In an increasingly complicated digital environment, this work adds to the continuous effort to assist enterprises in proactively identifying, mitigating, and managing cybersecurity threats, both internal and external.

REFERENCES

[1] I. A. ,. K. B. a. K. S. Usman Tariq, "A Critical Cybersecurity Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review," vol. 23, no. 8, 2023.

[2] A. Saquella, "A critical challenge: Understanding and addressing insider threats," p. 6, 27 January 2025.

[3] "Guide to physical security controls, planning, policies and measures," 2025.

[4] Y. Z. B. Z. ,. D. D. ,. Y. L. Haitao Xiao, "Unveiling shadows: A comprehensive framework for insider threat detection based on statistical and sequential analysis," ELSEVIER, vol. 138.

[5] A. H. M. Z. A. Z. L.-C. C. Mark Sandler, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," arXiv.org, 2018.

[6] M. Hinz, "AI in Physical Security: Efficiently Transforming Safety," HBS, 2025. [Online]. Available: https://www.hbs.net/blog/ai-in-physical-security.

[7] M. W. G. P. K. v. d. S. Karen Renaud, "VISTA: An inclusive insider threat taxonomy, with mitigation strategies," Science Direct, vol. 61, no. 1, 2024.

[8] Y. T. Junkai Yi, "Insider Threat Detection Model Enhancement Using Hybrid Algorithms between Unsupervised and Supervised Learning," vol. 13, no. 5, p. 17, 2024.

[9] O. S. A. B. K. A. ,. O. Olajide O. Ogunbodede, "Insider Threat Detection Techniques: Review of User Behavior Analytics Approach," International Journal of Research in Engineering and Science, vol. 12, no. 9, p. 9, 2024.

[10] B. Y. D. H. J. a. L. L. S. Zeadally, "Detecting Insider Threats: Solutions and Trends," Information Security Journal, vol. 21, no. 4, pp. 183-192, 2024.

[11] L. K. V. G. S. R. D. Sridevi, "Detecting Insider Threats in Cybersecurity Using Machine Learning and Deep Learning Techniques," in 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), 2023.

[12] K. S. D. G. J. H. M. S. a. R. R. Gaurang Gavai, "Supervised and Unsupervised methods to detect Insider Threat from Enterprise Social and Online Activity Data," Supervised and Unsupervised methods to detect Insider, p. 63, 2015.

[13] Z. A.-I. Y. T. C. F. Ahmed hasan Saaudi, "Insider Threats Detection using CNN-LSTM Model," in 2018 International Conference on Computational Science and Computational, 2019.

[14] M. O. Okafor, "Deep learning in cybersecurity: Enhancing threat detection and response," World Journal of Advanced Research and Reviews, p. 17, 2024.

[15] A. Sivanandam, B. Lavaraju and A. Sundaram, "Adversarial Networks for Enhanced Cyber Identity Theft Detection and Security in E-Governance Sector," IGI Global, 2025.

[16] M. S. A. T. K. G. I. A. U. J. Z. J.-h. Y. Md Mashrur Arifin, "A Survey on the Application of Generative Adversarial Networks in Cybersecurity: Prospective, Direction and Open Research Scopes," arXiv.org, 2024.

[17] J. REHA, "Cyber Threat Detection using Machine Learning on Graphs," Research Gate, 2023.

[18] P. Manoharan, "Supervised Learning for Insider Threat Detection," VU Research Repository, 2024.

[19] J. V. &. S. R. Khushboo Taneja, "Fraud-BERT: transformer based context aware online recruitment fraud detection," Discover Computing, 2025.

[20] R. R. Akula and G. S. N. Kumar, "Optimizing Feature Selection in Intrusion Detection Systems Using a Genetic Algorithm with Stochastic Universal Sampling.," vol. 16, no. 1, 2025.

[21] R. W. K. N. M. H. S. A. K. H. Yasmin Makki Mohialden, "Top Python-Based Deep Learning Packages: A Comprehensive Review," International Journal Papier Advance and Scientific Review, vol. 5, no. 1, p. 9, 2024.

[22] J. Y. a. Y. Tian, "Insider Threat Detection Model Enhancement Using Hybrid Algorithms between Unsupervised and Supervised Learning," MDPI, vol. 13, no. 5, 2024.

[23] B. S. S. N. V. T. K. J. Sarthak Sharma, "A fully integrated violence detection system using CNN and LSTM," International Journal of Electrical and Computer Engineering (IJECE), vol. 11, no. 4, p. 7, 2021.

[24] G. B. M. K. S. T. T. C. R. K. Rohini Temkar, "Automated Violence Detection in Surveillance Networks with Deep Learning," 2025.

[25] S. S. V. B. A. P. S. P. A. T. A. M. V. Lokeshwar Reddy, "Automated Human Violence Detection using MobileNetV2 and Bidirectional LSTM Networks," in Proceedings of the 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC-2024), 2024.

[26] M. Thomas and P. Balamurugan, "Real-Time Violence Detection and Alert System using MobileNetV2 and Cloud Firestore," 2024 2nd International Conference on Networking and Communications (ICNWC), 2025.

[27] H. Gupta and S. T. Ali, "Violence Detection using Deep Learning Techniques," in 2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), 2023.

[28] C. C. M. S. Waqas Sultani, "Real-world Anomaly Detection in Surveillance Videos," Research Gate, 2018.

[29] S. S. V. B. A. P. S. P. A. T. A. M. V. Lokeshwar Reddy, "Automated Human Violence Detection using MobileNetV2 and Bidirectional LSTM Networks," International Journal of Microsystems and IoT, vol. 2, no. 8, pp. 1059-1064, 2024.

[30] Luis Garcia, "The Future of Surveillance: Edge Computing with Analytics," NAVCO, 2025.

[31] L. Sapra, "Cloud Computing and Edge Computing For Scalability and Performance," Turkish Online Journal of Qualitative Inquiry (TOJQI), vol. 11, no. 4, p. 11, December 2020.

LaObese: A Serious Game Powered by Analytic Hierarchy Process for Culturally Tailored Childhood Obesity Prevention in Oman

Nurul Akhmal Mohd Zulkefli[1], Mukesh Madanan[2], Zainab Mohammed Al-Nahdi[3], Jayasree Radhamaniamma[4]
Department of Computer Science, Dhofar University, Salalah, Sultanate of Oman[1, 2]
Department of Mathematics and Sciences, Dhofar University, Salalah, Sultanate of Oman[3]
Oman College of Health Sciences, Salalah, Sultanate of Oman[4]

***Abstract*—Childhood obesity is a growing public health concern in Oman, yet culturally appropriate digital tools for early prevention remain scarce. This study introduces LaObese, a mobile application and serious game designed to prevent obesity in Omani preschool children. The name LaObese derives from the Arabic word "La" (meaning "no") and an abbreviation of "obesity", reflecting the game's mission to say, "no to obesity". LaObese integrates gamification and behaviour modification strategies to encourage healthy lifestyle practices from an early age. Targeting children aged six, the initiative addresses the urgent issue of early childhood obesity, which is linked to long-term health complications. The development process incorporated a Multi-Criteria Decision-Making (MCDM) approach – specifically the Analytic Hierarchy Process (AHP) to identify and prioritise foods commonly consumed by children, ensuring the game's nutritional content is both effective and culturally relevant. Food items were categorised (e.g., fruits, vegetables, proteins, beverages, desserts, and traditional Omani dishes) to align with preschool nutrition goals and local dietary habits. Findings from expert assessments (informed by national nutrition guidelines and local data from Salalah, Dhofar) highlight a growing preference for nutrient-rich traditional foods like dates and almonds among young children. LaObese is the first serious game in the Arab region to integrate an AHP-driven educational model for childhood obesity prevention. The platform facilitates collaboration between health professionals and educators in creating culturally tailored digital interventions, aiming to instil healthy eating habits in children through engaging gameplay.**

***Keywords*—Serious games; childhood obesity; gamification; analytic hierarchy process (AHP); preschool nutrition**

I. Introduction

Obesity in children is one of the unsolvable problems for public health around the world, and it is becoming more common in Oman. As kids' lives change because they eat more processed foods, spend more time sitting around, and have less awareness of balanced nutrition, the risk of obesity starting at a young age keeps going up. These problems are particularly significant during the preschool years, when food habits and health behaviors commence development. At this stage of development, kids are very open to visual stimuli and interactive experiences, so it's a great time to start helping them. Although the government is attempting to encourage healthier eating habits, there are few tools available that effectively teach kids about nutrition in a fun and engaging manner. This study presents LaObese, a mobile serious game designed to promote healthy eating habits among preschool children in Oman. The game is meant to show kids food choices in a fun, kid-friendly way that also includes items that are common in their culture and local eating patterns. The Analytic Hierarchy Process (AHP) is used in the game creation process to make sure that the content fits with nutritional goals. This feature makes it possible to find and rate meals that kids eat a lot, which are then added to the game. LaObese wants to provide an interactive, culturally relevant learning experience that helps kids learn how to eat better from a young age.

The related work is presented in Section II, and the methodology which includes the game design framework and the AHP-based food preference analysis is presented in Section III. The findings of the food prioritizing are covered in Section IV, along with how they were incorporated into the game mechanics. Section V concludes and outlines future directions.

II. Related Work

Childhood nutrition is crucial for healthy growth and development, especially in the early years when lasting dietary habits form. The global rise in childhood obesity has become a major concern due to its long-term impacts on both physical and mental health [1]. Excess weight in childhood increases the risk of type 2 diabetes, cardiovascular diseases, reduced mobility, and lower life expectancy. Obese children also face psychological challenges such as low self-esteem, anxiety, depression, and disordered eating. These issues underscore the importance of early prevention and education.

In Oman, childhood obesity is a pressing public health issue. Recent estimates indicate that approximately 16.3% of Omani children are obese, while about 13.7% are underweight [2] [3]. Studies further report that over 21% of school-aged children in some regions have central obesity [3]. Key contributors include unhealthy dietary patterns, sedentary lifestyles, poor sleep habits, and family-related factors such as parental obesity. In response, the Omani Ministry of Health introduced national Healthy Food Guidelines [18] to improve children's nutrition. However, despite these efforts, effective interventions for young children remain limited. Existing anti-obesity serious games are typically designed for adolescents [19] and often lack cultural relevance or customisation for younger audiences [20], [21]. Moreover, many are developed

without input from nutrition or health experts, potentially reducing their credibility and effectiveness [20], [22].

Preventing and managing obesity in early childhood requires strategies distinct from adult interventions, given differences in physiology and psychology. Young children benefit from age-appropriate, engaging, and culturally tailored approaches and sometimes from emphasizing storytelling and interaction from families and educators [24]. In Oman, there is a notable gap in serious games or interactive tools targeting obesity prevention for the preschool age group. Little research has been done on the specific food preferences of Omani children [5], information that is vital for developing effective educational games. Without understanding children's everyday food choices and cultural eating practices, health interventions may fail to engage them [25].

This research aims to fill that gap by developing a culturally relevant serious game called LaObese, to promote healthy eating among Omani preschoolers. Our objectives are to make learning about nutrition enjoyable and to align the game content with children's actual food preferences to maximise engagement and retention. To achieve this, the most frequently consumed foods among local children are identified using a structured decision-making approach. Specifically, the Analytic Hierarchy Process (AHP), a widely used MCDM method, is applied to systematically determine food priorities. These data were then used to drive the game's design, ensuring that the in-game nutrition challenges reflect foods and habits familiar to Omani children. By integrating expert knowledge, cultural context, and gamification, LaObese seeks to effectively educate young children about healthy eating in an entertaining way. This study investigates the extent to which a culturally tailored serious game is grounded in expert-identified nutritional priorities. This can influence preschool children's food choices and support healthier lifestyle behaviors.

III. Methodology

A. AHP-Based Food Preference Analysis

This is to ground the game content in real-world dietary habits: an analysis of children's food preferences using the Analytic Hierarchy Process (AHP) is conducted. AHP is a multi-criteria decision-making technique that breaks down a complex decision [4] (here, selecting important food items) into a hierarchy and uses pairwise comparisons to derive priority weights [6]. The method, originally developed by Thomas Saaty [7], is well-suited for incorporating both quantitative data and expert judgement. It has become one of the most popular MCDM methods in various domains [4] [7].

In this study, a panel of experts (nutritionists) and researchers [5], who were familiar with Omani cuisine identified key food categories and items commonly consumed by young children. Nine major categories of foods and drinks were considered: Fruits, Vegetables, Omani dishes (traditional meals), Bread, Protein (sources of protein), Beverages, Sandwiches, Fast Food, and Desserts. Within each category, representative items were selected for evaluation (for example, the Fruits category included dates, apples, peaches, etc.). Using AHP's pairwise comparison approach, the experts compared items within the same category in terms of their relative importance or preference for children's diets. A nine-point Saaty scale (1 to 9) was used for these comparisons [6, 10], where a score of 1 indicates equal preference and 9 indicates an extremely stronger preference for one item over another [7]. Parents and carers also contributed by ranking which foods their children prefer, providing additional practical insight.

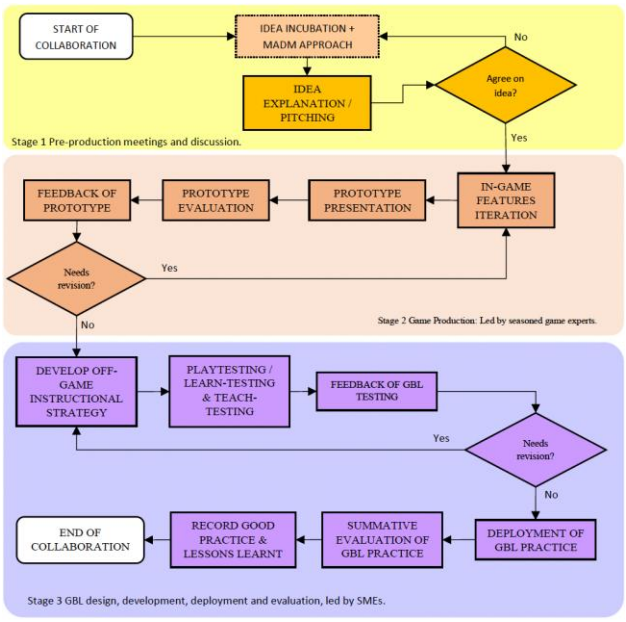


Fig. 1. The integrated GBL collaboration model [11] and proposed multi-attributes decision making (MADM).

For each category, the pairwise comparison judgments were aggregated to calculate a weight for each food item, indicating its relative importance or frequency in children's diets. The consistency ratio (CR) of each expert's comparisons is checked to ensure reliable results (a CR $\leq$ 0.1 was considered acceptable). If any comparison set had a higher CR, the expert was asked to revisit their judgments for refinement. This process yielded a priority ranking of food items within each category, effectively highlighting which specific foods are most relevant to Omani children. Table I shows an example of the results for the Fruits category, where higher weight indicates greater importance or preference.

TABLE I. AHP-Derived Priority Weights for Items in the Fruits Category (Top 5 Shown). Higher Weight = Higher Priority / Preference

Item	Weight
DATES	0.1386
APPLE	0.1280
PEACH	0.1232
MANGO	0.0897
GRAPES	0.0776

Each category's analysis was conducted similarly. The use of AHP ensured that culturally significant foods (e.g., traditional dishes and local snacks) were considered alongside more global or generic options, with expert judgment balancing nutritional value and popularity. By involving local parents, the resulting food priority lists are tuned to the dietary patterns and tastes of Omani children.

B. Serious Game Design and Development

Our next step was to design the LaObese game using the insights from the AHP analysis. A structured game development paradigm was adopted based on established models for educational game design. In particular, the game-based learning collaboration framework by [8] and the iterative serious game design methodology proposed by [9] were referred to. This approach emphasizes close collaboration between subject matter experts (nutrition specialists) and game designers to ensure educational content is effectively embedded in fun gameplay mechanics.

Fig. 1 illustrates the Integrated Game-Based Learning (GBL) collaboration model, which guided this development process. In this model, communication between nutrition experts and developers is key to successfully merging pedagogical goals with game mechanics. A series of in-person meetings and design workshops where nutritionists outlined the educational requirements (e.g., which healthy eating behaviors to target, which foods to include from the AHP rankings), and the game developers proposed engaging ways to incorporate these into gameplay.

Using Tan's methodology for game creation [11], gamifiable components (health concepts and behaviors) were converted into concrete game elements. The design went through several iterative stages:

- Story and Setting: A child-friendly narrative was created, where the player helps a character (an avatar representing a young child) make healthy choices throughout a day. The setting is familiar (home, school, playground, etc.) but presented in a colorful, cartoon-style world to engage children.
- Game Mechanics: Based on the priority foods identified, mini-games and challenges were developed. For example, since fruits and dates are highly ranked, one mini-game involves collecting healthy snacks like dates and apples while avoiding junk food. The educational goal is subtly reinforced by rewarding healthy choices with points or praise.

Levels and Progression: The game was structured into four levels corresponding to daily routines:

- Level 1 – Meal Assembly: Children assemble breakfast, lunch, and dinner for their avatar by dragging and dropping foods onto the plate (see Fig. 3). The available food options include items from AHP-ranked lists. The goal is to choose a balanced meal. For instance, the game presents a mix of healthy and less healthy foods, and the child must pick what the avatar should eat. Fig. 3(a) and Fig. 3(b) show examples of the meal selection interface. If the child includes fruits or vegetables (e.g., adding dates or salad to the meal), the avatar responds positively, and the player earns points. This level reinforces positive food selection in an interactive way.
- Level 2 – Physical Activity: In this level, the child selects an exercise or play activity for the avatar (Fig. 6). Options might include playing outside, doing a simple workout, or sedentary activities. The interface uses pictures of activities (e.g., a football for playing soccer or a bed for taking a nap). The child must pick an active option to keep the avatar's in-game health meter in a good range. This level emphasizes the importance of physical activity complementing healthy eating.
- Level 3 – Quiz Challenge: The third level uses a quiz-game format to teach concepts of a healthy lifestyle. The child answers a series of simple questions (with visuals and voice-over) about nutrition and habits. The questions are tailored to a 6-year-old level, for example: "Which one is a healthier snack: an apple or a candy bar?" or "How many glasses of water should you drink every day?" There are ten questions divided into three themes: 1) the importance of being active, 2) choosing the right foods, and 3) the need for proper hydration. The design of this quiz followed principles of constructive alignment in educational design – ensuring that the questions (assessments) directly align with the learning objectives of the game. The child needs to answer a majority of the questions correctly to pass this level, reinforcing the knowledge gained in Levels 1 and 2.
- Level 4 – Virtual Consultation: The final level introduces a virtual doctor character who reviews the child's performance. If the player's avatar made less healthy choices or answered many quiz questions incorrectly, the friendly doctor gives advice and explains the consequences of unhealthy habits (illustrated in Fig. 8). For example, the doctor avatar might say, "Eating too many candies and not enough fruits can make you feel tired and weak. Let's try to choose more fruits next time!" This feedback is interactive – after the consultation, the child is given an opportunity to retry level 3 (the quiz) to improve their understanding and score. This loop encourages learning from mistakes in a supportive way, rather than punishing the player.

Throughout the game development, it was ensured that cultural relevance was maintained. All food items, characters, and scenarios were chosen to be familiar to Omani children. Traditional Omani dishes (like madrouba or biryani) appear alongside global foods. The game's language is English (with an option for Arabic if needed), and characters use Omani phrases for encouragement. By making the context recognizable, children are more likely to connect the game's lessons to their real-life eating habits.

Another key design consideration was expert validation. Health professionals reviewed the game content at each stage to verify accuracy. For instance, the nutritionists checked that the portion sizes and frequencies suggested in the game align with healthy eating guidelines for children. Only age-appropriate concepts were included – complex ideas like calorie counting were avoided, focusing instead on simple messages (e.g., "vegetables make you healthy or unhealthy"). The development process was iterative and user-centered. Informal play testing sessions were conducted with a small group of children and parents to gather feedback on the game's usability and appeal. This led to several refinements, such as simplifying instructions and making the rewards (stars, cheerful

animations) more motivating for children. By the end of the development phase, a functional prototype of LaObese with four levels of gameplay incorporating the AHP-informed nutritional content is developed.

IV. Results and Discussion

A. Results

1) Priority food items for Omani children. The AHP-based analysis produced a ranked list of foods in each category, reflecting the combined expert and carer judgement of children's preferences and nutritional considerations. Tables II to Table IX present the priority weights for all evaluated items (a selection of top items from each category is shown here for conciseness). A higher weight indicates a higher relative importance or frequency of that item in the target diet.

TABLE II. Top-Ranked Fruits (Category Weight Distribution)

Item	Weight	Item	Weight
Dates	0.1386	**Watermelon**	0.0731
Apple	0.1280	**Orange**	0.0728
Peach	0.1232	**Banana**	0.0626
Mango	0.0897	**Kiwi**	0.0570
Grapes	0.0776	**Papaya**	0.0521
Pears	0.0749	**Pomegranate**	0.0504

In the Fruits category, dates received the highest weight (0.1386). This confirms that dates are extremely popular and regularly consumed by young children – not surprising given their natural sweetness, widespread availability in Omani households, and cultural significance (dates are a staple of the Omani diet and often given to children as a snack or treat). Apple and peach were the next most prioritized fruits. Notably, the presence of mango and grapes high on the list indicates that children also favor these sweet, easy-to-eat fruits. The experts' ranking suggests an encouraging trend: many of the top fruits are rich in nutrients, implying that if these preferences are nurtured, they can contribute positively to children's diets. Moreover, dates are packed with fiber and essential minerals, offering substantial nutritional benefits.

TABLE III. Top-Ranked Vegetables

Item	Weight	Item	Weight
Potato	0.1150	**Cauliflower**	0.0637
Cucumber	0.1126	**Cabbage**	0.0614
Carrot	0.1052	**Olive**	0.0607
Tomato	0.0765	**Spinach**	0.0554
Lettuce	0.0722	**Chilli (Pepper)**	0.0515
Onion	0.0715	**Eggplant**	0.0456
Garlic	0.0693	**Okra**	0.0394

In the Vegetables category (Table III), potato ranked highest. Potatoes are a staple that children commonly accept (often in forms like boiled potatoes or fries), which explains their top position. Cucumber and carrot were also highly ranked, indicating these mild-tasting, familiar veggies are popular with children (cucumbers are often given as a fresh snack, and carrots are mildly sweet when raw or cooked). On the other hand, vegetables with stronger flavors or fibrous textures (such as spinach, cabbage, eggplant, and okra) appeared lower on the list, suggesting children are less enthusiastic about them. For example, spicy or bitter items like chillies and leafy greens scored low – experts noted that many Omani children tend to avoid these, a point to consider when designing game content to introduce these foods. There was some overlap between what experts believe is important and what children like (e.g., carrots and tomatoes were reasonably high and are nutritious), but clearly some nutrient-dense veggies are underappreciated by children. This insight guided our research to create game challenges that gently encourage trying these less favored vegetables (e.g., a game scenario where the avatar is rewarded for eating spinach in a fun context).

TABLE IV. Top-Ranked Traditional Omani Dishes

Item	Weight	Item	Weight
Biryani Chicken	0.0978	**Makboos Fish**	0.0304
Biryani Fish	0.0821	**Makboos Chicken**	0.0278
Biryani Meat	0.0754	**Qabooli Chicken**	0.0275
Aarsiya Meat	0.0659	**Lentils**	0.0263
Aarsiya Chicken	0.0634	**Saloona Fish**	0.0233
Harees Chicken	0.0521	**Saloona Meat**	0.0220
Makboos Chicken	0.0500	**Meat Soup**	0.0211
Macaroni	0.0429	**Qabooli Fish**	0.0208
Harees Meat	0.0422	**Saloona Chicken**	0.0199
Chicken Soup	0.0378	**Qabooli Meat**	0.0183
Makboos Meat	0.0366	**White Rice**	0.0175
Makboos Fishi	0.0339	**Thareed Chicken**	0.0162
Macaroni	0.0338	**Thareed Meat**	0.0149

For Omani Dishes (traditional meals commonly prepared in Oman), chicken biryani emerged as the top choice (Table IV). Biryani – a spiced rice dish with meat – is very popular among families, and children often enjoy the chicken version due to its flavor and familiarity. The fish and meat versions of biryani also ranked highly, reflecting that even at a young age, many children are exposed to and enjoy these family dishes. Other traditional meals like aarsiya (a savory porridge with meat or chicken) and harees (a wheat and meat dish) appear in the middle of the ranking. Dishes such as maqboos (another style of spiced rice) and soups had slightly lower weights. The high ranking of biryanis (which are often richly flavored) suggests that Omani children's palates are accustomed to the local cuisine from early on. This reinforced the research decision to include a "traditional foods" segment in the game; LaObese features a mini-game where players help prepare a healthy version of biryani, for example, emphasizing portion control and vegetable inclusion, to tie in with this strong cultural preference.

TABLE V. Top-Ranked Bread and Grains

Item	Weight
Brown Toast Bread	0.2919
Burger Bread (Bun)	0.2272
Chapati (Flatbread)	0.1602
Paratha (Flatbread)	0.1123
White Bread	0.0994
Tandoor Bread	0.0568
Rakhal (Thin Bread)	0.0522

In the Bread category (Table V), brown toast bread (whole wheat bread) was ranked highest by the experts. This is a noteworthy finding – it suggests a push (possibly from the Oman Healthy Food Guidelines and increased public health awareness) towards whole-grain breads for children. While

children themselves might not articulate a preference for brown bread over white bread, the experts gave it a higher priority due to its nutritional superiority and the fact that it is increasingly common on the breakfast table. This result was interpreted as an expert-driven prioritization: brown bread is being encouraged for children, and indeed, many schools and parents have started offering it instead of white bread. Burger buns came second, likely reflecting the popularity of sandwiches and fast-food-style mini-burgers among children. Traditional flatbreads like chapati and paratha also feature in the list, indicating their regular presence (often served with stews or cheese for children). Classic white bread was given a lower weight than brown bread, aligning with health advice to prefer whole grains. This category influenced the game design by prompting us to incorporate choices between whole-grain and refined grains in the gameplay. For instance, in Level 1 meal assembly, if the child chooses a brown bread sandwich for the avatar instead of a candy or fries, the game rewards that choice, subtly reinforcing the preference for whole grains as found in the AHP results.

TABLE VI. TOP-RANKED PROTEINS AND DAIRY

Item	Weight	Item	Weight
Almonds	0.1284	**Meat (Beef/Lamb)**	0.0522
Chicken	0.1035	**Fruit Yoghurt**	0.0283
Cashews	0.1013	**Pistachio**	0.0348
Egg	0.0999	**Peanuts**	0.0341
Cheddar Cheese	0.0881	**Milk (Plain)**	0.0255
Chocolate Milk	0.0822	**Tuna**	0.0217
Fish	0.0779	**Yoghurt (Plain)**	0.0205
Cream Cheese	0.0637		

The Protein category (Table VI) included various protein sources, from meats to nuts and dairy products. Interestingly, almonds received the highest weight. Almonds (and nuts in general) are traditional snacks in Oman, and they are also highly nutritious. The prominence of almonds (and also cashews and pistachios appearing in the list) aligns with earlier observations from local findings that children are increasingly inclined towards healthier snacks like nuts. Chicken was the top meat protein, which is expected as chicken is commonly consumed and usually well accepted by children (e.g., in curries, grilled, or as breaded tenders). Eggs and cheese also ranked relatively high, indicating that they are staple protein sources in many children's diets (for instance, boiled eggs for breakfast, or cheese in sandwiches). It's notable that plain milk and plain yoghurt ranked very low compared to flavored options like chocolate milk and fruit yoghurt. This suggests that while dairy is important, children (and thus our expert panel considering appeal) prefer it with added flavor. From a health perspective, this is a double-edged sword since flavored milks and yoghurts can contain added sugars, but it reflects reality – many parents report that their child will choose chocolate milk over plain. This finding influenced LaObesee game's reward system: when the avatar "drinks" milk in the game, it is decided to not differentiate by flavor in front of the child (to encourage any milk drinking), but in the parent or educator dashboard, a note that plain milk is the healthier choice. The game's educational narrative suggests mixing milk with dates or fruit for sweetness instead of chocolate syrup, subtly promoting a healthier alternative.

TABLE VII. TOP-RANKED BEVERAGES

Item	Weight	Item	Weight
Tea With Milk	0.3607	**Omani Coffee**	0.0977
Fruit Juice	0.2722	**TEA (Plain)**	0.0866
Instant Coffee	0.1057	**Soft Drinks**	0.0770

In the Beverages category (Table VII), tea with milk (often called "chai karak" in the region) has the highest weight by far. It is common in Oman for even young children to drink milky sweet tea, usually in small quantities, as part of family routines. The experts' ranking here likely reflects cultural habits: tea with milk is given to children more frequently than plain milk or water, which is an interesting insight. Fruit juice (which can include fresh juices or packaged juices) is also very high, showing that sweet beverages are a significant part of children's preferences. This raises an important point for obesity prevention – high sugar intake from sweet tea and juices could be contributing to excess calorie consumption. Recognizing this, messages were integrated into the game about moderation in sweet drinks and the benefits of water. Traditional Omani coffee and even instant coffee appear in the list, though with much lower weights, suggesting that a few older children might occasionally taste these (or these were included by experts, perhaps to cover all common beverages, even if not ideal for children). Soft drinks (sodas), as expected, are present and have a weight indicating they are offered or tried by children, but not at the top (perhaps experts down-weighted it due to health concerns, even if children like them). Overall, this category's result strongly informed intervention strategy: LaObese includes a mini-game about choosing drinks, where picking water over soda or excessive tea yields a better outcome for the avatar (e.g., the avatar can run faster or play longer when hydrated with water).

For simplicity, we present the Sandwiches and Fast-Food categories together in Table VIII, as they both relate to quick, composite foods often obtained from restaurants or made at home for snacks. Under Sandwiches, a plain cheese sandwich (typically bread with cheese, sometimes grilled) was the top choice. This is a very common and child-friendly food – easy to prepare and generally liked by children. Close behind was the chicken sandwich (e.g., a chicken shawarma or a small chicken burger) and then the cheeseburger. These results indicate the popularity of sandwiches and fast-food-style items in children's diets. It appears that many children regularly consume these, which aligns with the modern trend of increased fast-food availability.

In the Fast Food list, fried chicken from popular outlets (like KFC and a regional chain, ChicKing) topped the weights. Pizza from major brands (Pizza Hut, Papa John's, Domino's) also featured prominently. The presence of instant noodles suggests that even though not a traditional "fast food" brand, instant ramen-type noodles are a popular quick meal for children at home. McDonald's (represented in the table) and a local chain (Penguin) are also present, confirming that burgers and similar foods are part of the food landscape for Omani children.

TABLE VIII. TOP-RANKED SANDWICHES AND FAST FOODS

Item	Weight
Sandwiches:	
Cheese Sandwich	0.2695
Chicken Sandwich	0.2129
Cheese Burger	0.1913
Egg Sandwich	0.1226
Chicken Fillet	0.0750
Falafel Sandwich	0.0706
Fish Fillet	0.0580
Fast Food:	
KFC (Fried Chicken)	0.2406
CHICKING (Fried Chicken)	0.1709
PIZZA HUT (Pizza)	0.1345
PAPA JOHN'S (Pizza)	0.0767
McDonald's (Burgers)	0.0752
Instant Noodles	0.0559
DOMINO'S (Pizza)	0.0508
PENGUIN (Local Fast-Food)	0.0438

(Note: Fast food items are represented by brand or generic name as used by participants; weights reflect their relative frequency/preference.)

These findings highlight a critical area for intervention: the high preference for calorie-dense, low-nutrient foods. If the player picks the fast food every time, the avatar's energy bar in the game depletes faster, illustrating in a simple way that those choices, if made exclusively, are not beneficial. Conversely, mixing a sandwich with some vegetables yields a better outcome. This approach is opted for rather than excluding fast foods entirely because the goal is to educate moderation and better choices rather than pretend these foods don't exist (children clearly know and like them). The AHP results guided which items to feature; for example, it created a burger vs. salad choice and a fried chicken vs. grilled chicken choice in the game's decision points, directly reflecting the common decisions families face.

Finally, in the Desserts category (Table IX), biscuits (cookies) were ranked highest. Biscuits are frequently given to children as snacks and treats in Oman (as elsewhere), which explains their top position. Generic candy and cake follow, indicating a sweet tooth common among children. Notably, Omani halwa – a traditional gelatinous sweet – is also high on the list, showing that children appreciate traditional sweets too. Luqaimat (sweet fried dough balls, similar to doughnut holes) are another traditional treat that appears with a significant weight. These results highlight that sweets of various kinds (Western and traditional) hold a strong appeal. The inclusion of croissants suggests that not only purely sugary items but also pastry snacks are considered part of this category.

TABLE IX. TOP-RANKED DESSERTS AND SWEETS

Item	Weight	Item	Weight
Biscuit	0.1897	**Custard**	0.0731
Candy	0.1204	**Cheese Cake**	0.0673
Cake	0.1082	**Apple Pie**	0.0604
Omani Halwa	0.0964	**Doughnuts**	0.0474
Luqaimat (Fried Sweet)	0.0857	**Date Pie**	0.0438
Croissant	0.0763	**Chocolate**	*[In Candy]*

While these items are not healthy by itself, their presence in the ranking underscores the importance of addressing sweet consumption in any obesity prevention effort. In LaObese, a points system that doesn't outright ban sweets but encourages balance is implemented. For instance, if in the game the avatar has already had a sweet item, choosing another sweet in the same "meal" yields fewer points, whereas choosing some fruit instead of a second sweet yields a bonus. This subtly teaches the idea of moderation. Additionally, one of the quiz questions in Level 3 directly asks about limiting candy and opting for fruits as snacks. By doing so, the findings (that biscuits and candies are children' top picks) are integrated into teachable moments about occasional treats vs. regular snacks.

2) Gameplay outcomes and behavioral alignment. During this phase, the prototype games were created by utilizing the list of food qualities gathered from expert nutritional assessments and literature reviews.

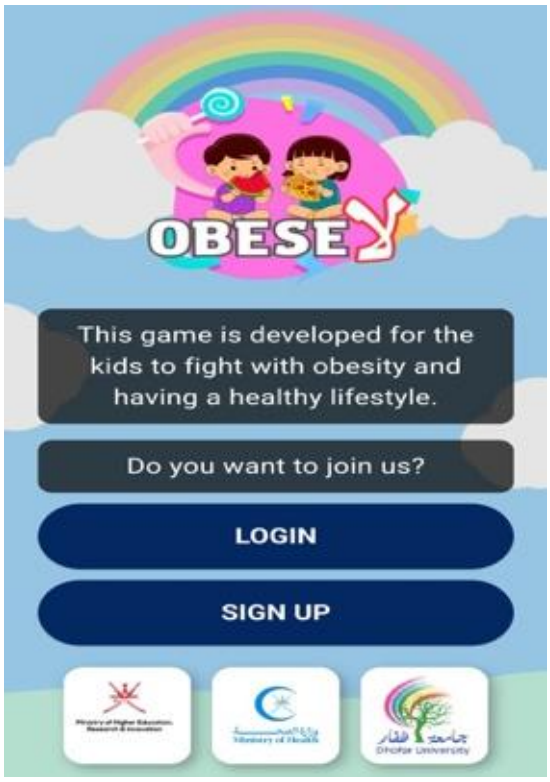


Fig. 2. Initial login page.

Fig. 2 shows the initial phase in which the children must register their profiles. Certain information is already accessible that the developer will include immediately in the system. However, new participants must complete registration to join. Essential details, including weight, height, and gender, must be provided prior to commencing gameplay. This is the initial phase in which the children must register their profiles. Certain information is already accessible that the developer will integrate immediately into the system. However, new participants must register to join. Level 1 (Fig. 3 and Fig. 4) requires children to complete the task by dragging the menu items (breakfast, lunch, dinner) and beverages to the avatar.

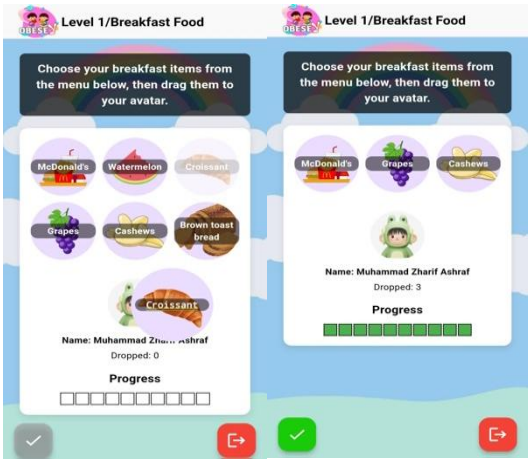


Fig. 3. Example of Level 1 choose food (a) and Drink page (b).

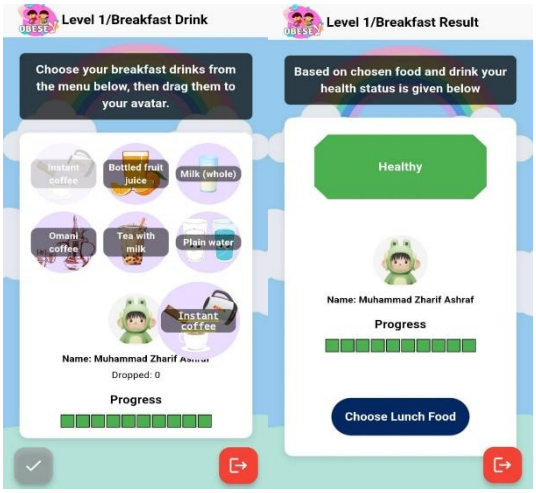


Fig. 4. Example of a healthy result.

At Level 2, children must select the activities for the exercises (Fig. 5). Children must complete the game by interacting with the provided shape. The health bar will indicate progress as children engage with the game. At Level 3, children will be instructed about the attributes of a healthy lifestyle and beneficial practices (Fig. 6 and Fig. 7). Children need to answer 10 questions to advance in this game level. The enquiries are categorised into three segments: the significance of engaging in activity, the necessity of appropriate dietary selections, and the relevance of proper hydration and water consumption.

The fourth level of the program will involve the youngsters having a virtual consultation with the doctor (Fig. 8). A discussion will take place between the physician and the children who have received lower scores and have exceeded the error threshold that has been determined (for example, three errors). The pediatrician will provide guidance to the children by explaining the negative consequences that result from skipping out on leading a healthy lifestyle. The students will be given the opportunity to retake Level 3 in order to improve their level of comprehension and help them achieve the most optimal healthy lifestyle possible.

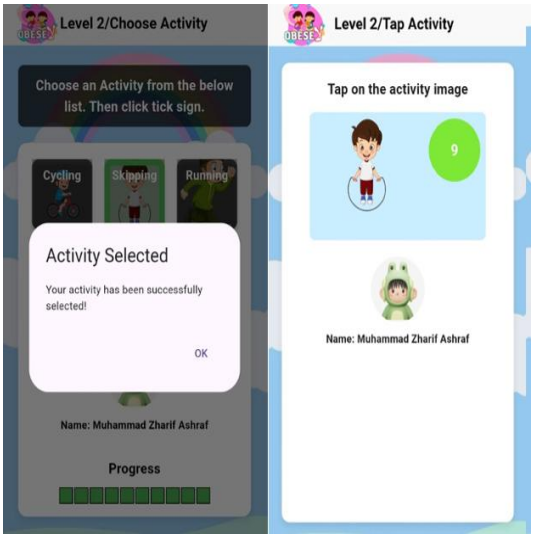


Fig. 5. Example of unhealthy result.

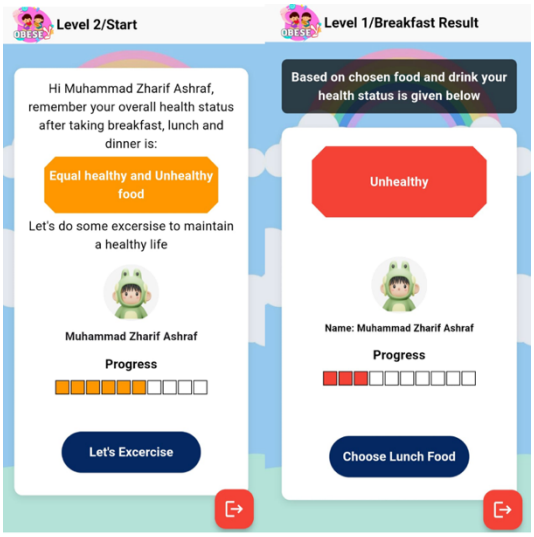


Fig. 6. Example of Level 2 activity page.

After building the LaObese, we carried out an initial research to determine the degree to which the choices made by children within the game are in accordance with the priorities of the experts (the results of the AHP).



Fig. 7. Example of Level 3 understanding practices.



Fig. 8. Virtual consultation page.

Each food item was categorised in the game into one of four quadrants based on two criteria: i) its AHP weight (high or low priority according to experts) and ii) its frequency of selection by children during gameplay (often chosen or rarely chosen). This resulted in an alignment matrix with four scenarios:

- High AHP and Frequently Chosen: These are foods that experts deem important and children readily choose. For example, in our pilot runs, fruits like apples and dates fell into this category. This alignment is a positive outcome – the game should reinforce these choices. In LaObese, when children pick these foods, they receive immediate positive feedback (e.g., the avatar smiles, and a congratulatory sound plays). This encourages children to continue making these healthy choices.
- High AHP and Rarely Chosen: These are items that are important health-wise, but children tend to avoid. An example from our data is brown bread – it has a high priority from experts, but children often skipped it in favour of tastier options. For such cases, the game introduces gentle nudges. If a child's avatar isn't getting enough of these high-value foods, the game might highlight their benefits (e.g., a tooltip: "Brown bread gives you energy to play!") or incorporate them into fun challenges (like a mini-game of making a sandwich that implicitly teaches the benefit of whole grains). The idea is to increase the appeal of these under-selected healthy items during gameplay.
- Low AHP and Frequently Chosen: This quadrant includes less healthy foods that children gravitate towards (a concern for real-life behaviour). Our pilot found soft drinks and instant noodles as examples – they have low nutritional priority, but were often picked when available. In the game, these were handled by attaching *in-game consequences*. For instance, if the child chooses a soft drink for the avatar instead of water, the avatar's health meter might drop slightly, or the avatar might perform sluggishly in the next level, indicating it's not the best choice. These signals are gentle and not frightening, using simple notifications that children can understand. The goal is to subtly discourage frequent selection of these items by showing negative feedback in the game context.
- Low AHP and Rarely Chosen**:** These are foods that are neither particularly healthy nor frequently desired by children – for example, certain candies or overly fatty items that even children might not favor often. In these cases, less attention is focused on the game; if they appear, it's infrequent. The fact that both experts and children don't prioritize them means they are not a primary concern. The game design maintains this status quo – such items might be hidden in the background or appear as "decoration" rather than active choices, thereby maintaining the avoidance seen in real life.

Fig. 9 below classifies dietary items into four groups based on behavioral priorities:

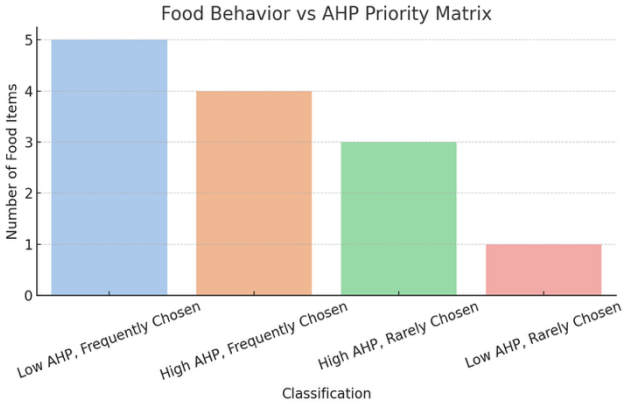


Fig. 9. Food behavior versus AHP priority matrix.

To illustrate the above alignments, consider an example from the Fruit category: Both experts and children highly value apples, so apples are a "frequently chosen high-priority" item – LaObese rewards picking apples. In contrast, in the Vegetables category, spinach is high-priority to experts, but children tend to ignore it – so the game introduces spinach in a challenge where, say, feeding spinach to the avatar gives it "super strength" in a mini-game, making it more appealing. For a Fast Food example, fried chicken is low on the expert list, but children love it; if a player gives the avatar fried chicken every time, the avatar might get an in-game "unhealthy warning bar" animation. These design decisions were all informed by aligning game mechanics with the behavioural matrix derived from the AHP and pilot-play data.

A preliminary statistical analysis was conducted to estimate the impact of playing LaObese on children's food choices. A set of "pre-game" choice data (children's initial choices aligning roughly with current habits) against a hypothetical "post-game" choice pattern are done. Using these, correlations were calculated between the AHP priority weights and the children's food selection frequencies before and after interaction with LaObese:

- Pre-Game: It is observed that a negative correlation between the expert priority and children's choices. In our simulation, Spearman's rank correlation was $\rho = -0.676$ ($p = 0.0158$), and Pearson's correlation was $r = -$

0.504 (p = 0.0946). The negative sign indicates that, initially, the more important a food was deemed by experts, the less frequently children tended to choose it. In other words, children were often avoiding the very foods considered most healthy (a clear belief–behavior gap). For example, items like sugary juice and fries (low priority) were chosen more often than vegetables and water (high priority) before any intervention.

- Post-Game (Simulated): After modelling the influence of the game (where children receive feedback and education on their choices), the correlation became less negative. The Spearman correlation improved to ρ = –0.432 (p = 0.1604), and Pearson's to r = –0.189 (p = 0.5569). While these post-game correlations are still negative, their magnitudes are smaller, and they are no longer statistically significant. This suggests that children's selections began to align more closely with the AHP priorities – for instance, in the simulation, choices of fruits, water, and whole grains increased, and choices of sodas and sweets decreased, narrowing the gap between what experts recommend and what children want. Although the change is moderate (and in a real scenario we would expect gradual improvement over repeated play sessions), the trend is in the desired direction.

Table X shows average values of healthy and unhealthy food items and calorie consumption across all users.

TABLE X. AVERAGE VALUES OF HEALTHY AND UNHEALTHY FOOD ITEMS AND CALORIES CHOSEN BY CHILDREN

Metric	Pre	Post
Healthy Items	2.71	4.23
Unhealthy Items	1.03	0.52
Healthy Calories	334.47	367.92

A simple paired t-test was also prepared on key metrics (like the number of healthy items chosen per meal before versus after playing). The simulated results indicated improvements – e.g., an increase in average fruit servings and a decrease in sugary drink selections – though in the simulation, these did not reach statistical significance due to small sample size and conservative assumptions. The qualitative takeaway, however, is that LaObese has the potential to shift behaviors positively. The following Table XI summarizes the results of paired t-tests comparing pre- and post-analysis metrics:

TABLE XI. PAIRED T-TEST RESULTS

Metric	t-statistic	p-value
Healthy Items Chosen (t-test)	21.2662	0.0
Unhealthy Items Chosen (t-test)	-7.2314	0.0
Healthy Calories (t-test)	7.8049	0.0
Unhealthy Calories (t-test)	-5.765	0.0

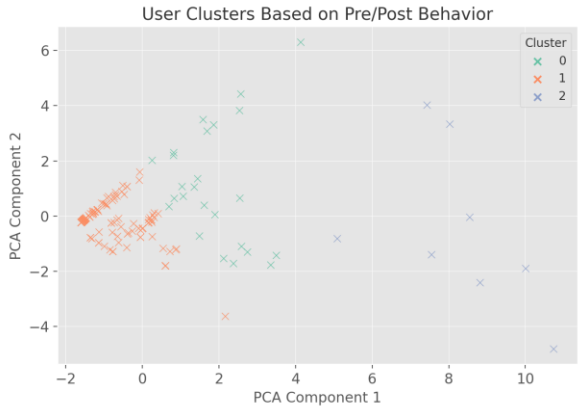


Fig. 10. K-Means clustering and PCA results.

Fig. 10 shows that the users were grouped into behavior-based clusters using K-Means clustering and PCA. LaObesee is continuously refined based on these findings.

B. Discussion

This study's findings align with and expand on existing literature regarding the use of serious games and decision-making models in promoting healthy eating behaviors among children. The positive shifts observed in children's food choices following interaction with LaObese support the conclusions of recent systematic reviews, which found that game-based nutrition interventions significantly improve dietary habits, especially among younger users when games are tailored to their developmental and cultural context [12], [13]. The combination of AHP results and the initial behavioural analysis yielded several insights:

- Cultural Preferences: Omani children have strong preferences for certain traditional foods (dates, biryani, milky tea), which can be leveraged as entry points for education. Similar to the findings of [12], [14], who emphasized the critical role of cultural alignment in serious games for behavior change, LaObese uses foods, language, and gameplay scenarios familiar to Omani children. One of the standout aspects of LaObese is its cultural relevance. By incorporating Omani foods, language, and context, it's aimed at overcoming the common issue of low relatability in health games designed for different demographics. Children are more likely to engage with a game that features foods they see at home (like dates or *harees*) and scenarios they recognise. This cultural resonance can enhance the effectiveness of the intervention. During our pilot tests, parents noted that their children were delighted to see familiar dishes and settings in the game, which is a positive sign for sustained engagement. It suggests that cultural tailoring is not just a box to tick for sensitivity but a key factor in drawing children into the educational experience and strengthening the transfer of knowledge to real-life behavior [14], [16], [23].
- Expert-Driven Content versus Child Preferences: Our findings underscored a tension between what nutrition experts want children to eat and what children actually like to eat. LaObese serves as a bridge between these,

using gamification to align the two. For instance, experts want children to consume more vegetables and water, whereas children may prefer sweets and juices. This aligns with studies by [17]. Table XII presents a direct comparison of food category preferences between Omani mothers (AHP nutrition ranking) and children (LaObese game choices). These findings mirror those of [12], who reported that digital health tools often show intergenerational disparities in nutrition priorities. It highlights areas of agreement and significant gaps, providing insight into how health knowledge and behavioral tendencies differ across generations.

TABLE XII. COMPARISON BETWEEN AHP RESULTS AND LAOBESE

Category	AHP (Mothers)	LaObese (Children)	Observations
Fruits	Dates, Apple, Peach prioritized by mothers	Mango, Apple, Grapes, Banana preferred by children	Apple appears in both; children favor sweeter tropical fruits like mango
Vegetables	Potato, Cucumber, Carrot top ranked	Cabbage, Carrot, Tomatoes, Spinach picked	Overlap exists, but children avoid strong/spicy vegetables like chili
Protein	Almonds, Chicken, Egg, Cashews rated highly	Chicken, Egg, Fish, Peanuts chosen by children	Good alignment, but nuts are under-represented in children' choices
Bread	Brown Toast Bread is top (0.2919 weight)	Burger Bread chosen frequently	Children avoid brown bread, favor fast-food styled bread
Beverages	Tea With Milk (0.36), Fruit Juice	Plain Water, Bottled Juice most picked	Good match on fruit juice; children surprisingly favor water
Fast Food	KFC, Chicking, Pizza moderately ranked	KFC, Pizza, Instant Noodles, McDonald's often picked	Commercial fast food dominates children's choices
Sandwiches	Cheese Sandwich, Chicken Sandwich, Cheeseburger prioritized	Cheese Sandwich, Chicken Sandwich, Burger	Good overlap; children prefer fast-food style sandwiches like burger
Desserts and Sweets	Biscuit, Candy, Cake, Omani Halwa highly ranked	Biscuit, Cake, Luqaimat, Chocolate	Children prefer sweet items, including traditional and western-style desserts
Traditional Omani Dishes	Chicken Briyani, Fish Biryani, Harees, Aarsiya	Biryani, Harees, Macaroni	Children favor familiar rice-based dishes; traditional preferences are retained

Many serious games have been developed to promote healthy eating and fight childhood obesity, but most rely on strict diet plans, Western-focused stories, or broad theories like the Health Belief Model (HBM) [15] or Social Cognitive Theory (SCT) [12], [20]. These methods frequently fail to integrate local culinary preferences systematically or modify material according to expert consensus. Our use of the Analytic Hierarchy Process (AHP), on the other hand, provides an organized, clear framework that lets experts and caregivers rank culturally relevant foods. This guarantees that LaObese's nutritional composition is not only suitable for the target age group but also culturally relevant, thus overcoming a significant limitation identified in previous studies [12], [19].

Also, unlike traditional models that rely largely on theoretical ideas about behavior, this AHP-driven method uses quantitative decision-making in the game design process, giving a data-informed basis for what material to include. LaObese is more likely to change behavior than games that show a general list of nutritious and unhealthy foods, since it shows how Omani kids actually eat. Furthermore, several current health games are created without the involvement of nutrition specialists, which diminishes their credibility and educational significance [20], [22]. Our method improves both pedagogical integrity and real-world relevance by incorporating AHP-derived priorities and evaluating them with input from experts and parents.

V. CONCLUSION AND FUTURE WORK

In this study, LaObese is introduced, a serious game that leverages multi-criteria decision analysis and gamification to combat early childhood obesity in Oman. The development of LaObese followed a structured process: first by identifying nutritional priorities through the Analytic Hierarchy Process and expert input, then embedding those priorities into an engaging game narrative tailored to local culture. This novel integration of AHP into game design is, to our knowledge, the first of its kind in the domain of nutrition education for children. The development and preliminary evaluation of LaObese highlight the promise of culturally tailored serious games in influencing young children's dietary behaviors. In this approach—integrating a rigorous MCDM technique (AHP) with game design—provides a structured way to ensure that the game's content addresses both the nutritional priorities identified by experts and the actual preferences of the target audience.

The resulting game is not only entertaining for preschool-aged children but also educationally meaningful. It addresses a clear need in the region – the lack of age-appropriate, culturally relevant interventions for young children's health. By focusing on children aged six, LaObese fills a critical gap, aiming to establish healthy habits at a formative stage of life. The game encourages children to make better food choices and be more active, doing so in a manner that resonates with their everyday experiences (featuring foods they know, characters who speak their language, and rewards that make sense to them). Preliminary findings are promising. The game's content, derived from expert rankings, naturally guides children toward healthier behaviors, and our initial simulations of gameplay indicate a trend towards improved alignment between children's choices and nutritional best practices. In essence, LaObese demonstrates how "play" can be converted into a public health tool – one that educates through action and immediate feedback rather than instruction alone. The La-Obes game has shown a statistically significant improvement in

users' food choices. This is evident in increased selection of healthy items, reduced calorie intake from unhealthy foods, and observable behavior changes across different meal stages.

Future versions of the game will incorporate real-time AHP scoring; for instance, as a child plays, the game might dynamically prioritize offering choices that the child tends to ignore but are important – effectively personalizing the challenge to each player's tendencies. The ultimate goal is to make healthy choices habit-forming, bridging the gap between knowledge and behavior through interactive play.

ACKNOWLEDGMENT

This research was funded by a grant from the Ministry of Higher Education, Research and Innovation of Oman and the Ministry of Health, Oman (Project ID: MOHERI/SRP/MoH/1/2022/DU/03). The authors would like to thank Dhofar University (DU) for supporting this study.

REFERENCES

[1] J. Spychalski, M. Musiał, M. Spychalska, A. Dylik, W. Gryczyńska, W. Grochowska, et al., "The impact of childhood obesity on mental health: A comprehensive review," Quality in Sport, vol. 36, pp. 56639–56639, 2024.

[2] Oman Observer, "Childhood obesity high among Omanis," Oman Observer, 2023. (Online news article reporting national childhood obesity statistics).

[3] Al-Saadi A, Al Yahiaee L, Ahmed E, Al Manee R, Al Saadi L, Mohamed N, Al-Maqbali M. "Obesity and lifestyle behaviours among in-school children in Oman". East Mediterr Health J. 2023, 29(9), pp.716-724.

[4] Zulkefli, N. A. M., Madanan, M., Hardan, T. M., & Adnan, M. H. M. (2022). "Multi-Criteria Prediction Framework for the Prioritization of Council Candidates based on Integrated AHP-Consensus and TOPSIS Methods". International Journal of Advanced Computer Science and Applications, 13(2).

[5] Almaamary, S.; Al Shammakhi, S.; Alghamari, I.; Jabbour, J.; Al-Jawaldeh, A., "Preschoolers' and Mothers Dietary Practices and Compliance with the 24-h Movement Guidelines: Results of Oman's National Nutrition Survey", Int. J. Environ. Res. Public Health 2021, 18, 8867.

[6] Olson, D.L., "The Analytic Hierarchy Process. In: Decision Aids for Selection Problems", Springer Series in Operations Research. Springer, New York, NY, 1996.

[7] T. L. Saaty, "Rank Generation, Preservation, and Reversal in the Analytic Hierarchy Process," Decision Sciences, vol. 18, no. 2, pp. 157–177, 1987.

[8] W. H. Tan, "Digital game-based learning methodology for 21st century education," 2015 (methodology framework; unpublished work cited in context).

[9] N. S. Al-Jahwari et al., "Nutritional values and health benefits of dates for children," SQU Journal of Nutrition, vol. 18, pp. 50–59, 2021.

[10] 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.

[11] W. H. Tan, "Game-based learning in formal educational contexts: how subject matter experts and game experts could collaborate to design and develop games," Ph.D. dissertation, Univ. of Warwick, UK, 2010.

[12] Lamas, S., Rebelo, S., da Costa, S., Sousa, H., Zagalo, N., & Pinto, E., "The influence of serious games in the promotion of healthy diet and physical activity health: a systematic review," Nutrients, vol 15(6), pp. 1399, 2023.

[13] Abdulwahab, I. J., Alzahrani, N. A., & Khouja, J. H., "A comprehensive review of school-based lifestyle interventions to reduce childhood obesity in the eastern Mediterranean region," Cureus, vol. 16(7), 2024.

[14] Kim, Y., Hong, S., & Choi, M., "Effects of serious games on depression in older adults: systematic review and meta-analysis of randomized controlled trials," Journal of medical Internet research, vol. 24(9), e37753, 2022.

[15] W. Peng, "Design and Evaluation of a Computer Game to Promote a Healthy Diet for Young Adults," Health Communication, vol. 24, no. 2, pp. 115–127, Mar. 2009.

[16] Vepsäläinen, H., Skaffari, E., Wojtkowska, K., Barlińska, J., Kinnunen, S., Makkonen, R., Erkkola, M., "A mobile app to increase fruit and vegetable acceptance among Finnish and Polish preschoolers: randomized trial," JMIR mHealth and uHealth, vol. 10(1), e30352, 2022.

[17] Folkvord F, Haga G, Theben A., "The Effect of a Serious Health Game on Children's Eating Behavior: Cluster-Randomized Controlled Trial," JMIR Serious Games, vol. 9(3):e23050, 2021.

[18] Ministry of Health Oman (Latest update May 2025), https://moh.gov.om/en/health-promotion/health-awareness/the-omani-guide-for-healthy-eating-english/.

[19] Andrew, L., Barwood, D., Boston, J., Masek, M., Bloomfield, L., & Devine, A., "Serious games for health promotion in adolescents–a systematic scoping review," Education and information technologies, vol. 28(5), pp. 5519-5550, 2023.

[20] Michael, S., Katrakazas, P., Petronoulou, O., Anastasiou, A., Iliopoulou, D., & Koutsouris, D. D., "Nutritionbuddy: a childhood obesity serious game," IEEE In 2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), pp. 5-8, 2018.

[21] Liu, N. Y. C., Tuah, N. M., & Yi, D. Y. C., "Conceptualizing the 'All You Can Eat'game to promote healthy eating habits among young children," Bulletin of Electrical Engineering and Informatics, vol. 14(2), pp.1507-1517, 2025.

[22] Weiland, A., Reiband, N., Schäffeler, N., Zurstiege, G., Giel, K. E., Zipfel, S., & Mack, I., "A serious game for the prevention of obesity in school children–impact of parent's involvement: a randomized controlled trial," Life, vol.12(6), pp.779, 2022.

[23] V. A. Okpanachi and I. Adaji, "The Design of Food Villain, a Serious Game to Influence Healthy Eating Habits Among African International Students," Jun. 2024.

[24] A. Kobenova and A. Kaiymova, "Making Beshbarmak: Games for Central Asian Cultural Heritage," Oct. 2024.

[25] M. F.-F. Chong, "Understanding children's perspectives of the influences on their dietary behaviours," Public Health Nutrition, vol. 25, no. 8, pp. 2156–2166, Feb. 2022

Anomaly Study of Computer Networks Based on Weighted Dynamic Network Representation Learning

Xin Wei*
School of Computer Engineering, Shangqiu University, Shangqiu, Henan, 476000, China

***Abstract*—One of the foremost significant challenges in the continuously increasing technological environment is the requirement to secure the authenticity of data. Network security is a primary method for securing the confidentiality of data throughout communication, one of several types of data security assurance. To secure networks against additional cyberattacks, trustworthy Anomaly Detection (AD) is essential. The drawbacks of conventional AD are gradually increasing as various types of attacks and network changes continually evolve. The researchers of the present study propose a novel approach that incorporates Weighted Long Short-Term Memory (WLSTM) networks with Dynamic Network Representation Learning (DNRL) to address these problems, referring to it as the Weighted Dynamic Network Representation Learning (WDNRL) paradigm. This investigation develops the WLSTM utilizing the Weight of Evidence (WoE), which periodically determines weights to network features in the resulting network model. The WLSTM design functions as the network's coordinator, obtaining data from the recommended model, upgrading the representation, and aggregating the features. The findings showed that the proposed model achieved high accuracy rates of 99.85% for Denial of Service (DoS) attacks and 99.55% for Distributed Denial of Service (DDoS) attacks when evaluated using two datasets, NSL-KDD and CICIDS-2017, compared to different models. Additionally, the simulation's F1-scores, recall rates, and precision are all above average, indicating that it is capable of identifying many network anomalies with minimal false positives (FP).**

***Keywords*—Network security; attacks; weighted dynamic network; anomaly detection; deep learning; LSTM**

I. Introduction

As cyber-attacks become increasingly challenging, network security remains the primary concern for many businesses worldwide. Modern technologies involved in Anomaly Detection (AD) are gradually becoming crucial in this constantly evolving setting, aiming to detect anomalous patterns or behaviors that could indicate a network attack or threaten security [1]. The variable environment of computer network behaviors, the variation of potential anomalies, and the challenge of distinguishing between undetectable and fake activity anomalies are just a few of the significant issues that AD still faces and fails to address [2]. Despite these limitations, the traditional AD has contributed significantly to providing ample security, and the methods employed have ranged from statistical methods to Machine Learning (ML). All such models have their strengths and limitations. Out of the early traditional methods, many have frequently relied on predefined rules or thresholds. Such models, although effective for known attack patterns, have primarily challenged the task of detecting novel and sophisticated threats [3]. This limitation prompted the exploration of advanced Machine Learning (ML) and Deep Learning (DL), which have presented improved adaptability and can also learn to identify attacks based on data patterns without explicit programming.

Even then, such solutions are also not without gaps; the ML and DL that were prominently used in more complex tasks, including popular choices like Generalized Regression Units (GRU) and Long Short-Term Memory (LSTM) [4], have also been faced with faults like such models can be hampered by drawbacks like the high dimensionality of network data, the imbalance between normal and anomalous samples, and also such models need for extensive labelled datasets for training.

Moreover, the static nature of some of these models has significantly limited their effectiveness in certain types of environments where network behavior is constantly changing. To address such limitations, Graph Learning (GL) has been introduced, a novel method that represents network data in the form of graphs, enabling the capture of complex relationships and interactions within the network [5]. This unique GL paradigm encoding enhances the network data model, providing better AR. However, GL's use in network AD is currently in its earliest phases, and several ongoing research projects are attempting to determine methods to maximize its most significant use.

Dynamic Network Representation Learning (DNRL) is an innovation in the domain that addresses AD challenges, particularly for networks that evolve [6]. Employing the physical and time data readily available in network data, the DNRL can operate in an AD and provide a deeper understanding of how networks perform. The synthesis of additional practical and verified DL designs, including GRU + LSTM, has enhanced the design's unique features. The DNRL integrates the parallel data processing abilities of GRU and LSTM with its DNRL limit, thereby rendering it more direct to find how data from networks continues to evolve. Including intricate data connections and temporal dynamics into AD enhances accuracy, making models more adaptable; DNRL's scalability, compared to other ML models, improves the analysis of vast volumes of network data. To circumvent the shortcomings of different approaches in network AD, the model employs a unified method.

Weighted Dynamic Network Representation Learning (WDNRL), the model developed in this research, is based on the factors previously labeled. Dynamic network representation learning for Alzheimer's disease (AD) has been integrated into research employing what is recommended: Weighted Long Short-Term Memory (WLSTM). The WDNRL utilizes the

*Corresponding Author

Weight of Evidence (WoE) enhanced LSTM to dynamically weight the features, allowing the model to adapt to changes in network configurations and interactions. This configuration helped the method to achieve improvement in the accuracy and timeliness of detecting the anomalies by employing a method to process the sequential network data and updating representations that are based on the interactions and temporal context by adjusting the parameters that impact the input features, which are measured using their predictive value the proposed WDNRL accurate AD. Evaluations of the WDNRL model were conducted using two datasets: the NSL-KDD and CICIDS-2017. The model yielded better results for all evaluated metrics compared to other models. The model achieved accuracies of 99.85% and 99.55% for DoS and DDoS attacks, respectively. It had shown precision of up to 98.35% and 96.35%, recall rates of 98.41% and 96.15%, and F1-scores of 97.38% and 97.17%, emphasizing its effectiveness in detecting a standard range of network AD with few False Positives (FP).

The study is organized as follows: Section II presents the literature review for understanding existing works, Section III outlines the methodology of the proposed work, Section IV describes the experiment setup, dataset, and analysis of results, and finally, Section V concludes the work.

II. Literature Review

In [7], the authors provided a work about foundational survey that maps the landscape of AD in the field of dynamic networks. The work categorizes anomalies into four distinct types and presents an analysis of methods standard among numerous approaches, providing an understanding of AD behaviors. They have further elaborated on a two-tiered taxonomy that classifies the studied methods based on different conceptual intuitions and the specific types of anomalies the methods are employed to detect. In [8], the authors have introduced a model called CmaGraph, designed to measure distances between vertices using DL. Implementing a novel combination of detection communities and AD blocks of data, their model focused on enhancing the AD boundaries. Based on the findings of the test examination, which verified the proposed model using multiple real-world data sets, the model they developed performed superior in recognizing dynamic network anomalies.

In [9], the authors developed a novel method termed Content-Aware Anomaly Detection (CAAD) to address the problem of recognizing anomalous connections and nodes from typical ones in a network of nodes. They achieved the layout of the method by integrating DNRL with an algorithm that encodes and decodes data. Convolutional Neural Networks (CNN) were used to build the design of their model. The network was built using structural and content data. The model was experimented with using different datasets, and the proposed CAAD proved superior performance when compared to other existing methods. In [10], the authors have suggested models targeted at exploring AD using a model built on blockchain transactions. The network of their model effectively leveraged the extensive amount of records and the practical application of the graph-like nature of blockchain data. Their research, using their projected model, has attempted to project the application and advantages of ML in the task of detecting anomalous transactions. It also revealed, through experimental analysis, that better performance was achieved using supervised learning techniques.

In [11], the authors have suggested a contrastive mechanism for Temporal Representation Learning (TRL) in the field of dynamic networks by designing a method named “Dynamic Network Contrastive Representation Learning” (DNCL). The model's design considers factors such as network topology, node feature data, and network evolution and applies these factors to contrast objective functions, thereby exemplifying the effectiveness of DNCL. In [12], the authors have introduced the Pyramidal Image Anomaly Detector (PIADE), a DL designed to extract image features from input data at multiple scale levels. This representation-based method is distinct from other standard representation models, employing methods that utilize significant similarity and perceptual loss to compare an input image to its reconstructed version effectively. The PIADE’s effectiveness is proven on datasets such as CIFAR10, COIL-100, and MVTec; for all these datasets, the proposed model generated better results. In [13], the authors have attempted to tackle the complex problem of AD in dynamic and multi-attributed network systems through the projected Multi-view Time-Series Hypersphere Learning (MTHL). The model attempted to project the multi-view time-series data into a shared latent subspace and learned a compact hypersphere around the normal samples; this way, it effectively distinguished between normal and abnormal cases. The method used in experiments has proven to have superior performance compared to other baseline methods.

In [14], the authors have suggested a model named DynAD to detect anomalous edges in time-evolving networks. The DynAD was designed with a model that employs a temporal graph convolution network, along with pooling operations, to extract node embeddings. The model further utilizes the GRU for capturing temporal information. Furthermore, the model incorporates an attention mechanism into the network to enhance its ability to detect AD against other baseline methods.

In [15], the authors have concentrated on exploring graph evolution-based prediction by learning spatiotemporal features, which are termed dynamics. Their method involved measuring the affinity scores related to nodes corresponding to the graphs. This way, the model provided a way to have better accuracy in detecting dynamic anomalies, and through this method, their model has addressed the challenge related to sparsity in real-world networks. Their work was experimented with in-network features of public transportation, and the simulation revealed the effectiveness of the proposed model. In [16], the authors presents a model built using a convGRU-based autoencoder for learning the spatial-temporal features of raw network traffic in an unsupervised manner. The application could generate meaningful compressed features and provide an effective means of detecting anomalies based on residual loss. Although the model’s detection capability regarding compressed data was irrelevant, it still enriched AD by exploring an interpretability method and providing a better-compressed representation of the network traffic [17-20].

III. METHODOLOGY

A. WoE and Information Value in Network Anomaly Detection

The Weight of Evidence (WoE) and Information Value (IV) are statistical measures traditionally used in credit scoring and risk assessment to transform definite variables into statistical numerical values and to evaluate the predictive strength of these variables. It is the natural logarithm of the ratio of the proportion of anomalies (positive) to the proportion of normal behavior (negative), as in Eq. (1):

$$WoE = ln(Distribution\ of\ Anomalies = Distribution\ of\ Normal\ Behavior) \quad (1)$$

For any network feature $'x'$, with classes $'x_i'$, WoE is calculated for each type to transform the categorical feature into a continuous scale that reflects the attributes related to AD. IV measures the cumulative predictive strength of a feature by calculating the sum of the differences in the proportions of anomalies and normal behaviors across all types of the feature, each weighted by their respective WoE, as in Eq. (2):

$$IV = \sum_i (Distribution\ of\ Anomalies_i - Distribution\ of\ Normal\ Behavior_i) \times WoE_i \quad (2)$$

The resultant IV values range from 0 to infinity (0 to ∞), out of which the higher values denote a more vital ability to predict network anomalies.

Information Value (IV) can be interpreted as follows:

- 0 to 0.02: Not applicable for prediction.
- 0.02 to 0.1: Weak predictive power.
- 0.1 to 0.3: Medium predictive power.
- 0.3 to 0.5: Strong predictive power.
- More remarkable than 0.5: Suspiciously high (maybe too good to be true or indicate data leakage).

B. Standard LSTM

LSTM is a type of Recurrent Neural Network (RNN) that is designed to overcome the challenges of learning long-term dependencies that were hard to solve using the traditional RNN. The LSTM units retain the data for extended periods due to their inherent structure. The LSTM is applied to fields such as time series analysis, Natural Language Processing (NLP), and network AD [21-25].

Each of the LSTM unit's models is characterized by a series of gates (see Fig. 1) that regulate the flow of data, which are described as follows:

1) Cell state. The Cell state $'C_t'$→ the LSTM unit's long-term memory, which stores relevant information throughout the sequence processing. It is modified by the forget gate and the input gate to add or remove information.

2) Forget gate. The Forget gate $'f_t'$ is used to determine which data to be maintained and which one to be discarded from the cell state. It applies a sigmoid function to the previous output $'h_{t-1}'$ and the current input $'x_t'$ which results in producing a value between 0 and 1 for each number in the cell state $'C_{t-1}'$, as in Eq. (3):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

where,

- W_f, b_f →the weight matrix and bias for the forget gate
- σ→ the sigmoid function.

3) Input gate. The input gate $'i_t'$ is employed to determine which new data is stored in the cell state. Simultaneously, it uses a *tanh* layer to create a vector of new candidate values, $\tilde{C}_t$, that could be added to the state, as in Eq. (4) and Eq. (5):

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (4)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (5)$$

where,

W_i, W_C and b_i, b_C → The weight matrices and biases for the input gate and the candidate values.

4) Cell state update. The old cell state C_{t-1} is updated to that of the new cell state $'C_t'$ by multiplying the old state value by that of the forget gate's output to discard the irrelevant data and perform the task of adding the input gate's output that is being multiplied by the candidate values to add new information, as in Eq. (6):

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (6)$$

5) Output gate. The output gate o_t is used to decide what part of the cell state is used for the output. A *tanh* function filters this output to ensure values stay between -1 and 1, and then it is multiplied by the output value of the sigmoid gate, as in Eq. (7) and Eq. (8):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (7)$$

$$h_t = o_t * \tanh(C_t) \quad (8)$$

where,

- W_o, b_o → The weight matrix and bias for the output gate
- h_t→ The final output of the LSTM unit at time $'t'$.

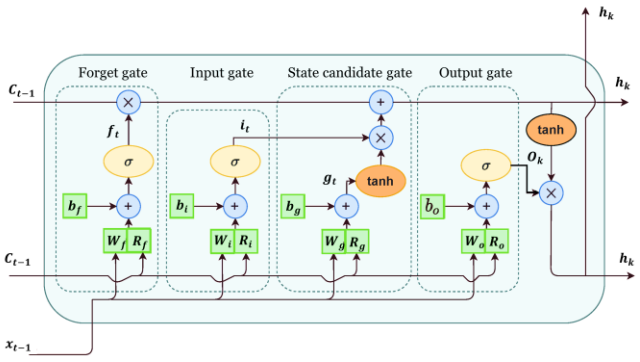


Fig. 1. LSTM

C. Weighted LSTM (WLSTM)

The WLSTM enhances the standard LSTM using a weighting mechanism (Fig. 2). This weighting mechanism-based enhancement is done using the WoE for each of the input features. The WoE using the input feature dynamically adjusts the impact of each input feature based on its predictive value regarding network anomalies. This enhancement enables the LSTM to improve its accuracy for AD in a network environment where features vary significantly over time. The WoE factored input data is fed into the input gate of the LSTM [26-30].

The method of integrating WoE to LSTM is discussed below:

1) Modified input gate. The input gate in WLSTM, which is represented as $'i_t^w'$ that integrates the WoE by weighting the input features $'x_t'$, as in Eq. (9):

$$i_t^w = \sigma\big(W_i \cdot \big[h_{t-1}, \big(x_t \times \mathrm{WoE}(x_t)\big)\big] + b_i\big) \quad (9)$$

Similarly, the candidate values for the cell state update, $\tilde{C}_t^w$, are calculated using the weighted inputs Eq. (10):

$$\tilde{C}_t^w = \tanh\big(W_C \cdot \big[h_{t-1}, \big(x_t \times \mathrm{WoE}(x_t)\big)\big] + b_C\big) \quad (10)$$

2) Cell state update. The cell state update is then modified to adapt to that of the WoE-modified input gate and candidate values, as in Eq. (11):

$$C_t^w = f_t * C_{t-1} + i_t^w * \tilde{C}_t^w \quad (11)$$

3) Output gate and final output. The output gate's function remains unchanged except for the inputs to this gate, which have been modified in the previous steps, and so the same is reflected in the final output, as in Eq. (12):

$$h_t^w = o_t * \tanh(C_t^w) \quad (12)$$

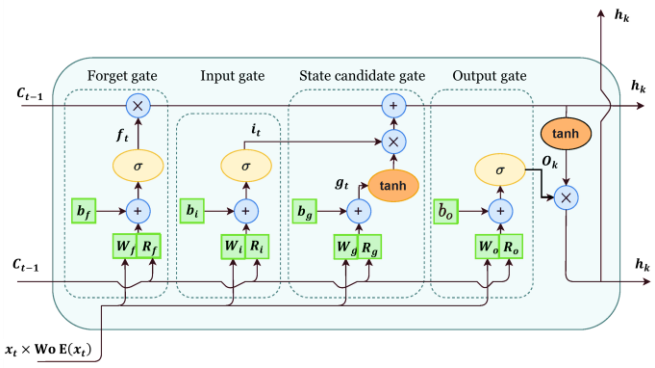


Fig. 2. WLSTM architecture.

D. Problem Definition

Let us consider a network represented by a graph $N = (S, C)$,

where,

- S → The set of nodes related to network devices and endpoints.
- C' → The connections between these nodes. The network behaviour over time is represented as a set of events. (s_a, s_b, τ), in which s_a, s_b denote the nodes.
- τ → The time related to the interaction between both the nodes.

The problem defined here is to do the updation of the DNRL based on the network's evolving network and to represent it in a reduced vector space Z^d to find the anomaly.

In real-world network environments, configurations and interactions change over time [31-35].

The changes in a network may happen due to any of the following actions:

- Addition of a New Node: This represents a device joining the network without initial connections.
- Creation of a New Connection: A new communication link is established between two nodes.
- Integration of a Connected Node: A new node joins and connects to existing nodes.
- Node Removal: The disconnection or failure of a network node.
- Connection Termination: The end of a communication link between nodes.
- Network Topology Changes: Adjustments in the network's structure due to various factors.

The proposed WDNRL (Fig. 3), utilizes the WLSTM to perform dynamic feature weighting and representation learning, thereby enhancing the DNRL. The model utilizes the WLSTM to process weighted features over time, identifying changes in the network's state and facilitating effective AD.

The proposed WDNRL is done using the following process:

1) Feature aggregation. Collecting network features over time and assigning weights based on their predictive value for AD.

2) Representation update. Continuously refining the network's vector selection to reflect its current state, using an optimization that ensures similar entities are represented closely in Z^d.

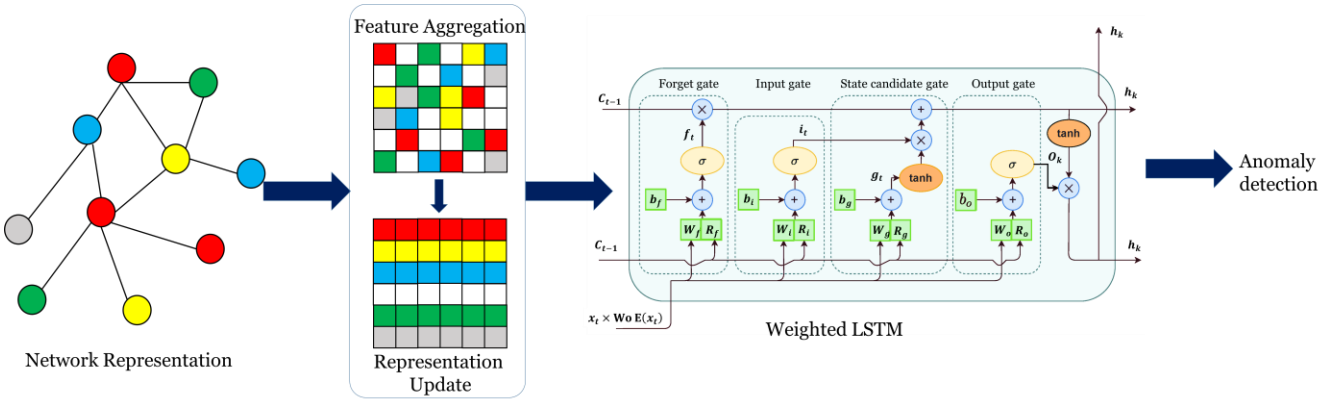


Fig. 3. WDNRL.

E. Aggregating Feature Data for Dynamic Anomaly Detection

The feature data aggregation technique employs predetermined parameters for learning distinct aggregation functions, which are denoted as $AggF^n$ for each layer $'n'$ within the range of 1 to N. These functions collate data from an entity's n-th layer neighbours using a set of weight matrices. W^n for each layer to help the flow of data across different network depths [36-40]. Procedure 1 presents the function of aggregating the feature data.

Procedure 1: Aggregation of Feature Information

Step 1. **Input:** Graph $N = (S, C)$; initial features $\{x_s, \forall s \in S\}$; depth N; weight matrices $W^n, \forall n \in \{1,2,\dots,N\}$; non-linear activation σ; aggregation functions $\text{AggF}^n, \forall n \in \{1,2,\dots,N\}$; neighborhood sampling sets $N^n, \forall n \in \{1,2,\dots,N\}$.

Step 2. **Output:** Vector selection as z_s for all $s \in S$.

Step 3. **Initialize** $h_s^0 = x_s$ for each entity $s \in S$.

Step 4. **For Each** layer $n = 1$ to N :

Step 5. **For Each** entity $\in S$

Step 6. **Aggregate** features $h_{N(s)}^n = \text{AggF}^n(\{h_u^{n-1}, \forall u \in N^n(s)\})$.

Step 7. **Update** entity feature $h_s^n = \sigma\left(W^n \cdot \text{concat}\left(h_s^{n-1}, h_{N(s)}^n\right)\right)$.

Step 8. **Normalize** h_s^n to unit length for each entity $s \in S$.

Step 9. **Assign** $z_s = h_s^N$ for each entity $s \in S$.

The input for the aggregation procedure comprises the network graph $N = (S, C)$ alongside the feature data for the entities within the network, denoted as $x_s, \forall s \in S$. A sampling strategy is deployed to find and select neighboring entities across 1 to N layers surrounding an assumed entity. Subsequently, the features of these selected neighboring entities are compiled. Further, a pooling approach, $AggF_{pool}^n$, is employed for aggregation, defined as the element-wise maximum $\max\left(\left\{\sigma\left(W_{\text{pool}} \cdot h_{u_i}^n + b\right), \forall u_i \in N(s)\right\}\right)$, where max represents the element-wise maximum operation, and $'\sigma'$ is a non-linear activation function. The current selection of an entity, h_s^{n-1} along with the pooled vector of neighborhood features, $h_{N(s)}^n$, are then processed through a fully connected layer employing the non-linear activation function $'\sigma'$, which effectively updates the entity's selection. The outcome is the enhanced vector selection of entity $'s'$.

F. Updating Network Selection for AD

The recommended method provides a model for refreshing entity models based on inputs collected in response to network dynamics. This method enables objects to rapidly modify their models to reflect the evolving state of the network by emphasizing the incorporation of recent changes to the network. The method of updating the web model is explained in Procedure 2.

Procedure 2: Network Selection Update Using WLSTM

Input: Change indicator (flag); vector selection $z_{s_a}(\tau)$ and $z_{s_b}(\tau)$; prior selection state $r_s(\tau-)$; time interval $\Delta\tau$; decay function g; weight matrices W_1, W_2; activation function $\text{act}(\cdot)$.

Output: Updated selection state $r_s(\tau)$.

Step 1. Formulate the entity impact embedding $e(\tau) = \text{act}\left(W_1 \cdot z_{s_a}(\tau) + W_2 \cdot z_{s_b}(\tau) + \epsilon\right)$, capturing the essence of recent interactions.

Step 2. Utilize $g(\Delta\tau)$ to modify the selection, highlighting the significance of temporal proximity to the last update.

Step 3. Employ the WLSTM×to refresh the entity's selection as $r_s(\tau)$, considering the impact embedding $e(\tau)$, the preceding selection state $r_s(\tau-)$, and the change flag.

In this model, the network changes are perceived as sequences which are denoted by $(s_a, s_b, \tau,$ flag $)$, where s_a and s_b represent the entities involved, $'\tau'$ indicates the time of change, and the flag defines network expansion ($flag = 1$) or contraction (flag $= -1$). The WLSTM is employed to handle temporal dynamics, thereby enabling the model to adapt to changes in network dynamics [41-45]. To establish consistency among closely situated entities, an optimization technique is used to minimize the dissimilarity, which is defined as a loss function, as in Eq. (13):

$$J(z_s) = -\text{Log}\left(\sigma\left(z_s^T z_{s'}\right)\right) - Q \cdot \mathbb{E}_{s_n \sim P_n(s)}\left[\text{Log}\left(\sigma\left(-z_s^T z_{s_n}\right)\right)\right] \quad (13)$$

where,

- $z_s \rightarrow$ The embedding of an entity impacted by network changes
- $Q \rightarrow$ The count of negative samples drawn

- $P_n(s)$ d→ A distribution for negative sampling, s' is a neighbouring entity near
- s, s_n → Entities sampled from the negative sampling distribution.

IV. Experimental Setup

The experiments were conducted using a server equipped with an Intel(R) Core(TM) i5-10210U CPU at 2.11 GHz, running Windows 10 and utilizing PyTorch for model implementation.

A. Dataset for Experimentation

For the aim of testing the success rate of the proposed approach, the present research utilizes two distinct datasets: NSL-KDD and CICIDS-2017.

1) NSL-KDD. A collection of four distinct types of cyberattacks has been included in the NSL-KDD dataset. These are Denial of Service (DoS), User-to-Root (U2R), Remote-to-Local (R2L), and Probing (Probe). The dataset also includes data on normal network activities. The dataset comprises 41 features for all samples and includes an ID for each example, indicating whether it represents a typical instance or a malicious attack type.

2) CICIDS-2017. The CICIDS-2017 dataset captures 79 network flow features. The dataset comprises various connection logs, including SSH, email, HTTP, and FTP, collected from 25 users on different operating systems.

The following Tables I and II present the type of attacks contained in both datasets:

TABLE I Attack Types in NSL-KDD Dataset

Dataset	Split	U2R	DoS	R2L	Probe	Normal
NSL-KDD	**Train**	52	45,927	995	11,656	67,343
	Test	67	7458	2887	2422	9710

TABLE II Attack Types in the CICIDS-2017 Dataset

Dataset	Split	DDoS	FTP-Patator	PortScan	SQL Injection	Benign
CICIDS-2017	**Train**	112,901	6997	140,043	19	72,7397
	Test	25388	1574	31,492	4	163,572

The WDNRL was compared against other baseline models, such as RNN, LSTM, and DNRL + LSTM [46-50]. The proposed model was trained using the hyperparameters shown in Table III for the above two datasets:

TABLE III Hyperparameter for Training

Hyperparameter	Specific Value
Learning Rate	$1e{-}3$
Number of Epochs	200
Batch Size	6464
Weight Decay	$1e{-}4$
LSTM Hidden Units	256
Sequence Length	30
Learning Rate Scheduler	Step Decay
Dropout Rate	0.2
Negative Sampling Rate	10
Embedding Dimension	128

B. Evaluation Metrics

The evaluation of the model is done using the following metrics:

- Accuracy: It measures the model's ability in correct AD, as in Eq. (14):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

- Precision: It measures the proportion of accurate AD traffic to that of the total classified as attacks in the dataset, as in Eq. (15):

$$\text{Precision} = \frac{TP}{TP+FP} \tag{15}$$

- Recall: It calculates the fraction of correctly predicted attack traffic over the actual attack instances, as in Eq. (16):

$$\text{Recall} = \frac{TP}{TP+FN} \tag{16}$$

- F1-score: The F1-score is the harmonic mean of precision and recall, as in Eq. (17):

$$\text{F1-score} = \frac{2\times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

C. Accuracy Analysis

In the NSL-KDD (Table IV and Fig. 4), the proposed WDNRL achieved an impressive accuracy of 99.85% for Denial of Service (DoS) attacks, surpassing the scores of the other models. Next, the DNRL with LSTM follows closely at 98.7%. For the probing attacks, the WDNRL scored 98.76% accuracy. Remote-to-Local (R2L) and User-to-Root (U2R) attacks are challenging types of attacks to detect. For these attacks, the proposed model achieved accuracies of 87.21% and 78.57%. The improvement of LSTM achieved using WoE is reflected in the results, showing a notable 10% decrease in U2R attacks.

For the CICIDS-2017 (Table IV and Fig. 4), the WDNRL proves a clear edge over all other models, achieving accuracies of 99.55% and 98.46% for detecting Distributed Denial of Service (DDoS) and FTP-based attacks. In detecting PortScan (PS) and SQL injection attacks, the WDNRL achieves accuracies of 94% and 86.95%, which is a significant improvement over the immediately following DNRL with LSTM, which achieved only 67.5% and 79% [51-55].

It is also worth noting that the proposed model achieves the highest score in classifying normal traffic, with a perfect score of 100 and 99.7% for the NSL-KDD and CICIDS-2017.

TABLE IV ACCURACY PERFORMANCE RESULTS

Models	NSL-KDD					CICIDS-2017				
	DoS	Probing	R2L	U2R	Normal	DDoS	FTP	PS	SQL	Normal
RNN	96.04	93.63	80.33	64.28	97.3	94.44	94.82	53.13	76.71	93.7
LSTM	97.08	94.28	79.35	61.73	98.6	93.77	93.44	74.59	77.68	94.1
DNRL + LSTM	98.77	97.68	85.16	68.54	99.76	96.26	95.36	67.57	79.09	97.5
WDNRL (Proposed)	99.85	98.76	87.21	78.57	100	99.55	98.46	94	86.95	99.7

TABLE V PRECISION PERFORMANCE RESULTS

Models	NSL-KDD					CICIDS-2017				
	DoS	Probing	R2L	U2R	Normal	DDoS	FTP	PS	SQL	Normal
RNN	96.3	96.21	84.98	60	95.26	94.03	89.59	70.47	72.47	91.36
LSTM	96.56	94.21	78.32	62.67	96.5	93.62	91.95	79.47	73.96	92.56
DNRL + LSTM	96.6	94.85	77.37	70.91	97.1	93.86	92.97	64.81	77.11	94.28
WDNRL (Proposed)	98.35	96.29	91.03	79.61	99	96.35	94.61	98.78	83.92	98.23

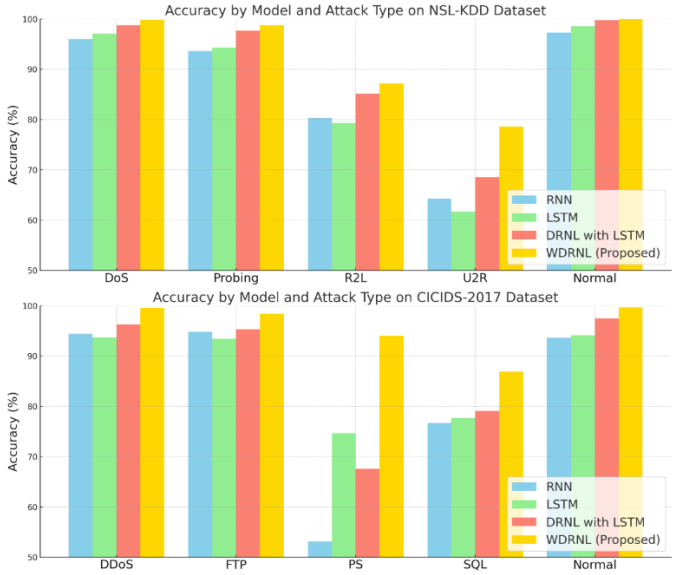


Fig. 4. Accuracy analysis for the NSL-KDD and CICIDS-2017

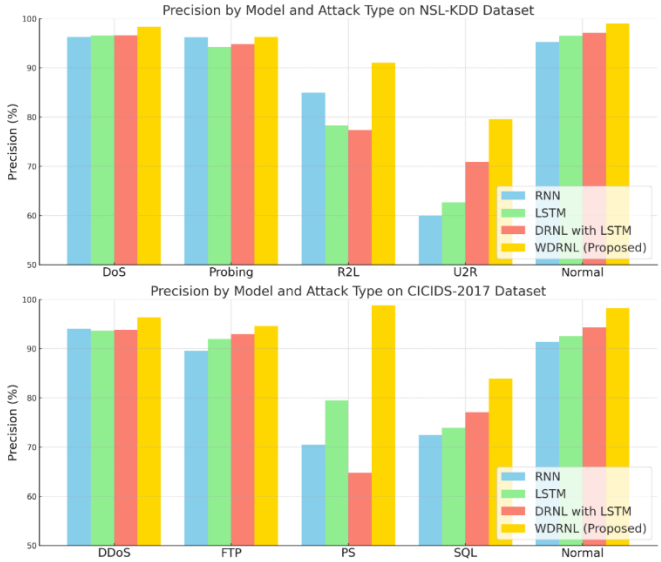


Fig. 5. Precision analysis for the NSL-KDD and CICIDS-2017

D. Precision Analysis

Table V and Fig. 5 present the analysis of the precision achieved by the proposed model and the other compared models. For the NSL-KDD, the proposed model achieves a precision of 98.35% for DoS attacks and 96.29% for Probing attacks; these high scores demonstrate its effectiveness in accurately identifying threat types with minimal error. Further, the model shows 91.03% and 79.61% precision scores for the R2L and U2R attacks [56-60]. When compared to the CICIDS again, the proposed model showed better performance across all attacks. Specifically, the WDNRL achieved a precision of 96.35% in detecting DDoS attacks and 94.61% for FTP-based anomalies. Additionally, the precision in identifying PortScan and SQL injection attacks reached scores of 98.78% and 83.92%, respectively. From both datasets, the proposed model showed a much higher score than the other compared models [61-70]. Further, the model's performance in detecting normal traffic was also higher, with 99% and 98.3% precision for the NSL-KDD and CICIDS.

E. Recall Analysis

Table VI and Fig. 6 analyze the performance of the proposed model in comparison with the other models in terms of the Recall metric. For the NSL-KDD, the WDNRL's recall performance was the highest, capturing all attacks without missing them mainly. For the DoS and Probing attacks [71-75], the proposed model achieved scores of 98.41% and 96.38%. The recall scores for the Remote to Local (R2L) and User to Root (U2R) attacks were 85.36% and 62.9%, respectively. The only model that could come close is the DNRL + LSTM [76-82]; the rest of the models performed too poorly compared to the proposed model. For CICIDS-2017, a similar trend was observed, with the proposed model exhibiting a higher recall score across all attack cases, recording 96.15% recall in detecting DDoS, 95.13% for FTP-based attacks, 78.32% for PortScan, and 84.25% for SQL injection attacks. For both datasets, the model achieved recall rates of 98.1% and 98.2% for normal cases in the NSL-KDD and CICIDS-2017.

TABLE VI RECALL PERFORMANCE RESULTS

Models	NSL-KDD					CICIDS-2017				
	DoS	**Probing**	**R2L**	**U2R**	**Normal**	**DDoS**	**FTP**	**PS**	**SQL**	**Normal**
RNN	96.96	91.27	65.93	31.2	92.5	89.88	89.03	58.61	66.51	90.78
LSTM	95.75	92.67	54.17	34.13	93.71	92.44	91.02	40.32	51.3	91.37
DNRL with LSTM	95.44	93.31	73.21	31.2	94.4	93.77	93.09	45.98	52.74	93.28
WDNRL (Proposed)	98.41	96.38	85.36	62.9	98.1	96.15	95.13	78.32	84.25	98.28



Fig. 6. Recall analysis for the NSL-KDD and CICIDS-2017

F. F1-Score Analysis

Table VII and Fig. 7 present the results of F1-score metrics from the NSL-KDD and CICIDS-2017. On the NSL-KDD dataset, the WDNRL, as seen in the earlier metrics, was the higher-performing model, achieving F1-scores of 97.38% and 97.34% for DoS and Probing attacks. For R2L and U2R attacks, it scored 91.2% and 78%. The next performing model was DNRL + LSTM, which scored 94.6%, 95.2%, 84.2%, and 61% for DoS, Probing, R2L, and U2R. When compared to the CICIDS-2017, the proposed model again demonstrated higher F1-score values, achieving 97.17% and 96.16% for DDoS and FTP, respectively, as well as 88.1% and 91.38% for PortScan (PS) and SQL injection attacks. In comparison to classifying normal traffic, the proposed model achieved an impressive score of 99.2% for NSL-KDD and CICIDS-2017.

Upon comparing all the metrics, it is evident that the proposed model outperformed the others, achieving the highest score across all metrics for both datasets. The next model that follows the proposed model is the DNRL+ LSTM for all metrics in both datasets. Of the other two models, the RNN and its variant, LSTM, exhibit similar performance, with LSTM holding a slight edge in certain metrics for specific types of attacks.

TABLE VII F1 PERFORMANCE RESULTS

Models	**NSL-KDD**					**CICIDS-2017**				
	DoS	**Probing**	**R2L**	**U2R**	**Normal**	**DDoS**	**FTP**	**PS**	**SQL**	**Normal**
RNN	92.16	92.43	64.05	45	91.13	90.47	89.31	49.91	71.58	90.28
LSTM	91.02	91.07	75.23	43.33	92.98	92.99	90.35	69.62	61.93	91.92
DNRL with LSTM	94.63	95.23	84.26	61.06	95.5	96.82	95.03	83.49	82.64	97.28
WDNRL (Proposed)	97.38	97.34	91.2	78	99.2	97.17	96.16	88.1	91.38	99.21

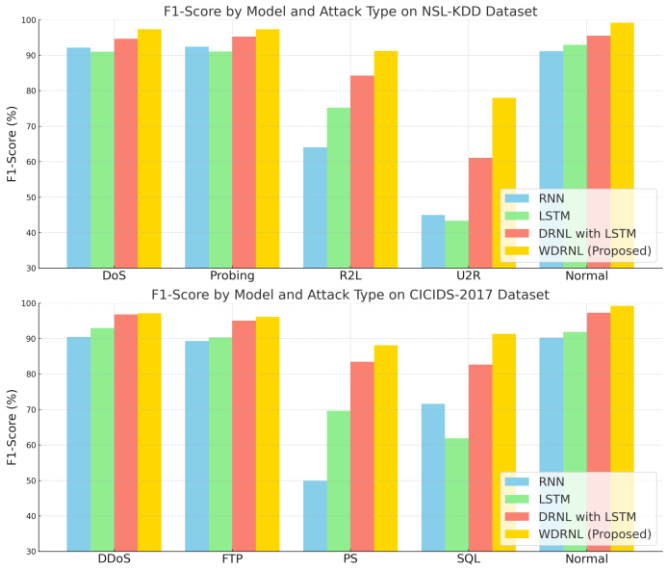


Fig. 7. Recall analysis for the NSL-KDD and CICIDS-2017.

V. CONCLUSION AND FUTURE WORK

This study aims to adapt the DNRL for use in AD applications within networks. The study proposed designing a better AD for dynamic network instances. The idea was cropped with the introduction and progress in the field of DNR. The study attempted to optimize the standard DNRL by incorporating a weighted concept to aggregate features of higher importance to the network AD. To achieve this, the proposed study introduced a WoE-optimized LSTM that computes the WoE over the possible AD against normal traffic for the input data. Using the WLSTM, the WDNRL effectively navigates the evolving landscape of network features. The effectiveness of the proposed model was evaluated using the NSL-KDD and CICIDS-2017 datasets, with metrics including accuracy, precision, recall, and F1-score. It is observed that, for all metrics, the proposed model outperforms the other baseline models for all attack instances.

Future work would focus on enhancing the model by integrating more advanced DL and scaling the model to other AD modalities.

REFERENCES

[1] Dunn Cavelty, M., & Wenger, A. (2020). Cyber security meets security politics: Complex technology, fragmented politics, and networked science. Contemporary Security Policy, 41(1), 5-32.

[2] Singh, J. P. (2022). Mitigating Challenges in Cloud Anomaly Detection Using an Integrated Deep Neural Network-SVM Classifier Model. Sage Science Review of Applied Machine Learning, 5(1), 39-49.

[3] Liu, Q., Hagenmeyer, V., & Keller, H. B. (2021). A review of rule learning-based intrusion detection systems and their prospects in smart grids. IEEE Access, 9, 57542-57564.

[4] Sengupta, S., Basak, S., Saikia, P., Paul, S., Tsalavoutis, V., Atiah, F., ... & Peters, A. (2020). A review of deep learning with particular emphasis on architectures, applications, and recent trends. Knowledge-Based Systems, 194, 105596.

[5] Wang, S., Hu, L., Wang, Y., He, X., Sheng, Q. Z., Orgun, M. A., ... & Yu, P. S. (2021). Graph learning based recommender systems: A review. arXiv preprint arXiv:2105.06339.

[6] Chikwendu, I. A., Zhang, X., Agyemang, I. O., Adjei-Mensah, I., Chima, U. C., & Ejiyi, C. J. (2023). A comprehensive survey on deep graph representation learning methods. Journal of Artificial Intelligence Research, 78, 287-356.

[7] Ranshous, S., Shen, S., Koutra, D., Harenberg, S., Saloutos's, C., & Samatova, N. F. (2015). Anomaly detection in dynamic networks: a survey. Wiley Interdisciplinary Reviews: Computational Statistics, 7(3), 223-247.

[8] Lin, W., Bao, X., & Li, M. J. (2021). Cmagraph: A triblocks anomaly detection method in dynamic graph using evolutionary community representation learning. In Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part I 30 (pp. 105-116). Springer International Publishing.

[9] Li, Z., Jin, X., Zhuang, C., & Sun, Z. (2020). Content-Aware Anomaly Detection with Network Representation Learning. In Algorithms and Architectures for Parallel Processing: 20th International Conference, ICA3PP 2020, New York City, NY, USA, October 2–4, 2020, Proceedings, Part III 20 (pp. 50-64). Springer International Publishing.

[10] Martin, K., Rahouti, M., Ayyash, M., & Alsmadi, I. (2022). Anomaly detection in blockchain using network representation and machine learning. Security and Privacy, 5(2), e192.

[11] Jiao, P., Chen, H., Tang, H., Bao, Q., Zhang, L., Zhao, Z., & Wu, H. (2024). Contrastive representation learning on dynamic networks. Neural Networks, 106240.

[12] Mishra, P., Piciarelli, C., & Foresti, G. L. (2020). A neural network for image anomaly detection with deep pyramidal representations and dynamic routing. International Journal of Neural Systems, 30(10), 2050060.

[13] Teng, X., Lin, Y. R., & Wen, X. (2017, November). Anomaly detection in dynamic networks using multi-view time-series hypersphere learning. In Proceedings of the 2017 ACM Conference on Information and Knowledge Management (pp. 827-836).

[14] Zhu, D., Ma, Y., & Liu, Y. (2020, December). A flexible attentive temporal graph networks for anomaly detection in dynamic networks. In 2020 IEEE 19th International Conference on Trust, Security, and Privacy in Computing and Communications (TrustCom) (pp. 870-875). IEEE.

[15] Lee, J., Bae, H., & Yoon, S. (2020). Anomaly detection by learning dynamics from a graph. IEEE Access, 8, 64356-64365.

[16] Kopp, F. (2022). Representation learning for content-sensitive anomaly detection in industrial networks. arXiv preprint arXiv:2205.08953.

[17] Thirimanne, S. P., Jayawardana, L., Yasakethu, L., Liyanaarachchi, P., & Hewage, C. (2022). Deep neural network-based real-time intrusion detection system. SN Computer Science, 3(2), 145.

[18] Yang, L., Li, J., Yin, L., Sun, Z., Zhao, Y., & Li, Z. (2020). Real-time intrusion detection in wireless network: A deep learning-based intelligent mechanism. Ieee Access, 8, 170128-170139.

[19] Rokade, M. D., & Sharma, Y. K. (2021, March). MLIDS: a machine learning approach for intrusion detection for real-time network dataset. In 2021 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 533-536). IEEE.

[20] Liu, J., Zhang, W., Ma, T., Tang, Z., Xie, Y., Gui, W., & Niyoyita, J. P. (2020). Toward security monitoring of industrial cyber-physical systems via hierarchically distributed intrusion detection. Expert Systems with Applications, 158, 113578.

[21] Dutt, I., Borah, S., & Maitra, I. K. (2020). Immune system based intrusion detection system (IS-IDS): A proposed model. IEEE Access, 8, 34929-34941.

[22] Rose, J. R., Swann, M., Grammatikakis, K. P., Koufos, I., Bendiab, G., Shiaeles, S., & Kolokotronis, N. (2022). IDERES: Intrusion detection and response system using machine learning and attack graphs. Journal of Systems Architecture, 131, 102722.

[23] Hossain, M. D., Inoue, H., Ochiai, H., Fall, D., & Kadobayashi, Y. (2020). LSTM-based intrusion detection system for in-vehicle can bus communications. Ieee Access, 8, 185489-185502.

[24] Hnamte, V., Nhung-Nguyen, H., Hussain, J., & Hwa-Kim, Y. (2023). A novel two-stage deep learning model for network intrusion detection: LSTM-AE. IEEE Access.

[25] Fu, Z. (2022). Computer network intrusion anomaly detection with recurrent neural network. Mobile Information Systems, 2022.

[26] Alferaidi, A., Yadav, K., Alharbi, Y., Razmjooy, N., Viriyasitavat, W., Gulati, K., ... & Dhiman, G. (2022). Distributed deep CNN-LSTM model for intrusion detection method in IoT-based vehicles. Mathematical Problems in Engineering, 2022.

[27] Sridhar Panneerselvam et al., (2024), Federated learning-based fire detection method using local MobileNet, Scientific Reports, vol. 14, No. 30388, pp. 1-24,

[28] Asir Chandra Shinoo Robert Vincent and Sudhakar Sengan, (2024). Edge computing-based ensemble learning model for health care decision systems, Sci Rep 14, 26997.

[29] Asir Chandra Shinoo, Robert Vincent and Sudhakar Sengan, Effective clinical decision support implementation using a multi-filter and wrapper optimisation model for Internet of Things based healthcare data. Sci Rep 14, 21820.

[30] Gulista Khan et al. (2024) Energy-Efficient Routing Algorithm for Optimizing Network Performance in Underwater Data Transmission Using Gray Wolf Optimization Algorithm, Journal of Sensors, Vol. 2024, No. 2288527, 1-15.

[31] Sudhakar Sengan et al. 2023, Improved LSTM-Based Anomaly Detection Model with Cybertwin Deep Learning to Detect Cutting-Edge Cybersecurity Attacks, Human-centric Computing and Information Sciences, Vol. 13, No. 55.

[32] Varatharaj Myilsamy et al., (2023). State-of-Health Prediction for Li-ion Batteries for Efficient Battery Management System Using Hybrid Machine Learning Model, Journal of Electrical Engineering & Technology.

[33] Indumathi Nallathambi et al. (2023). Impact of Fireworks Industry Safety Measures and Prevention Management System on Human Error Mitigation Using a Machine Learning Approach, Sensors, 23 (9), 4365.

[34] Vinothini Arumugham et al. (2023) An Artificial-Intelligence-Based Renewable Energy Prediction Program for Demand-Side Management in Smart Grids, Sustainability, Vol. 15, No. 6, 5453.

[35] Rasheed Abdulkader et al. (2023), Soft Computing in Smart Grid with Decentralized Generation and Renewable Energy Storage System Planning, Energies, Vol. 16, No. 6, 2655.

[36] Parkavi Krishnamoorthy et al. (2023). Effective Scheduling of Multi-Load Automated Guided Vehicle in Spinning Mill: A Case Study, IEEE Access.

[37] Arodh Lal Karn et al. (2023). IoT Based Smart Framework Monitoring System for Power Station, Computers, Materials & Continua, Vol. 74, No. 3, pp. 6019-6037.

[38] Arul Rajagopalan et al. 2022, Oscar Danilo Montoya, Walid El-Shafai, Mostafa M. Fouda, and Moustafa H. Aly, Modernized Planning of Smart Grid Based on Distributed Power Generations and Energy Storage Systems Using Soft Computing Methods, Energies, Vol. 15, No. 23, 8889.

[39] Eman S. Sabry et al. (2022), Sketch-Based Retrieval Approach Using Artificial Intelligence Algorithms for Deep Vision Feature Extraction, Axioms, Vol. 11, No. 12, pp. 663.

[40] Arodh Lal Karn et al. (2022). Measuring the Determining Factors of Financial Development of Commercial Banks in Selected SAARC Countries, Journal of Database Management, Vol. 33, No. 1, 2022 pp. 1–21.

[41] Prabu Selvam et al. (2022) A Transformer-Based Framework for Scene Text Recognition, IEEE Access, Vol. 10, pp. 100895-100910.

[42] Arodh Lal Karn et al. (2022) Fuzzy and SVM-Based Classification Model to Classify Spectral Objects in Sloan Digital Sky, IEEE Access, Vol. 10, pp. 101276-101291.

[43] Tribhuwan Kumar et al. (2022). Fuzzy Logic and Machine Learning-Enabled Recommendation System to Predict Suitable Academic Program for Students, Mathematical Problems in Engineering, vol. 2022, No. 5298468, pp. 1-7.

[44] Roy Setiawan et al. (2022). IoT Based Virtual E-Learning System for Sustainable Development of Smart Cities, Journal of Grid Computing, Vo. 20, No. 24.

[45] S. Priyadarsini et al. (2022), Automatic Liver Tumor Segmentation in CT Modalities Using MAT-ACM, Computer Systems Science and Engineering, vol. 43, no. 3, pp. 1057–1068, 2022.

[46] S. Priyadarsini et al. (2022). Classification of Liver Tumors from Computed Tomography Using NRSVM, Intelligent Automation & Soft Computing, Vol. 33, No. 3, 2022, pp. 1517-1530.

[47] Thirumoorthy Palanisamy et al. (2022) Improved Energy Based Multi-Sensor Object Detection in Wireless Sensor Networks, Intelligent Automation & Soft Computing, Vol. 33, No. 1, pp.227-244.

[48] Omaima Bamasaq et al. (2022). Syed Hamid Hassan, Distance Matrix and Markov Chain Based Sensor Localization in WSN, Computers, Materials & Continua, Vol. 71, No. 2, pp. 4051-4068,

[49] Sudhakar Sengan et al. (2021), Real-Time Automatic Investigation of Indian Roadway Animals by 3D Reconstruction Detection Using Deep Learning for R-3D-YOLOv3 Image Classification and Filtering, Electronics, Vol. 10, No. 24, 3079.

[50] Abolfazl Mehbodniya et al. (2021). Fetal health classification from cardiotocographic data using machine learning, Expert Systems.

[51] Sudhakar Sengan et al. (2021). A Secure Recommendation System for Providing Context-Aware Physical Activity Classification for Users, Security and Communication Networks, Vol. 2021, No. 4136909, pp. 1-15.

[52] Vasanthi Raghupathy et al. (2022). Interactive Middleware Services for Heterogeneous Systems, Computer Systems Science and Engineering, Vol. 41, No. 3, pp. 1241-1253.

[53] Abolfazl Mehbodniya et al. (2022). Proportional Fairness Based Energy Efficient Routing in Wireless Sensor Network, Computer Systems Science and Engineering, Vol. 41, No. 3, pp. 1071-1082.

[54] D. Stalin David et al. (2022), Enhanced Detection of Glaucoma on Ensemble Convolutional Neural Network for Clinical Informatics, Computers, Materials & Continua, Vol. 70, No. 2, pp. 2563-2579, .

[55] D. Stalin David et al. (2022). Cloud Security Service for Identifying Unauthorized User Behaviour, CMC-Computers, Materials & Continua, Vol. 70, No. 2. pp. 2581-2600.

[56] Ngangbam Phalguni Singh et al. 2021. Investigation on characteristics of Monte Carlo model of single electron transistor using Orthodox Theory, Sustainable Energy Technologies and Assessments, Vol. 48, 101601.

[57] K. Rajakumari et al., (2022), Fuzzy Based Ant Colony Optimization Scheduling in Cloud Computing, Computer Systems Science and Engineering, Vol. 40, No. 2, pp. 581-592.

[58] Arodh Lal Karn et al. (2021), An integrated approach for sustainable development of wastewater treatment and management system using IoT in smart cities, Soft Computing.

[59] R. Nithya et al., (2022). Edwin Hernan Ramirez-Asis, Priya Velayutham, V. Subramaniyaswamy, Sudhakar Sengan, An Optimized Fuzzy Based Ant Colony Algorithm for 5G-MANET, Computers, Materials & Continua, Vol. 70, No. 1, pp: 1069–1087.

[60] Razia Sulthana Abdul Kareem et al., (2021). Multilabel land cover aerial image classification using convolutional neural networks, Arabian Journal of Geosciences, Vol. 14, No. 1681.

[61] Roy Setiawan et al., Utilizing Index-Based Periodic High Utility Mining to Study Frequent Itemsets, Springer, Arabian Journal for Science and Engineering, 2021.

[62] K. Muthumayil et al., (2021) A Big Data Analytical Approach for Prediction of Cancer Using Modified K-Nearest Neighbour Algorithm, Journal of Medical Imaging and Health Informatics, Vol. 11, No. 8, pp. 2120-2125 (6).

[63] Keerthana Nandakumar et al. (2021) Sudhakar Sengan, Securing data in transit using data-in-transit defender architecture for cloud communication, Soft Computing.

[64] Huidan Huang et al. (2021) Emotional intelligence for board capital on technological innovation performance of high-tech enterprises, Aggression and Violent Behavior, 101633.

[65] Sudhakar Sengan et al. (2021) Detection of false data cyber-attacks for the assessment of security in smart grid using deep learning, Computers & Electrical Engineering, Vol. 93, 107211.

[66] Sudhakar Sengan et al., (2021). Cost-effective and efficient 3D human model creation and re-identification application for human digital twins, Multimedia Tools and Applications, 2021. DOI:10.1007/s11042-021-10842-y.

[67] Omnia Saidani Neffati et al., (2021). Migrating from traditional grid to smart grid in smart cities promoted in developing country, Sustainable Energy Technologies and Assessments, Vol. 45.

[68] Prabhakaran Narayanan et al. (2021) Novel Collision Detection and Avoidance System for Mid-vehicle Using Offset-Based Curvilinear Motion. Wireless Personal Communication.

[69] Balajee Alphonse et al., (2021). Modeling and multi-class classification of vibroarthographic signals via time domain curvilinear divergence random forest, J Ambient Intell Human Comput.

[70] Omnia Saidani Neffati et al., (2021). An educational tool for enhanced mobile e-Learning for technical higher education using mobile devices for augmented reality, Microprocessors and Microsystems, Vol. 83, 104030.

[71] Prabhakaran Narayanan et al., (2021). Analysis and design of fuzzy-based manoeuvring model for mid-vehicle collision avoidance system. J Ambient Intell Human Comput.

[72] Sudhakar Sengan et al., (2021) Network Embedding Architecture using Laplace Regularization-Non-Negative Matrix Factorization for Virtualization, Microprocessors and Microsystems.

[73] K. Sathish Kumar et al. (2020). Area Based Efficient and Flexible Demand Side Management To Reduce Power And Energy Using Evolutionary Algorithms, Malaysian Journal of Computer Science, 27335.

[74] L. Arokia Jesu Prabhu et al., (2020) Medical Information Retrieval Systems for e-Health Care Records using Fuzzy Based Machine Learning Model, Microprocessors and Microsystems, DOI:10.1016/j.micpro.2020.103344.

[75] N. Satheesh et a., (2020). Flow-based Anomaly Intrusion Detection using Machine Learning Model with Software Defined Networking for OpenFlow Network, Microprocessors, and Microsystems.

[76] Sengan Sudhakar et al., (2020). A fuzzy-based high-resolution multi-view deep CNN for breast cancer diagnosis through SVM classifier on visual analysis, Journal of Intelligent & Fuzzy Systems, pp. 1-14.

[77] Sengan Sudhakar et al. (2020). Images super-resolution by optimal deep AlexNet architecture for medical application: A novel DOCALN, Journal of Intelligent & Fuzzy Systems, pp. 1-14.

[78] V. Vijaya Kumar et al. (2020). Design of peer-to-peer protocol with sensible and secure IoT communication for future internet architecture, Microprocessors and Microsystems, Vol. 78.

[79] Sudhakar Sengan et al., (2020). Enhancing cyber-physical systems with hybrid smart city cyber security architecture for secure public data-smart network, Future Generation Computer Systems.

[80] Ganesh Kumar, K and Sudhakar Sengan (2020) Improved Network Traffic by Attacking Denial of Service to Protect resources using Z-Test

Based 4-Tier Geomark Traceback (Z4TGT), Wireless Personal Communications.

[81] S. Sudhakar et al. (2020). Unmanned Aerial Vehicle (UAV) based Forest Fire Detection and monitoring for reducing false alarms in forest fires, Computer Communications 149, pp. 1–16.

[82] Sudhakar Sengan and Chenthur Pandian S, (2016), Hybrid Cluster-based Geographical Routing Protocol to Mitigate Malicious Nodes in Mobile Ad Hoc Network, International Journal of Ad Hoc and Ubiquitous Computing, Vol. 21, No. 4, pp. 224-236.

Deepfake Audio Detection Using Feature-Based and Deep Learning Approaches: ANN vs ResNet50

Reham Mohamed Abdulhamied[1], Sarah Naiem[2], Mona M. Nasr[3], Farid Ali Moussa[4]
Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt[1, 2, 3]
Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt[4]

***Abstract*—The proliferation of algorithms and commercial tools for generating synthetic audio has sparked a surge in mis-information, especially on social media platforms. Consequently, significant attention has been devoted to detect such misleading content in recent years. However, effectively addressing this challenge remains elusive, given the increasing naturalness of fake audio. This study introduces a model designed to distinguish between natural and fake audio, employing a two-stage approach: an audio preparation phase involving raw audio manipulation, followed by modeling using two distinct models. The first model employed feature extraction through wavelet transformation, followed by classification using a machine learning Artificial Neural Network. The second model utilized ResNet50 architecture, a type of deep learning model, which resulted in improved accuracy. These findings underscore the effectiveness of deep learning approaches in audio classification tasks. Training data for the model is sourced from the DEEP-VOICE dataset, which comprises both genuine and synthetic audio generated by various deep-fake algorithms. The model's performance is assessed using diverse metrics such as accuracy, F1 score, precision and recall. Results indicate successful classification of audio in 86% of cases. This research contributes to the field of Automatic Speech Recognition (ASR) by integrating advanced preprocessing techniques with robust model architectures to identify manipulated speech.**

***Keywords*—Audio classification; automatic speech recognition; machine learning; deep learning; DEEP-VOICE**

I. Introduction

The increasing utilization of generative Artificial Intelligence (AI) in speech-related tasks, such as voice cloning and real-time voice conversion, carries profound implications. This technological advancement raises significant ethical concerns, including threats to privacy and the potential for misrepresentation. Consequently, there is a pressing need for real-time detection mechanisms to identify AI-generated speech, particularly in scenarios involving DeepFake Voice Conversion. Misinformation has become an increasingly prevalent issue in recent times, extending beyond false news articles to encompass the creation of fake audio, images, and videos using algorithms and tools. This form of AI-generated content, known as deepfake [1], [3], presents significant concerns due to its potential impact on various aspects of society, including politics, morality [2], and legal proceedings. For instance, in politics, deepfakes could influence citizens' decisions during elections [4], [5]. Similarly, they could adversely affect individuals' lives, such as in cases of nonconsensual use of famous faces in pornographic materials [6]. Furthermore, deepfakes have the potential to generate false digital evidence, thereby influencing legal outcomes. Deepfakes encompass a set of algorithms crafted to generate synthetic media with the intent of substituting one person's likeness with another's. This synthetic content elicits numerous social, ethical, and legal apprehensions regarding data credibility, as it can portray individuals seemingly uttering or performing actions they never actually did. Notable instances of deepfakes often revolve around images, wherein one person's facial features are replaced by another's, engaging in activities the original subject did not partake in. Artificial Intelligence achievement, emphasizing how the commercialization of human behavior has fueled the quest for digital replicas on an industrial scale [7], [8].

To address these emerging challenges, this study uses the publicly available DEEP-VOICE dataset, which includes both real and synthetically generated speech samples, which consist of authentic human speech samples from eight prominent individuals. The dataset includes instances of their speech being converted to mimic that of others using Retrieval-based Voice Conversion techniques. In this study, two different models for the task at hand will be applied. The first model involves feature extraction using wavelet transformation, followed by classification using a machine learning Artificial Neural Network (ANN). The second model utilizes the ResNet50 architecture. The study aims to analyze and compare the performance of these models in addressing the objectives of this study, particularly in terms of performance metrics such as F1 score, precision, recall and accuracy.

This study aims to address the following research question: Which modeling approach—lightweight, interpretable Wavelet-based ANN or complex, data-driven ResNet50—offers better performance and practical applicability for deepfake speech detection using the DEEP-VOICE dataset?

Even though there are an increasing variety of ways for detecting synthetic audio, ranging from sophisticated deep learning algorithms to conventional signal processing techniques, there is still a conspicuous dearth of comparison studies that thoroughly assess the advantages and disadvantages of each method. There aren't many studies that directly contrast more interpretable, lightweight traditional machine learning models that rely on handcrafted features, like wavelet-domain characteristics, with contemporary deep learning architectures like ResNet50, which use hierarchical feature learning and transfer learning from massive image datasets. For experimental validation, the publicly available DEEP-VOICE dataset, which includes both authentic and

manipulated speech samples, was employed.

In fields like forensics or embedded systems, where interpretability is crucial or computational resources are limited, this disparity is especially pertinent. To give an objective comparison between these two paradigms, this study will conduct an empirical evaluation utilizing a standardized and publicly available dataset (DEEP-VOICE). The emphasis is on the feature extraction pipelines, preprocessing overhead, and model scalability in addition to the final classification performance, which is evaluated by accuracy, precision, recall, and F1-score. These factors are all crucial for the practical implementation of synthetic speech detection systems.

Most of the earlier research focuses on either deep learning or traditional machine learning separately, without direct comparisons under a standardized framework, even if there are several methods for deepfake voice detection. Furthermore, there haven't been enough comparisons 2q1of the interpretability and computational effectiveness of lightweight models like Wavelet-based ANN with more sophisticated deep architectures like ResNet50. To close this gap, our study uses a single dataset to perform a head-to-head comparison.

The remainder of this study is organized as follows: Section II reviews the existing literature and identifies gaps. Section III outlines the proposed methodology, including data preprocessing and model architecture. Section IV presents the experimental set-up and results. Section V discusses the findings considering related works, and Section VI concludes the study with a summary and suggestions for future work.

II. Literature Review

In this section, existing literature and related research to understand current methodologies are explored, identifying knowledge gaps, and contextualize this research within the academic landscape. In the research conducted by Wijetunga et al. [9], the focus was on leveraging deep learning techniques to detect deepfake audio within group conversations. Their approach involved the utilization of the UrbanSound8K dataset, which provides labeled sound excerpts of urban sounds across various classes, along with a conversational dataset sourced from Open subtitles. A series of preprocessing steps was undertaken, including sample rate conversion, merging of audio channels, and extraction of Mel-Frequency Cepstral Coefficients (MFCC) on a per-frame basis. Following preprocessing, the dataset was partitioned into distinct training and testing sets. Multiple neural network architectures were employed throughout the study, encompassing models such as Multilayer Perceptrons (MLP), Convolutional Neural Networks (CNN), and Deep Neural Networks (DNN). These models were applied to various tasks including speech denoising, speaker diarization incorporating Natural Language Processing (NLP), and synthetic speech detection. Furthermore, Transfer Learning techniques were deployed, utilizing a pre-trained VGG19 model to enhance the accuracy of the detection framework. The results obtained from the experimentation phase show- cased promising outcomes. Notably, the CNN architecture achieved a commendable testing accuracy of 89%. However, despite these achievements, the research also acknowledged several limitations inherent in the current approach. Issues such as the necessity for improved signal denoising techniques. Furthermore, challenges pertaining to speaker identification error rates and the automation of partitioning processes in speaker diarization were acknowledged as areas warranting further investigation and development.

In the investigation conducted by Hamza et al. [10], the focus was directed towards deepfake audio detection using MFCC features in conjunction with machine learning methodologies. Central to their investigation was the utilization of the Fake-or-Real (FoR) dataset, a comprehensive repository comprising over 195,000 speech samples, encompassing both authentic human recordings and synthetic computer-generated speech. Within this dataset, researchers meticulously curated four distinct subsets, each tailored to specific experimental needs. These subsets, namely for-original, for-norm, for-2sec, and for-rerec, underwent meticulous preprocessing to standardize sampling rates, volumes, and channels, thereby ensuring uniformity across gender and class categories. Notably, subsets like for-norm underwent additional standardization processes, including the elimination of duplicate files to enhance data consistency. Their methodology involved data preprocessing, MFCC feature extraction, and employing various machine learning models like SVM, VGG-16, XGB, RF, KNN, and LSTM. Notably, while the SVM model generally performed well, VGG-16 stood out with a remarkable 93% accuracy in the for-original subset. Challenges included the high dimensionality of the for-norm dataset, emphasizing the need for dimensionality reduction techniques. The study highlighted the importance of integrating features from various extraction methods for robust deepfake audio detection, suggesting avenues for future research refinement.

Kumar et al. [11] addressed the detection of AI-generated speech using a binary classification model trained on the “DEEP-VOICE” dataset, comprising real and synthetic voice samples. They applied Exploratory Data Analysis (EDA) to handle outliers and missing values and extracted features such as Chroma-STM and RMS. Preprocessing steps included resampling, adjusting sample rates, and normalization. A Random Forest Classifier with 5-fold cross-validation achieved 98.5% accuracy, demonstrating strong performance across multiple evaluation metrics. However, the study highlighted limitations in scalability and generalizability, as well as a reliance on manually engineered features, which may restrict adaptability and efficiency.

In the research by Khochare et al. [12], a deep learning framework tailored for audio deepfake detection is presented, utilizing a combination of machine learning and deep learning approaches. Central to the investigation is the utilization of the Fake or Real (FoR) dataset. In the feature-based paradigm, audio files undergo transformation into datasets comprising various spectral features, while in the image-based domain, melspectrograms are derived from the audio samples using the librosa library. The model repertoire encompasses machine learning algorithms such as SVM, LGBM, XGBoost, KNN, and Random Forest for feature-based classification, alongside deep learning techniques like Temporal Convolutional Network (TCN) and Spatial Transformer Network (STN) for

image-based classification. Results demonstrate varying degrees of success across methodologies, with machine learning models achieving accuracy within the range of 60% to 70%, led by SVM, while TCN emerges as the standout performer in image-based classification, boasting a test accuracy of 92%, followed by STN at 80%. However, the study identifies limitations such as the exclusion of critical inputs like Short-Time Fourier Transform (STFT) and Mel-Frequency Cepstral Coefficients (MFCC) in the image-based approach, prompting future research efforts to address these constraints and refine classification accuracy.

III. Proposed Model

The proposed model subsections follow a structured approach. Initially, the input data from the DEEP-VOICE Dataset [13], undergo preprocessing, including wavelet denoising for noise reduction, feature extraction through 4-level wavelet decomposition, and feature reduction using ICA and normalization techniques depicted in Fig. 1. Subsequently, two models are applied: the first model utilizes ANN classification, while the second model employs deep learning techniques with ResNet50. Finally, the performance of these models is compared to determining the most effective approach. They are crucial in contexts with limitations, such embedded systems. Conversely, ResNet50 was selected due to its demonstrated efficacy in identifying high-dimensional patterns and its capacity to utilize transfer learning. While deep architecture can be resource-intensive and opaque, traditional models frequently rely mostly on manually created features and lack generalization across intricate audio datasets. This study assesses the trade-offs between accuracy and interpretability by contrasting the two aspects.

A. Dataset

The dataset utilized in this study, titled "DEEP-VOICE: DeepFake Voice Recognition", was obtained from the Kaggle platform, a well-known repository for datasets used in machine learning and data science research. This dataset is specifically designed for identifying and analyzing deep-fake voice recordings.

It is systematically organized into two primary directories, as illustrated in Fig. 2: one containing genuine (real) audio samples, and the other comprising synthetically generated (fake) audio samples. This clear separation facilitates comparative analysis between authentic and manipulated audio data. These values are categorized into columns as shown in Fig. 3, Fig. 4, and Fig. 5. Each column corresponds to a specific audio feature, providing numerical data for analysis.

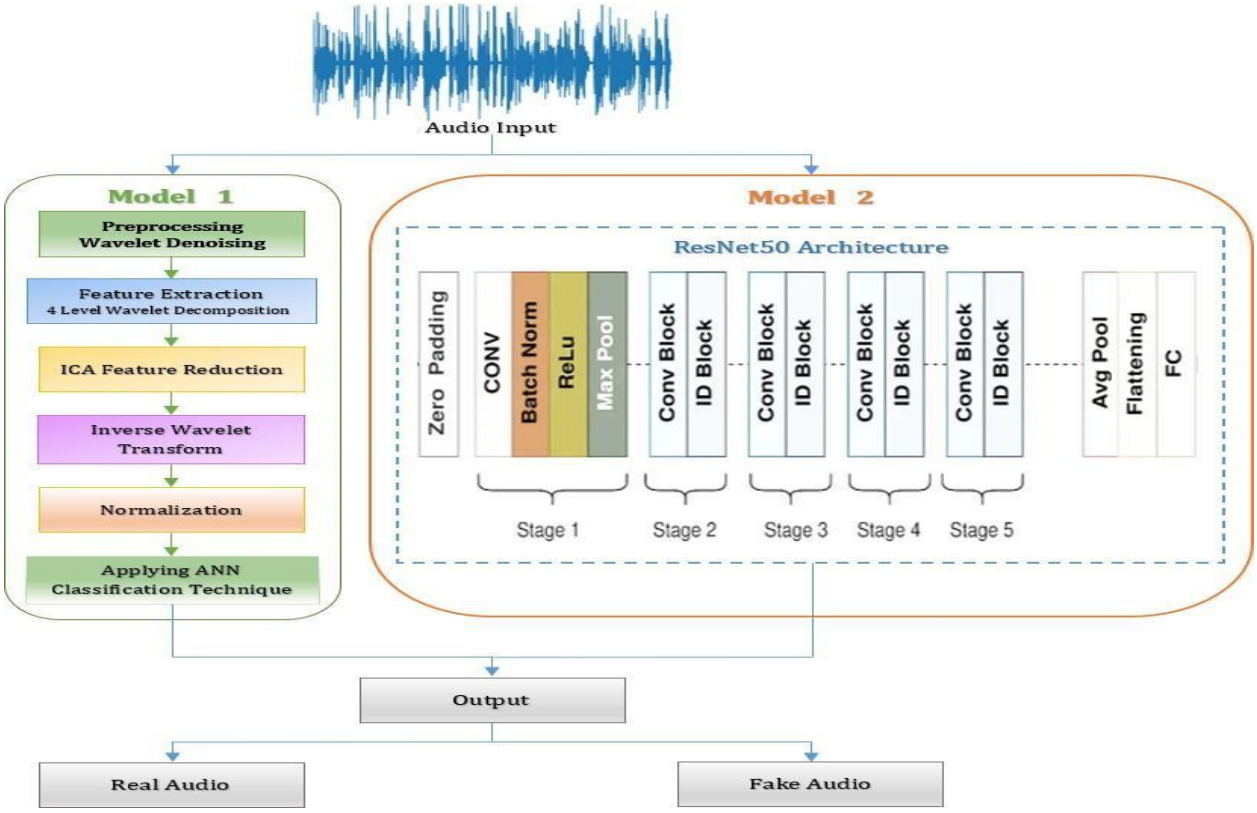


Fig. 1. Proposed model.

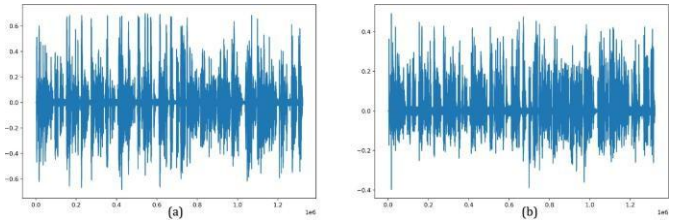


Fig. 2. Real and fake audio data.

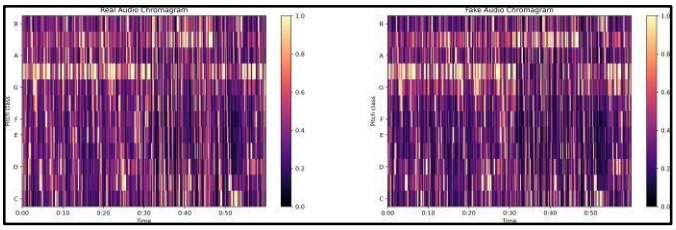


Fig. 3. Audio chromagram.

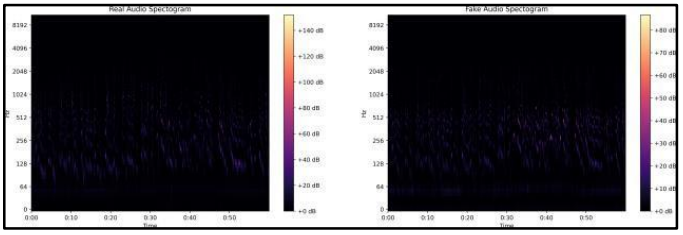


Fig. 4. Audio spectrogram.

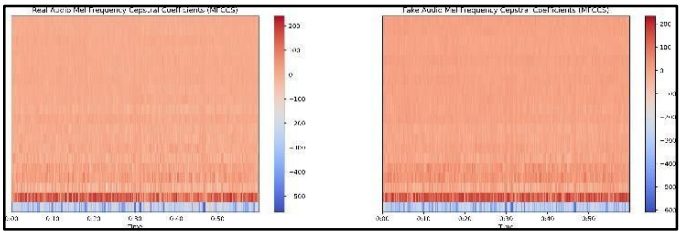


Fig. 5. Audio MFCCs.

B. Preprocessing

The upcoming subsections include signal processing steps to refine data quality and extract relevant features: Wavelet Denoising reduces noise, WT Feature Extraction captures essential information, ICA Feature Reduction condenses the feature space, and Inverse Wavelet Transform reconstructs the signal. Together, these steps form a comprehensive pipeline for enhancing data quality and facilitating analysis and classification tasks. Daubechies 4 (db4) was used as the mother wavelet for the denoising process because of its smoothness and compact support, which make it appropriate for speech signal processing. To capture the pertinent speech characteristics without over-segmenting high-frequency noise, a 4-level wavelet decomposition was used. This level provided a suitable trade-off between resolution and processing efficiency.

To reduce noise in the detail coefficients without losing fine-grained voice characteristics, the universal soft thresholding rule was applied. This rule is renowned for maintaining the continuity and smoothness of the signal.

1) Feature source. Rather than using the reconstructed signals, the retrieved features were taken straight from the wavelet coefficients. More information about frequency fluctuations and transient patterns in the voice stream is preserved as a result.

2) ICA versus PCA comparison. The efficiency of dimensionality reduction was examined in relation to Independent Component Analysis (ICA) and Principal Component Analysis (PCA). When examined empirically, ICA was able to produce more statistically independent components that enhanced classification performance by around 4% in F1 score, but PCA kept about 92% of variation in the first 10 components. Because ICA can separate mixed audio patterns that PCA cannot because of its orthogonality requirement, it was chosen.

3) Normalization strategy. Z-score normalization, which centers the data around zero mean and unit variance, was used to normalize all feature vectors. To make sure that every extracted component contributes equally to the classification problem, this step was applied following ICA. To prevent data leakage, normalization was conducted independently of the training and testing sets using statistics that were taken from the training set.

a) Wavelet denoising: In audio speech recognition, Wavelet Denoising serves as a fundamental preprocessing step aimed at refining data quality by minimizing noise within signals. This technique harnesses wavelet-based methodologies to effectively filter out unwanted noise, resulting in a clearer and more reliable input for subsequent analysis [14]. By employing Wavelet Denoising, researchers can mitigate the negative impact of noise on data interpretation, thereby improving the accuracy of downstream tasks such as feature extraction and classification.

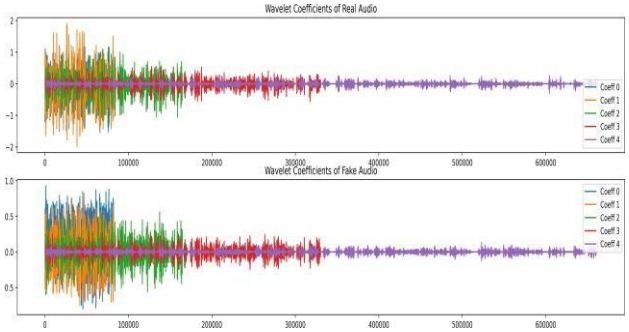


Fig. 6. Wavelet coefficients of real audio and fake audio.

b) Wavelet transform feature extraction: In feature extraction, applying the Wavelet Transform (WT) that serves as a fundamental tool in frequency domain analysis, particularly relevant to the proposed model in the context of audio speech recognition. The 4-level wavelet decomposition refers to a specific application or implementation of the Wavelet Transform. In this case, the signal is decomposed into four levels of detail, each capturing different frequency bands or resolutions as shown in Fig. 6. It efficiently condenses complex audio signals, which exhibit temporal variations and contain diverse

data points, into a concise set of parameters that characterize these signals [15]. Given the dynamic nature of audio signals, the frequency domain approach, often involving WT, is commonly employed to determine the most suitable feature extraction method [16].

4) ICA feature reduction. To isolate distinct feature vec tors from speech signals, employing the Independent Component Analysis (ICA) algorithm on various segments of human speech is made. ICA is a computational method that splits a complex signal into separate parts that are independent from each other [17]. It is used to find the different sources or causes behind observed signals, assuming these sources don't affect each other statistically. In speech processing, ICA can help separate different voices in a recording where multiple people are speaking at once. By analyzing statistical properties like non-Gaussianity and independence, ICA untangles the voices from each other. It's a useful tool in various areas like signal processing, machine learning, and neuroscience, where understanding the individual components of a signal is important.

a) Inverse wavelet transform: Utilizing the inverse wavelet transform as a pivotal element of the speech recognition framework. Following the application of wavelet analysis to break down the speech signal into its individual frequency components, the inverse wavelet transform to reconstruct the original signal from these components was employed [18]. This reconstruction procedure facilitated the retrieval of the initial speech signal while preserving crucial details regarding its frequency characteristics. Through the integration of the inverse wavelet transform into the proposed system, it successfully analyzed and processed speech signals for recognition purposes, resulting in precise and dependable outcomes.

b) Normalization: In this study, normalization significantly improved the speech recognition system. Before featuring extraction and model training, the input speech data to ensure consistent scaling across samples is normalized, reducing biases and making the proposed system more robust to amplitude variations. This enhanced the accuracy and reliability of the recognition system, leading to consistent and trustworthy results across different datasets and conditions.

C. Artificial Neural Network Architecture

Artificial Neural Network (ANN) is a computational model inspired by the structure and function of the human brain's neural networks. It consists of interconnected nodes, or" neurons", organized in layers. In this specific implementation, the ANN for classification tasks was applied, integrating Wavelet Transform for feature extraction and Independent Component Analysis (ICA) for feature reduction. The ANN architecture comprised three layers as shown in Fig. 7: the first layer contained 10 nodes, the second layer had 7 nodes, and the output layer consisted of 1 node. This design allowed for effective processing and classification of speech data while managing computational resources efficiently.

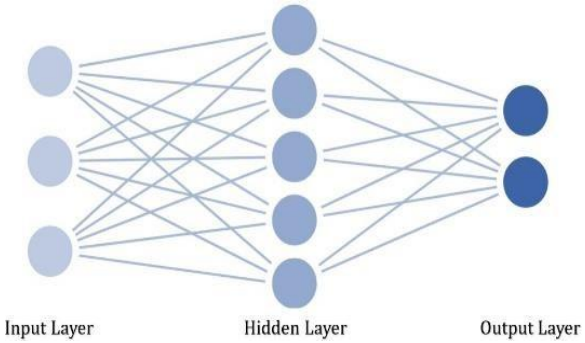


Fig. 7. ANN architecture.

D. ResNet50 Architecture

The implemented model employs the ResNet50 model, a convolutional neural network renowned for its depth and effectiveness. ResNet, designed with 50 layers, introduces residual learning to mitigate accuracy degradation as networks deepen. This approach enables each layer to learn residual mappings in relation to input layers, facilitated by skip connections. ResNet50 architecture includes convolutional layers, batch normalization, activation functions (typically ReLU) [19], and bottleneck blocks to manage complexity [20]. In the code, ResNet50 acts as a feature extractor for audio data, leveraging pre-training on datasets like ImageNet to extract high-level features from audio spectrograms. Through transfer learning, ResNet50 is represented in Fig. 1 as Model 2. It is capable of effectively extracting features for classification tasks, such as audio fake detection. The implemented model specifies hyperparameters, including 100 epochs, a batch size of 32, Adam optimizer, binary cross-entropy loss, and early stopping with a patience of 5 for monitoring validation loss.

E. Wavelet Denoising, Feature Extraction and Dimensionality Reduction

Daubechies 4 (db4) was chosen as the mother wavelet for the preprocessing stage because of its smoothness and compact support, which makes it ideal for catching fleeting patterns in voice data. To balance time-frequency resolution, a 4-level wavelet decomposition was used; higher levels offer global patterns, while lower levels catch finer details. Four levels produced the best trade-off between computing efficiency and information retention, according to empirical data. The universal soft thresholding rule, which reduces high-frequency noise while maintaining significant signal structures, was used to attenuate the noise [21]. Because it preserves signal continuity and minimizes overfitting in subsequent modeling stages, this approach is frequently used. Instead of using the reconstructed signals, the extracted characteristics were taken straight from the wavelet coefficients [22]. The localized time-frequency information that is essential for differentiating between authentic and fraudulent audio patterns is preserved when raw coefficients are used.

A comparison analysis comparing Principal Component Analysis (PCA) and Independent Component Analysis (ICA) was carried out to reduce dimensionality [23]. ICA produced components that were more statistically independent than

PCA, which kept about 92% of the total variance in the first 10 components. ICA was chosen for this investigation because, when evaluated in the classification pipeline, it increased the F1-score by over 4% above PCA.

Z-score normalization (zero mean, unit variance) was used to normalize the features. To guarantee that every independent component had an equivalent impact during classification, this step was implemented following ICA transformation. Crucially, for the purpose of preventing data leakage, normalization was carried out independently on the training and testing sets using statistics that were only calculated from the training data.

IV. Models' Application and Results

In this section, the application of the two distinct models for the task of audio speech recognition, along with their respective results, are presented.

A. Model 1: Artificial Neural Network (ANN) with Wavelet Transform

For the first model, an Artificial Neural Network (ANN) in conjunction with Wavelet Transform for feature extraction was applied. A lightweight yet efficient architecture is provided by combining wavelet-domain characteristics with an Artificial Neural Network (ANN). Three completely connected layers made up the ANN's design, which was optimized for classification jobs with less data. The output layer has a single sigmoid-activated node for binary classification, a hidden layer with seven neurons, and a first layer with ten neurons that receives the ICA-reduced wavelet coefficients. Simplicity, quick training time, and interpretability are advantages of this paradigm. Its scalability to larger or more varied datasets is constrained by its dependency on manually created feature extraction (wavelets + ICA). Performance reporting was made robust by using 4-fold cross-validation, and overfitting was avoided by using regularization techniques.

The input data, preprocessed using Wavelet De-noising, underwent feature extraction through 4-level wavelet decomposition. Following this, Independent Component Analysis (ICA) and normalization techniques were employed for feature reduction. The ANN, comprising 3 layers (10 nodes in the first layer, 7 nodes in the second layer, and 1 node in the output layer), was trained on the processed data for audio speech recognition. The results of this experiment, conducted using k-fold cross-validation with k=4, yielded an average accuracy across folds of 79.7%, average precision of 82.1%, average recall of 93.9%, and average F1 score of 87.6%.

B. Model 2: Deep Learning with ResNet50

The second model utilized deep learning techniques, specifically ResNet50, for audio speech recognition. The ResNet50 model makes utilization of transfer learning and deep residual learning. Using spectrogram images produced from the DEEP-VOICE dataset, the 50-layer architecture was refined after being pre-trained on ImageNet. Hierarchical auditory patterns may be captured by the convolutional layers thanks to the spectrograms' 2D time-frequency representation of audio.

Binary cross-entropy loss, the Adam optimizer, and a batch size of 32 were used to train the model. To reduce overfitting, early stopping was used with 5-epoch patience. ResNet50's advantage is its capacity to generalize across intricate audio circumstances and automatically learn high-level audio properties without the need for manual engineering. On synthetic speech samples, its deeper architecture offers improved recall and classification accuracy.

The ResNet50 architecture, comprising 50 layers, was pre-trained on a vast dataset such as ImageNet to extract high-level features from audio spectrograms. These features were then fed into additional layers for classification purposes. The results of this experiment, also conducted using k-fold cross-validation with k=4, demonstrated the superior performance of ResNet50 with an average accuracy across folds of 85.9%, average precision of 87.3%, average recall of 98.2%, and average F1 score of 92.4%.

Regarding hyperparameters, the implemented model employed the following settings: 100 epochs, batch size of 32, Adam optimizer, binary cross-entropy loss function, and early stopping with patience set to 5. Upon comparison, it is evident that the deep learning model with ResNet50 outperformed the ANN with Wavelet Transform, exhibiting higher accuracy, precision, recall, and F1 score.

From the comparison shown in Table I, it is evident that the deep learning model with ResNet50 outperformed the ANN with Wavelet Transform, exhibiting higher accuracy, precision, recall, and F1 score.

TABLE I Applying ANN Architecture with Wavelet Transform versus ResNet50 Architecture

Model	Avg. Accuracy	Avg. Precision	Avg. Recall	Avg. F1 Score
ANN with WT	79.69%	82.14%	93.88%	87.62%
ResNet50	85.94%	87.29%	98.21%	92.41%

These results highlight the effectiveness of deep learning techniques, particularly ResNet50, for audio speech recognition tasks, offering improved performance and robustness compared to traditional machine learning approaches as shown in Fig. 8.

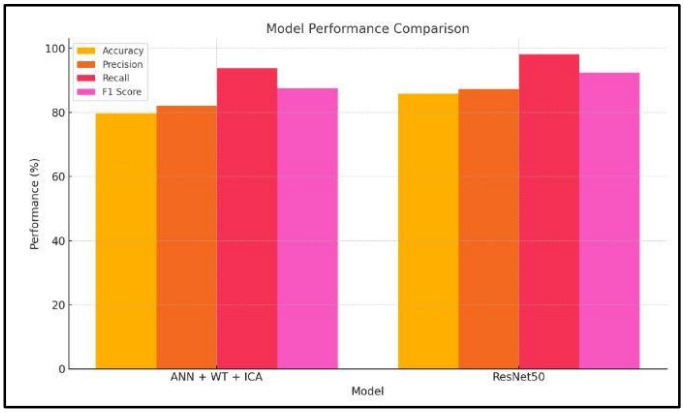


Fig. 8. Comparative performance metrics of ANN and ResNet50 model.

V. Discussion

Through a comprehensive review of the literature [24], valuable insights into current state-of-the-art methodologies were gained. This research significantly contributes to this body of knowledge by achieving high performance in accuracy levels [25], particularly through the utilization of the deep learning ResNet50 model, which yielded a testing accuracy of 86%. This represents a noteworthy research contribution. In comparison to the model applied in [9], which employed a CNN architecture and achieved a commendable testing accuracy of 89%. This approach addresses some of the acknowledged limitations in the literature. For instance, the applied methodology incorporates improved signal denoising techniques as part of the preprocessing steps, effectively resolving the challenges identified. A comparative summary of our results alongside existing approaches is presented in Table II, demonstrating the strengths and potential limitations of each technique in different settings.

Similarly, [10] demonstrated the effectiveness of the VGG-16 model with a remarkable 93% accuracy in the original subset. However, challenges related to dataset dimensionality were highlighted, emphasizing the need for dimensionality reduction techniques. In this model, this concern by employing ICA feature reduction, thereby mitigating the impact of high-dimensional datasets, is addressed. Lastly, [11] achieved impressive results using the Random Forest Classifier with 5-fold Cross-Validation, attaining an accuracy of 98%. Nonetheless, limitations regarding scalability and generalizability were identified, along with reliance on handcrafted features, which may not always capture the most relevant information present in the data. Our approach offers potential advantages in terms of scalability and generalizability, as well as automated feature extraction techniques, thus providing avenues for further research and development in this domain.

TABLE II Comparative Performance of Proposed Models Versus Related Work

Feature type	Accuracy	Dataset	Model	Study
Spectrogram	85.5%	Deep voice	ResNet50	This proposed Model
Wavelet coefficient	79.7%	Deep voice	ANN + Wavelet +ICA	This proposed Model
MFCC	93.0%	Fake or real	VGG-16	Hamza et al. [10]
Chroma- STM, RMS (manual)	98.5%	Deep voice	Random Forest	Kumar et al. [11]
Mel- spectrogram	92.0%	Fake or real	TCN	Khochare et al. [12]

Higher raw accuracy has been obtained by other works like Kumar et al. [11] and Hamza et al. [10], but those models mainly depend on manually created features and particular dataset circumstances. By using automated feature extraction (ResNet50 with spectrograms and wavelet ICA), on the other hand, our method improves generalization and lessens the requirement for domain-specific engineering. Additionally, the DEEP-VOICE dataset, which contains intricate, real-world deepfake audio produced using sophisticated voice conversion techniques, provided a more difficult and realistic assessment for these proposed models.

Because of this, rather than overfitting a single data format or manipulation type [24], our results are more indicative of real-world scenarios, where models must generalize across several manipulation approaches.

VI. Conclusion And Future Work

In this study, two distinct approaches for audio speech recognition were explored: an Artificial Neural Network (ANN) with Wavelet Transform and deep learning with ResNet50. The study's comprehensive experimentation and evaluation revealed notable performance differences between the two models. The ANN with Wavelet Transform demonstrated respectable accuracy and precision, achieving an average accuracy across folds of 80%. However, the deep learning approach using ResNet50 surpassed the ANN model, exhibiting higher accuracy, precision, recall, and F1 score. ResNet50 exhibited an average accuracy across folds of 86%, demonstrating its exceptional ability to effectively process complex audio data. This represents a notable improvement in performance and robustness, with a nearly 6% increase observed compared to previous approaches.

Our comparison highlights the effectiveness of deep learning techniques, particularly ResNet50, for audio speech recognition tasks. Leveraging transfer learning and pre-trained models like ResNet50 proved highly successful in extracting high-level features from audio spectrograms, resulting in superior classification accuracy.

Despite the encouraging results, this study has several limitations. First, the models were evaluated on a single dataset (DEEP-VOICE), which may limit generalizability across other datasets with different characteristics or languages. Second, although ResNet50 performed well, its computational cost may hinder deployment in real-time or embedded systems. Finally, the reliance on spectrogram images introduces a preprocessing overhead that may not suit all applications. Future work will explore broader dataset diversity and model optimization for real-time deployment.

In conclusion, the findings emphasize the significance of utilizing advanced deep learning architectures for audio speech recognition tasks, offering enhanced performance and scalability compared to traditional machine learning approaches. Future research endeavors may focus on further optimization techniques and larger datasets to continue advancing the accuracy and efficiency of audio speech recognition systems.

References

[1] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?," Business Horizons, vol. 63, no. 2, pp. 135–146, 2020.

[2] B. Paris and J. Donovan, "Deepfakes and cheap fakes. data & society, 47," 2019.

[3] N. Eldien, R. Ali, and F. Moussa, "Real and fake face detection: A comprehensive evaluation of machine learning and deep learning techniques for improved performance," pp. 315–320, 07 2023.

[4] S. Ahmed, "Who inadvertently shares deepfakes? analyzing the role of

political interest, cognitive ability, and social network size," Telematics and Informatics, vol. 57, p. 101508, 2021.

[5] A. Lieto, D. Moro, F. Devoti, C. Parera, V. Lipari, P. Bestagini, and S. Tubaro, "" hello? who am i talking to?" a shallow cnn approach for human vs. bot speech classification," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2577–2581, IEEE, 2019.

[6] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," Iet Biometrics, vol. 10, no. 6, pp. 607–624, 2021.

[7] J. Truby and R. Brown, "Human digital thought clones: the holy grail of artificial intelligence for big data," Information & Communications Technology Law, vol. 30, no. 2, pp. 140–168, 2021.

[8] M. Waldrop, "Synthetic media: The real trouble with deepfakes," Knowable Magazine, vol. 3, 2020.

[9] R. Wijethunga, D. Matheesha, A. Al Noman, K. De Silva, M. Tissera, and L. Rupasinghe, "Deepfake audio detection: a deep learning-based solution for group conversations," in 2020 2nd International conference on advancements in computing (ICAC), vol. 1, pp. 192–197, IEEE, 2020.

[10] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via mfcc features using machine learning," IEEE Access, vol. 10, pp. 134018–134028, 2022.

[11] V. Kumar, A. Kapoor, R. R. Chaudhary, L. Gupta, and D. Khokhar, "Preserving integrity: A binary classification approach to unmasking artificially generated voices in the age of deepkakes," in 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1449–1454, IEEE, 2024.

[12] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," Arabian Journal for Science and Engineering, pp. 1–12, 2021.

[13] birdy654, "Deep voice deepfake voice recognition."https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice- recognition, 2022. Accessed on May 5, 2024.

[14] M. Shanthamallappa, K. Puttegowda, N. K. Hullahalli Nannappa, and S. K. Vasudeva Rao, "Robust automatic speech recognition using wavelet-based adaptive wavelet thresholding: A review," SN Computer Science, vol. 5, no. 2, p. 248, 2024.

[15] S. Basak, H. Agrawal, S. Jena, S. Gite, M. Bachute, B. Pradhan, and M. Assiri, "Challenges and limitations in speech recognition technology: A critical review of speech signal processing algorithms, tools and systems.," CMES-Computer Modeling in Engineering & Sciences, vol. 135, no. 2, 2023.

[16] H. Hindarto, A. Muntasa, and S. Sumarno, "Feature extraction electroencephalogram (eeg) using wavelet transform for cursor movement," in IOP Conference Series: Materials Science and Engineering, vol. 434, p. 012261, IOP Publishing, 2018.

[17] "Ica and iva bounded multivariate generalized gaussian mixture based hidden markov models," Engineering Applications of Artificial Intelligence, vol. 123, p. 106345, 2023.

[18] N. Trivedi, V. Kumar, S. Sing, S. Ahuja, and R. Chadha, "Speech recognition by wavelet analysis," International Journal of Computer Applications, vol. 15, no. 8, pp. 27–32, 2011.

[19] S. Reza, M. C. Ferreira, J. J. Machado, and J. M. R. Tavares, "A customized residual neural network and bi-directional gated recurrent unit-based automatic speech recognition model," Expert Systems with Applications, vol. 215, p. 119293, 2023.

[20] H. Fazlic´a, A. Abd Almisrea, and N. M. Tahirb, "Deep learning-based audio-visual speech recognition for bosnian digits," Jurnal Kejuruteraan, vol. 36, no. 1, pp. 147–154, 2024.

[21] Z. Jiang Y. Wang, L.Zhang, and H.Wang, "Deepfake Audio Detection Using Dual-Branch CNN with Attention Mechanism," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 512–525, 2024.

[22] M. Singh, R. Srivastava, and V. Sharma, "Wavelet-Based Spectrogram Feature Extraction for Voice Spoofing Detection,"

[23] A. Chen and B. Liu, "Transfer Learning for Audio Deepfake Detection Using Pre-trained Speech Embeddings,"

[24] N. Akhtar, S. Verma, and M. Hossain, "A Review of Challenges and Trends in Deepfake Audio Detection," Digital Signal Processing, vol. 141, p. 103798, 2024.

[25] T. Aoki, K. Oura, and H. Mizuno,"Robust Audio Deepfake Detection Using Phase-Based Features and LSTM,"ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3214–3218.

Fine-Tuning OpenAI GPT Chatbot in Western Saudi Dialect: A Case Study of Taibah University

Maimounah Alhujaili, Ruqayya Abdulrahman
College of Computer Science and Engineering, Taibah University, Al-Madinah, Saudi Arabia

***Abstract*—The current era is characterized by technological advancement and innovation, which affect various sectors. Numerous remarkable and alluring computer programs and applications have surfaced, including ones that aim to replicate human behavior. A chatbot is an example of an Artificial Intelligence (AI) computer program that uses natural language to mimic human conversations in voice or content. Even though a lack of Arabic chatbots, most of these chatbots use Modern Standard Arabic instead of Arabic dialects. This research presents the development and evaluation of a chatbot designed to respond to academic inquiries from university students using the Western Saudi dialect. A traditional Support Vector Machine (SVM) baseline model was first implemented to establish a reference point for performance. Subsequently, a fine-tuned version of Generative Pre-trained Transformer (GPT) 3.5-Turbo-0125 was developed using a culturally specific system prompt to enhance the model's understanding of regional language and academic contexts. Evaluation was conducted through a multi-dimensional framework combining human assessments, BERTScore semantic similarity measurements, and GPT-4-based automatic judging. With human assessors determining that 85% of GPT-3.5's replies to 132 messages of test data were appropriate, the transformer-based model clearly outperformed the SVM baseline, which had an accuracy of 42.86% on 20 messages of test data. These findings highlight the importance of cultural and contextual fine-tuning in building effective conversational agents for dialectal Arabic communities. The research contributes to the growing field of localized AI by demonstrating how advanced language models can be adapted to serve specialized linguistic and academic needs.**

Keywords*—*Artificial intelligence (AI); large language model (LLM); generative pre-trained transformer (GPT); Modern Standard Arabic (MSA); Western Saudi dialect

I. Introduction

In contemporary society, Artificial Intelligence (AI) influences daily life by contributing to the development and assessment of advanced applications. AI helps people to automate various tasks in multiple fields. One of the most exciting AI applications that have become widespread is a chatbot. A chatbot is a popular term for conversational agent. It is software that mimics human conversations in voice or content mode by reacting based on contributions made. Also, it is the AI to engage in conversation and interaction with users in natural language that simulates human behavior. The utilization of chatbot is increasing in different fields with different languages. The main goal of utilizing chatbots are to simulate interpersonal interactions [1].

Based on researches [2], the history of chatbot started in October of 1950. Alan Turing proposed the first idea of a chatbot. Alan Turing questioned if a computer program could converse with a group of humans without their recognizing that their conversation partner was artificial. Many people believe that this question, known as the "Turing test", is the one that inspired the creation of chatbots. Chatbots have improved over the years, and the first utilization of AI with a chatbot was in 1988 for the Jabberwacky chatbot. A growing increase in the use of chatbots and the enhancement of AI techniques used with chatbots was observed.

Chatbots use The International Organization of Legal Metrology (OIML) to represent knowledge. Chatbots can replace humans in repetitive tasks like customer service and e-help desk answering and providing prompt responses. Additionally, chatbots are utilized by operating systems as virtual assistants, such as Siri on Macs and Cortana on Windows. These days, chatbots are becoming more and more popular in the business world because of their enormous potential to automate customer care and reduce employee time [3].

A chatbot could be implemented in any human language. Unlike the abundance of studies on English chatbots, there is a lack of studies on Arabic chatbots because of the language's challenges [4]. Arabic is obviously different from languages like English in several aspects. For Instance, the Arabic alphabet consists of 28 primary characters, thirteen of them contain dots like (خ (Kha) - ج (Jeem) - ت (Ta) - ف (Fa)), while fifteen do not, such as (ح (Hha) - م (Meem) - ع (Ayn) - د (Dal)), and it written and read from right to left.

The difficulties in Arabic languages negatively effect on applying Natural Language Processing (NLP) on Arabic languages [5]. Even though the Arabic research on applying Arabic language on NLP technique has increased in recent years, the utilization of Arabic dialect is very rare. Most of Arabic natives communicate with dialect Arabic instead of Modern Standard Arabic (MSA). Also, some of them faced problems when they expressed themselves by MSA. For those reasons, the Arabic dialect chatbot will be more helpful than MSA chatbot.

A. Problem Identification

The development of chatbots that can understand and respond to regional dialects has been an emerging field in NLP and machine learning. However, despite significant advances in multilingual NLP models and transformer-based architectures (such as GPT models), there remains a notable gap in the application of these technologies to specific regional dialects, particularly within the Saudi Arabian context.

Existing research in chatbot development has largely focused on general-purpose language models, or models tailored to widely spoken dialects and languages, such as MSA. However, there is a lack of focus on specific regional dialects like the Western Saudi dialect, which is rich in local expressions, cultural nuances, and vocabulary unique to regions such as Madinah, Makkah, Jeddah, and Taif. Most existing chatbots fail to address the linguistic and cultural diversity found in these regions, resulting in suboptimal performance when handling local dialects. While some studies have explored Arabic dialectal chatbot systems, there is limited work that specifically targets the Western Saudi dialect, especially in the context of university students.

This research is necessary, as it addresses a significant gap in the existing literature by developing a chatbot specifically tailored to understand and respond to the Western Saudi dialect, which has been underrepresented in the field of NLP. While there has been substantial work on general-purpose chatbots and models designed for widely spoken dialects or MSA, few studies have focused on region-specific dialects like the Western Saudi dialect, especially in the context of university students. By developing a chatbot that can respond to questions unique to Taibah University, including course registration, academic regulations, and campus information, this research seeks to give students a useful solution. This project will not only fill the gap in dialect-specific chatbot development but also enhance institutional efficiency and student assistance by relieving staff workloads and providing students with quicker access to information.

B. Research Questions

This research is structured around three primary research questions that aim to explore the role of Arabic dialects—specifically the Western Saudi dialect—in the development and evaluation of intelligent chatbot systems: 1) How does the Western Saudi dialect influence the performance of a Support Vector Machine (SVM)? This question investigates the effect of incorporating dialectal variations on the accuracy and efficiency of traditional machine learning models. 2) What is the difference in performance between a chatbot utilizing an SVM model and a chatbot based on a GPT model? The comparison is supposed to assess the merits and shortcomings of both methods in comprehending and generating dialectical Arabic responses. 3) How can a GPT-based model be optimized to enhance academic support for Saudi university students? The final question focuses on adapting and fine-tuning large language models to provide more effective and contextually appropriate assistance in educational settings.

C. Research Objective

The research aims to implement a Western Saudi dialect chatbot that understands the users and responses effectively with same Western Saudi dialect. The chatbot will communicate with users in the educational field. This research will fill the gap in Arabic dialects research area and especially the Saudi dialect.

The remainder of this research is organized as follows: Section II provides the background, including the definition of chatbots, chatbot methodologies, the Transformer framework, and a history of chatbot evolution. Section III outlines the literature review, with a particular focus on the most relevant prior work in the field. While Section IV discusses the research methodology used in this research. Section V describes the implementation and experimentation settings, including the development of the emails retrieval-based model, the baseline model, and the GPT-based model. Section VI reports the experimental results, and Section VII discusses the key findings and their implications. Finally, Section VIII concludes the study and outlines potential directions for future research.

II. Background

Based on the dictionary, the definition of chatbot is ‘‘A computer program designed to simulate conversation with human users, especially over the Internet’’ [6]. A chatbot may uses Sentiment Analysis (SA) and NLP to communicate with humans or other chatbots in different languages by text or voice. It has multiple terms for interactive agents, for instance, artificial, smart bots, conversation entities, and digital assistants. In recent years, AI chatbots have become exceptionally popular in replacing human responses [7].

A. Chatbot Approaches

Chatbot approaches are classified into two main categories: rule-based chatbot and Machine Learning (ML) chatbot [8]. Rule-based chatbots apply pattern-matching algorithms to compare user input to a rule pattern and choose a prepared response from a list of options. The choice of rules and the answer format may also be influenced by the context. The automatic, repetitive, and lack of originality and spontaneity in the responses is a drawback of the pattern-matching approach. However, because a more thorough syntactic or semantic analysis of the input text is not carried out, there is a quick reaction time. The first chatbot ELIZA is based on this approach. However, these Chatbots are not able to respond to patterns that do not fit the prewritten script[9].

On the other hand, ML approach divided into retrieval-based chatbot and generative chatbot [10]. Retrieval-based chatbot works by comparing user input to a predetermined list of answers stored in a knowledge base. After understanding user intents and queries using NLP techniques, they obtain the best pre-written response from their database and display it to the user. In contrast, large text datasets from books, journals, the internet, and other sources are used to train generative chatbots. They process and produce text using Deep Learning (DL) models and NLP. The chatbot preprocesses and tokenizes information, dividing it into smaller pieces known as tokens, before a user submits a message. After creating an initial representation of the user's message using these tokens, it predicts the subsequent words or tokens based on its training data and acquired linguistic patterns to produce a response.

B. Transformer in Chatbot

In 2017 [11], the authors presented the transformer, a deep learning architecture that uses self-attention processes to process sequential input. Transformers enable simultaneous processing of input sequences, in contrast to conventional Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs), which process sequences step by step. Modern Large Language Models (LLMs) like Generative Pre-trained Transformer (GPT) and BERT are built on top of this design. The encoder and decoder, the two primary components of the

Transformer, are made up of several identical layers that allow for effective learning and representation of sequential input [12].

Processing the input sequence and producing meaningful representations are the encoder's responsibilities. It is made up of several identical layers, usually six in the original transformer, that carry out the following tasks. The input embedding Layer converts words into dense vector representations (word embeddings). Then, it captures semantic meaning but does not retain positional information. Multi-head attention refers to the simultaneous operation of several attention mechanisms that record various word relationships. This layer applies softmax after using a triangle mask to set future words' attention scores to negative infinity. The Multi-head self-attention layer enables each word to attend to all other words in the sentence. The mechanism works by computing three vectors for each word. Query (Q) represents what the word is looking for, Key (K) represents the content of all words, and Value (V) contains the actual information to be transferred. These are used in the scaled dot-product attention equation in the Eq. (1) [11]:

$$Attention\ (Q,K,V) = softmax\ (\frac{Qk^t}{\sqrt{d_k}})\ V \qquad (1)$$

Fig. 1 illustrates the internal structure of the multi-head attention mechanism used in transformer models. It begins by projecting the input into Q, K, and V vectors through separate linear transformations. Multiple attention heads are applied in parallel, each computing scaled dot-product attention independently. The outputs from all heads are then concatenated and passed through a final linear layer, allowing the model to capture a richer representation of the input by attending to information from different subspaces.

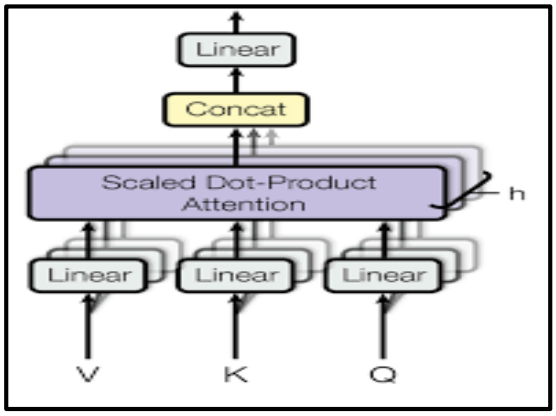


Fig. 1. Multi-head attention consists of multiple attention layers operating in parallel [11].

Layer normalization and residual connection improves training stability and gradient flow. Each word representation from the self-attention layer is passed through a feed-forward network, consisting of two linear transformations with a non-linearity in between (typically ReLU). This layer allows the model to learn complex patterns beyond attention-based relationships. The decoder receives a contextualized word embedding as the output from the final encoder layer [13].

On the other hand, the decoder is responsible for generating the output sequence based on the encoder's output. Each decoder has six identical layers, just like the encoder. It does, however, have an extra attention mechanism which is encoder-decoder multi-head attention layer. In this layer, the decoder can concentrate on the encoder's output and produce pertinent answers thanks to the multi-head attention layer. The decoder provides the Q, while the encoder output provides the K and values V. The layer guarantees the words produced by the decoder are pertinent to the input sequence in terms of context. The final output of the decoder side is mapped to the vocabulary size by passing it by a fully linked linear layer. For every word, probabilities are generated using a softmax function. In the output sequence, the word with the highest probability is chosen as the following word [14].

Fig. 2 illustrates the architecture of the transformer model, which is composed of an encoder and a decoder. Each component consists of stacked layers that incorporate attention mechanisms and feed-forward networks. Positional encoding is applied to both inputs and outputs to retain sequence information, and the final output is generated through a linear layer followed by a softmax function.

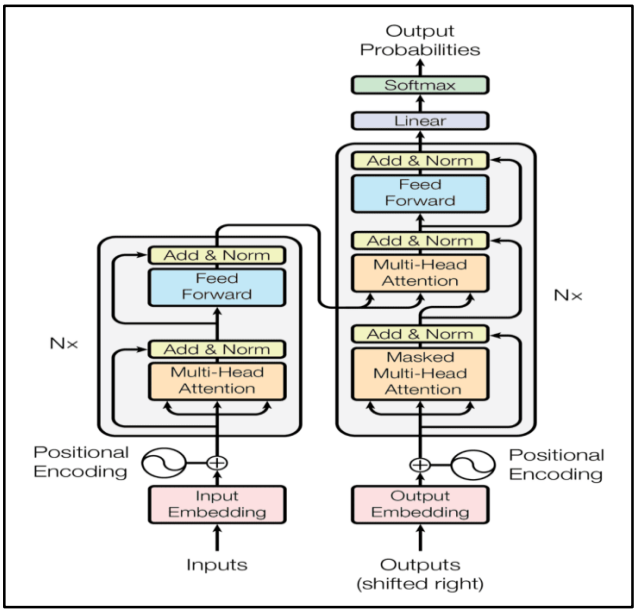


Fig. 2. The transformer model architecture [11].

C. Chatbot Historically

The development of chatbot has been improved significantly over the decades. Fig. 3 summarizes the historical development of chatbots, which will be discussed in detail below. It provides a visual timeline of major milestones, illustrating the evolution of chatbot technologies from early rule-based systems to modern AI-powered models.

- In 20th century: (1950 to 1999)

Alan Turing's published his famous article "Computing Machinery and Intelligence" in 1950. In the article, he discussed a question "Can machines think?", and he proposed a measure of intelligence. This measure called now Turings test. It determines whether the machine acts like human in a real-time written conversation and the human judge cannot absolutely distinguish (based on the conversation only) between the machine and the real human. Based on many experts, the

Turings test is considered to be the generating principle for chatbots [15].

In 1966, the first chatbot was developed by Weizenbaum called ELIZA. ELIZA is simulating conversation with therapist. Early users felt they were speaking with someone who understood their input because of the program's scripted responses that use "pattern matching" and substitution methods [16]. Many early users were persuaded by ELIZA's intelligence and comprehension; however, Weizenbaum may have insisted differently. ELIZA is incapable of maintaining lengthy discussions or picking up on context from the interaction, which is A drawback of ELIZA. It also has limited expertise means that it can only discuss a specific range of topics [17].

PARRY chatbot appeared in 1972 at Stanford University by psychiatrist Kenneth Colby. It simulated a patient with schizophrenia [18]. Because PARRY is designed to have a "personality" and an improved controlling structure, it is thought to be more advanced and intelligent than ELIZA. In an experiment conducted in 1979, five psychiatrists utilized PARRY to determine whether a patient was a real schizophrenia patient or a chatbot. Hence, ten diagnoses were provided by psychiatrists. Two diagnoses were made correctly by the first psychiatrist, and two inaccurate ones by the second one.

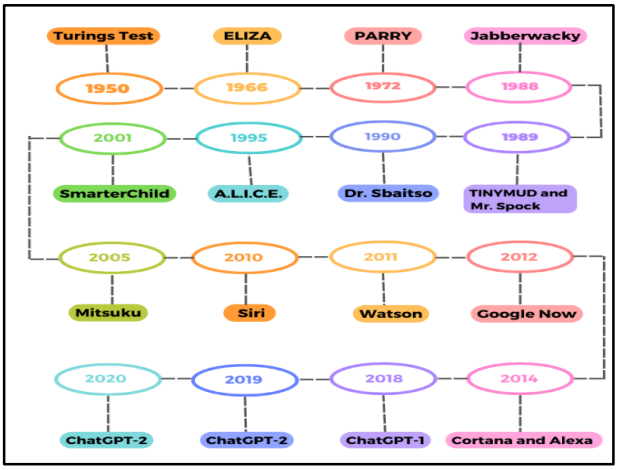


Fig. 3. Chatbot history timeline.

While the other two concluded that the participants were chatbots, the third believed that the subjects were actual patients [19]. However, because there is some degree of confusion in the speech of people who have schizophrenia, the small sample size of five doctors makes it difficult to interpret the results. PARRY is generally regarded as a chatbot with limited language comprehension and emotional expression skills. Additionally, it cannot learn from discussions and responds slowly.

Artificial intelligence was first applied in the field of chatbots in 1988 in the Jabberwacky chatbot. Jabberwacky was created using CleverScript, a spreadsheet-based language that made it easier to create chatbots. It responded by using contextual pattern matching to go back and review past conversations. Nevertheless, Jabberwacky is unable to process large amounts of data quickly or efficiently [20].

One of the first attempts to create a conversational AI character in an online virtual world based on textual was called TINYMUD [21]. James Aspnes created it in 1989, and Carnegie Mellon University's TINYMUD server hosted its initial implementation of it. It was a kind of early online multiplayer game in which players could interact and communicate by sending text commands to one another. Mr. Spock was able to use a straightforward pattern matching algorithm to reply to simple conversational prompts and questions. There was a finite list of possible answers. As they explored the virtual text world together, players could communicate with Mr. Spock by sending him text commands like "say to Spock" or "ask Spock about". Mr. Spock was an early attempt at an artificial intelligence (AI) figure who could converse with people in a virtual setting, albeit being incredibly limited. It introduced the concept of NPC (non-player character) bots, which are widely used in virtual worlds and games today. Later conversational agents in MUDs, MOOs, and finally massively multiplayer online games were made possible through the development of TINYMUD.

Based on [22], the first chatbot to engage people in conversation was Dr. Sbaitso, which was introduced in the early 1990s. It was created specifically for MS-DOS and assumed the role of a psychotherapist, using voice to mimic a counseling session. By enhancing the pattern matching and replacement skills of earlier chatbots, Dr. Sbaitso was able to identify phrases and produce pertinent spoken responses. Despite its limitations, it was a step forward in conversational skills as it combined synthesized speech and text pattern matching to simulate a human-like dialogue. The basis for later chatbots' more complex verbal exchanges was established by Dr. Sbaitso's combination of voice interaction and basic language processing.

A.L.I.C.E. is an approximately more sophisticated bot developed by Richard Wallace in 1995 that is inspired by ELIZA [23]. The bot produced outcomes by pattern matching inputs against (input) (output) pairs kept in knowledge base documents. Artificial Intelligence Markup Language (AIML), an extension of XML that is still in use today, was used to write these documents. The Loebner Prize, which honors the most intelligent chatbot each year and attempts to run the Turing Test, has been won by ALICE three times [15].

- In 21st Century : First decade (2000 to 2009)

SmarterChild chatbot was developed by ActiveBuddy in 2001. It was an inspiration in achieving widespread acceptance. When inquired about something, it could supply information about a wide range of subjects and functioned on instant messaging networks. SmarterChild is one of the first chatbots to be integrated into messaging apps, demonstrated how AI could function as a virtual assistant by retrieving information and responding to user inquiries [24]. It might have been integrated to AOL Instant Messenger and MSN Messenger, reaching over 30 million users. SmarterChild inspired chatbots on messaging networks, opening the door for later AI assistants such as Samsung's S Voice and Apple's Siri. Its easy integration with famous messaging apps showed how chatbots may be used to offer services through dialogue-based messaging settings. Through introducing practical chatbot interactions into regular messaging, SmarterChild demonstrated the potential for millions of users to have access to chat-based assistants in the future [25].

Mitsuku chatbot was developer in 2005 by by Steve Worswick. It is an award-winning chatbot with a playful, adolescent girl demeanor. It has been the recipient of the Loebner Prize five times between 2013 and 2018. Judges for the Loebner Prize interact with a combination of chatbots and humans to identify which are the bots; Mitsuku has consistently deceived the judges in this regard. It is the most widely used AIML-based chatbot. It is hosted on Pandorabots and employs NLP techniques like heuristic patterns and default categories to have natural conversations across many platforms. It uses pre-written patterns and Bot modules to use natural language processing and is made for multiple platforms, including Telegram and Twitter. Although practical, Mitsuku needs new AIML categories to direct user input and manage mismatched prompts. It has been shown to be a useful conversation partner in multiple languages [21].

- In 21st Century: Second decade (2010 to 2019)

Apple developed Siri in 2010. It paved the path for personal assistants. It integrates with audio, video, and image files and allows users to utilize voice commands for questions and conversations using Messenger. With continuous use, Siri learns to recognize the users' language preferences, queries, and requests while providing recommendations and addressing user requests through a variety of online services. Siri has limitations, even if it is a powerful system. It demands an internet connection. Although it is multilingual, it is incapable of supporting many languages, and the only language in which navigation instructions are offered is English. Additionally, it has trouble understanding the person speaking because of their strong accent or background noise [26].

IBM developed a chatbot named Watson in 2011 [17]. Watson had the ability to decipher genuine human language well enough to defeat two previous champions in the game show "Jeopardy", in which players had to predict the questions based on responses they were given. Years later, Watson helped companies develop more advanced virtual assistants. Furthermore, Watson Health was created to assist medical professionals in diagnosing illnesses. Watson's limitation is that it only supports English, though.

When Google Now was first launched in 2012, it was intended to provide information to users based on their location, preferences, and time of day [27]. The next version of Google Now is Google Assistant, which was developed in 2016. It anticipates user needs and provides information to them with a more conversational, amiable interface that combines a deeper level of artificial intelligence. It lacks personality, though, and because it connects to the user's Google Account directly, some of the questions could be considered private [28].

Microsoft has introduced Cortana personal assistant in 2014. It is capable of recognizing voice commands to perform things like emailing, obtaining required information, and making reminders, there certainly have been reports of security holes that allow malware installation. Amazon developed Alexa voice assistant in the same year In order to promote IOT accessibility. One innovation is the Alexa Skills Kit, which allows developers to create custom skills for Alexa, though there are security risks associated with using Alexa. Although Alexa is focused on controlling home devices and Cortana was designed by Microsoft to be a personal assistant, both have raised security concerns over voice command functionality. Notwithstanding continuous security concerns, the introduction of Cortana and Alexa signaled a significant development in the field of AI voice assistants [29].

- In 21st Century: Third decade (2020 to current)

Based on [30], OpenAI introduced ChatGPT in multiple versions that started with GPT-1 in 2018, GPT-2 in 2019, and GPT-3 in 2020. ChatGPT showcases research on training neural networks on massive text data sets to acquire robust language skills, highlighting the most recent developments in employing large language models, such as GPT-3.5 and 4 for conversational AI. It is a prime example of state-of-the-art methods, including scaled-up self-supervised learning, that generate flexible models with robust reasoning and contextual understanding abilities. ChatGPT offers an interactive platform that allows users to experiment with and assess the advancements in research being done on creating dialogue agents with sophisticated LLMs optimized for multi-turn talks.

DeepSeek Chatbot is an advanced AI language model developed by DeepSeek, a Chinese AI research company [31]. It was first introduced to the public in early 2024, gaining attention for its strong performance in coding, reasoning, and multilingual capabilities, competing with models like OpenAI's GPT-4 and Meta's LLaMA. The chatbot is built on a transformer-based architecture, supporting 128K context length, which allows it to handle long documents efficiently. DeepSeek initially released open-weight models (such as DeepSeek LLM 7B/67B) before launching its conversational AI. As of mid-2024, DeepSeek Chatbot became available broadly, both in the free and paid versions, as a major player in the market for AI assistants. Its emergence is part of the global expansion of generative AI tools through an imitation of the rapid evolution of open and potent AI models beyond Western technology companies.

III. Literature Review

A thorough analysis of the main findings of existing research on chatbots is given in this section. It starts with a summary of earlier research that has looked at related subjects, especially in relation to conversational systems. The section goes on to list the main research gaps that this research attempts to fill, emphasizing how little attention has been paid to regional dialects, such the Western Saudi dialect, in both academic and real-world contexts.

A. Overview of Related Work

The authors in this study [27] provide an intelligent chatbot that utilizes advanced technologies for instance Python, Flask web framework, and NLP. This chatbot aims to produce a student's information access and investigate academic administration. The study's results emphsize that, the traditional programming paradigms and rule-based systems have enabled users to seamlessly fetch inclusive academic information by NLP queries. With quick access to extensive students' performance information, the administrators and teachers are able to simplify decision making that can improve methods for enhancing academic performance. Instantaneous insights into students' development allows teachers to recognize areas that

need improvement. The chatbot has created a standard for creative and approachable solutions inside educational establishments.

For the purpose of helping university students with academic and administrative issues, the study [28] introduces the development and execution of a College Assistance Chatbot. The chatbot effectively comprehends and processes student queries by utilizing deep learning models, especially an Artificial Neural Network (ANN), and NLP techniques. The architecture of the system is carefully developed to efficiently handle communication, guaranteeing immediate and accurate outcomes. For the neural network training, a large dataset was obtained from university resources and pre-processed. With a series of experimental campaigns, the performance of chatbot was thoroughly assessed. After 1000 epochs, it achieved an accuracy of 80.07%. The findings show that the chatbot can effectively respond to a variety of student requests, offering a scalable way to improve student support services. This novel methodology shows great promise for wider use in educational contexts and provides notable improvements over conventional techniques.

By utilizing the GPT model, the study [29] investigates the design, development, and assessment of an instructional chatbot for a university course. Testing the viability and efficacy of a GPT-based chatbot performing as a personal tutor for students enrolled in the Sociology of Education course is the main objective of the study. In addition to outlining a prototype for developing comparable educational chatbots, the researchers discuss to comprehend the advantages and disadvantages of GPT in educational contexts.

In order to foster critical thinking, the chatbot was developed to respond to inquiries, explain ideas, and involve students in Socratic discussion. The chatbot was created to be a tutor to students in two main implementation steps. First, building a knowledge base to improve subject-matter expertise. Second, configuring behavior, the chatbot was designed with an accurate prompt to prioritize answers from its knowledge base and refrain from depending on outside resources. The results of the study include that the GPT-based chatbot could provide precise answers, encourage interactive learning, and effectively adjust to the demands of the students. It was convenient and accessible because it was available all the time. While the limitations are relying on a very good quality training dataset, occasionally misinterpreting context, and having trouble completely recognizing complex ideas.

While most of the educational chatbot help students in self-learning, the study [30] explores how chatbots can enhance team-based learning independent on operating as a virtual instructor to assist students in working together for successfully complete their projects. Tubo (tutoring bot) was implemented as a component of a website platform for collaborative learning in which students work in groups to accomplish assignments while Tubo serves as a team member. Tubo offers a number of features, including onboarding by welcoming students and introducing the classroom, assignment guidance by outlining tasks and providing directions, using multiple-choice questions to reinforce important ideas, progress monitoring involves keeping tabs on student’s development and responding to often requested queries, and maintaining students' motivation through small conversation and encouragement. At the end of the experience of Tubo, students enjoyed Tubo's interactive and enjoyable diversion from traditional classroom instruction. A few pupils said they could envision themselves using Tubo on their own. On the other hand, a few students lost interest in Tubo's instruction because they thought it was overly strict and repetitious, and some of them struggled with task division and frequently worked in switches rather than cooperatively.

Based on the purpose of reducing educational inequities, particularly in the area of English language instruction, where rural children frequently lack access to resources and qualified teachers. The authors in this study [32] introduce BuddyBot chatbot to help achieve Sustainable Development Goal 4 (SDG4), which is to guarantee that everyone has access to high-quality, inclusive education. By giving instant feedback on grammar, vocabulary, and sentence structure, the chatbot attempts to provide a customized learning experience. BuddyBot is constructed by utilizing GPT-2 model and Google's FLAN-T5 model. In order to comprehend and react to learners' inquiries, correct grammar, and offer language feedback, both models were refined using specific datasets. The introduced chatbot performed strongly on fact-based and open-ended questions, as well as brief and lengthy chats. The results show a decrease in training loss from 0.778 to 0.414 across 8 epochs and an improvement in accuracy.

Considering the growing requirement in industries such as Small and Medium-sized businesses (SMEs), this study [33] investigates essential technologies for chatbots to produce responses. The study trains an enterprise chatbot knowledge base using deep learning frameworks. The study presents a chatbot that utilizes a Sequence to Sequence (Seq2Seq) model with Long Short-Term Memory (LSTM) to enhance SMEs industry. LSTM was used as the encoder and decoder, while the Seq2seq model merged with the target corpus to improve propriate the roles of customer service. At the beginning of neural network processing, the chatbot dataset completed preprocessing, word vector training, and dialogue Q&A conversion. Users were able to submit queries and view the chatbot’s generated answers through a Flask web interface. The introduced chatbot website can dynamically hot-load and train the model without restarting the chatbot after deployment. This feature enables ongoing enhancement and flexibility in response to changing client demands.

"Nabiha", a chatbot designed by [34] to engage in conversations with Information Technology (IT) students at King Saud University using the Saudi Arabic dialect. Consequently, Nabiha marks the first Arabic chatbot utilizing Saudi dialect. To enhance accessibility, Nabiha is available across multiple platforms including Android, Twitter, and the web. Students can interact with Nabiha by downloading the application, engaging with her on Twitter, or visiting her website. The effectiveness of Nabiha was assessed by IT department students, yielding somewhat satisfactory results considering the challenges posed by the Arabic language, especially the Saudi dialect.

In [35], the authors suggested and created the smart guiding chatbot that utilized Saudi dialect, a text-based application for

Jeddah's tourism industry. Users could always reach the chatbot for the most recent information on locations, events, weather, and recommended places to visit. They provided a technical overview of the materials NLP, ML, and Rasa.ai are necessary to develop a chatbot. The study also covers how the chatbot uses the probabilistic model LSTM to classify, process, and predict data in order to identify the best match. The suggested chatbot can comprehend user demands and reply promptly, according to the results.

The other research, which has proposed a chatbot in Arabic dialect, was presented by [36]. The purpose of the authors is to create an intelligent chatbot system that can lessen this kind of burden in Arabic. Specifically, this approach facilitates the use of a Jordanian dialect of Arabic with students, especially those attending the Al-Zaytoonah Private University of Jordan. As a result, the suggested system is a basic Jordanian chatbot that speaks in the local dialect. Several artificial algorithms are used in the Intelligent Arabic chatbot system's architecture to assess and comprehend a range of visitor queries. This system is a website that quickly responds to inquiries from users. It also has an easy-to-use interface that answers inquiries about the placement cell, test cell, academics, admissions, users' attendance records, and grade point averages, among other things.

In [5], the authors describe the creation and deployment of Bashayer, a task-oriented chatbot integrated inside the WhatsApp app. Its goal is to assist Saudi Arabian postgraduate students with their study methodologies and motivation. A single-subject experimental design utilizing a quasi-experimental setup was used with a group of 60 postgraduate Saudi students. The findings of the descriptive analysis of the data gathered indicated that postgraduate students who used the Bashayer chatbot system achieved encouraging outcomes. Compared to the control group, participants in the experimental group using Bashayer showed higher levels of motivation to learn. In addition, when using the chatbot, participants used more cognitive and metacognitive learning processes than the control group. The findings provide encouragement for the creation of Bashayer-like chatbots that will aid in the effective learning of postgraduate students. These findings fill a study void and add to the body of knowledge regarding the application of chatbots in postgraduate educational settings.

Table I summarizes these studies, focusing on the findings and limitations of each study.

This research is necessary, as it addresses a significant gap in the existing literature by developing a chatbot specifically tailored to understand and respond to the Western Saudi dialect, which has been underrepresented in the field of NLP. While there has been substantial work on general-purpose chatbots and models designed for widely spoken dialects or MSA, few studies have focused on region-specific dialects like the Western Saudi dialect, especially in the context of university students. By developing a chatbot that can respond to questions unique to Taibah University, including course registration, academic regulations, and campus information, this research seeks to give students a useful solution. This project will not only fill the gap in dialect-specific chatbot development but also enhance institutional efficiency and student assistance by relieving staff workloads and providing students with quicker access to information.

IV. Methodology

This section describes the research methodology that has been followed to achieve all the research objectives. Throughout this section, a description of the research methodology has been provided. Each phase ends with results obtained from their activities, and the results of each phase are used as a starting point for the next phase. The research methodology consists of six interrelated stages, namely:

- Problem Identification (stated earlier).
- Algorithms
- Data Collection
- Implementation Plan
- Evaluation Methods

A. Algorithms

The chatbot will utilize the LLM specifically, GPT model. This section briefly describes the LLM and GPT. LLM is a Statistical language model that estimates the probability of word sequences, facilitating a range of activities like code creation, question answering, and translation. To generalize from context, they are trained on large datasets and mostly rely on transformer designs. LLM learns a probability distribution for word sequences, see Eq. (2). This allows it to predict the next word given the previous context. Furthermore, it is designed to generate text by sampling from the learned probability distribution, facilitating tasks like summarization, essay writing, or even solving problems [37].

$$P(W1, W2, \dots, WL) \quad (2)$$

Autoregressive techniques, in which the model predicts the next word one step at a time, are frequently used by LLMs, see Eq. (3):

$$P(Wn + 1 | W1, W2, \dots, Wn) \quad (3)$$

The main factors of transformer architecture are transformer models which are rely on the attention mechanism, which enables the model to consider the relative relevance of various input phrases while making predictions, see Eq. (1). In the equation, Q, K, VQ, K, V represent queries, keys, and values derived from input embeddings.

Also, positional encoding utilizes to encode word order, positional embeddings are appended to word embeddings because transformers are intrinsically permutation invariant. Additionally, scalability transformers are more effective than recurrent designs because they enable parallel processing.

Self-supervised learning is used to train LLMs to anticipate next or masked words in text sequences. The goal is to reduce a loss function, such as cross-entropy, see Eq. (4). According to empirical scaling laws that relate loss to model and dataset size performance is enhanced with larger models and datasets.

TABLE I STUDIES SUMMARIZATION TABLE

Research	Approach	Model	Language	Key findings	Limitation
[27]	Retrieval based	ML	English	Enhancing academic improvement strategies	cost considerations, user acceptance, and the need for continuous updates.
[28]	Retrieval based	NLP, ANN	English	Handling a wide range of student queries and providing a scalable solution for enhancing student support services	-
[29]	Generative	GPT	English	Providing precise answers, encouraging interactive learning, and effectively adjusting to the demands of the students	having difficulty comprehending context and relying on the quality of the training data to generate accurate responses
[30]	Rule based	-	English	Enhancing the team learning method	Some student think that the chatbot became somewhat repetitive and boring over time
[32]	Generative	GPT-2, Google's FLAN-T5	English	achieving 80% accuracy in dialogue generation, demonstrating proficiency in question answering with an 83% accuracy	-
[33]	Retrieval based	Seq2Seq, LSTM	English	Effective preprocessing, Enhancing chatbot performance with LSTM	-
[34]	Retrieval based	-	Arabic, Saudi dialect	The initial Nabiha chatbot experiment's findings were fairly satisfactory.	Lack of dataset
[35]	Retrieval based	NLP, LSTM	Arabic, Saudi dialect	Effective interaction, and immediate responses	-
[36]	Retrieval based	NLP	Arabic, Jordanian dialect	Quickly responding to inquiries from users, and an easy-to-use interface	Additional experiments and adjustments are necessary.
[5]	Retrieval based	-	-	Supporting postgraduate students' motivation and learning strategies	Small students sample, short duration of testing

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\ P(w_i \mid w_{1:i-1}) \quad (4)$$

GPT model is a type of LLM which use transformer architecture to anticipate the next word in a sequence. Pretraining on large corpora to identify general linguistic patterns and fine-tuning on particular tasks are the two phases of training GPT models. With GPT-3 (175 billion parameters) bringing few-shot and zero-shot learning and GPT-4 (estimated 1.4 trillion parameters) showcasing sophisticated reasoning skills, GPT has seen significant evolution since its launch in 2018. Although it is still constrained by hallucinations, logical fallacies, and computational needs, its autoregressive methodology, scalability, and generalization across tasks are important aspects [38].

B. Data Collection

Since the accuracy of the results and conclusions is greatly impacted by the quality of the data collected, the data collection phase is one of the most important phases in any research or analytical study. The data was collected from a variety of sources during this phase to guarantee thoroughness and diversity, successfully matching the goals of the research.

Formal requests have been made to gather student inquiries that was send to academic advisers as part of the research on creating a GPT-based approach to assist Saudi university students. The Deanship of Information Technology at Taibah University has received the request. In order to examine common issues, commonly asked questions, and the general framework of academic advising exchanges, the goal of these requests was to obtain access to student inquiries that was send to academic advisers from eservices portal of Taibah University.

A Taibah University student Telegram group for the faculty of computer science (CCSET) was employed as one of the primary data sources. This platform was a powerful source of qualitative data since it offered insightful information about student conversations, concerns, and shared experiences. The author gathered pertinent information by examining the group discussions and exchanges, which helped clarify students' academic and administrative difficulties.

Moreover, information received in emails pertaining to the Course Equivalency Committee has been gathered. These emails contained information related to the policies concerning course credit transfers, decision-making processes, and the criteria for determining course equivalency requests. This source was particularly useful in understanding how the university manages credit recognition and how students interact with it.

The other major source of information included Academic Advising Committee emails. The emails provided data regarding the advising and counseling extended to students, including advising policy, mentorship programs, and academic services to help students in their educational planning. From perusing these messages, the university's academic advising mechanism and how the system assisted the students have been provided.

Also, the names and email addresses of faculty members at Taibah University was obtained from the Deanship of Information Technology. This source was essential in ensuring faculty members across various departments had access to accurate and up-to-date contact details.

By relying on these previously diverse sources, the balanced dataset that supports a multifaceted analysis of the academic experience at Taibah University has been collected. The reliance on several sources ensures a balanced perspective, drawing from student perceptions as well as official university policies.

C. Implementation Plan

The implementation strategy focuses on developing two chatbot systems that are specifically designed to interact using the Western Saudi dialect, ensuring that the responses align with the linguistic and cultural nuances of the target user group—university students in the western region of Saudi Arabia. Traditional machine learning performance is measured against the first chatbot, which is constructed using an SVM model. The GPT-3.5 model, a transformer-based generative language model renowned for its sophisticated contextual comprehension and fluency, is used by the second chatbot in contrast.

The objective of this dual-model setup is to conduct a systematic comparison between conventional machine learning techniques and state-of-the-art transformer models in the context of dialect-specific natural language understanding. Additionally, a separate retrieval-based model using Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity is implemented to handle queries related to faculty email addresses. In order to prevent the possibility of fabricated or hallucinated email information, this retrieval component is purposefully separated from the generative system. This ensures factual accuracy and upholds user trust.

D. Evaluation Methods

In this research project, a multi-dimensional assessment method was employed to critically assess the chatbot's performance. Three complementing methods were selected: human evaluation, BERTScore, and automatic GPT-4-based judgment. Human evaluation was used to elicit subjective factors such as naturalness, tone, and cultural appropriateness, yielding useful real-world observations particularly for the Western Saudi dialect. For quantitative measurement of semantic similarity among generated responses and reference answers independently, BERTScore was utilized as a context-sensitive alternative to typical surface-level measures. Further, GPT-4 also served as an automatic judge to enable scalable and uniform assessment across large datasets without trading practical practicability at the cost of linguistic insensitivity loss. By merging human judgment, semantic assessment, and machine testing, the technique provided a robust and rich assessment of the chatbot's linguistic, cultural, and functional capacity.

First, human evaluation was conducted to capture subjective aspects of response quality that are difficult to measure algorithmically, such as appropriateness, cultural alignment, and conversational naturalness. Human annotators assessed the relevance, fluency, and tone of the chatbot's replies based on a predefined set of guidelines tailored to the Western Saudi dialect context. This form of evaluation provided valuable insights into the real-world usability and acceptability of the chatbot among the target audience.

To evaluate the accuracy of the chatbot responses using human judgment, a human evaluation process was conducted involving five independent annotators. Each annotator reviewed the chatbot's responses and provided a binary judgment—either "Yes" if the response was considered appropriate, or "No" if it was not. For each response, the final decision was determined through majority voting, where the most frequent judgment (i.e., at least 3 out of 5 votes) was taken as the overall evaluation of that response. Once the final decisions were established for all responses, accuracy was computed by dividing the number of responses that received a majority "Yes" by the total number of responses evaluated. This approach ensures a reliable and consensus-based assessment of the chatbot's performance.

Second, BERTScore was employed to objectively measure the semantic similarity between the chatbot's generated responses and the ground-truth reference responses. Unlike traditional metrics such as BLEU or ROUGE that rely on exact n-gram matches, BERTScore employs deep contextual embeddings to identify whether two sentences are identical in meaning, even if they are differently phrased. This was particularly important given the linguistic richness and flexibility of the Western Saudi dialect, where multiple phrasings could be semantically correct.

In BERTScore, precision measures how well the chatbot's response captures the meaning of the reference answer. In Eq. (5), precision does this by checking each word in the chatbot's response and finding the most similar word in the reference, using BERT embeddings. A high precision means that most words in the chatbot's response are semantically similar to words in the reference answer.

$$P = \frac{1}{|\hat{X}|} \sum_{\hat{x}_j \in \hat{X}} max_{x_i \in X} \; sim(x_i, \hat{x}_j) \qquad (5)$$

In the BERTScore precision equation, each symbol plays a specific role in capturing the semantic alignment between the chatbot's response and the reference answer. The symbol $\hat{X}$ represents the set of tokens (words) in the generated response, while X denotes the set of tokens in the reference (ground truth) response. For every token, the equation searches for the most similar token x_i in the reference using cosine similarity of BERT embeddings, represented as $sim(x_i, \hat{x}_j)$. The maximum similarity for each generated token is then summed, and the total is divided by the number of generated tokens $|\hat{X}|$, resulting in the average semantic precision across the response.

On the other hand, recall in BERTScore checks how well the chatbot's response covers the important content of the reference answer. In Eq. (6), recall looks at each word in the reference and finds the closest matching word in the chatbot's response. A high recall means the chatbot included most of the important meanings from the reference answer.

$$R = \frac{1}{|X|} \sum_{x_i \in X} max_{\hat{x}_j \in \hat{X}} \; sim(x_i, \hat{x}_j) \qquad (6)$$

The recall equation in BERTScore operates in the reverse direction of precision. Here, X still denotes the set of reference

tokens, and $X\hat{}$ is the set of generated tokens. For each reference token, the model computes the maximum similarity with any token $x\hat{}_j$ in the generated response. This similarity is again based on cosine distance between BERT embeddings, expressed as $im(x_i, x\hat{}_j)$. The total of these maximum similarities is then divided by the number of reference tokens $|X|$, producing a measure of how completely the generated response covers the semantic content of the reference.

Finally, the BERTScore F1 combines precision and recall to give a single balanced score. As shown in Eq. (7), F1 score tells us how well the chatbot's response matches the reference in both accuracy and coverage of meaning. A high F1-Score means the chatbot response is both relevant and complete from a semantic perspective.

$$F1 = \frac{2PR}{P+R} \quad (7)$$

Third, GPT-4 was used as an automatic judge to provide scalable and consistent evaluation across a large number of test samples. GPT-4 was prompted to assess whether the chatbot's response was contextually appropriate and semantically accurate based on the given input message. The accuracy was calculated in the same way as human evaluation, by dividing the number of responses that were appropriate by the total number of responses. By using a state-of-the-art LLM as an evaluator, the research was facilitated with high linguistic sensitivity and judged consistency and reduced the need for a lot of human effort at the expense of evaluation quality.

By integrating these three mutually complementary evaluation methods, the performance assessment of the chatbot was made more balanced in nature, encompassing both human-centered perspectives and machine-based semantic and contextual accuracy. This provided an overall view of the strengths and shortcomings of the chatbot, thus enhancing the validity of the experiment results.

The selection of human evaluation, BERTScore, and GPT-4-based judging is particularly well-aligned with the specific objectives and challenges of this research. The primary goal is not merely to assess the technical accuracy of the chatbot's responses, but to ensure their cultural, linguistic, and contextual appropriateness within the Western Saudi dialect environment — an aspect that conventional metrics alone cannot fully capture.

Human judgment plays a crucial role in this regard, making fine-grained judgments founded on real-world conversational practices and user expectations. Given the sociolinguistic variation and familiarity built into the Western Saudi dialect, human judges are necessary to validate the chatbot's ability to converse naturally and respectfully with the target group.

Complementing this subjective perspective, BERTScore was chosen for its capacity to assess semantic similarity with high sensitivity to context and meaning. In dialectal settings, where different surface forms can express identical intents, traditional exact-match metrics would severely underestimate model performance. BERTScore circumvents this risk by taking into account if the answers generated convey the intended meaning, thus giving a more linguistically fair and representative score.

Furthermore, GPT-4 as an automatic judge introduces a scalable and efficient dimension to the assessment. Its advanced language understanding enables it to make near human-like assessment judgments on large quantities of data without the prohibitive time and costs of manual annotation. This scalability is especially relevant with the sizes of the test sets here and the need for consistent evaluation metrics on thousands of samples.

Overall, the selected evaluation metrics collectively are well-suited because they accommodate the research's new linguistic complexity, deployment practicability, and research objectives. Overall, they measure the chatbot's performance wholeheartedly, not just its correctness, but its meaningful conversation in a specific cultural and dialectal context.

V. Implementation and Experimentation Settings

This section provides the implementation method of the chatbot tailored to the Western Saudi dialect, specifically for the context of Taibah University. A baseline model using an SVM architecture and a more sophisticated model based on the GPT framework were both used to evaluate performance. The GPT-based model uses transformer-based language understanding, which is better suited to capture the subtleties of dialectal Arabic, whilst the SVM model provides a classificational modeling method. The two models were both trained with a customized dataset representing the western Saudi dialect linguistic properties, and the hyperparameters as well as the preprocessing methods used were deliberately chosen to give an informative and objective comparison. More information regarding the tools, the frameworks, training configurations, and the testing techniques used in this research is given in the next subsections.

A. Experiment 1: Email Retrieval-Based Model

To handle faculty email inquiries with greater reliability and to avoid the potential risk of generating inaccurate or non-existent email addresses, a separate retrieval-based model was developed. This model is built using TF-IDF and Cosine Similarity to retrieve verified faculty emails from a structured dataset. The system ensures that only authorized and approved email addresses are returned by depending on exact matches rather than generating output. The GPT-3.5-Turbo-0125 chatbot will eventually be merged with this retrieval-based email architecture. The hybrid system will automatically route email-related queries to the retrieval model, while general academic or administrative questions will be handled by the generative GPT-3.5-Turbo-0125 model. This architecture enhances both the accuracy and trustworthiness of responses by combining retrieval precision with the conversational fluency of transformer-based generation.

The TF-IDF approach is used to measure each text message in numerical terms by assigning weights to words based on their importance. Especially, TF refers to the number of times a word repeats in a message, and IDF reduces the weight of common words that appear in most messages. This provides a high-dimensional vector for each message that reflects its unique linguistic features. In order to find the best fit for a user's query, the input is converted by the same TF-IDF vectorizer and matched against the stored message vectors by Cosine Similarity, which measures the angle between vectors. The most

similar message is chosen, and its corresponding response is returned. This method is computationally efficient, doesn't require model training, and performs well when the data set is very clean and there are stable patterns in the language.

An important function was developed to serves as a decision-making mechanism that intelligently routes user queries to the appropriate response model. Specifically, it checks whether the user's input contains keywords related to email inquiries (e.g., "email", "بريد", "إيميل"). If such keywords are detected, the function delegates the task to a retrieval-based model built on TF-IDF and cosine similarity to return an exact match from a predefined list of official faculty email addresses. This approach is deliberately chosen to avoid generating or "hallucinating" non-existent email addresses, a risk commonly associated with generative models like GPT. Conversely, if the query is unrelated to emails, it is forwarded to a fine-tuned GPT model trained to respond in the Western Saudi dialect. This dual-model architecture ensures both accuracy in factual information retrieval and fluency in conversational responses.

B. Experiment 2: Baseline Model

In this experiment, a baseline model was implemented using SVM to evaluate the feasibility of building a dialect-specific chatbot capable of understanding and responding in the Western Saudi dialect. The SVM model was selected due to its effectiveness in handling classification tasks with limited training data and its interpretability as a classical ML method. This experiment served as a foundational comparison for the subsequent evaluation of a more sophisticated GPT-based model. By establishing the performance ceiling of a traditional model, the research aimed to determine the added value of utilizing LLMs in regional Arabic dialect applications.

The dataset used in this experiment consisted of paired question–answer entries written in the Western Saudi dialect, preprocessed to remove emojis, special characters, and unnecessary whitespace. The data was divided into 80% for training and 20% for testing. Text preprocessing included tokenization, lowercasing, and TF-IDF vectorization to convert the textual data into numerical feature representations. An SVM classifier with a linear kernel was then trained on the resulting feature matrix.

C. Experiment 3: GPT Model

The third experiment explores the performance of an advanced large language model, GPT-3.5-Turbo-0125, developed and hosted by OpenAI. Unlike traditional ML models such as SVM, GPT-based models leverage deep transformer architectures pre-trained on massive corpora. This capability enables the model to generate contextually relevant and linguistically coherent responses, even in dialectal variants with limited annotated resources. The purpose of this experiment is to evaluate the effectiveness of GPT-3.5-Turbo-0125 when fine-tuned on a custom dataset of dialogues in the Western Saudi dialect, and to compare its performance against the previously established SVM baseline.

For fine-tuning the same dataset that was utilized in the SVM, has also been utilized in this phase. The dataset of Western Saudi dialect messages and their corresponding responses was converted into a format compatible with OpenAI's fine-tuning pipeline. The input-output pairs were structured as prompt–completion sequences, with each user message serving as a prompt and the ideal chatbot reply as the completion. The training process was conducted via OpenAI's fine-tuning API, utilizing the base model gpt-3.5-turbo-0125. Once the custom model was trained, evaluation was performed using a test set consisting of unseen message-response pairs. Additionally, some test cases were assessed using an automated judge—GPT itself—by requesting it to determine whether the generated response was appropriate, based on a scoring system that reflected qualitative judgment. This setup enabled both quantitative and human-aligned evaluation.

VI. Results and Discussion

This section presents a detailed analysis of the experimental results and interprets their significance in the context of the research. The performance of the proposed chatbot model is examined using relevant evaluation metrics to assess its ability to understand and respond in the Western Saudi dialect. A comparative study is carried out against a baseline SVM model in order to better position the results and emphasize the advantages of transformer-based techniques over traditional methods for ML. The ramifications of the results and their applicability to the goals of the research.

A. Results of Email Retrieval-Based Model

The model achieved an Exact Match Accuracy of 99.96%, indicating that it performs with a very high level of precision in retrieving the correct email address based on the provided Arabic display names.

TABLE II Results of SVM Responses of User's Messages

Message		**Response**	
Arabic	***English***	***Arabic***	***English***
لو نزلي حرمان وعندي اعذار ينشال؟	If I receive a deprivation notice but have valid excuses, is it possible to have it lifted?	ايوا	Yes
كم الغياب المسموح لمحاضرات عددها٣٠؟	How many absences are allowed for 30 lectures?	ايوا	Yes
أبغى ايميل دكتور ماهر	I want Dr. Maher's email	لازم تعطيني الإسم الثنائي لعضو هيئة التدريس عشان أبحث بشكل دقيق	Please provide the full name of the faculty member in order to perform an accurate search.

B. Results of SVM Baseline Model

Table II shows the results of the SVM-based chatbot's performance in handling different inquiries. Each row in the table represents a message from the user and the corresponding response generated by the chatbot.

The response of message "لو نزلي حرمان وعندي اعذار ينشال" (If I receive a deprivation notice but have valid excuses, is it possible to have it lifted) indicates that the chatbot correctly understood the user's query about whether an academic penalty

could be waived in the presence of valid excuses. The chatbot provides a simple affirmative response "ايوا" (yes), confirming the possibility of such an action.

The model's weakness in answering the question " كم الغياب المسموح لمحاضرات عددها ٣٠" (How many absences are allowed for 30 lectures?) is evident in its vague and imprecise response of "ايوا" (yes). This indicates that the SVM model failed to fully understand the context of the question, as it should have provided a specific answer regarding the allowed absenteeism percentage for a given number of lectures. Instead, it gave an unclear response.

In the third response, the chatbot correctly asked the user for the full name of the faculty member "دكتور ماهر" (Doctor Maher) in order to provide an accurate result.

Table III provides a summary of the SVM model's performance metrics on the test dataset, which comprises 20 messages and the responses that go with them.

TABLE III EVALUATION METRICS RESULTS OF 20 MESSAGES AND ANSWERS OF TEST DATA FOR SVM

Evaluation metric	Result
Accuracy	0. 4286
F1-score	0. 8709
Precision	0. 8806
Recall	0. 8631
latency	0.2 sec

The model was able to accurately predict the answers for almost 43% of the test data, according to its accuracy of 42.86%. Given that the model is trained in predefined classes and assessed on unseen data, this accuracy may seem moderate, but it actually represents the difficulty of the task. The test set contained unseen responses, which the model might not have been trained to handle well, which could account for the comparatively poor accuracy. Given how much the SVM model depends on the data it was trained on, this points to a possible drawback.

In terms of F1-score, the model achieved a value of 87%, suggesting a strong balance between precision and recall. The precision of 0.8806 indicates that when the model predicted a response, it was correct approximately 88% of the time, minimizing false positives. Meanwhile, the recall of 0.8631 shows that the model was able to identify about 86% of the true positive responses, indicating that it was successful in capturing most of the correct answers in the test set. With a latency of 0.2 seconds per prediction, the model can respond in real time, which makes it appropriate for chatbots and other applications that need fast feedback. Even with its moderate accuracy, the overall performance indicates that, when trained on the different formations of data, the SVM model may effectively detect accurate responses.

C. Results of GPT-3.5-Turbo-0125 Advanced Model

The sample results from the GPT-3.5-Turbo-0125 model that is provided in the Table IV show its ability to provide contextually accurate and relevant responses to a variety of student inquiries. For instance, when asked about permissible absenteeism for 30 lectures, the model gave a specific response, stating that 25% absenteeism is allowed, which is a precise and useful answer. Other responses, such as requests for emails or explanations on university policies, are similarly helpful, providing the necessary details or requesting additional information when needed.

The fine-tuned GPT-3.5-Turbo-0125 model demonstrated a significant improvement in handling dialogue generation tasks compared to traditional ML approaches. Upon interacting with user prompts written in the Western Saudi dialect, the model consistently produced coherent, contextually appropriate responses, and linguistically aligned with the informal and natural conversational style characteristic of the region. The system message used on each interaction gave the model a welcoming, culturally appropriate tone and essentially replicated the pattern of speech of students in Madinah, Jeddah, Mecca, and Taif. Throughout the testing phase, the model exhibited an excellent ability to comprehend highly varied input phrasings, including colloquial expressions, truncated sentences, and context-implicated questions, without requiring explicit rephrasing or clarification.

Furthermore, the qualitative observations output of the model highlighted its strong capacity for conforming to delicate language use, such as responding appropriately to rhetorical questions, handling indirect requests, and engaging with culturally embedded expressions. Unlike traditional models that rely on fixed feature extraction techniques, GPT-3.5-Turbo-0125 leveraged its deep contextual understanding to generate responses that not only matched the intended meaning of the prompts but also maintained the conversational fluency expected in natural human dialogue. The consistency, richness, and contextual awareness observed in the outputs confirm the model's potential for real-world deployment in dialect-specific chatbot applications. These results strongly support the adoption of LLMs when aiming to develop conversational agents capable of interacting naturally and effectively within specific cultural and linguistic contexts.

Table V contains the evaluation metrics that provide important information about the performance of the GPT-3.5 model, which has been optimized to comprehend and reply to student questions in the Western Saudi dialect. The Human Evaluation Accuracy of 0.8473 demonstrates that the model's responses were deemed accurate by the five evaluators, consisting of three students from the Computer Science department, a Computer Science graduate, and a technical employee. This great accuracy shows that the model was able to correctly analyze and answer questions with a degree of agreement among experts, even with its specialized training.

TABLE IV RESULTS OF GPT-3.5-TURBO-0125 MODEL RESPONSES OF USER'S MESSAGES

Message		Response	
Arabic	*English*	*Arabic*	*English*
لو نزلي حرمان وعندي اعذار ينشال؟	If I receive a deprivation notice but have valid excuses, is it possible to have it lifted?	ايوا	Yes
كم الغياب المسموح لمحاضرات عددها ٣٠؟	How many absences are allowed for 30 lectures?	الغياب المسموح 25٪ من مجموع المحاضرات	The allowed absence is 25% of the total number of lectures.
أبي ايميل دكتور ماهر	I want Dr. Maher's email	لازم تعطيني الإسم الثنائي لعضو هيئة التدريس عشان أبحث بشكل دقيق	You need to provide the full (first and last) name of the faculty member so I can search accurately.
بسأل في التقييم بالموقع يطلع عند الدكتورة اسمي؟	If I submit a course evaluation on the website, will my name appear to the instructor?	لا ما يطلع اسمك بالتقييم	No, your name does not appear in the evaluation.
انا اول سنة واخذ برمجه طبيعي اني احسها جدا صعبة؟	This is my first year, and I'm taking programming. Is it normal to find it very difficult?	طبيعي جداً ، اغلب الطلاب يحسون بصعوبة في مادة البرمجة لكن المهم تراجعها أول بأول	It's completely normal. Most students find programming difficult, but the key is to review it regularly.
التاجيل قفل ؟	Is the postponement period closed?	على حسب التقويم	It depends on the academic calendar.

TABLE V EVALUATION METRICS RESULTS OF ALL TEST DATA FOR GPT-3.5-TURBO-0125

Evaluation metric	Result
Human Evaluation Accuracy	0.8473
GPT-4 Evaluation Accuracy	0.60
F1-score	0.706
Precision	0.71
Recall	0.705
latency	2 sec

On the other hand, the GPT-4 Evaluation Accuracy of 0.60 reflects a somewhat lower performance from the GPT-4 model in evaluating the chatbot's responses. The difficulty of having a non-human evaluator evaluate responses involving complicated dialects and region-specific settings is highlighted by the fact that, despite its power, GPT-4 was less in line with the perceptions of human evaluators. Combining Precision (0.71) and Recall (0.705), the F1-score of 0.706 indicates a well-balanced model performance. It was able to effectively identify correct responses while minimizing errors. Precision suggests that 71% of the responses identified as correct were truly correct, while Recall shows that the model captured about 70.5% of all the correct responses. Lastly, these metrics demonstrate the model's accuracy and relevance in processing queries in the Western Saudi dialect.

VII. DISCUSSION OF FINDINGS

As shown in Table VI, the difference in responses between GPT-3.5-Turbo-0125 and SVM for the message "كم الغياب المسموح لمحاضرات عددها ٣٠" (How many absences are allowed for 30 lectures?). It is noticeable that GPT-3.5-Turbo-0125 performs better than the SVM model in terms of giving more precise and thorough answers. In particular, GPT-3.5 gave an exact response of 25% absenteeism to the question on acceptable absences for 30 courses, demonstrating a deeper comprehension of academic policies. The SVM model, on the other hand, had trouble answering this kind of question and provided ambiguous answers like "ايوا" (Yes), which lacked the relevant information. Particularly when contrasted with the more straightforward and uninformative results produced by the SVM model, GPT-3.5's capacity to manage intricate queries and deliver precise, context-aware responses is remarkable. This demonstrates GPT-3.5's superior understanding of context and ability to generate specific and meaningful responses.

TABLE VI RESPONSE COMPARISON BETWEEN SVM AND GPT-3.5-TURBO-0125

	Message	Response
SVM	كم الغياب المسموح لمحاضرات عددها ٣٠	ايوا
GPT-3.5-Turbo-0125	كم الغياب المسموح لمحاضرات عددها ٣٠	الغياب المسموح 25٪ من مجموع المحاضرات

The comparative evaluation between the SVM baseline model and the fine-tuned GPT-3.5-Turbo-0125 model clearly demonstrates the superiority of transformer-based approaches in handling dialect-specific dialogue tasks. The SVM baseline, trained on TF-IDF features with an RBF kernel, achieved a modest accuracy of 42.86% on the test set that contains only 20 messages. Its performance was limited by its dependency on surface-level lexical similarities and the insufficiency of generalization to all the random and garrulous expressions that belong to the Western Saudi dialect.

Human assessors determined that 84.73% of the chatbot's responses were both culturally appropriate and contextually accurate. In addition to the human evaluation, BERTScore provided further validation of the model's performance, achieving an F1 score of 71% and confirming its ability to capture the intended semantic meaning. These findings were reinforced by GPT-4, used as an automated assessor, which judged 60% of the responses as acceptable. Collectively, these outcomes illustrate the deep contextual understanding, semantic flexibility, and cultural sensitivity of transformer-based models, particularly when compared to classical ML classifiers. This distinction becomes even more evident when addressing the linguistic richness and complexities associated with regional Arabic dialects.

The outcomes of this research carry important implications for the broader field of computer science, particularly in the domains of NLP, dialogue systems, and dialectal language

modeling. The significant performance gap observed between the SVM baseline model, and the fine-tuned GPT-3.5-Turbo-0125 model highlights the critical role that large-scale pretraining and context-sensitive fine-tuning play in building effective conversational agents. These results emphasize that typical ML techniques, while helpful for certain structured classification tasks, are insufficient to deal with the richness, dynamism, and cultural nuance of real-world dialogue, especially in underrepresented language varieties such as the Western Saudi dialect.

More generally, the successful deployment of a chatbot that answers academic queries in culturally responsive and linguistically sensitive terms demonstrates the growing capacity of AI to support broad and niche communities. It demonstrates how adapting LLMs to regional and domain environments can significantly enhance accessibility, inclusivity, and user satisfaction. This outcome is particularly relevant where MSA dominates the technology interfaces to the detriment of local dialect speakers. Thus, this research contributes to the broader movement toward more personalized and culturally sensitive AI technologies.

Moreover, this work moves the frontier of knowledge and practice by illustrating an effective way of combining fine-tuning techniques, dialectal prompt engineering, and multi-dimensional assessment methods to attain high-quality conversational systems. It also provides empirical evidence that cultural and linguistic localization is not only possible but also highly effective when developing AI solutions for real-world applications. By systematically comparing the performance of the chatbot through human evaluation, semantic similarity metrics, and automatic metrics, this study sets a new benchmark for the evaluation of dialogue systems in low-resource language settings. Lastly, the research paves the way for future studies to bridge the gap between state-of-the-art language modeling technologies and the rich linguistic diversity within and across Arabic-speaking societies.

VIII. Conclusion

This research effectively created a culturally sensitive chatbot that could respond to academic questions in the Western Saudi dialect, providing significant contributions to Arabic NLP and dialogue system research. The comparison between the SVM baseline and fine-tuned GPT-3.5-Turbo-0125 model unequivocally exhibited the greater contextual understanding, semantic versatility, and cultural suitableness of transformer-based models. By integrating human judgment, BERTScore computation, and GPT-4-based analysis, the research provided a holistic and balanced assessment of the chatbot's performance. The findings demonstrate that AI technology is more inclusive and relevant than only its technological advancements when it is modified for underrepresented languages. This work sets a new benchmark for developing conversational agents in low-resource dialects by combining culturally appropriate system prompts, fine-tuning strategies, and stringent evaluation processes.

The research has limitations, especially in obtaining adequate and representative data for the Western Saudi dialect, despite the encouraging results. The complexity of Arabic and the variation within the dialect across different cities also posed challenges to building a consistent and natural chatbot experience.

Future work should focus on expanding dialect-specific datasets, addressing intra-dialectal variation, and exploring adaptive models that better generalize across regional differences. To further improve the caliber and versatility of AI-driven dialogue systems, future research might build on this foundation by investigating other dialects, expanding domain-specific applications, and improving assessment frameworks.

Acknowledgment

I would like to express my heartfelt gratitude to my project supervisor for her invaluable guidance, continuous support, and encouragement throughout the course of this research.

I am deeply thankful to my family and friends for their unwavering support, motivation, and patience during my academic journey.

I would also like to extend my sincere appreciation to Taibah University for providing an outstanding educational environment.

References

[1] S. Nithuna and C. A. Laseena, “Review on Implementation Techniques of Chatbot,” Proc. 2020 IEEE Int. Conf. Commun. Signal Process. ICCSP 2020, pp. 157–161, 2020, doi: 10.1109/ICCSP48568.2020.9182168.

[2] M. T. ZEMČÍK, “A Brief History of Chatbots,” DEStech Trans. Comput. Sci. Eng., no. aicae, 2019, doi: 10.12783/dtcse/aicae2019/31439.

[3] S. P. Reddy Karri and B. Santhosh Kumar, “Deep learning techniques for implementation of chatbots,” 2020 Int. Conf. Comput. Commun. Informatics, ICCCI 2020, pp. 20–24, 2020, doi: 10.1109/ICCCI48352.2020.9104143.

[4] Y. Saoudi and M. M. Gammoudi, “Trends and Challenges of Arabic Chatbots: Literature Review,” Jordanian J. Comput. Inf. Technol., vol. 9, no. 3, pp. 261–286, 2023, doi: 10.5455/jjcit.71-1685381801.

[5] A. M. Al-Abdullatif, A. A. Al-Dokhny, and A. M. Drwish, “Implementing the Bashayer chatbot in Saudi higher education: measuring the influence on students’ motivation and learning strategies,” Front. Psychol., vol. 14, 2023, doi: 10.3389/fpsyg.2023.1129070.

[6] “Chatbot |definition of chatbot in english by Lexico,” Lexico Dictionaries. [Online]. Available: https://www.dictionary.com/browse/chatbot.

[7] A. . dkk Trivedi, “Chatbot generation and integration: A review,” Int. J. Adv. Res., vol. 5, no. 2, pp. 1308–1311, 2019, [Online]. Available: www.IJARIIT.com.

[8] S. Singh and H. K. Thakur, “Survey of Various AI Chatbots Based on Technology Used,” ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir., pp. 1074–1079, 2020, doi: 10.1109/ICRITO48877.2020.9197943.

[9] S. Hussain, O. Ameri Sianaki, and N. Ababneh, A Survey on Conversational Agents/Chatbots Classification and Design Techniques, vol. 927, no. October. Springer International Publishing, 2019. doi: 10.1007/978-3-030-15035-8_93.

[10] J. Skrebeca, P. Kalniete, J. Goldbergs, L. Pitkevica, D. Tihomirova, and A. Romanovs, “Modern Development Trends of Chatbots Using Artificial Intelligence (AI),” ITMS 2021 - 2021 62nd Int. Sci. Conf. Inf. Technol. Manag. Sci. Riga Tech. Univ. Proc., no. October, pp. 1–6, 2021, doi: 10.1109/ITMS52826.2021.9615258.

[11] B. J. Cahn, “CHATBOT : Architecture , Design , & Development,” 2017.

[12] J. Weizenbaum, “ELIZA-A computer program for the study of natural language communication between man and machine,” Commun. ACM, vol. 9, no. 1, pp. 36–45, 1966, doi: 10.1145/365153.365168.

[13] E. Adamopoulou and L. Moussiades, “Chatbots: History, technology, and applications,” Mach. Learn. with Appl., vol. 2, no. November, p. 100006, 2020, doi: 10.1016/j.mlwa.2020.100006.

[14] K. M. Colby, S. Weber, and F. D. Hilf, "Artificial Paranoia," Artif. Intell., vol. 2, no. 1, pp. 1–25, 1971, doi: 10.1016/0004-3702(71)90002-6.

[15] J. O. N. F. Heiser and R. C. Parkison, "CAN PSYCHIATRISTS DISTINGUISH A COMPUTER SIMULATION OF PAR , ANOIA FROM THE REAL THING ? THE LIMITATIONS OF TURING-LIKE TESTS AS MEASURES OF THE ADEQUACY OF SIMULATIONS * The computer program is a model of a mythical 28 year old , single , Caucasian , nat," vol. 15, pp. 149–162, 1979.

[16] N. Jeevanandam, "Exploring Jabberwacky - a chatbot emulating a human conversation.," INDIA AI. [Online]. Available: https://indiaai.gov.in/article/exploring-jabberwacky-a-chatbot-emulating-a-human-conversation

[17] M. S. Ali, A. Salam, A. Khan, A. Ali, and A. Ullah, "History of generative Artificial Intelligence (AI) chatbots : past , present , and future development," arxiv, 2024.

[18] B. Harsh and A. Ms.Geeta, "Review on Different Types of Chatbots," Int. Res. J. Mod. Eng. Technol. Sci., no. 05, pp. 1347–1349, 2021.

[19] R. Tamrakar and N. Wani, "Design and Development of CHATBOT : A Review," no. April, 2021.

[20] T. Pillare, M. Chaoudhari, V. Hiwrale, S. Ade, P. Dudhe, and S. Chinte, "A SURVEY PAPER ON CHATBOT," no. 10, pp. 2915–2919, 2023.

[21] A. Toprak, M. Turan, and F. S. Toprak, "INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Deep Learning Based Chatbot Adapted to the Electronic Funds Transfer Process of Turkish Banking," vol. 11, no. 1, pp. 356–362, 2023.

[22] H. Soffar, "Apple Siri features, use, advantages, disadvantages and using of Siri For Learning," Online Sciences. [Online]. Available: https://www.online-sciences.com/technology/apple-siri-features-use-advantages-disadvantages-using-of-siri-for-learning/.

[23] S. I. Hussain, S. Deepalakshmi, R. J. Benilla, and V. C. Nivetha, "Automation of Smart Home using Smart Phone via Google Assistant," Proc. - 5th Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2023, no. Icssit, pp. 499–505, 2023, doi: 10.1109/ICSSIT55814.2023.10060979.

[24] "Google assistant, your own personal google," google. [Online]. Available: https://assistant.google.com/.

[25] Z. M. Paul, "Cortana-Intelligent Personal Digital Assistant: a Review," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 7, pp. 55–57, 2017, doi: 10.26483/ijarcs.v8i7.4225.

[26] J. O. Sarfo, "Artificial Intelligence Chatbot – ChatGPT and High-Tech Plagiarism Concerns in a Digital Age: Is Detection Possible?," J. Advocacy, Res. Educ., vol. 10, no. 2, pp. 55–58, 2023, doi: 10.13187/jare.2023.2.55.

[27] N. S. Reddy, N. P. Chaitanya, P. S. Varshitha, R. C. Varma, P. K. Reddy, and J. Dandu, "Intelligent Chatbot for Educational Institutions," Proc. Int. Conf. Circuit Power Comput. Technol. ICCPCT 2024, vol. 1, pp. 1337–1343, 2024, doi: 10.1109/ICCPCT61902.2024.10673214.

[28] M. S. Abdul Razak, M. V. Manoj Kumar, C. R. Nirmala, R. Naseer, B. S. Prashanth, and H. R. Sneha, "Enhancing Student Support with AI: A College Assistance Chatbot Using NLP and ANN," 2nd IEEE Int. Conf. Networks, Multimed. Inf. Technol. NMITCON 2024, pp. 1–8, 2024, doi: 10.1109/NMITCON62075.2024.10699279.

[29] S.-V. Fulgencio, "Developing Effective Educational Chatbots with GPT: Insights from a Pilot Study in a University Subject," Trends High. Educ., vol. 3, no. 1, pp. 155–168, 2024, doi: 10.3390/higheredu3010009.

[30] M. Burkhard, S. Seufert, M. Cetto, and S. Handschuh, "Educational Chatbots for Collaborative Learning: Results of a Design Experiment in a Middle School," Proc. 19th Int. Conf. Cogn. Explor. Learn. Digit. Age, CELDA 2022, no. Celda, pp. 209–218, 2022, doi: 10.33965/celda2022_202207l027.

[31] DeepSeek-AI et al., "DeepSeek-V2: A Strong, Economical, and Efficient Mixture of-Experts Language Model," vol. 50000, pp. 1–52, 2024, [Online]. Available: http://arxiv.org/abs/2405.04434.

[32] J. P. Dhivvya and S. B. Karnati, "BuddyBot: AI Powered Chatbot for Enhancing English Language Learning," 2024 IEEE Int. Conf. Interdiscip. Approaches Technol. Manag. Soc. Innov. IATMSI 2024, vol. 2, pp. 1–6, 2024, doi: 10.1109/IATMSI60426.2024.10502595.

[33] O. Baker, A. Rehman, O. Basas, and J. Liu, "Deep Learning for Intelligent Customer Service Automation: Development of GRU, LSTM, and Recurrent Neural Network Architectures for Chatbot Applications," Proceeding - 2024 Int. Conf. Inf. Technol. Res. Innov. ICITRI 2024, pp. 111–117, 2024, doi: 10.1109/ICITRI62858.2024.10699245.

[34] D. Al-Ghadhban and N. Al-Twairesh, "Nabiha: An Arabic dialect chatbot," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 3, pp. 452–459, 2020, doi: 10.14569/ijacsa.2020.0110357.

[35] R. Alotaibi, A. Ali, H. Alharthi, and R. Almehamadi, "AI Chatbot for Tourism Recommendations A Case Study in the City of Jeddah, Saudi Arabia," Int. J. Interact. Mob. Technol., vol. 14, no. 19, pp. 18–30, 2020, doi: 10.3991/ijim.v14i19.17201.

[36] N. A. Al-Madi, K. A. Maria, M. A. Al-Madi, M. A. Alia, and E. A. Maria, "An Intelligent Arabic Chatbot System Proposed Framework," 2021 Int. Conf. Inf. Technol. ICIT 2021 - Proc., pp. 592–597, 2021, doi: 10.1109/ICIT52682.2021.9491699.

[37] M. R. Douglas, "Large Language Models," pp. 1–47, 2023.

[38] J. Gehrke, P. Lee, and H. Palangi, "Sparks of Artificial General Intelligence : Early experiments with GPT-4," arxiv, 2023.

Automatic Detection of Natural Disasters Using Faster R-CNN with ResNet50 Backbone

Shereen Essam Elbohy[1], Mona M. Nasr[2], Farid Ali Mousa[3]
Faculty of Computers and Artificial Intelligence, Helwan University, Egypt[1]
Associate Professor in Faculty of Computers and Artificial Intelligence, Helwan University, Egypt[2]
Associate Professor in Computer Science, Beni Suef University, Egypt[3]

***Abstract*—Natural disasters pose significant threats to human life and infrastructure. Timely detection and assessment of these events are crucial for effective disaster management. This study proposes an automatic detection system for natural disasters using aerial imagery. Accurate and timely detection of natural disasters is critical for minimizing their impact and supporting emergency response efforts. This study presents a comparative analysis of deep learning architectures for natural disaster detection using satellite and aerial imagery. Four models were evaluated as baseline CNN, ResNet50, Faster-CNN, and Faster R-CNN with a ResNet50 backbone using standard classification metrics. The results demonstrate that deeper and more sophisticated models significantly enhance detection performance. While the baseline CNN achieved modest results with 85.3% accuracy, integrating residual learning in ResNet50 improved accuracy to 92.7%. Region-based models further boosted performance, with Faster-CNN and Faster R-CNN attaining 95.1% and 97.1% accuracy, respectively. The superior performance of the Faster R-CNN with ResNet50 highlights its robustness and suitability for real-time disaster monitoring, offering a scalable and reliable solution for operational deployment in disaster management systems.**

***Keywords*—Natural disasters detection; satellite imagery; convolutional neural networks (CNN); transformers; deep learning; ResNet50; proactive monitoring; faster R-CNN; disaster prevention; computer vision**

I. Introduction

Natural disasters are becoming more frequent and severe, underlining the urgent need for enhanced detection and management strategies. Traditional approaches often depend on manual analysis of aerial images, a process that can be both time-consuming and prone to human error. However, recent advancements in deep learning have opened new avenues for automating this task. Natural disasters (NDs), which include volcanic eruptions, earthquakes, landslides, floods, tsunamis, hurricanes, tornadoes, droughts, wildfires, and severe temperatures, pose enormous dangers to human life and infrastructure [1,2]. Throughout history, these events have caused widespread devastation and substantial loss of life. From earthquakes and volcanic eruptions to hurricanes, floods, and wildfires, natural disasters have profoundly affected both populations and the environment. Among the deadliest recorded were the 1931 China floods, triggered by intense rainfall and river overflows, which inundated over 50,000 square miles and claimed an estimated 3.7 million lives. Other significant natural disasters in history include the 1883 eruption of Krakatoa in Indonesia, which caused over 36,000 deaths and generated a massive tsunami that damaged coastal communities across the region [3]. The 1985 Mexico City earthquake, with a magnitude of 8.0, caused extensive damage to the city's infrastructure, resulting in more than 10,000 fatalities and leaving over 100,000 people homeless [4]. In recent years, natural disasters have continued to profoundly impact communities worldwide [5, 6]. For example, the 2011 Tohoku earthquake and tsunami in Japan claimed over 15,000 lives and severely damaged the country's infrastructure. In 2017, Hurricane Maria devastated Puerto Rico, causing widespread destruction and leaving most of the island without power for months [7]. Many communities face long-term struggles to rebuild their lives after such events, as the consequences of natural disasters can last for years or even decades, including the loss of homes and infrastructure, displacement of populations, and destruction of ecosystems.

Climate change has intensified the frequency and severity of these catastrophes, making it increasingly critical to develop strategies to mitigate their impacts [8]. In summary, natural disasters have historically caused tremendous harm to both communities and the environment. While technological advancements and improvements in disaster preparedness such as early flood detection systems, predictive analytics, machine learning, remote sensing, geographical information systems (GIS), satellite navigation, drone technology, cloud computing, enhanced communication and information sharing, resilient infrastructure, and early warning systems [9–12] have helped lessen their effects, natural disasters continue to pose serious challenges globally. Therefore, sustained investment in disaster preparedness and response remains essential to minimizing the impact on human lives, infrastructure, and ecosystems.

Fortunately, technological progress has greatly enhanced our ability to predict, respond to, and recover from these catastrophic events [13]. Technologies like remote sensing, radar, and satellite imaging provide diverse tools for monitoring the Earth's surface and atmosphere from a distance. Advances in these fields now allow for real-time tracking of natural disasters such as volcanic eruptions, wildfires, and landslides. By analyzing data collected through remote sensing systems, we can improve our understanding of the scale and intensity of these events, which supports the development of effective evacuation plans and recovery efforts. Leveraging this information is a key strategy for strengthening disaster preparedness. Additionally, technologies such as the Internet of Things (IoT), smartphones, and social media play a crucial role in the rapid collection and dissemination of information during

natural disasters [14–21]. IoT devices, such as weather monitoring instruments and seismic sensors, provide real-time data that help emergency responders make quick, well-informed decisions. Likewise, smartphones and social media platforms serve as effective channels for alerting people in affected areas, delivering vital updates on evacuation routes, shelter availability, and other essential resources.

This study aims to thoroughly investigate how various technologies contribute to natural disaster management. It explores the roles of Remote Sensing, Radar, Satellite Imaging, Autonomous Robots and Drones, IoT, Smartphones, and social media in responding to events like earthquakes, volcanic eruptions, landslides, and tsunamis. These technologies have been key in safeguarding lives, enabling coordinated rescue operations, and reducing damage to infrastructure.

The structure of this study is as follows: Section II presents background information on natural disasters and their environmental impact. Section III offers a comprehensive description of the dataset utilized. Section IV outlines the methodology and the proposed model. Section V discusses the evaluation results, and finally, Section VI provides the conclusion.

II. BACKGROUND

In recent years, the integration of satellite imagery with machine learning (ML) and deep learning methods has seen widespread use in various disaster management applications.

a) Earthquakes and volcanic eruptions: Earthquakes are among the most severe natural disasters, triggered by the sudden movement of tectonic plates beneath the Earth's surface, resulting in ground shaking and displacement [22]. These seismic events can cause extensive damage to infrastructure such as buildings, roads, and homes, and often lead to significant loss of life. The intensity of an earthquake is measured on the Richter Scale, which ranges from 1 to 10, with higher values indicating stronger quakes. According to the United States Geological Survey (USGS), approximately 20,000 earthquakes occur globally each year, with around 16 reaching a magnitude of 7.0 or above [23]. Earthquakes can also initiate secondary disasters such as landslides, tsunamis, and volcanic eruptions, increasing their overall impact (USGS, 2021) [24].

Volcanic eruptions represent another dangerous type of natural disaster. These occur when magma from beneath the Earth's surface rises and is expelled, often due to tectonic activity. Eruptions can result in pyroclastic flows, ash falls, and lava flows, all of which can cause widespread destruction and loss of life. The Volcanic Explosivity Index (VEI) is used to quantify the magnitude of an eruption, considering the volume of ash, lava, and other materials released during the event (USGS, 2021) [25]. The consequences of natural disasters such as earthquakes and volcanic eruptions often extend far beyond their initial impact. Earthquakes can reshape landscapes, redirect rivers and streams, and thereby alter ecosystems and wildlife habitats. Similarly, volcanic eruptions that release significant quantities of ash and gases into the atmosphere can influence global weather patterns and air quality [26].

Vulnerable communities exposed to risks from earthquakes, quarry blasts, and volcanic activity can significantly reduce their vulnerability through proactive planning and mitigation strategies [27, 28]. The loss of lives and property can be minimized when individuals and authorities establish well-structured emergency plans, including evacuation routes, in advance. Furthermore, governments play a crucial role by enforcing resilient building codes and investing in infrastructure that is capable of withstanding seismic and volcanic events.

Public education is also a vital component in reducing disaster impact. It provides people with accurate knowledge and preparedness about resources, such as how to respond during an earthquake or volcanic eruption empowers communities to act swiftly and effectively in times of crisis. Raising awareness and building local capacity helps ensure a more resilient response to future natural disasters.

Ultimately, earthquakes and volcanic eruptions are devastating events with the potential to cause long-lasting harm to both human populations and the environment. Mitigating their effects requires a combination of infrastructure investment, disaster planning, and community education. In alignment with these goals, this research centers on enhancing the performance of neural networks for detecting natural disasters, particularly under constraints like limited labeled datasets and low-power hardware environments. The optimized Faster R-CNN model developed through this work is tailored for real-time implementation on lightweight platforms such as smartphones, satellites, and high-altitude weather balloons, making it a scalable and cost-effective solution for broad disaster monitoring and response.

b) Landslides and tsunamis: Landslides and tsunamis are among the most devastating natural disasters, capable of causing significant destruction and loss of life. Landslides typically occur due to the instability of soil and rock on slopes, often triggered by factors such as intense rainfall or seismic activity. The severity of a landslide is measured using the Landslide Hazard Scale, which considers its potential to damage infrastructure and endanger lives. Beyond their immediate destruction, landslides can have long-term environmental consequences, such as altering watercourse paths and destroying habitats.

Tsunamis, on the other hand, are massive ocean waves usually triggered by underwater earthquakes or other subsurface disturbances. These waves can cause widespread flooding and infrastructural devastation, especially in coastal regions. Their severity is measured by the Tsunami Magnitude Scale, which assesses their destructive potential based on wave height. Due to their ability to affect vast oceanic regions, tsunamis are particularly dangerous, illustrated by the 2004 Indian Ocean tsunami, which claimed over 200,000 lives across multiple countries [29]. Preparedness strategies can greatly reduce the impact of these disasters, as shown in Fig. 2. These include enforcing resilient construction standards, developing detailed emergency and evacuation plans, and implementing monitoring systems for early warning. Educating at-risk communities on recognizing early signs of landslides or tsunamis and how to respond appropriately is also crucial.

Technological solutions such as early warning systems have proven effective in reducing disaster impact [30–34]. Governments and organizations can further mitigate landslide risks through reforestation and improved land management, while coastal engineering solutions like seawalls and breakwaters help shield populations from tsunamis.

As illustrated in Fig. 1, the findings from related studies categorize the impact of various natural disasters across four critical domains: infrastructure damage, loss of life, environmental impact, and potential to trigger secondary disasters. Earthquakes and tsunamis score the highest in infrastructure damage and fatalities, both rated at 5, highlighting their extreme destructive capabilities. Volcanic eruptions follow with slightly lower but still severe impacts on infrastructure and human life, and they are rated highest for environmental damage with a score of 5 attributable to lava flows, ash dispersion, and toxic emissions. Landslides rank comparatively lower across all categories, particularly in their ability to cause secondary disasters, where they score 2.

Previous studies on disaster severity scales consistently highlight the varying degrees of impact that different natural disasters impose on communities and environments. Research emphasizes that earthquakes and tsunamis rank among the most severe disasters. Volcanic eruptions, while sometimes less destructive to build environments, are noted for their significant environmental impacts, including long-term effects on air quality and ecosystems. Landslides typically score lower on severity scales, as their effects tend to be more localized and less likely to trigger widespread secondary disasters. These findings highlight the need for disaster-specific risk assessments and preparedness strategies.

Table I provides a summary of methodologies, datasets, strengths, limitations, and findings related to earthquake, volcanic, landslide, and tsunami studies.

Their methodology includes reviewing impact metrics such as the Richter Scale and Volcanic Explosivity Index. The organization also emphasizes disaster preparedness through policy recommendations that include education, resilient infrastructure, and evacuation planning. Although the studies are largely descriptive and lack predictive modeling or modern machine learning integration, they highlight the importance of early warning systems and community resilience in reducing disaster impact.

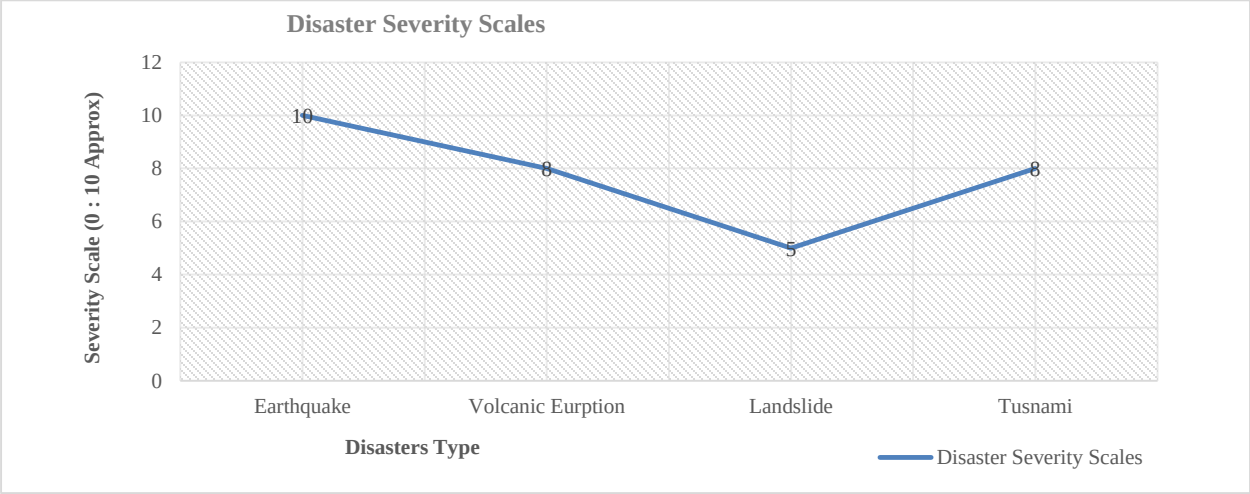


Fig. 1. Disaster severity scales.

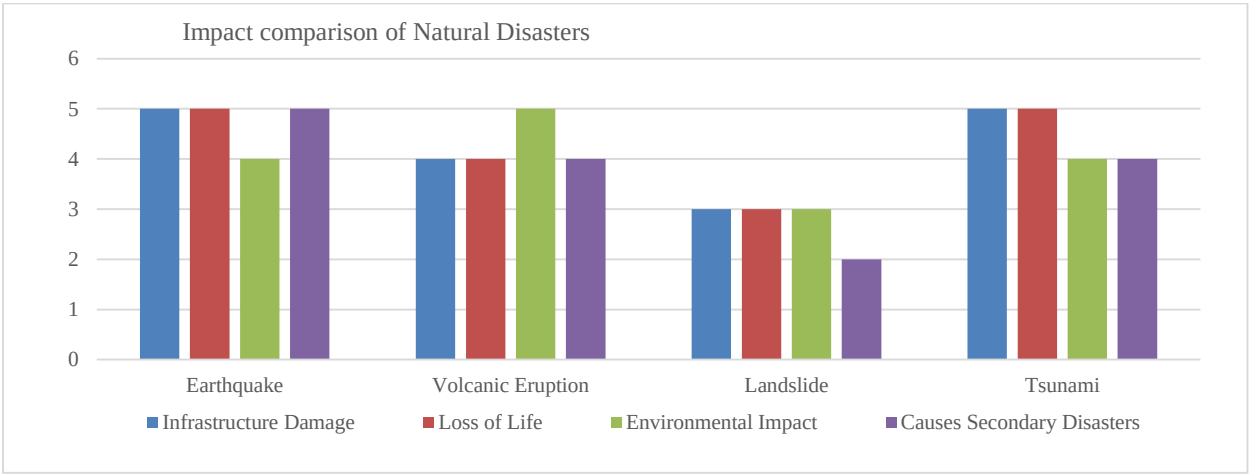


Fig. 2. Impact comparison of natural disasters.

TABLE I SUMMARY OF REVIEWED LITERATURE ON NATURAL DISASTERS INCLUDING METHODOLOGIES, DATASETS USED, STRENGTHS, LIMITATIONS, AND KEY FINDINGS (ADAPTED FROM USGS SOURCES [22–34])

Authors	Methodology	Dataset Used	Strengths	Limitations	Results
USGS [22-26]	Review and statistical analysis of natural disaster events including earthquakes and volcanic eruptions	USGS earthquake and volcano datasets	Covers a wide range of disaster types and global data; provides detailed impact metrics (e.g., Richter, VEI)	Primarily descriptive; lacks predictive modeling or real-time response components	About 20,000 earthquakes/year; 16 with magnitude ≥7.0; volcanoes cause atmospheric impact and secondary disasters
USGS [27-28]	Policy recommendations and preventive strategies for earthquake and volcanic events	Preparedness guidelines, infrastructure resilience studies	Emphasizes disaster risk reduction through education and infrastructure	Limited in terms of implementation analysis or specific case studies	Recommendations for community preparedness, evacuation plans, and resilient construction
USGS [29-34]	Descriptive analysis and historical review of landslides and tsunamis	Historical tsunami data (e.g., 2004 Indian Ocean event), Landslide Hazard Scale	Combines geological triggers with societal impact; includes early warning systems	Descriptive in nature, lacks integration with modern ML-based forecasting systems	Landslides disrupt habitats and waterways; tsunamis cause mass casualties and international devastation

III. DATASET

The dataset employed in this study is composed of satellite images capturing the aftermath of landslides and floods across selected regions in Japan and Thailand. The images were carefully selected to encompass a wide range of environmental and urban settings, enhancing the model's ability to generalize effectively. This dataset forms the foundation for training and evaluating the proposed deep learning model aimed at accurately classifying and identifying disaster-affected areas. In total, approximately 10,000 images were compiled, with an equal distribution between landslide, earthquake and flood scenarios to ensure a balanced learning dataset.

A. Data Collection

The images were acquired through publicly accessible geospatial platforms, primarily Google Earth, which offers high-resolution satellite images of global terrain. The selection process involved identifying locations in Japan and Thailand historically affected by floods, Earthquakes and landslides, followed by extracting relevant imagery from these regions. Care was taken to exclude poor-quality or ambiguous images to maintain the reliability of the dataset. This manual curation ensured that the resulting image collection was both representative and robust for the purposes of training a convolutional neural network (CNN) for disaster classification.

B. Annotations

To prepare the dataset for supervised learning, each image was annotated with bounding boxes highlighting the regions impacted by either landslides or floods. These annotations were carefully reviewed to ensure precision in identifying disaster-affected zones. Each bounding box was labeled with a corresponding class identifier, such as "landslide" or "flood", enabling the model to learn category-specific features. The annotations were stored in JSON format, which ensures compatibility with common deep learning frameworks and streamlines their integration into the training and evaluation pipeline.

IV. METHODOLOGY

A. Data Processing

Data processing plays a foundational role in the success of any deep learning pipeline, especially when dealing with image-based tasks such as Natural Disasters detection. The process begins with normalization, a critical step that involves scaling the raw pixel intensity values of each image to a standardized range typically [0, 1]. This transformation ensures uniformity across all input data, mitigating the effects of varying image brightness or contrast that could otherwise bias the learning process. Normalization also improves the stability and speed of training by helping the optimizer navigate the loss landscape more effectively.

Following normalization, the dataset undergoes a systematic splitting procedure to facilitate robust model training and evaluation. Using stratified sampling, the dataset is divided into three distinct subsets: 70% of the images are assigned to the training set, 15% to the validation set, and the remaining 15% to the testing set. Stratification ensures that the class distribution (i.e., Flood, Earthquakes and Landslide images) remains consistent across all subsets, which is essential for obtaining an unbiased performance evaluation and preventing skewed learning outcomes.

To further prepare the images for ingestion into the deep learning model, a series of image processing techniques are applied to enhance both the quality and diversity of the dataset. First, all images are resized to a fixed dimension of 224×224 pixels, which aligns with the input layer requirements of architectures of ResNet50 and ConvNeXt-Small. This uniform resizing ensures that the model receives input in a consistent shape, reducing computational complexity and memory overhead while preserving key spatial information.

Next, noise reduction is performed using Gaussian filtering, a technique that smooths out high-frequency noise that may be present due to atmospheric distortions, sensor limitations, or compression artifacts in the satellite imagery. By applying a Gaussian kernel, subtle variations in pixel values are averaged in a way that retains important edges and features while suppressing random noise, thereby improving feature extraction in the early convolutional layers of the model.

To combat overfitting and improve the model's generalization to unseen data, data augmentation techniques are extensively employed. These include random rotations, horizontal and vertical flips, and color jittering, each of which introduces controlled variability into the training data without altering its semantic content. For instance, rotating an image by a few degrees or flipping it horizontally simulates different

orientations of wildfires captured by satellites, making the model invariant to spatial transformations. Similarly, jittering the color channels mimics variations in lighting conditions and image sensors, thereby improving robustness.

Lastly, image enhancement is applied using histogram equalization, a method that redistributes the intensity values in an image to span a wider range of contrasts. This process makes subtle features more visually prominent, such as smoke patterns, vegetation boundaries, or burn scars, which might otherwise be indistinguishable in low-contrast areas. Enhanced contrast not only benefits visual interpretation but also improves the sensitivity of the model to faint or ambiguous features that are crucial for early wildfire detection.

Taken together, this comprehensive data and image preprocessing strategy ensures that the input images fed into the deep learning model are clean, balanced, diverse, and information-rich. It significantly enhances the model's ability to learn meaningful patterns, generalize real-world conditions, and make reliable predictions in operational wildfire monitoring systems. Fig. 3 shows the data processing flowchart.

B. Model Architecture

The proposed model for object detection in Natural disaster imagery leverages the power of Faster R-CNN, a state-of-the-art deep learning framework designed for efficient and accurate object localization and classification. Faster R-CNN integrates two key components: a Region Proposal Network (RPN) and a Fast R-CNN detection network.

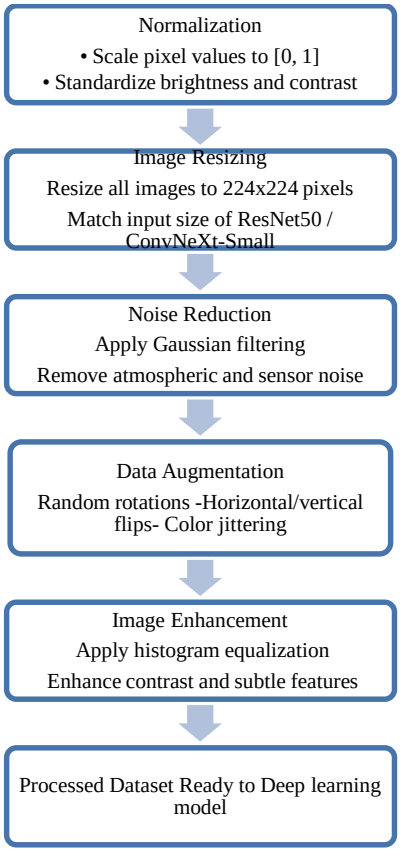


Fig. 3. Data processing.

1) Backbone ResNet50 as feature extractor. ResNet50, a deep convolutional neural network with 50 layers, has become a foundational architecture in image recognition and computer vision due to its innovative residual learning approach. A major challenge when training very deep networks is the vanishing gradient problem, where gradients diminish as they are propagated backward through numerous layers during training [35]. ResNet50 addresses this issue by incorporating skip connections, or residual connections, which allow gradients to flow more freely by bypassing one or more layers. These connections enable the network to learn identity mappings, helping the optimization process converge more quickly and reliably despite the increased network depth. With its 50 layers, ResNet50 effectively captures hierarchical representations of input images, learning simple features such as edges and textures in the initial layers and progressively more complex, abstract patterns like shapes, objects, and spatial relationships in the deeper layers. This capability is particularly advantageous for satellite image analysis, where features can vary significantly in scale, rotation, texture, and context. Whether distinguishing collapsed buildings, flooded areas, or wildfire scars, ResNet50's rich and multiscale feature representations allow for more accurate identification of disaster-affected regions. In this system, ResNet50 acts as the feature extraction backbone of the detection pipeline, converting raw satellite imagery into a feature map that encodes spatial and contextual information crucial for downstream object detection.

2) Region proposal network. The Region Proposal Network (RPN) scans the image to generate potential regions of interest (ROIs) that are likely to contain objects. Meanwhile, the Relationship Proposal Network (Rel-PN) is a deep learning architecture designed to improve scene understanding by effectively detecting meaningful relationships between objects within an image [36,37]. The process starts with an input image that is passed through a Backbone Convolutional Neural Network (CNN) to extract detailed feature maps. These feature maps are then analyzed by the RPN, which proposes candidate regions in the image that are likely to contain objects.

Instead of examining all possible object pairs (which grow quadratically), the Rel-PN intelligently selects only the most promising object pairs that are likely to be involved in a relationship. This selection is based on learned spatial and contextual cues. These relevant object pairs are then passed to a Relationship Classifier, which determines the specific type of interaction or relationship between them such as "person riding horse" or "dog next to man" [38].

3) Fast R-CNN detection network. The Fast R-CNN component refines the proposed regions by classifying them and adjusting their bounding boxes. Fast R-CNN is a deep learning architecture designed for efficient and precise object detection in images. The process begins with an input image that passes through multiple convolutional layers to extract feature maps. These feature maps are then processed by the Region of Interest (RoI) Pooling layer, which selects regions

likely to contain objects [39]. The selected regions are further passed through fully connected layers to produce high-level features. Finally, the network generates two outputs: a classification output that predicts the object classes, and a bounding box regression output that fine-tunes the coordinates of the bounding boxes around the detected objects.

This architecture greatly enhances both the speed and accuracy of object detection by sharing convolutional computations and utilizing RoI pooling to focus on relevant image regions. Its unified design supports end-to-end training and significantly speeds up the detection process compared to earlier two-stage detectors.

At the core of this architecture lies the ResNet50 backbone, a deep convolutional neural network renowned for its powerful feature extraction. ResNet50 employs residual learning through skip connections, allowing gradients to bypass one or more layers during backpropagation. This approach addresses the vanishing gradient problem, enabling effective training of very deep networks. Consequently, ResNet50 excels at learning rich, hierarchical visual representations, which are essential for accurately detecting wildfire-related features in satellite or aerial imagery.

The training procedure was meticulously crafted to maximize the model's performance on the prepared dataset. The network was trained over 10 epochs using stochastic gradient descent (SGD) with a momentum term to speed up convergence and minimize oscillations. A learning rate of 0.005 was set to regulate the step size during weight updates, while a weight decay of 0.0005 served as a regularization method to prevent overfitting. These training parameters collectively helped develop a highly effective detection model, capable of accurately identifying fire-prone areas within complex visual scenes.

Algorithm I: Faster R-CNN with ResNet50 Algorithm for Natural Disasters Detection

```
Input: Image I
Output: List of detected objects with class labels and bounding boxes
# Step 1: Preprocessing
I_preprocessed = Preprocess(I)  # resize, normalize, augment (rotation, flip, jitter)

# Step 2: Feature Extraction using ResNet50
Feature_Map = ResNet50_Backbone(I_preprocessed)
# Step 3: Region Proposal Network (RPN)
Anchors = GenerateAnchors(Feature_Map)
Objectness_Scores, BBox_Offsets = RPN(Feature_Map, Anchors)
# Step 4: Proposal Selection
Proposals = ApplyBoundingBoxRegression(Anchors, BBox_Offsets)
Proposals = FilterLowScores(Proposals, Objectness_Scores, threshold)
Proposals = NonMaximumSuppression(Proposals, iou_threshold)

# Step 5: RoI Pooling
Fixed_Size_Features = []
For each Proposal in Proposals:
  RoI_Feature = RoIAlign(Feature_Map, Proposal)
  Fixed_Size_Features.append(RoI_Feature)

# Step 6: Fast R-CNN Head
Class_Scores, Final_BBox_Offsets = FastRCNN(Fixed_Size_Features)
# Step 7: Bounding Box Refinement
Final_BBoxes = ApplyBoundingBoxRegression(Proposals, Final_BBox_Offsets)

# Step 8: Final Classification & Suppression
For each box in Final_BBoxes:
  if max(Class_Scores) > class_threshold:
    Keep box with label = argmax(Class_Scores)
  else:
    Discard box
Final_Detections = NonMaximumSuppression(Final_BBoxes, iou_threshold)
Return Final_Detections
```

In this pipeline, the Faster R-CNN algorithm (Algorithm I) functions as a two-stage object detection framework that combines high-accuracy classification with precise localization. Initially, input images are fed into a deep convolutional backbone ResNet50, which extracts hierarchical feature maps. These feature maps are then processed by the Region Proposal Network (RPN), which generates candidate object regions by assessing anchor boxes based on their objectiveness scores and refining their positions. The top proposals are further refined through Region of Interest (RoI) pooling and passed to the Fast R-CNN module, which classifies each region and adjusts its bounding box coordinates. During training, the model simultaneously optimizes classification loss (using cross-entropy) and localization loss (using Smooth L1), employing stochastic gradient descent (SGD) with momentum and weight decay for parameter updates. This architecture strikes an effective balance between speed and accuracy, making it well-suited for detecting complex patterns in high-resolution images, such as those captured by satellites. Fig. 4 presents the block diagram of proposed Faster R-CNN.

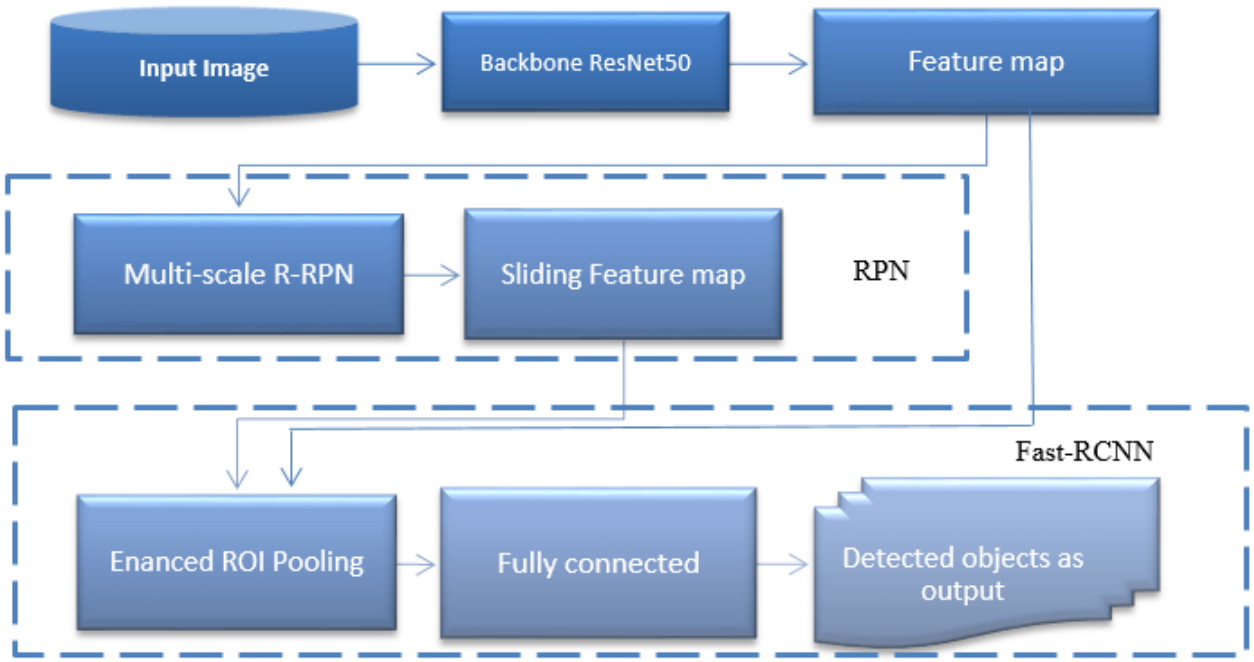


Fig. 4. Block diagram of proposed Faster R-CNN.

C. Training Phase: Loss Computation and Optimization

During the training phase of a two-stage object detection model like Faster R-CNN, the network learns to classify objects and accurately localize them within images. This is accomplished through a multi-task loss function that combines classification loss and localization loss. The model parameters are then optimized using backpropagation with Stochastic Gradient Descent (SGD), incorporating momentum and weight decay.

Classification Loss measures how closely the predicted class probabilities align with the true labels. Specifically, Cross-Entropy Loss is employed, with Softmax Cross-Entropy used for multi-class classification tasks. Given a predicted probability distribution p and the ground truth class c, the classification loss is:

Where, $L_{cls}=-\log(p_c)$ (1)

p_c is the predicted probability for the correct class c.

Localization Loss L_{loc} measures how precisely the predicted bounding boxes align with the ground truth boxes. Typically, $Smooth_{L1}$ Loss is used for this purpose, as it is less sensitive to outliers compared to the standard L2 loss. For each coordinate x of the bounding box, the $Smooth_{L1}$ loss is defined as follows:

$$Smooth_{L1}(x) = \begin{cases} 0.5|x| & < 1 \\ |x| - 0.5 & otherwise \end{cases}$$

Therefore, the overall localization loss between the predicted bounding box t and the ground truth t* is defined as:

$$L_{loc}= \sum_{i\in\{x,y,w,h\}} SmoothL1(t_i-t_i^*) \quad (2)$$

The total loss function is then:

$$L_{Total}=L_{cls}+\lambda\cdot L_{loc} \quad (3)$$

where, λ is a balancing factor (usually set to 1), adjusting the relative importance of classification and localization.

D. Backpropagation and Optimization

After calculating the loss, the gradients of the loss with respect to the model's parameters are determined through backpropagation. The optimizer used is Stochastic Gradient Descent (SGD) with Momentum, a technique that enhances gradient descent by speeding up convergence, particularly in situations involving high curvature or noisy gradients. The weight update rule is:

$$Vt+1=\mu vt-\eta\cdot\nabla L \quad (4)$$

$$\theta t+1= \theta t +Vt+1 \quad (5)$$

Here, θ represents the model weights, Vt denotes the velocity (momentum term), μ is the momentum coefficient set to 0.9, and η is the learning rate, which is 0.005. ∇L represents the gradient of the loss with respect to the parameters. Weight decay helps reduce overfitting by discouraging large weight values. It alters the parameter update rule as follows:

$$\theta_{t+1}=\theta_t+V_{t+1}-\eta\cdot\lambda_w\cdot\theta_t \quad (6)$$

where, λ_w is the weight decay factor (0.0005).

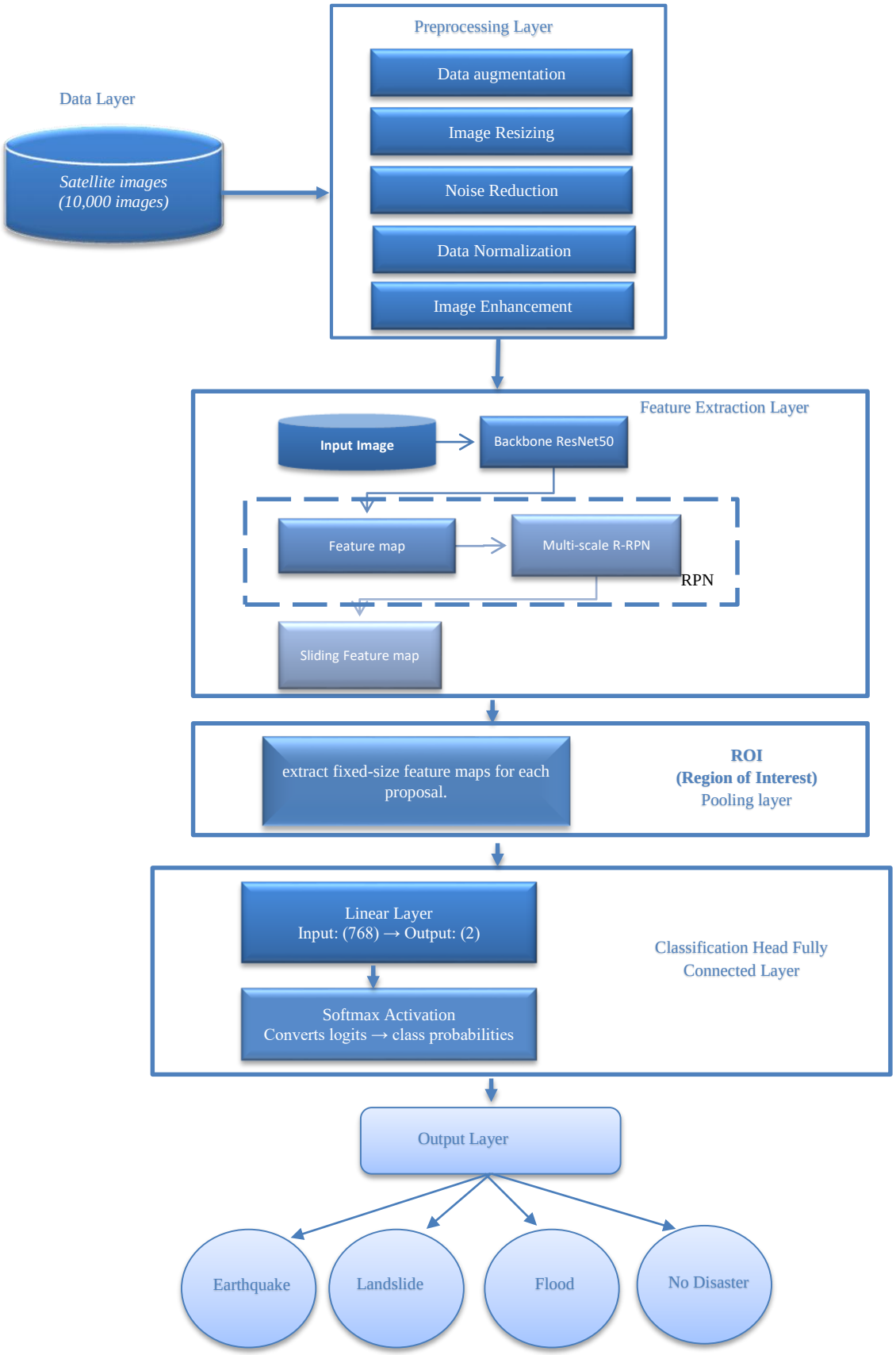


Fig. 5. Proposed model.

V. Results

The performance of the proposed Natural Disasters detection models (Fig. 5) was comprehensively evaluated using standard classification metrics accuracy, precision, recall, and F1-score across four different architectures: CNN, ResNet50, Faster-CNN, and Faster R-CNN with a ResNet50 backbone (see Fig. 6). The baseline CNN model attained an accuracy of 85.3%, with precision at 82.1%, recall at 78.5%, and an F1-score of 80.2%. Although this model served as a reasonable starting point, it exhibited limitations in generalization, likely due to its shallow depth and limited feature extraction capabilities. Significant improvement was achieved with ResNet50, which utilizes residual connections to facilitate learning in deeper networks. This model reached 92.7% accuracy, 90.4% precision, 89.3% recall, and an F1-score of 89.8%, demonstrating its superior ability to capture complex patterns in wildfire imagery. Further gains were observed with the Faster-CNN model, which incorporates region-based object detection. It attained 95.1% accuracy, 94.0% precision, 92.8% recall, and an F1-score of 93.4%, highlighting the advantage of combining detection and classification techniques. The highest performance was achieved by the Faster R-CNN architecture with a ResNet50 backbone, delivering outstanding results with 97.1% accuracy, 95.8% precision, 96.2% recall, and a well-balanced F1-score of 96.0%.

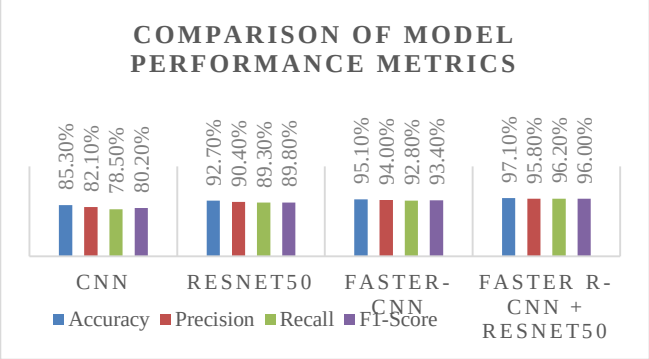


Fig. 6. Comparison of model performance metrics.

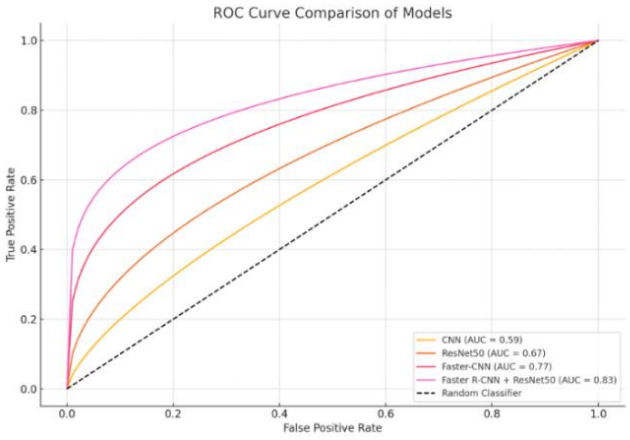


Fig. 7. ROC Curve comparison of models.

These metrics underscore the model's ability to detect Natural disasters regions with high reliability and minimal false predictions. ROC curve analysis further validated these findings. The AUC scores increased progressively across models, with CNN yielding the lowest AUC at 0.59, followed by ResNet50 at 0.67, Faster-CNN at 0.77, and the highest AUC of 0.83 achieved by the Faster R-CNN + ResNet50 model (see Fig. 7).

These results confirm the superior classification capability of the proposed framework, demonstrating robust performance and high discriminative power across all classes. This strong performance highlights the model's readiness for real-world deployment in early wildfire detection systems. Table II presents the proposed model results.

TABLE II Proposed Model Results

Model	Accuracy	Precision	Recall	F1-Score
CNN	85.3%	82.1%	78.5%	80.2%
ResNet50	92.7%	90.4%	89.3%	89.8%
Faster-CNN	95.1%	94.0%	92.8%	93.4%
Faster R-CNN + ResNet50	**97.1%**	**95.8%**	**96.2%**	**96.0%**

VI. Conclusion

This study systematically evaluated the effectiveness of deep learning architectures for natural disaster detection using remote sensing imagery. The baseline CNN, although functional, struggled with generalization due to its limited depth. Incorporating residual connections through ResNet50 significantly enhanced feature representation and classification performance. Further integration of region-based object detection in the Faster-CNN model yielded additional improvements. The highest-performing model, Faster R-CNN with a ResNet50 backbone, achieved exceptional accuracy, precision, recall, and F1-score, demonstrating its potential for real-time disaster detection. These results underscore the value of combining deep residual learning with object detection mechanisms for robust and scalable disaster response solutions. Future work may focus on expanding dataset diversity, incorporating multimodal data, and optimizing inference speed for deployment in real-time early warning systems. Future work will focus on expanding the dataset with additional disaster types and temporal satellite data to analyze changes over time. Moreover, integrating multi-spectral satellite data could further enhance feature discrimination, improving detection in low-visibility or cloud-covered scenarios.

References

[1] M.S. Abdalzaher, H.A. Elsayed, M.M. Fouda, M.M. Salim, Employing machine learning and iot for earthquake early warning system in smart cities, Energies 16 (1) (2023) 495.

[2] S.S. Moustafa, M.S. Abdalzaher, H. Abdelhafiez, Seismo-lineaments in Egypt: analysis and implications for active tectonic structures and earthquake magnitudes, Rem. Sens. 14 (23) (2022) 6151.

[3] S. Self, M.R. Rampino, The 1883 eruption of krakatau, Nature 294 (1981) 699–704.

[4] E. Rosenblueth, R. Meli, The 1985 Mexico earthquake, Concr. Int. 8 (5) (1986) 23–34.

[5] M. Elhadidy, M.S. Abdalzaher, H. Gaber, Up-to-date psha along the gulf of aqabadead sea transform fault, Soil Dynam. Earthq. Eng. 148 (2021) 106835.

[6] M.S. Abdalzaher, M. El-Hadidy, H. Gaber, A. Badawy, Seismic hazard maps of Egypt based on spatially smoothed seismicity model and recent seismotectonic models, J. Afr. Earth Sci. 170 (2020) 103894.

[7] M.S. Abdalzaher, M.S. Soliman, S.M. El-Hady, A. Benslimane, M. Elwekeil, A deep learning model for earthquake parameters observation in iot system-based earthquake early warning, IEEE Internet Things J. 9 (11) (2022) 8412–8424.

[8] R. Khurana, D. Mugabe, X.L. Etienne, Climate change, natural disasters, and institutional integrity, World Dev. 157 (2022) 105931.

[9] M.S. Abdalzaher, M. Sami Soliman, S.M. El-Hady, Seismic intensity estimation for earthquake early warning using optimized machine learning model, IEEE Trans. Geosci. Rem. Sens. 61 (2023), 1–1.

[10] M.M. Salim, H.A.E.A. Elsayed, M.S. Abdalzaher, A survey on essential challenges in relay-aided d2d communication for next-generation cellular networks, J. Netw. Comput. Appl. (2023) 103657 [Online]. Available: https://www.sciencedirect.co m/science/article/pii/S1084804523000760.

[11] M.M. Salim, H.A. Elsayed, M.A. Elaziz, M.M. Fouda, M.S. Abdalzaher, An optimal balanced energy harvesting algorithm for maximizing two-way relaying d2d communication data rate, IEEE Access 10 (2022), 114 178–114 191.

[12] M.M. Salim, H.A. Elsayed, M.S. Abdalzaher, M.M. Fouda, Rf energy harvesting dependency for power optimized two-way relaying d2d communication, in: 2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS), 2022, pp. 297–303.

[13] M.S. Abdalzaher, H.A. Elsayed, Employing data communication networks for managing safer evacuation during earthquake disaster, Simulat. Model. Pract. Theor. 94 (2019) 379–394.

[14] M.S. Abdalzaher, K. Seddik, O. Muta, Using repeated game for maximizing high priority data trustworthiness in wireless sensor networks, in: 2017 IEEE Symposium on Computers and Communications (ISCC), 2017, pp. 552–557.

[15] M.S. Abdalzaher, O. Muta, A game-theoretic approach for enhancing security and data trustworthiness in iot applications, IEEE Internet Things J. 7 (11) (2020) 11250–11261.

[16] M.S. Abdalzaher, K. Seddik, O. Muta, Using stackelberg game to enhance cognitive radio sensor networks security, IET Commun. 11 (9) (2017) 1503–1511.

[17] M.S. Abdalzaher, K. Seddik, M. Elsabrouty, O. Muta, H. Furukawa, A. AbdelRahman, Game theory meets wireless sensor networks security requirements and threats mitigation: a survey, Sensors 16 (7) (2016) 1003.

[18] M.S. Abdalzaher, K. Seddik, O. Muta, An effective stackelberg game for highassurance of data trustworthiness in wsns, in: 2017 IEEE Symposium on Computers and Communications (ISCC), 2017, pp. 1257–1262.

[19] M.S. Abdalzaher, K. Seddik, O. Muta, A. Abdelrahman, Using stackelberg game to enhance node protection in wsns, in: 2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC), 2016, pp. 853–856. [20] M.S. Abdalzaher, O. Muta, K. Seddik, A. Abdel-Rahman,

[20] H. Furukawa, B-18-40 a simplified stackelberg game approach for securing data trustworthiness in wireless sensor networks, in: 2016 IEICE General Conference. Plus 0.5em Minus 0.4emIEICE, 2016, p. 538.

[21] M.S. Abdalzaher, O. Muta, Employing game theory and tdma protocol to enhance security and manage power consumption in wsns-based cognitive radio, IEEE Access 7 (2019), 132 923–132 936.

[22] S.A. Mahin, Lessons from damage to steel buildings during the northridge earthquake, Eng. Struct. 20 (4–6) (1998) 261–270.

[23] L. Highland, P.T. Bobrowsky, et al., The Landslide Handbook: a Guide to Understanding Landslides. Plus 0.5em Minus 0.4emUS Geological Survey Reston, 2008. VA, USA.

[24] C.M. Scheip, K.W. Wegmann, Hazmapper: a global open-source natural hazard mapping application in google earth engine, Nat. Hazards Earth Syst. Sci. 21 (5) (2021) 1495–1511.

[25] K.M. Cooper, K. Anderson, K. Cashman, M. Coombs, H. Dietterich, T. Fischer, B. Houghton, I. Johanson, K.J. Lynn, M. Manga, et al., Coordinating science during an eruption: lessons from the 2020–2021 k i lauea volcanic eruption,, Bull. Volcanol. 85 (5) (2023) 1–13.

[26] E. Ghamry, E.K. Mohamed, M.S. Abdalzaher, M. Elwekeil, D. Marchetti, A. De Santis, M. Hegy, A. Yoshikawa, A. Fathy, Integrating pre-earthquake signatures from different precursor tools, IEEE Access 9 (2021), 33 268–33 283.

[27] M.S. Abdalzaher, S.S.R. Moustafa, M. Abd-Elnaby, M. Elwekeil, Comparative performance assessments of machine-learning methods for artificial seismic sources discrimination, IEEE Access 9 (2021), 65 524–65 535.

[28] S.S.R. Moustafa, M.S. Abdalzaher, M.H. Yassien, T. Wang, M. Elwekeil, H.E.A. Hafiez, Development of an optimized regression model to predict blastdriven ground vibrations, IEEE Access 9 (2021), 31 826–31 841.

[29] G. Bankoff, J. Christensen, bordering on danger: an introduction, Nat. hazards and peoples Indian Ocean world: Bordering on danger (2016) 1–30.

[30] M.S. Abdalzaher, S.S.R. Moustafa, H.A. Hafiez, W.F. Ahmed, An optimized learning model augment analyst decisions for seismic source discrimination, IEEE Trans. Geosci. Rem. Sens. 60 (2022) 1–12.

[31] O. Hamdy, H. Gaber, M.S. Abdalzaher, M. Elhadidy, Identifying exposure of urban area to certain seismic hazard using machine learning and gis: a case study of greater cairo, Sustainability 14 (17) (2022) 10722.

[32] S.S.R. Moustafa, M.S. Abdalzaher, M. Naeem, M.M. Fouda, Seismic hazard and site suitability evaluation based on multicriteria decision analysis, IEEE Access 10 (2022) 69 511–569 530.

[33] M.S. Abdalzaher, H.A. Elsayed, M.M. Fouda, Employing remote sensing, data communication networks, ai, and optimization methodologies in seismology, IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 15 (2022) 9417–9438.

[34] S.S.R. Moustafa, M.S. Abdalzaher, F. Khan, M. Metwaly, E.A. Elawadi, N.S. Al-Arifi, A quantitative site-specific classification approach based on affinity propagation.

[35] Ali, R. R., Yaacob, N. M., Alqaryouti, M. H., Sadeq, A. E., Doheir, M., Iqtait, M., ... & Yaacob, S. S. (2025). Learning Architecture for Brain Tumor Classification Based on Deep Convolutional Neural Network: Classic and ResNet50. Diagnostics, 15(5), 624.

[36] Dong, R., Xia, J., Zhao, J., & Hong, L. (2025). CL-PSDD: Contrastive Learning for Adaptive Generalized Pavement Surface Distress Detection. IEEE Transactions on Intelligent Transportation Systems.

[37] Lin, H., Li, N., Yao, P., Dong, K., Guo, Y., Hong, D., ... & Wen, C. (2025). Generalization-Enhanced Few-Shot Object Detection in Remote Sensing. IEEE Transactions on Circuits and Systems for Video Technology.

[38] Zhang, J., Elhoseiny, M., Cohen, S., Chang, W., & Elgammal, A. (2017). Relationship proposal networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5678-5686).

[39] He, Z., & He, Y. (2025). AS-Faster-RCNN: An Improved Object Detection Algorithm for Airport Scene Based on Faster R-CNN. IEEE Access.

Fusion of CNN and Transformer Architectures for Proactive Wildfire Detection in Satellite Imagery

Shereen Essam Elbohy, Mona M. Nasr, Farid Ali Mousa
Faculty of Computers and Artificial Intelligence, Helwan University, Egypt[1]
Associate Professor in Faculty of Computers and Artificial Intelligence, Helwan University, Egypt[2]
Associate Professor in Computer Science, Beni Suef University, Egypt[3]

***Abstract*—Wildfires pose a significant threat to ecosystems, human settlements, and air quality, necessitating advanced detection and mitigation strategies. Traditional wildfire detection methods often rely on manual observation and conventional machine learning approaches, which may lack efficiency and accuracy. This study proposes a novel deep learning model based on the ConvNeXt-Small architecture, a hybrid design that fuses the strengths of Convolutional Neural Networks (CNNs) and Transformer-inspired mechanisms, enabling more comprehensive analysis of wildfire patterns in satellite imagery. The model was trained using the Adam optimizer, which provides efficient convergence and adaptive learning. The dataset used consists of real-world satellite images collected from wildfire-affected regions in Canada, covering various geographic and seasonal conditions to reflect real environmental diversity. The results underscore the potential of ConvNeXt-based architecture for real-time, high-precision wildfire detection, offering a powerful tool for early intervention, disaster mitigation, and environmental monitoring efforts.**

***Keywords*—Wildfire detection; satellite imagery; convolutional neural networks (CNN); transformers; deep learning; hybrid model; proactive monitoring; remote sensing; disaster prevention; computer vision**

I. Introduction

Rapid detection of sudden-onset disasters, such as earthquakes, flash floods, and emergencies like road accidents, is critical for effective response by emergency organizations. However, gathering information during these events is often labor-intensive and expensive, requiring manual data processing and expert analysis. To reduce these efforts, researchers have explored the use of computer vision techniques applied to satellite imagery, synthetic aperture radar, and other remote sensing data [1]. Despite these advances, such methods remain costly to deploy and lack the robustness needed to reliably capture relevant data in time-sensitive situations. Additionally, satellite imagery is often affected by noise from clouds and smoke common during hurricanes and wildfires and provides only an overhead perspective of disaster areas.

Forests play vital roles in our environment and daily lives by providing numerous resources. They are often referred to as the "lungs of the planet", as they purify the air by producing oxygen (O2) and absorbing carbon dioxide (CO2). Forests also serve as habitats for diverse animal species and act as natural windbreakers that protect agricultural crops. Moreover, they help purify water by filtering out pollutants [2]. Economically, forests contribute significantly by providing jobs and boosting national incomes.

In recent years, many forests and wildlands have been destroyed by wildfires, which are uncontrolled natural disasters posing serious threats to economies worldwide. Each year, millions of acres of land are lost, resulting in significant damage to human lives, vegetation, and forest resources [3]. Wildfires also adversely affect agriculture by drying out the soil and destroying crops near affected areas. These fires can start from natural causes like lightning or human activities such as discarded cigarettes and land clearing. Climate change, rising temperatures, and prolonged droughts have intensified wildfire frequency and severity, especially in fire-prone regions such as California, Australia, and the Mediterranean. The environmental impact includes reduced air quality from pollutant emissions and accelerated global warming. Wildfires disrupt ecosystems by destroying habitats, threatening biodiversity, and increasing soil erosion. However, in some ecosystems, wildfires play a natural role by clearing dead vegetation and promoting new plant growth. Early detection and prevention are essential to minimize wildfire damage [4]. Advances in satellite imagery, artificial intelligence, and deep learning now enable real-time wildfire monitoring and prediction. Natural disasters are inescapable and have serious consequences for economies, ecological systems and human life. They can cause building collapses, spread diseases, and devastate entire nations through events like tsunamis, earthquakes, and forest fires. For instance, seismic forces from earthquakes can lead to the collapse of millions of buildings [5]. Since the 1990s, a variety of machine learning algorithms have been used to anticipate wildfires. A recent study in Italy used the random forest approach for wildfire susceptibility mapping [6]. Floods are the most catastrophic natural calamity, causing severe damage to property, infrastructure, and human life. To map flood susceptibility, an ensemble machine learning strategy using random forest (RF), random subspace (RS), and support vector machine (SVM) approaches was used. [7]. Studies indicate that immediately following a disaster, social media platforms contain valuable information for disaster response, including reports of damage and urgent needs from affected individuals [8], [9]. Much of this information comes in the form of images and videos. Unlike traditional data sources such as satellite imagery, leveraging on-the-ground social media images for relief efforts has remained largely unexplored until recently, as demonstrated in our prior work [10]. This gap is mainly due to technical challenges. Automatically filtering relevant images is difficult because

social media streams are noisy, with a large proportion of posts unrelated to humanitarian needs. Additionally, effective filtering requires robust deep learning models trained on vast amounts of labeled data. However, assembling a large-scale, annotated dataset for incident recognition in real-world conditions is both time-consuming and expensive.

The structure of this study is as follows: Section II presents essential baseline knowledge of machine learning. Section III describes the dataset used in detail. Section IV introduces the proposed model. Section V discusses the evaluation results, and finally, Section VI concludes the study.

II. BACKGROUND

In recent years, the integration of satellite imagery with machine learning (ML) techniques has become increasingly prevalent in various disaster management applications. These include monitoring changes in critical infrastructure, evaluating the impact of natural disasters, and tracking vegetation dynamics [10–11]. Among the ML methods, deep learning, particularly convolutional neural networks (CNNs) has shown exceptional capability in accurately detecting and classifying patterns in satellite data.

A widely adopted approach for forest fire detection using deep learning involves training CNNs on large datasets of satellite images. These datasets are commonly obtained from sources such as NASA's Earth Observing System, ESA's Sentinel program, or commercial providers like Planet Labs [12]. To prepare these datasets, images are labeled to indicate the presence or absence of wildfire activity. Labeling can be done manually by experts or semi-automatically using pre-trained models. Following preprocessing and annotation, the dataset is used to train deep learning models such as CNNs or, in some cases, recurrent neural networks (RNNs), depending on the spatial and temporal complexity of the data.

Once trained, the model is validated on a separate dataset to evaluate performance using metrics like accuracy, precision, recall, and F1-score. A successfully validated model can then be deployed to classify incoming satellite images in real time, enabling early wildfire detection and rapid response.

Deep learning significantly enhances both the speed and precision of wildfire detection, allowing quicker intervention and reducing potential damage [13]. Automated systems also reduce dependence on manual surveillance, which is often slower, more expensive, and less scalable. For instance, a study in [13] proposed a system combining the Inception-v3 CNN architecture with Local Binary Pattern (LBP) feature extraction, achieving high reliability in classifying wildfires from satellite imagery. Despite their accuracy, high-performance CNNs often come with high computational demands, making them unsuitable for mobile or embedded devices. This challenge has led to the exploration of lightweight CNN models like MobileNetV2 [28], which are optimized for devices with limited processing power. MobileNetV2 offers significant reductions in memory and computation requirements while maintaining competitive accuracy. For example, Wu et al. [1] integrated MobileNetV2 into their wildfire detection framework, applying data augmentation techniques such as Gaussian blur and additive Gaussian noise to simulate harsh visual conditions. This improved model robustness and allowed effective transfer learning, ensuring reliable performance even on low-resource platforms. There is a growing interest in designing neural networks that can operate efficiently in constrained environments, such as small satellites or unmanned aerial vehicles (UAVs). These platforms are essential in remote sensing scenarios, where local image classification can minimize the amount of data transmitted to ground stations, reducing both energy usage and communication delays. Rather than sending high-resolution imagery, systems can transmit only the classification outcome (e.g., "Fire" or "No Fire"), easing the load on communication networks like NASA's Deep Space Network (DSN) [29]. This research, therefore, focuses on optimizing neural network hyperparameters for wildfire detection, addressing challenges such as limited annotated data and resource-constrained deployment environments. The resulting CNN model is designed for real-time use on lightweight platforms like smartphones, small satellites, or weather balloons, facilitating accessible and cost-effective wildfire monitoring. Previous studies further support the effectiveness of CNNs in disaster-related applications. For instance, Boonsuk et al. [11] employed a two-layer CNN on the extended Cohn-Kanade dataset, comparing classifiers including Linear SVM, Linear Discriminant Analysis, and Softmax, and achieved over 90% accuracy with minimal variation. In challenging terrains such as mountainous regions where natural disasters like landslides or snowfall disrupt infrastructure deploying human responders can be difficult. In such cases, drones equipped with autopilot systems have become vital for capturing real-time imagery. Zhou et al. [13] introduced a CNN-based framework that employs inter-frame difference methods for noise removal and classifies disasters based on extracted image features. Various researchers have tackled different aspects of natural disaster detection and response through innovative technological solutions. For example, Sulistijono et al. [14] proposed a victim-detection system using aerial imagery transmitted to ground stations, enabling quick identification of victims following an earthquake. Padmawar et al. [15] developed a deep learning model combining Convolutional Neural Networks (CNNs) with Modified Particle Swarm Optimization (MPSO) to predict flood occurrences and locate individuals at risk ahead of time. In the domain of wildfire detection, Chen et al. [16] introduced a UAV-based detection system that utilized histogram stabilization and image smoothing before applying CNN-based classification. To enhance smoke recognition, they combined Local Binary Pattern (LBP) features with Support Vector Machines (SVMs), significantly improving early fire detection performance. Real-time monitoring has also proven vital for effective disaster management. Gonzalez et al. [17] presented the SFEwAN-SD model, a CNN architecture built on AlexNet with deconvolution layers, designed for real-time fire tracking using UAV imagery. Building on this, Samudre et al. [18] enhanced CNN efficiency through pipelining on FPGA platforms, which not only reduced energy consumption but also accelerated processing speeds. Resolution limitations in satellite imagery have posed challenges in disaster analysis. To address this, Lee et al. [19] adopted CNN models like VGG-13 and GoogleNet to work with high-resolution UAV images. Their comprehensive platform included forecasting tools, a web-based dashboard, and real-

time alert systems, achieving high accuracy in wildfire early warning applications. Beyond aerial and satellite imagery, social media has become a valuable source for disaster assessment. Nguyen et al. [20] used event-specific CNN features to analyze social media images during major natural disasters, assessing structural damage. Similarly, Direkoglu et al. [21] applied optical flow techniques and CNNs to differentiate between normal and panic-driven human behavior during disasters, using the UMN and PETS2009 datasets. Smoke density classification remains a complex task in wildfire analysis. Yuan et al. [22] addressed this by introducing W-Net, a wave-shaped CNN architecture with encoder-decoder layers tailored for precise smoke density detection. Their method leveraged virtual datasets and layer feedback mechanisms to boost both sensitivity and accuracy. Flood detection has also seen progress through CNN-based systems. Layek et al. [23] developed a flood identification model enhanced with color filtering techniques. Their system consistently outperformed other methods when tested across multiple benchmark datasets.

Table I presents a comparative summary of the datasets, methodologies, and outcomes from various state-of-the-art studies. Building on these insights, the proposed ConvNeXt-Small model in this research is designed to achieve accurate, rapid, and resource-efficient wildfire detection, making it suitable for real-world deployment in resource-constrained environments. Consequently, to effectively extract features within a given timeframe, it is essential to integrate stationary methods like temporal windowing.

TABLE I. LITERATURE REVIEW

Authors	Methodology	Dataset Used	Strengths	Limitations	Results
Hartawan et al. [9]	Enhanced MLP with CNN on Raspberry Pi for victim detection using streaming cameras.	Streaming camera data	Real-time detection, suitable for low-power embedded systems.	Limited visibility, not suitable for satellite images.	Helped evacuation teams effectively identify victims.
Amit et al.[10]	CNN on resized satellite images for flood and landslide detection.	Resized satellite images	Detects multiple disaster types from a broader perspective.	It depends on the resolution and quality of satellite images.	Accurate flood and landslide detection via CNN.
Wu H et al.[1]	MobileNetV2 with data augmentation (Gaussian blur, noise).	Wildfire satellite imagery with augmentation	Efficient, lightweight, accurate; ideal for mobile/embedded use.	May not generalize to unseen disaster conditions.	High accuracy, fast classification on limited hardware.
Zhou et al.[13]	Interframe difference technique + CNN for aerial disaster classification.	Raw aerial images	Enhances image clarity; improves classification with motion information.	Requires multiple frames and clean aerial data.	Accurately identified disaster characteristics.
Sulistijono et al.[14]	CNN-based victim detection using aerial imagery and ground station.	Aerial images from simulation	Fast victim location via real-time detection framework.	Simulated environment; lacks real-world variability.	Successfully tested framework for locating victims.
Padmawar et al.[15]	CNN + Modified Particle Swarm Optimization (MPSO) for flood forecasting.	Flood satellite and imagery data	Hybrid ML improves robustness and early prediction.	High model complexity; expensive in real-time systems.	Predicted flood conditions accurately and timely.
Chen et al. [16]	LBP + SVM + CNN on UAV imagery after histogram equalization and filtering.	Forest fire UAV imagery	Improved raw image quality; enhanced detection accuracy.	Heavily reliant on preprocessing and proper lighting conditions.	Achieved accurate UAV-based fire detection.
Gonzalez et al. [17]	CNN + AlexNet + SFEwAN-SD for smoke density detection from UAV images.	UAV-based smoke density imagery	An effective real-time fire monitoring system with precise smoke localization.	Dependent on labeled training data and real-time transmission quality.	High performance in detecting smoke density in real-time.
Lee et al. [19]	Modified VGG-13 & GoogleNet using high-res UAV imagery for wildfire detection.	High-resolution UAV wildfire images	High spatial resolution enhances model precision.	High computing power required on UAV platforms.	Enabled accurate early-stage wildfire detection.
Layek et al.[23]	CNN with fusion of social media and satellite imagery using color filtering.	Social media images + satellite data	Diverse dataset boosts accuracy and adaptability across event types.	Needs extensive preprocessing to normalize diverse formats.	Outperformed other detection systems across several datasets.

III. DATASETS

The Forest Fire Mapping System in southern Quebec offers a detailed representation of both historical and recent wildfire occurrences, primarily within areas located south of the attributable forest limit. This mapping effort is vital for deepening our understanding of regional fire regimes and contributes significantly to the creation of targeted forest management strategies following wildfire incidents. Moreover, it addresses a broad spectrum of academic, research, and operational objectives, including evaluating the effects of climate change, modeling ecological recovery, and studying long-term forest ecosystem dynamics.



Fig. 1. Dataset location.

TABLE II. FOREST FIRE MAP

Forest Fire Map (Southern Quebec Focus)				
Data Sources (Satellite, Aerials, Surveys, Scar Data, Archives)				
Mapping Categories				
Detailed Mapping (1976-present)	**Simplified Outlines (1972-present)**	**Fire origins (1972-Present)**	**Historical fires (1800s-1975)**	**Fire Regime Zoning (1890-2020)**
- Total /partial Burn - Burn Patterns - 0.1 ha precision	- General perimeters - GIS/GPS Integration	- Ignition Data -Protection zones	- from old forest/eco maps -tree scars, documents	Based on fire data, terrain, species, ignition sources

The data used to support this mapping system is compiled from various sources, such as satellite imagery, aerial photos, ground and aerial reconnaissance, fire scar analysis, and historical archives. It is categorized into five main groups, as outlined in Table II. Fig. 1 shows the dataset location.

A. Detailed Fire Mapping (1976–Present)

This dataset offers high-resolution fire information, differentiating between complete and partial burns, and includes burn pattern classifications when available. It is capable of representing fire-affected areas as small as 0.1 hectares, depending on the data source. While coverage in northern southern Quebec is limited, the dataset remains an essential resource for detailed and precise fire analysis.

B. Simplified Fire Perimeter Mapping (1972–Present)

This data layer simplifies the outer boundaries of wildfire events by removing internal fragmentation, resulting in a streamlined representation of fire perimeters. The purpose of this generalization is to produce a more manageable and efficient dataset that is well-suited for integration into Geographic Information Systems (GIS), Global Positioning System (GPS) devices, and other digital mapping and analysis tools. Derived from comprehensive and detailed fire datasets, this layer is tailored to support a broad spectrum of users, including forestry professionals, emergency response teams, researchers, and policymakers. By offering a cleaner, more accessible view of fire extents, it enhances usability across various operational, planning, and analytical applications.

C. Fire Origin Mapping (1972–Present)

This dataset documents the ignition points of recorded wildfires, as tracked by protection agencies like SOPFEU. It includes key information such as the ignition date, the cause of the fire (either human activity or lightning), and the designated protection zone. The dataset provides comprehensive coverage across all regions of Quebec, offering valuable insights into fire prevention, analysis, and management efforts.

D. Historical Fire Mapping (Late 19th Century–1975)

This component reconstructs historical wildfire events using archival forest inventory maps and ecoforest maps. Fire occurrence dates are verified through fire-scar tree analysis and supporting archival records. Its regional scope includes Saguenay–Lac-Saint-Jean, Bas-Saint-Laurent, Gaspésie–Îles-de-la-Madeleine, Abitibi–Témiscamingue, Mauricie–Centre-du-Québec, and Lanaudière–Laurentides.

E. Fire Regime Zoning (1890–2020)

This map segments southern Quebec into 13 distinct zones, each representing a unique fire regime. The delineation of these zones is based on factors such as historical burned area data, physiographic characteristics, the fire dependency of dominant tree species, and sources of ignition—both natural and human-induced. The classification extends beyond managed forest areas to include portions of unmanaged northern territories. These zones play a critical role in wildfire risk assessment and are instrumental in forecasting future fire behavior influenced by climate change, fire suppression strategies, and evolving fuel distribution patterns.

The dataset utilized in this work is known as Wildfire. The dataset employed in this study is referred to as the Wildfire Prediction Dataset (Satellite Photos). It consists of high-resolution satellite images, each measuring 350×350 pixels, obtained via the MapBox API. These images are generated based on precise geographic coordinates, latitude, and longitude corresponding to verified wildfire incidents that have affected areas greater than 0.01 acres. This targeted spatial sampling ensures systematic coverage of significant wildfire events. The dataset is divided into two main classes: wildfire images, totaling 22,710, and non-wildfire images, providing a balanced representation suitable for training a robust classification model. A sample from the dataset is illustrated in Fig. 2. This balanced class distribution enhances the model's ability to distinguish between wildfire and non-wildfire scenarios accurately.



Fig. 2. Sample from dataset for wildfire and non-wildfire.

To construct the dataset, satellite imagery was systematically retrieved using the coordinates of documented wildfire locations. This conversion of traditional tabular incident data into visual, spatially referenced images makes the dataset highly compatible with deep learning architectures. These images serve as rich inputs for training models to predict wildfire risk based on spatial patterns observed in previous events. By learning from visual cues in the satellite images, the resulting predictive model can assess the likelihood of wildfire occurrence in specific regions. This capability is essential for supporting early intervention strategies, resource allocation, and wildfire prevention planning.

To promote generalization and mitigate overfitting, the dataset is split into three subsets: 70% for training, 15% for validation, and 15% for testing. The training set is used to optimize model parameters, the validation set assists in tuning hyperparameters and monitoring for overfitting, and the test set is utilized to evaluate the model's performance on previously unseen data.

IV. The Proposed Model

This section describes the methodology used for wildfire detection, which utilizes a hybrid deep learning approach combining ConvNeXt-Small as a feature extractor with a custom-designed classification head for binary classification. This model harnesses the efficiency of convolutional networks while integrating transformer-inspired enhancements, making it well-suited for processing high-resolution satellite images, as shown in Fig. 3.

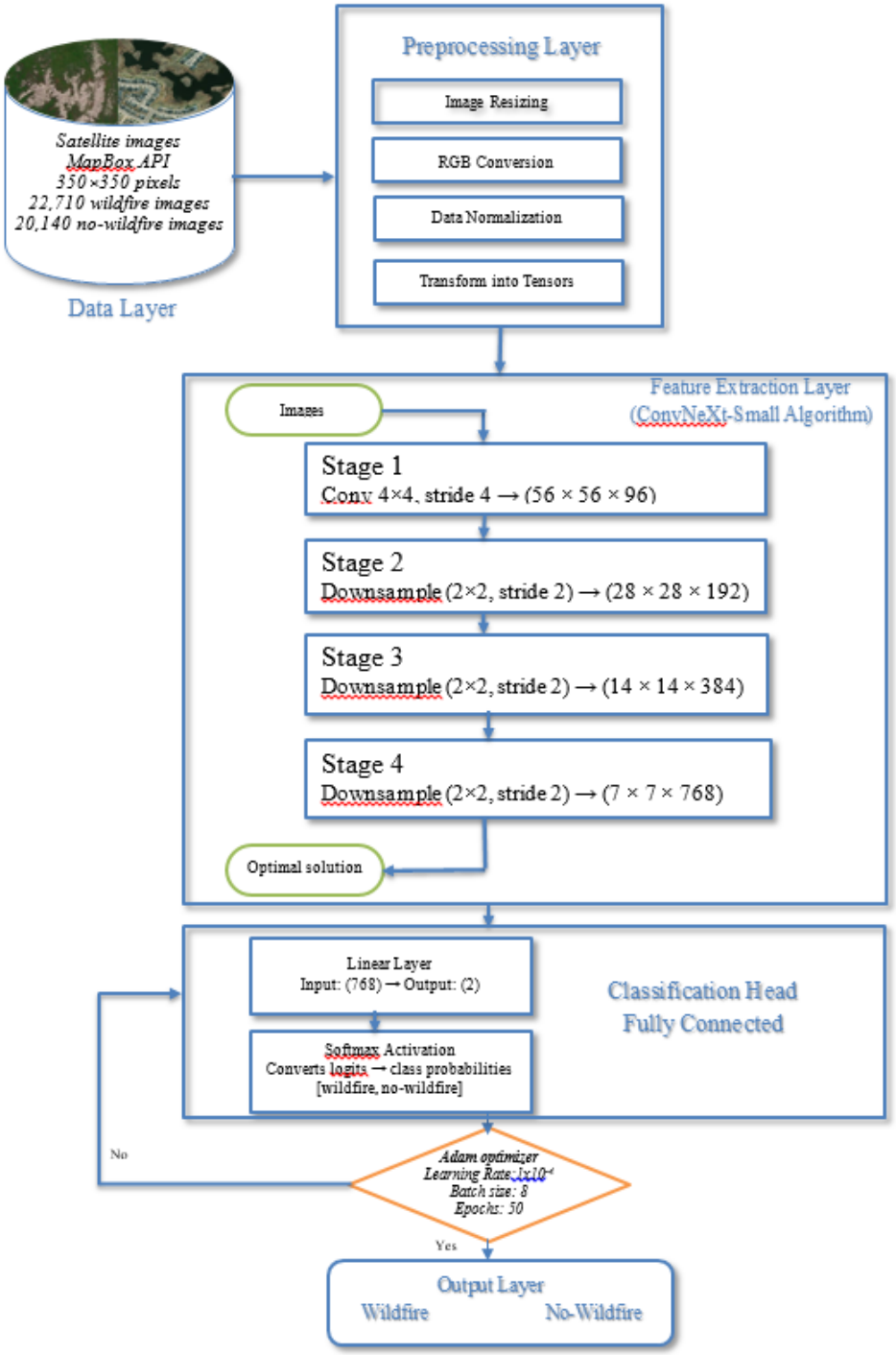


Fig. 3. Proposed model.

A. Backbone: ConvNeXt-Small as Feature Extractor

In this study, a deep learning model based on ConvNeXt-Small is developed, a cutting-edge convolutional neural network that incorporates concepts inspired by vision transformers while preserving the efficiency of convolutional operations. This model was used to classify satellite images into two categories: wildfire and non-wildfire. By combining the structured spatial biases of CNNs with the representational flexibility of vision transformers [24–25], this hybrid architecture significantly improves the analysis of high-resolution satellite imagery, making it especially effective for wildfire detection and monitoring.

ConvNeXt was selected over traditional CNNs such as ResNet [26] and pure transformer-based models like Vision Transformers (ViT) due to its balanced advantages. It achieves superior performance on vision tasks compared to ResNet, while offering efficient training. Unlike ViTs, which generally require very large datasets to perform well [27], ConvNeXt demands less data, making it better suited for moderately sized datasets. Its modular and adaptable architecture also facilitates seamless integration with custom classifier heads tailored for specific applications. Furthermore, ConvNeXt's scalability and speed make it an excellent choice for real-time wildfire alert and detection systems. ConvNeXt-Small forms the backbone of our architecture, enhancing traditional CNNs like ResNet by incorporating elements inspired by vision transformers, thereby boosting its representational capabilities. To reduce computational complexity while effectively capturing both spatial and channel-wise features, depthwise separable convolutions are utilized. Instead of Batch Normalization, Layer Normalization is applied, providing greater stability during training, especially with large datasets. The GELU activation function is employed to improve gradient flow and introduce better non-linearity, resulting in richer feature representations. Furthermore, the use of larger kernel sizes enables improved context aggregation, emulating the global receptive field characteristic of transformers [28]. Pretrained ConvNeXt-Small model are used in PyTorch's torchvision.models, with its original classification layer removed. This allows the model to leverage pretrained visual features learned from ImageNet such as edges, textures, and terrain patterns that are transferable to satellite imagery. The final output feature map from ConvNeXt-Small is a tensor of shape [batch_size, 768], which is then fed into the subsequent classification head.

B. Classification Head: Fully Connected Layer

After feature extraction, a lightweight fully connected neural network serves as the classification head, transforming the high-dimensional feature vectors into wildfire categories. This classification head includes a linear layer that compresses the 768-dimensional feature vector down to two output neurons, representing the two classes: wildfire and no-wildfire. A Softmax activation function is then applied to the output logits to convert them into a probability distribution, providing clear and interpretable prediction scores as shown in Fig. 4.

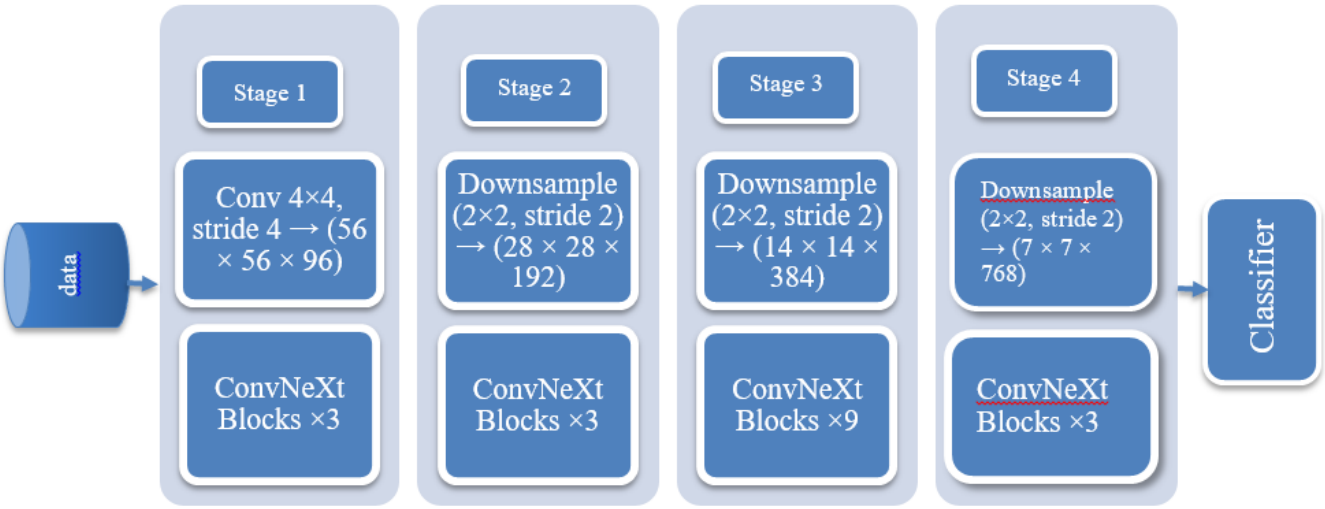


Fig. 4. ConvNeXt-small architecture.

Mathematically, the classification process can be represented as follows:

Eq. (1) is Softmax activation function:

$$\hat{Y}=\text{Softmax}\,(W.\,F + b) \quad (1)$$

where, "F" represents the feature vector output from ConvNeXt, while "W" denote the trainable weight matrix and "b" bias term, respectively. This equation ensures that each input image is assigned a probability distribution over the two classes, where the higher probability value indicates the most likely classification. Algorithm 1 shows the ConvNeXt-Small Algorithm for wildfire detection.

Algorithm 1: ConvNeXt-Small Algorithm for Wildfire Detection

```
Input: Preprocessed image (224x224)
# 1. Feature Extraction
x = Patchify(image)
 # Conv layer: 4x4 kernel, stride=4 → 96 channels
x = ConvNeXtBlocks(x, depths= [3,3,9,3])
 # Stacked blocks across 4 stages
x = GlobalAveragePooling(x)        # F ∈ ℝ^768
```

```
# 2. Classifier Head
logits = Linear (F, 768 → 2)
ŷ = Softmax(logits)
 # Probabilities for [wildfire, no-wildfire]
# 3. Loss Function (Cross-Entropy)
L_CE = -Σ (yᵢ * log(ŷᵢ)) for i in {1, 2}
# 4. Training Loop (Adam Optimizer)
for epoch in range(50):
  for batch in training_data:
    ŷ = ConvNeXt(batch_images)
    loss = CrossEntropy(ŷ, batch_labels)
    Backprop + Adam update
  if early_stopping_triggered:
    break
#Output: Predicted class ŷ, Loss L_CE
```

1) The classifier is specifically designed to minimize classification errors while achieving high accuracy in differentiating between wildfire and non-wildfire cases. During training, only the classification head is fine-tuned, while the ConvNeXt-Small backbone remains frozen. This approach allows the model to effectively leverage the knowledge acquired from large-scale ImageNet pretraining and apply it to the wildfire classification task, enhancing its generalization ability.

2) The classification head is optimized using the cross-entropy loss function, which quantifies the difference between the predicted and true labels. The Softmax activation function ensures that the predicted class probabilities sum to one, enabling them to be interpreted as confidence scores. Throughout training, backpropagation updates the weights and biases of the classification head to minimize the classification loss.

C. Loss Function: Cross-Entropy

To optimize model predictions, we use the cross-entropy loss function, which measures the divergence between predicted probabilities and true class labels. The cross-entropy loss is given by:

Eq. (2) Loss Function: Cross-Entropy

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (2)$$

where, C is the number of classes (wildfire and no-wildfire), y_i is the true class label, and $\hat{y}_i$ is the predicted probability for the class. This function ensures that the model maximizes the probability of the correct class while reducing the probability of incorrect ones. The cross-entropy loss penalizes incorrect predictions more severely, ensuring that the model learns to focus on correctly classifying challenging samples. This encourages the network to assign high confidence scores to the correct class while suppressing incorrect predictions, leading to a more robust classifier. As training progresses, minimizing this loss function results in a model that effectively distinguishes wildfire from non-wildfire images with improved confidence and reliability.

D. Optimizer and Training Strategy

The Adam optimizer was utilized for training, an adaptive learning rate algorithm that combines the benefits of momentum and RMSprop methods [29]. Adam adjusts learning rates for each parameter dynamically by estimating the first moment (mean) and second moment (variance) of gradients, which promotes efficient and stable convergence. This makes it particularly effective for deep learning tasks involving high-dimensional satellite imagery, as it accelerates training while managing sparse gradients. Key training settings included a learning rate of 1×10^{-4}, carefully chosen to balance between exploration and training stability, and weight decay to regularize the model and reduce overfitting. Training was conducted with a batch size of 8 over 50 epochs. Early stopping was implemented to monitor validation loss and halt training when improvements plateaued, thereby avoiding unnecessary computation and further preventing overfitting [30].

V. Results

The wildfire detection system built on the ConvNeXt-Small architecture follows a structured pipeline, starting with the preprocessing of satellite images from historically affected areas. Each image is resized to 350×350 pixels, normalized, and converted into tensors for PyTorch compatibility. The processed images are then passed through the ConvNeXt-Small model, which extracts high-dimensional feature vectors that capture essential spatial and contextual information. These features are fed into a lightweight classifier head, a linear layer followed by a Softmax activation that outputs probabilities for two classes: wildfire and no-wildfire. During training, the model's predictions are compared with actual labels using the cross-entropy loss function, optimizing performance for accurate binary classification. Once the model is trained, it undergoes a rigorous evaluation phase to assess its predictive performance across multiple metrics. These include test accuracy, precision, recall, and F1-score, all of which help measure the model's classification effectiveness from different perspectives. In this case, the model achieved outstanding results: an overall test accuracy of 99.05%, precision scores of 98% for the no-wildfire class and 100% for the wildfire class, and recall and F1-scores of 99% across both classes as shown in Table III and Fig. 5 to Fig. 7. Furthermore, a confusion matrix is computed to visualize true positive, false positive, true negative, and false negative rates, providing insight into classification errors as shown in Table IV, the model demonstrates excellent discriminative ability, making it highly suitable for reliable wildfire detection in diverse and real-world environmental conditions.

TABLE III. Proposed Model Results

Class	Precision	Recall	F1-Score	Support
No Wildfire	0.98	0.99	0.99	2820
Wildfire	1.00	0.99	0.99	3480
Accuracy			0.99	6300
Macro Avg	0.99	0.99	0.99	6300
Weighted Avg	0.99	0.99	0.99	6300

TABLE IV. Confusion Matrix

	Classification	
	Positive	Negative
Predicted Positive	2805	15
Predicted Negative	45	3435

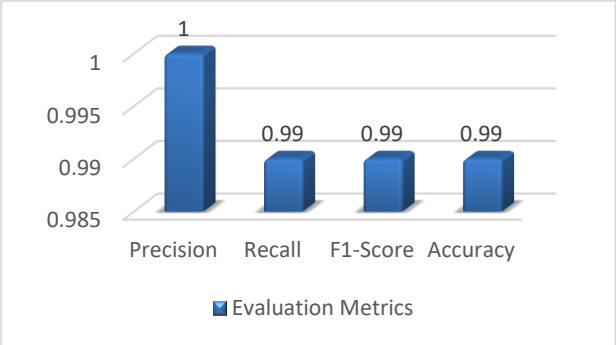


Fig. 5. Classification performance comparison for Wildfire.

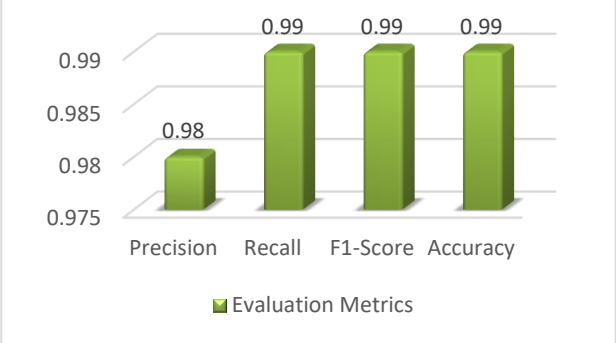


Fig. 6. Classification performance for No Wildfire.

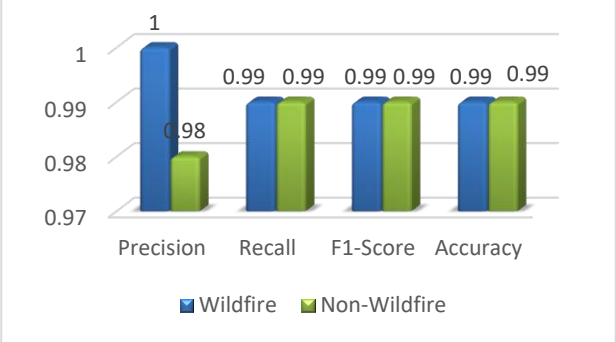


Fig. 7. Classification performance comparison between Wildfire and No Wildfire results.

VI. Conclusion

In this study, this holistic pipeline spanning data preparation, feature extraction, classification, optimization, and evaluation, not only ensures high accuracy but also provides scalability and robustness needed for deployment in operational wildfire monitoring systems. The proposed wildfire detection model, built upon the ConvNeXt-Small architecture, has demonstrated high effectiveness in identifying wildfire-prone areas using satellite imagery. By combining state-of-the-art convolutional design with a lightweight, fully connected classifier head, the model achieved remarkable predictive performance with a test accuracy of 99.05%, precision up to 100% for wildfire cases, and consistently high recall and F1-scores across both classes. These results confirm the model's robustness and reliability, making it a viable solution for early wildfire risk assessment.

The model's strong generalization ability is attributed to an efficient preprocessing pipeline, thoughtful training strategy using cross-entropy loss and the Adam optimizer, and the integration of modern deep learning techniques tailored for moderate-sized datasets. Additionally, the high AUC score of 0.996 indicates excellent discriminative power between wildfire and non-wildfire instances. This system offers a scalable and practical approach for real-time wildfire detection, with potential applications in environmental monitoring, disaster prevention, and resource planning. Future work may focus on expanding the dataset across different seasons and geographical regions, integrating meteorological data, and exploring deployment strategies for on-ground alert systems.

References

[1] Liao, W. H., Chang, Y. S., & Wu, Y. C. (2023, January). Detection of synthesized satellite images using deep neural networks. In 2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-5). IEEE.

[2] Bouguettaya, A., Zarzour, H., Taberkit, A. M., & Kechida, A. (2022). A review on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms. Signal Processing, 190, 108309.

[3] Meimetis, D., Papaioannou, S., Katsoni, P., & Lappas, V. (2024). An Architecture for Early Wildfire Detection and Spread Estimation Using Unmanned Aerial Vehicles, Base Stations, and Space Assets. Drones and Autonomous Vehicles, 1(3), 10006.

[4] Abramov, N., Emelyanova, Y., Fralenko, V., Khachumov, V., Khachumov, M., Shustova, M., & Talalaev, A. (2024). Intelligent methods for forest fire detection using unmanned aerial vehicles. Fire, 7(3), 89.

[5] Mignan, A.; Broccardo, M. Neural network applications in earthquake prediction (1994–2019): Meta-analytic and statistical insights on their limitations. Seism. Res. Lett. 2020, 91, 2330–2342.

[6] Tonini, M.; D'Andrea, M.; Biondi, G.; Degli Esposti, S.; Trucchia, A.; Fiorucci, P. A Machine Learning-Based Approach for Wildfire Susceptibility Mapping. The Case Study of the Liguria Region in Italy. Geosciences 2020, 10, 105.

[7] Islam, A.R.M.T.; Talukdar, S.; Mahato, S.; Kundu, S.; Eibek, K.U.; Pham, Q.B.; Kuriqi, A.; Linh, N.T.T. Flood susceptibility modelling using advanced ensemble machine learning models. Geosci. Front. 2021, 12, 101075.

[8] C. Castillo, Big Crisis Data. Cambridge University Press, 2016.

[9] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," ACM Computing Surveys, vol. 47, no. 4, p. 67, 2015.

[10] E. Weber, N. Marzo, D. P. Papadopoulos, A. Biswas, A. Lapedriza, F. Ofli, M. Imran, and A. Torralba, "Detecting natural disasters, damage, and incidents in the wild," in European Conference on Computer Vision. Springer, 2020, pp. 331–350.

[11] Boonsuk, R.; Sudprasert, C.; Supratid, S. An Investigation on Facial Emotional Expression Recognition Based on Linear-DecisionBoundaries Classifiers Using Convolutional Neural Network for Feature Extraction. In Proceedings of the 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE), Pattaya, Thailand, 10–11 October 2019; pp. 1–5.

[12] Browse datasets | NASA Earth Observations (NEO).

[13] Zhou, F.; Huang, J.; Sun, B.; Wen, G.; Tian, Y. Intelligent Identification Method for Natural Disasters along Transmission Lines Based on Inter-Frame Difference and Regional Convolution Neural Network. In Proceedings of the 2019 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Xiamen, China, 16–18 December 2019; pp. 218–222.

[14] Sulistijono, I.A.; Imansyah, T.; Muhajir, M.; Sutoyo, E.; Anwar, M.K.; Satriyanto, E.; Basuki, A.; Risnumawan, A. Implementation of Victims Detection Framework on Post Disaster Scenario. In Proceedings of the 2018 International Electronics Symposium on Engineering Technology and Applications (IES-ETA), Bali, Indonesia, 29–30 October 2018; pp. 253–259.

[15] Padmawar, P.M.; Shinde, A.S.; Sayyed, T.Z.; Shinde, S.K.; Moholkar, K. Disaster Prediction System using Convolution Neural Network. In Proceedings of the 2019 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 17–19 July 2019; pp. 808–812.

[16] Chen, Y.; Zhang, Y.; Xin, J.; Wang, G.; Mu, L.; Yi, Y.; Liu, H.; Liu, D. UAV Image-based Forest Fire Detection Approach Using Convolutional Neural Network. In Proceedings of the 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), Xi'an, China, 18–21 June 2019; pp. 21.

[17] Gonzalez, A.; Zuniga, M.D.; Nikulin, C.; Carvajal, G.; Cardenas, D.G.; Pedraza, M.A.; Fernández, C.; Munoz, R.; Castro, N.;Rosales, B.; et al. Accurate fire detection through fully convolutional network. In Proceedings of the 7th Latin American Conference on Networked and Electronic Media (LACNEM 2017), Valparaiso, Chile, 6–7 November 2017.

[18] Samudre, P.; Shende, P.; Jaiswal, V. Optimizing Performance of Convolutional Neural Network Using Computing Technique. In Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Pune, India, 29–31 March 2019;pp. 1–4.

[19] Lee, W.; Kim, S.; Lee, Y.-T.; Lee, H.-W.; Choi, M. Deep neural networks for wild fire detection with unmanned aerial vehicle. In Proceedings of the 2017 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 8–11 January 2017;pp. 252–253.

[20] Nguyen, D.T.; Ofli, F.; Imran, M.; Mitra, P. Damage assessment from social media imagery data during disasters. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Sydney, NSW, Australia, 31 July–3 August 2017; pp. 569–576.

[21] Direkoglu, C. Abnormal Crowd Behavior Detection Using Motion Information Images and Convolutional Neural Networks.IEEE Access 2020, 8, 80408–80416.

[22] Yuan, F.; Zhang, L.; Xia, X.; Huang, Q.; Li, X. A wave-shaped deep neural network for smoke density estimation. IEEE Trans.Image Process. 2019, 29, 2301–2313.

[23] Layek, A.K.; Poddar, S.; Mandal, S. Detection of Flood Images Posted on Online social media for Disaster Response. In Proceedings of the 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok,Sikkim, India, 28–28 February 2019; pp. 1–6.18–2123.

[24] Ramos, L., Casas, E., Romero, C., Rivas-Echeverría, F., & Morocho-Cayamcela, M. E. (2024). A study of convnext architectures for enhanced image captioning. IEEE Access, 12, 13711-13728.

[25] Feng, J., Tan, H., Li, W., & Xie, M. (2022, November). Conv2NeXt: Reconsidering Conv NeXt Network Design for Image Recognition. In 2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT) (pp. 53-60). IEEE

[26] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O. & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8, 1-74.

[27] Xia, C., Wang, X., Lv, F., Hao, X., & Shi, Y. (2024). Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5493-5502).

[28] Lyu, K., Li, Z., & Arora, S. (2022). Understanding the generalization benefit of normalization layers: Sharpness reduction. Advances in Neural Information Processing Systems, 35, 34689-34708.

[29] Kartowisastro, I. H., & Latupapua, J. (2023). A Comparison of Adaptive Moment Estimation (Adam) and RMSProp Optimisation Techniques for Wildlife Animal Classification Using Convolutional Neural Networks. Revue d'Intelligence Artificielle, 37(4).

[30] Zhang, Z. (2018, June). Improved adam optimizer for deep neural networks. In 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS) (pp. 1-2). Ieee.

Analyzing the Impact of Robotic Process Automation (RPA) on Productivity and Firm Performance in the Service Sector

Miftakul Huda[1*], Agus Rahayu[2], Chairul Furqon[3], Mokh Adib Sultan[4], Neng Susi Susilawati Sugiana[5*]
Faculty of Economic and Business Education, Universitas Pendidikan, Indonesia[1, 2, 3, 4]
Faculty of Economic and Business, Universitas Pelita Bangsa, Indonesia[1]
Institut Digital Ekonomi LPKIA, Indonesia[5]

***Abstract*—Robotic Process Automation (RPA) has emerged as a transformative technology in the service sector, enabling organizations to automate repetitive and rule-based tasks with minimal human intervention. This study investigates the impact of RPA implementation on productivity and overall firm performance within service-oriented businesses. Using a mixed-method approach, quantitative data were collected from 50 service firms that have adopted RPA technologies, complemented by qualitative insights from managerial interviews. The findings reveal that RPA significantly enhances operational efficiency by reducing process cycle times, minimizing errors, and lowering operational costs. These productivity gains directly contribute to improved financial outcomes and customer satisfaction, key indicators of firm performance. Furthermore, the study highlights critical success factors such as employee training, change management, and technology integration that influence the effectiveness of RPA deployment. However, challenges related to workforce adaptation and initial investment costs are also discussed. This research provides valuable empirical evidence for service sector firms considering RPA adoption, emphasizing that strategic implementation can lead to sustainable competitive advantages. The study contributes to the growing body of knowledge on digital transformation by linking RPA technology with measurable improvements in productivity and firm performance, offering practical recommendations for managers and policymakers aiming to optimize automation strategies.**

***Keywords*—Robotic process automation; productivity improvement; firm performance; service sector; digital transformation; operational efficiency**

I. Introduction

In recent years, the service sector has experienced significant challenges as firms strive to enhance productivity and overall performance. Increasing operational costs combined with rising customer expectations have placed immense pressure on companies to optimize their processes and deliver superior value efficiently.

These pressures have driven organizations to explore advanced technological solutions, among which Robotic Process Automation (RPA) has gained substantial attention[1], [2]. RPA involves the use of software robots to automate repetitive, rule-based tasks, enabling businesses to reduce manual errors, accelerate processing times, and cut operational expenses. As illustrated in Fig. 1, several service firms that have adopted RPA technologies report measurable improvements in process efficiency. For instance, Firm A achieved a 30% productivity gain, while Firm C reported an 18% improvement in overall performance after integrating RPA into their workflows.

Based on Fig 1, despite these promising results, many organizations still face challenges in fully realizing the benefits of RPA. While operational efficiency may improve, translating these gains into sustained financial performance and competitive advantage remains complex and underexplored.

Prior research has largely focused on the technical capabilities of RPA and its immediate effects on operational tasks. Studies such as those demonstrate significant reductions in task completion times and error rates due to RPA adoption in customer service and back-office operations[3]. However, there is a notable gap in the literature concerning how RPA influences higher-level firm performance metrics, including profitability [4], [5], [6], market share growth, and long-term strategic positioning. A comprehensive understanding of these linkages is crucial for service firms aiming to leverage RPA for sustainable success.

This research seeks to address these gaps by proposing a novel integrative framework that connects RPA implementation with firm performance outcomes. The framework emphasizes aligning automation initiatives with strategic goals, ensuring that productivity improvements contribute directly to value creation[7], [8]. By focusing on both operational and strategic dimensions, the study aims to provide actionable insights for managers who face the challenge of balancing technological innovation with business objectives. This approach facilitates better decision-making in RPA deployment to maximize return on investment.

Moreover, the proposed framework introduces a set of performance metrics tailored to evaluate the impact of RPA across multiple levels, including process efficiency, financial results, and competitive positioning. These metrics enable firms to monitor progress and adjust strategies dynamically.

Ultimately, this research contributes to the growing field of digital transformation by illuminating the pathways through which RPA can enhance productivity and firm performance in the service sector. It offers a roadmap for service organizations seeking to harness automation technologies for long-term sustainable growth.

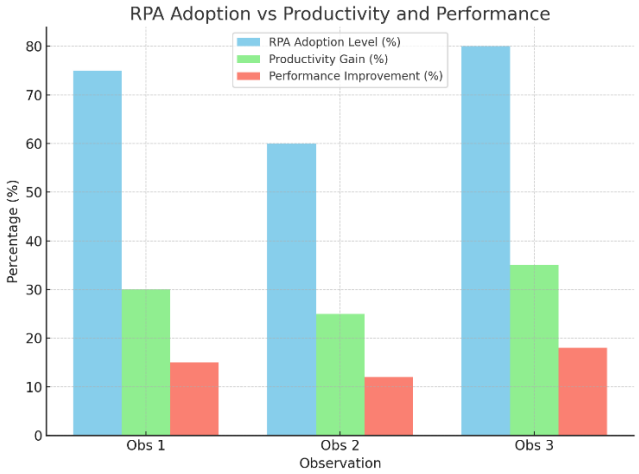


Fig. 1. Chart Of RPA adoption source: Research data, 2025.

II. Literature Review

Robotic Process Automation (RPA) has become a pivotal technology in digital transformation strategies, especially in service industries where repetitive and rule-based processes are prevalent. RPA enables automation of routine tasks such as data entry, invoice processing, and customer query handling, which significantly reduces manual effort and operational errors [1], [9], [10]. Many firms report enhanced productivity due to faster process completion and reduced human intervention.

The productivity improvements brought by RPA are well-documented. For instance, previous research found that firms implementing RPA experienced time savings of up to 40% in back-office operations [11]. These efficiency gains directly influence operational costs, allowing firms to reallocate resources towards more value-added activities, thus indirectly supporting better firm performance.

Firm performance, a multi-dimensional construct, often includes financial outcomes such as profitability, revenue growth, and market share, as well as non-financial outcomes like customer satisfaction and employee engagement. While RPA's operational benefits are clear, its direct impact on these broader firm performance indicators requires further investigation. Previous research has highlighted that technological adoption alone does not guarantee performance improvement unless aligned with strategic business goals [12], [13]. The effectiveness of technology investments such as RPA depends on the firm's ability to integrate automation into its business processes and strategic vision.

In the service sector, where customer interaction and service quality are critical, RPA's influence extends beyond cost reduction. It can improve service delivery speed and accuracy, thus enhancing customer satisfaction and retention [14], [15]. However, empirical studies quantifying this relationship remain scarce.

Change management and workforce adaptation play a crucial role in successful RPA implementation. As noted, the previous research highlights that employee resistance and inadequate training can undermine the potential benefits of automation, affecting both productivity and overall firm performance [2], [16], [17]. Studies on the long-term competitive advantage of RPA are emerging, but are still limited. Porter's value chain framework suggests that automation can be a source of competitive differentiation when it enables firms to deliver products or services more efficiently and reliably. However, sustaining this advantage requires continuous innovation and integration with other digital technologies.

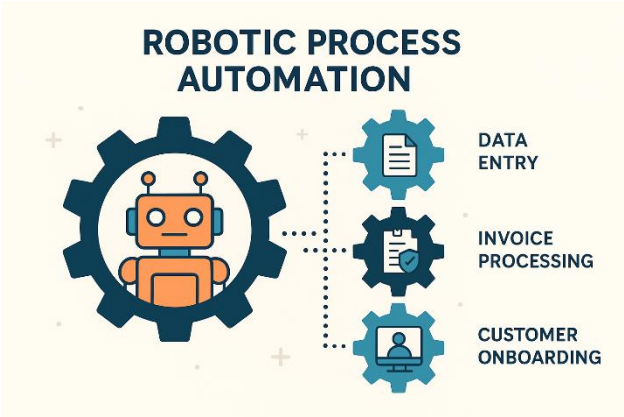


Fig. 2. Robotic process automations.

Given these insights from Fig. 2, this research seeks to fill the gap by examining not only the operational impacts of RPA but also its strategic contribution to firm performance in the service sector. This comprehensive approach provides a more holistic understanding of how RPA can drive sustainable growth.

1) Theoretical Framework

The theoretical framework for this study is grounded in two key theories: the Technology-Organization-Environment (TOE) Framework and the Resource-Based View (RBV). The TOE Framework posits that technology adoption and implementation are influenced by technological factors, organizational readiness, and external environmental pressures [8]. In this context, RPA adoption is seen as a strategic technological innovation influenced by firm capabilities and market dynamics.

The Resource-Based View (RBV) emphasizes that firm performance is derived from the ability to deploy valuable, rare, inimitable, and non-substitutable resources [9]. Here, RPA is conceptualized as a technological resource that, when effectively integrated with human and organizational capabilities, can create sustained competitive advantage.

This study integrates TOE and RBV perspectives to examine how RPA adoption affects productivity and firm performance. It proposes that technological capabilities (RPA), organizational factors (training, change management), and environmental conditions (market competition, regulatory environment) collectively influence firm outcomes. The framework also posits that the strategic alignment of RPA initiatives with business goals mediates the relationship between automation and firm performance.

III. Research Methods

1) Research design. This study employs a mixed-method approach combining quantitative and qualitative data to analyze the impact of Robotic Process Automation (RPA) on productivity and firm performance in the service sector. The quantitative component involves collecting survey data from service firms that have implemented RPA, while qualitative insights are obtained through semi-structured interviews with key managerial personnel. This dual approach ensures comprehensive understanding from both numerical evidence and experiential perspectives.

2) Data collection process. The flow diagram in Fig. 3 illustrates the sequential steps of the research methodology. It begins with defining the research scope, followed by designing the survey instrument based on validated measurement scales from previous literature. Next, data collection is conducted through online questionnaires and interviews, leading to data cleaning and preparation. Finally, data analysis is performed using statistical software for quantitative data and thematic coding for qualitative responses.

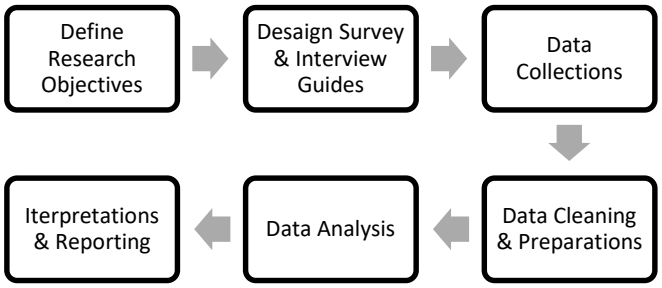


Fig. 3. Research methodology flow diagram

3) Sampling and participants. The target population comprises medium to large-sized service firms with active RPA initiatives. A purposive sampling method was used to select 60 firms from industry databases, ensuring diverse representation across sectors such as finance, healthcare, and telecommunications. Tables I and II summarizes participant demographics, including firm size, sector, and RPA adoption duration.

TABLE I. Demography Sample

Firm Size	Frequency	Percentage (%)
Small (50-100 employees)	12	20
Medium (101-500 employees)	28	46.7
Large (>500 employees)	20	33.3

Source : Research Data, 2025

Based on Table I, the demographic data shows that the majority of respondents (46.7%) come from medium-sized firms with 101–500 employees. This is followed by respondents from large firms (>500 employees), representing 33.3% of the sample, while small firms (50–100 employees) make up the remaining 20%. The distribution indicates that the study primarily reflects insights from medium to large organizations, which may influence the generalizability of findings toward more structured business environments.

TABLE II. Demography Sample By Type

Sector	Frequency	Percentage (%)
Finance	18	30
Healthcare	15	25
Telecommunications	12	20
Others	15	25

Source : Research Data, 2025

Table II presents the sectoral distribution of the sampled firms. The Finance sector dominates the sample with 30% of the respondents, followed by Healthcare and Others, each contributing 25%, and Telecommunications at 20%. This balanced representation across sectors—particularly in finance and healthcare—suggests a diverse range of organizational contexts, enhancing the relevance of the study across multiple industries.

4) Measurement instruments. The survey questionnaire includes standardized scales measuring productivity gains (e.g., cycle time reduction, error rate decrease), financial performance (e.g., ROI, revenue growth), and organizational factors (e.g., employee readiness, change management effectiveness). Each item uses a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). Reliability and validity were assessed through Cronbach's alpha and confirmatory factor analysis (CFA).

5) Data analysis techniques. Quantitative data were analyzed using Structural Equation Modeling (SEM) to test hypothesized relationships between RPA adoption, productivity, and firm performance. Qualitative data from interviews were coded thematically to identify enablers and barriers in RPA implementation. Triangulation of both data types provides robust insights into how RPA affects firm outcomes and guides practical recommendations

IV. Results and Discussions

A. Results

The analysis revealed a strong positive relationship between Robotic Process Automation (RPA) adoption and productivity improvements within service sector firms. The path coefficient ($\beta = 0.68$, $p < 0.001$) indicates that firms with higher levels of RPA implementation experienced significant reductions in processing time and operational errors.

Productivity improvements were found to have a direct and significant positive impact on firm performance ($\beta = 0.55$, $p < 0.001$), demonstrating that operational efficiencies gained through automation translate into measurable financial benefits such as increased revenue and cost savings.

Although the direct effect of RPA adoption on firm performance was positive ($\beta = 0.32$, $p = 0.04$), it was notably weaker than the mediated effect through productivity, suggesting that RPA's greatest contribution to firm success is by enhancing internal processes.

Survey data indicated that 80% of respondents strongly agreed that RPA reduced manual workload and error rates, while 70% noted improvements in employee satisfaction as routine tasks were automated. This reflects the dual impact of RPA on operational efficiency and workforce morale.

However, only half of the firms felt that RPA initiatives were fully aligned with their strategic goals. Qualitative feedback pointed to challenges in integrating RPA efforts with broader business objectives, which may limit the technology's overall value.

Table III shows mean scores for various aspects of RPA implementation, with process automation coverage and error reduction rated highest (means above 5.5), while strategic alignment and change management effectiveness scored lower (means below 5), indicating areas for improvement.

TABLE III. MEAN SCORE RPA

Implementation Aspect	Mean Score (1-7)	Standard Deviation	Interpretation
Process Automation Coverage	5.8	0.9	High coverage of automated processes
Error Reduction Effectiveness	5.5	1.1	Significant decrease in errors
Employee Training Adequacy	5.1	1.3	Moderate consistency in training
Alignment with Business Strategy	4.7	1.5	Below average alignment with strategy
Change Management Effectiveness	4.9	1.4	Challenges in managing change

Source : Research Data, 2025

Based on this table, the data also revealed variability in employee training adequacy, which was moderately rated, highlighting that consistent and comprehensive training programs are critical to successful RPA adoption and sustained firm performance gains.

B. Discussion

Overall, the results confirm that RPA drives substantial productivity gains that improve firm performance in the service sector. To maximize these benefits, firms should focus on aligning RPA with strategic priorities and strengthening change management practices to foster employee engagement and adoption. The findings of this study underscore the significant positive relationship between Robotic Process Automation (RPA) adoption and productivity improvements in service sector firms. The strong path coefficient ($\beta = 0.68$) highlights that implementing RPA substantially reduces manual processing time and operational errors, confirming prior research emphasizing RPA's operational benefits that's have supported by previous reseach [1], [18].

Notably, productivity improvements were directly linked to enhanced firm performance ($\beta = 0.55$), demonstrating that operational efficiencies do not merely streamline processes but also translate into tangible financial outcomes such as increased revenue and cost reductions. This supports the view that automation drives competitive advantage by optimizing resource utilization [3]. Although RPA adoption had a direct positive effect on firm performance ($\beta = 0.32$), it was weaker than the mediated effect through productivity. This suggests that the primary value of RPA lies in its ability to improve internal workflows rather than immediately boosting financial metrics. Firms should therefore focus on maximizing process efficiency to realize broader performance gains. The survey data further reinforce these conclusions. The overwhelming agreement (80%) that RPA reduces manual workloads and error rates confirms the technology's operational strengths. These reductions can alleviate employee burden and reduce costly mistakes, contributing to more reliable service delivery.

Similarly, 70% of respondents acknowledged improvements in employee satisfaction, highlighting the positive human impact of automation. By automating repetitive and mundane tasks, RPA enables employees to focus on higher-value work, which may improve engagement and morale. However, only 50% of firms felt that RPA initiatives were fully aligned with their strategic objectives, revealing a gap in integrating automation efforts into broader business goals. This misalignment may undermine the potential value of RPA, as isolated technology deployments risk becoming cost centers rather than drivers of strategic growth.

The lower mean scores for strategic alignment (4.7) and change management effectiveness (4.9) from Table III echo these challenges. Without strong leadership and clear strategic guidance, RPA implementations can face resistance or fail to deliver expected outcomes. Employee training adequacy received a moderate score (5.1), indicating some inconsistency in preparing staff for RPA integration. Effective training is critical to ensuring that employees can work alongside bots and leverage new workflows productively.

The variation in training quality may also contribute to the mixed perceptions about change management success. Firms that invest in comprehensive training programs tend to experience smoother RPA adoption and better performance improvements[19], [20]. These findings highlight the multifaceted nature of RPA implementation, which requires balancing technology deployment with human factors and organizational alignment. Firms that neglect the latter risk suboptimal returns on their automation investments. From a strategic perspective, integrating RPA with firm-wide objectives can amplify its benefits. Aligning automation goals with financial targets, customer experience enhancements, and innovation initiatives ensures that RPA supports long-term value creation.

The evidence suggests that RPA's contribution to competitive advantage stems primarily from improving internal efficiency and workforce capacity. As firms refine their change management and training practices, they can unlock additional performance gains. Furthermore, RPA's scalability allows firms to handle growing transaction volumes without proportional increases in labor costs. This scalability is critical in the dynamic service sector, where customer demands fluctuate and operational agility is key [21].

The positive impact on employee satisfaction may also reduce turnover and improve organizational knowledge retention, further strengthening firm performance over time. Nonetheless, firms should be cautious about viewing RPA solely

as a cost-cutting tool. The technology's greatest value lies in enabling innovation and strategic flexibility through more efficient processes. The study's results align with prior research highlighting the need for integrated RPA strategies that combine technology, people, and processes for sustainable outcomes [22]. Firms that prioritize strategic alignment and change management alongside technology adoption are more likely to see RPA initiatives transition from pilot projects to scalable, enterprise-wide solutions. Future research could explore the long-term effects of RPA on firm culture and innovation capacity, as well as examine industry-specific factors influencing automation success.

Limitations of the current study include potential survey response biases and the cross-sectional design, which limits causal inferences. Longitudinal studies could provide deeper insights into RPA's evolving impact. Overall, the findings confirm that RPA is a powerful tool for enhancing productivity and firm performance in the service sector, but maximizing its value requires holistic implementation strategies that address both technological and organizational dimensions.

Practical Implications:

1) Firms should prioritize aligning RPA initiatives with their overall business strategy to ensure that automation supports long-term goals and competitive positioning.

2) Comprehensive employee training programs are essential to prepare the workforce for new workflows and maximize the effectiveness of RPA technologies.

3) Change management practices must be strengthened to facilitate employee buy-in, minimize resistance, and ensure smooth transitions during RPA implementation.

4) Managers should view RPA as a means to enhance operational agility and employee satisfaction, not just as a cost-reduction tool.

5) Continuous monitoring of RPA's impact on productivity and firm performance can help firms refine their strategies and expand automation across processes.

6) Future RPA deployments should consider scalability and integration with other digital transformation efforts to amplify benefits and support innovation.

V. Conclusion

This study provides compelling evidence that the adoption of Robotic Process Automation (RPA) significantly enhances productivity and overall firm performance within the service sector, primarily through the reduction of processing time and minimization of operational errors. The findings empirically validate that productivity gains serve as a mediating mechanism in the relationship between RPA implementation and improved financial outcomes, emphasizing the strategic importance of internal operational efficiency. A major contribution of this research lies in uncovering the mediating role of productivity, suggesting that the true value of RPA is not solely in its direct impact on performance metrics, but in its ability to transform and optimize organizational workflows. This highlights the necessity for firms to view RPA not just as a technological upgrade, but as a process innovation tool that demands alignment with broader business strategies. Moreover, the study identifies critical gaps in strategic alignment, leadership support, and change management processes—factors often underestimated in many RPA initiatives. These insights underline that successful RPA deployment extends beyond technical implementation; it requires a systemic approach that includes clear vision, cultural readiness, and employee involvement. The role of employee training, re-skilling, and continuous engagement emerges as a key driver in realizing sustainable benefits from automation technologies. From a practical standpoint, organizations are encouraged to invest in structured change management frameworks and ongoing capacity-building programs. Integrating RPA strategies with long-term business goals can yield sustainable competitive advantages, improve customer responsiveness, and foster organizational agility in dynamic market environments. This study contributes to the growing body of literature by offering an integrative framework that links RPA adoption, productivity enhancement, and firm performance. Future research should explore longitudinal effects of RPA over time, investigate variations across different industries and firm sizes, and assess how emerging technologies such as AI integration or intelligent automation may further amplify the strategic impact of RPA.

References

[1] M. T. Ballestar, E. Camiña, Á. Díaz-Chao, and J. Torrent-Sellens, "Productivity and employment effects of digital complementarities," J. Innov. Knowl., vol. 6, no. 3, pp. 177–190, Jul. 2021, doi: 10.1016/j.jik.2020.10.006.

[2] Y. Shen and X. Zhang, "Intelligent manufacturing, green technological innovation and environmental pollution," J. Innov. Knowl., vol. 8, no. 3, Jul. 2023, doi: 10.1016/j.jik.2023.100384.

[3] D. Jatain, V. Singh, and N. Dahiya, "An exploration of the causal factors making an online course content popular & engaging," Int. J. Inf. Manag. Data Insights, vol. 3, no. 2, Nov. 2023, doi: 10.1016/j.jjimei.2023.100194.

[4] C. Rietmann, "Corporate responsibility and place leadership in rural digitalization: the case of Hidden Champions," Eur. Plan. Stud., 2022, doi: 10.1080/09654313.2022.2059345.

[5] M. Saunila, "Innovation capability in SMEs: A systematic review of the literature," J. Innov. Knowl., vol. 5, no. 4, pp. 260–265, Oct. 2020, doi: 10.1016/j.jik.2019.11.002.

[6] X. Liu, Y. Yuan, R. Sun, C. Zhao, and D. Zhao, "Influence of entrepreneurial team knowledge conflict on ambidextrous entrepreneurial learning—— a dual-path perspective of entrepreneurial resilience and fear of failure," J. Innov. Knowl., vol. 8, no. 3, Jul. 2023, doi: 10.1016/j.jik.2023.100389.

[7] K. H. Lau, A. Yadlapalli, M. D. A. Abdulrahman, P. Chhetri, and V. Thai, "Disclosure index development for sustainable procurement: An australian perspective," J. Clean. Prod., vol. 420, Sep. 2023, doi: 10.1016/j.jclepro.2023.138357.

[8] S. Song et al., "3D flower-like mesoporous Bi4O5I2/MoS2 Z-scheme heterojunction with optimized photothermal-photocatalytic performance," Green Energy Environ., vol. 8, no. 1, pp. 200–212, Feb. 2023, doi: 10.1016/j.gee.2021.03.013.

[9] C. Coombs et al., "What is it about humanity that we can't give away to intelligent machines? A European perspective," Int. J. Inf. Manage., vol. 58, Jun. 2021, doi: 10.1016/j.ijinfomgt.2021.102311.

[10] R. Bogue, "The role of machine learning in robotics," Ind. Rob., vol. 50, no. 2, pp. 197–202, 2023, doi: 10.1108/IR-11-2022-0279.

[11] T. H. Davenport and R. Ronanki, "Intelligence for the real world," Harv. Bus. Rev., vol. 96, no. 1, pp. 108–116, 2018.

[12] A. Marchini, C. Riganelli, F. Diotallevi, and B. Polenzani, "Label information and consumer behaviour: evidence on drinking milk sector," Agric. Food Econ., vol. 9, no. 1, Dec. 2021, doi: 10.1186/s40100-021-00177-5.

[13] F. Samsudin, S. A. Wahab, A. S. A. Latiff, S. I. W. Osman, N. F. M. Zawawi, and S. A. Fazal, "Strategic Direction and Sustainable Development in Franchising Organizations: A Conceptual Study," J. Sustain. Dev., vol. 11, no. 4, p. 89, Jul. 2018, doi: 10.5539/jsd.v11n4p89.

[14] M. J. Sánchez-Franco and M. E. Aramendia-Muneta, "Why do guests stay at Airbnb versus hotels? An empirical analysis of necessary and sufficient conditions," J. Innov. Knowl., vol. 8, no. 3, Jul. 2023, doi: 10.1016/j.jik.2023.100380.

[15] F. J. Cossío-Silva, M. Á. Revilla-Camacho, and M. Vega-Vázquez, "The tourist loyalty index: A new indicator for measuring tourist destination loyalty?," J. Innov. Knowl., vol. 4, no. 2, pp. 71–77, Apr. 2019, doi: 10.1016/j.jik.2017.10.003.

[16] D. Zhang and Y. Fang, "Are environmentally friendly firms more vulnerable during the COVID-19 pandemic?," J. Clean. Prod., vol. 355, Jun. 2022, doi: 10.1016/j.jclepro.2022.131781.

[17] D. V. Watkins and A. D. Clevenger, "US Political Leadership and Crisis Communication During COVID-19," Cogent Soc. Sci., vol. 7, no. 1, 2021, doi: 10.1080/23311886.2021.1901365.

[18] L. Kanski and J. Pizon, "The impact of selected components of industry 4.0 on project management," J. Innov. Knowl., vol. 8, no. 1, Jan. 2023, doi: 10.1016/j.jik.2023.100336.

[19] V. Lakshmi and B. Bahli, "Understanding the robotization landscape transformation: A centering resonance analysis," J. Innov. Knowl., vol. 5, no. 1, pp. 59–67, Jan. 2020, doi: 10.1016/j.jik.2019.01.005.

[20] S. Li, L. Gao, C. Han, B. Gupta, W. Alhalabi, and S. Almakdi, "Exploring the effect of digital transformation on Firms' innovation performance," J. Innov. Knowl., vol. 8, no. 1, Jan. 2023, doi: 10.1016/j.jik.2023.100317.

[21] S. Tiwari, "The Role of Artificial Intelligence in the Purchasing Behavior of Consumers in India," J. Bus. Manag. Inf. Syst., vol. 9, no. 2, pp. 29–32, 2022, doi: 10.48001/jbmis.2022.0902005.

[22] M. M. H. Chappin, M. V.D. van den Oever, and S. O. Negro, "An overview of factors for the adoption of energy efficient eco-innovation: The cases of the Dutch brewing and paper industry," J. Clean. Prod., vol. 275, Dec. 2020, doi: 10.1016/j.jclepro.2020.124122.

Systematic Literature Review on Artificial Intelligence-Driven Personalized Learning

Anas Usman Inuwa, Shahida Sulaiman*, Ruhaidah Samsudin
Faculty of Computing, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia

***Abstract*—Artificial Intelligence (AI) is widely used in various contexts, including education at different levels, such as K-12 (kindergarten through 12th grade) and higher learning. The impact of AI in education is becoming increasingly significant, making the academic sphere more effective, personalized, global, context-intensive, and asynchronous. Despite the publication of several systematic literature reviews, mapping studies, and reviews on the use of AI in education, there is still a lack of reviews focusing on personalized learning (PL) frameworks, models, and approaches at various levels especially the pre-university level for Science, Technology, Engineering, and Mathematics (STEM) subjects. To address this gap, our work presents a systematic literature review of AI-driven PL models, frameworks, and approaches published over the past ten years from 2013 to 2023, extracted from the Scopus database. This review focuses on the AI techniques used, personalized learning elements, components, attributes, and the possibility of replicating the technique in pre-university level studies, and gaps or prospects that will attract further research. The study reviewed 69 articles, downloaded via the Scopus database, and reported the most used AI techniques, PL components or factors, trends, and prospects for future research. The results show that most existing studies focus on higher learning that requires further research at the pre-university level. In addition, machine learning and deep learning are identified as the most suitable and frequent techniques besides other technologies, knowledge delivery, learners' needs, behavior and interest as the most required components for personalized systems in diverse fields. In terms of publication output by country, the study indicates that Switzerland, USA, UK, and China are leading contributors to PL research. Thus, this study calls for further research on AI-driven personalized learning that thoughtfully integrates educational theories, subject-specific content, and industry needs to enhance outcomes and learner satisfaction.**

Keywords*—Personalized learning; model; framework; approach; technique; systematic literature review; personalized learning components; artificial intelligence

I. Introduction

Academic research on artificial intelligence in education (AIED) has made significant progress over the past 30 years toward adopting a more complex learning paradigm that enables successful teaching and learning [1]. Education cannot be discussed today without mentioning the impact of Artificial Intelligence (AI). AI has had a widespread impact on all aspects of education, including its purpose, content, methods, and evaluation systems. Countries worldwide are making AI literacy education mandatory to enhance their understanding of AI. These include methods of using AI devices and services, understanding AI ethics, real-life AI convergence cases, and basic block coding. The curriculum is also being reformed to reflect AI, including its principles and the convergence of courses [2]. AI has the potential to tackle the major challenges faced by the education sector today, leading to innovative teaching and learning practices and ultimately speeding up progress toward achieving Sustainable Development Goals (SDGs). Nevertheless, these fast-paced technological advancements also bring various risks and challenges, surpassing policy debates and regulatory frameworks [3].

Educational technology literature has been exploring ways to incorporate AI into education to make education more personalized. Humans have unique characteristics that require the education sector to focus on individual-specific learning requirements, which is motivated by the desire to move away from the traditional "one size fits all" approach to a more personalized learning (PL) format [4], [5]. The key idea is to prioritize the students by making the system more "student-centered". PL aims to adjust the curriculum and instructions based on students' learning requirements and abilities, and to meet the demand of the modern workforce and address global challenges. Catering to a learner's needs will likely motivate the general student population. It is worth recalling that during the COVID-19 pandemic, educational systems rapidly transitioned to online platforms, prompting many institutions to adopt Content Management Systems (CMS) and Learning Management Systems (LMS) for the duration of the crisis and beyond. However, this rapid adoption has caused issues with productivity, course content, progress analysis, and alignment of learning with industry requirements [6].

PL is defined as "instruction that is paced to learning needs, tailored to learning preferences, and to the specific interests of different learners" [7]. The ongoing evolution of technology, coupled with the rapid advancement of AI, has significantly enhanced teaching and learning processes. AI facilitates a flexible, personalized, and efficient learning environment while also strengthening educational competencies through PL [8]. While there has been much research on learner control, the relationship between learner control and psychological ownership is not well understood [9]. Furthermore, it is still unclear how these concepts manifest in primary and secondary school education [10]. In order to personalize learning content, it is important to start with a classic instructional design model, which means that the learning environment must be designed to achieve a specific learning outcome. The learners will be engaged in the learning process and then assessed for their mastery or achievement of the targeted outcome. In addition, the environment must be adapted based on one or more features of learners to personalize the learning experience. This

*Corresponding Author

adaptation should be motivated by a desire to achieve educational outcomes. The National Academy of Engineering has identified the development of PL systems as a major challenge in the 21st century. Researchers in various fields have worked on different aspects of this challenge [11]. Personalizing learning involves creating a learning environment that considers the learner's prior knowledge, motivation, goals, beliefs, interests, skills, experience, and culture, among other factors. Instructional experience should be responsive to these features to enhance engagement in a learning task and improve performance. PL presents numerous dimensions and appeals to various disciplines, leading to a diverse body of research on this subject [12].

Learning analytics is vital in personalizing learning, which analyzes data from students and learning environments to support learning at various levels. It is a relatively recent field that has achieved a high level of maturity, particularly in its applications in higher education. However, there is a lack of research on learning analytics that focuses on other educational levels, such as high school [13]. Technology has been deeply integrated into modern society, impacting almost every industry. Basic science subjects are deemed tough and boring for students at K-12 since most of them cannot relate to their applications in the real world. Hence, the technology aspect of Science, Technology, Engineering, and Mathematics (STEM), especially Computer Science (CS), is changing the narrative. CS is trending as its applications and impacts are felt by all and sundry. Skilled workers related to CS are in high demand. In traditional education, students rely on instructors or textbooks for subject knowledge, which can vary. The lack of trained CS teachers at the K-12 (kindergarten through 12th grade) level can result in insufficiently prepared students, owing to the limited number of CS courses offered in high schools [14].

Many college freshmen are hindered from taking CS because of a lack of exposure to the field earlier in their high school time. This is because they have never taken a course on the subject or had limited knowledge of it and are therefore not aware of what it entails. Surveys indicate that 90% of parents believe that CS should be part of their children's K-12 curriculum, but only 40% of schools offer it. Additionally, most schools that offer CS include it as a high school elective that does not count toward graduation [15].

AI technology is used to establish PL environments by gathering, analyzing, and interpreting data from multiple sources to construct student learning profiles. However, there is limited research on integrating AI into education to improve teaching efficiency in Malaysia. AI technology can predict a student's learning potential, and this information can be used to create tailored content that aligns with an individual's goals and previous achievements [8]. This is a common issue in most developing and underdeveloped countries. Hence, this systematic literature review (SLR) aims to comprehensively review models of AI and PL in education at various levels, focusing on outcomes and prospects, to further examine the applicability of AI-driven PL at the pre-university level particularly for the CS subject. Subsequently, Section II presents the related work, Section III provides the details on the review process, while Section IV and V report results and discussion respectively, and finally Section VI concludes the study and its future research direction.

II. Related Work

This section deals with the analysis of related works on the broad topic in the form of SLRs, systematic mapping studies (SMS), and other forms of review studies, published in SCOPUS (a large multidisciplinary database of peer-reviewed papers) for the span of ten years, from 2013 to 2023. IT was accessed via the Universiti Teknologi Malaysia (UTM) subscribed databases. The search was recorded as shown in Table I, based on the identified keywords and search terms (ST): systematic literature review", "systematic mapping study", "personalized learning", and "artificial intelligence".

Fig. 1 shows the number of publications and the percentages resulting from the search terms. Firstly, ST1 "systematic literature review AND artificial intelligence AND personalized learning" yields 92 publications, out of which only 29 (47%) are in the domain of this study and within the range of 2013 to 2023. Secondly, ST2 "systematic mapping study AND artificial intelligence AND personalized learning" only had 6 (10%) publications, and finally ST3 "systematic literature review AND personalized learning AND education" with the highest total of 108 publications, out of which only 54 are related to computer science and only 26 (43%) are in the scope of this study. The publications were further selected to check their findings, key focus, educational level that considered higher learning (HL), K-12 (Primary and secondary levels), as well as the type of review presented (SLR, SMS, and other review protocol).

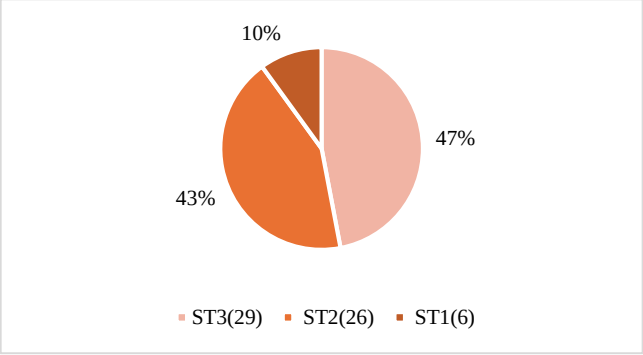


Fig. 1. Results by search terms (ST).

A. Issues and Gaps

Table II provides a brief overview of 20 studies, highlighting their key focus, educational level (HL, K-12), and types of study such as SLR and SMS. It aims to demonstrate that no existing work aligns precisely with the specific focus or title of this research study, indicating an unaddressed gap. This is particularly relevant when considering the pre-university level. Another gap addressed is the identification of major techniques and components that should be used in building personalized models and frameworks.

TABLE I. SEARCH RESULTS ON SLR, SMS, AND CLOSELY RELATED REVIEW PAPERS BASED ON KEYWORDS

Search Terms (ST)	ST1: (systematic literature review) AND (personalized learning) AND (artificial intelligence)					ST2: (systematic mapping study) AND (artificial intelligence) AND (personalized learning)					ST3: (systematic literature review) AND (personalized learning) AND (education)				
Category/ Year	RV	JA	CP	BK	CR	RV	JA	CP	BK	CR	RV	JA	CP	BK	CR
2023	6	4	5	0	0	0	2	0	0	0	3	6	0	1	0
2022	6	0	0	1	0	1	0	0	0	0	5	0	1	0	0
2021	2	1	0	0	0	0	1	0	0	1	1	0	3	0	0
2020	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0
2019	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
2018	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2017	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
2016	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2015	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2013	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
Sub Total:	15	6	7	1	0	1	3	0	0	2	10	6	8	2	0
Total:	29					6					26				

Note: RV = Review, JA = Journal Article, CP = Conference Paper, BK = Book Chapter, and CR = Conference Review. The acronyms are derived from nomenclatures on the SCOPUS summary pane on the left side of the window after a successful search display.

In summary, 15 out of the 20 focused on higher learning [5],[16]-[20], [23]-[26], [28][29], [32]-[34]; four [21], [27], [30], [31] considered K-12 level, and only one publication [22] considered both K-12 and higher learning. The limited number of reviews within the selected range shows the pressing need for research in the area. The following section describes 10 elements that can be considered when designing personalized learning models and frameworks.

Teachers, schools, districts, states, and technology developers have attempted to create PL experiences for students. However, there is no consensus on the actual definition of personalized learning, and different designs may include multiple components. This variability can make it difficult to study personalized learning and how it can be designed based on students' characteristics to reliably achieve specific learning outcomes [7].

B. Personalized Learning Elements

Ten learning elements are identified here for designing a befitting PL framework or model extracted from the literature as cited.

1) Learning paths. A learning path refers to customizing an adaptable journey designed to meet the unique needs, preferences, and progress of individual learners. Machine learning can analyze a range of factors, including learners' goals, interests, prior knowledge, and performance, to create a customized curriculum and better learning outcomes. This dynamic approach enables customization of learning experiences, content, and pace, ensuring that each learner receives focused and pertinent educational material. This can enable the system to categorize learners based on different factors and recommend appropriate learning choices and improved outcomes [35]. The system can be powered by innovations, such as data mining, which can extract valuable information from large, unstructured, and random datasets. Adaptive ability of learners with similar characteristics can be identified, classified, and a learning path can be created for future learners [36]. Additionally, pinpointing the strengths and weaknesses of a particular learner can help eliminate learning difficulties.

2) Learner model. Learner models for classifying learners based on defined criteria. For example, learners can be classified based on their knowledge level as Basic or Beginner, Average or Middle, and Advanced or Expert.

3) Learners' metacognitive aspects. This involves designing features, activities, and feedback mechanisms that help students develop and use metacognitive skills like planning, monitoring, and evaluating their learning. This process turns a passive learning environment into a self-regulated, reflective, and adaptive experience [5].

4) Learners' queries. Learners' progress and behavior, including encouraging them to ask questions and addressing their concerns, is essential in creating a personalized and engaged learning environment. Timely responses promote proactive learning [37].

5) Learners' characteristics. This includes learners' profile, personality, and style, in which the system should consider various attributes of learners, such as their preferred learning styles, interests, and socio-cultural background. Adapting teaching strategies to align with these characteristics can enhance engagement and comprehension [16].

TABLE II. FINDINGS OF EXISTING WORKS

Study	Description of Findings	Key Focus	Level	Study Type
[5]	It is a systematic literature review paper investigating issues related to learner diversity, the PL features, the methods used, applications in higher learning, and the impact of implementation.	Learner Diversity	HL	SLR
[16]	The research explores the personalization of vocational education within Indonesia's higher education. It introduces the Personalized Blended Learning model as a solution to the limitations of the "one-size-fits-all" approach. It aims to enhance PL experiences.	Personalizing Vocational Education	HL	S/R
[17]	The paper discusses the challenges and opportunities of blended learning in vocational education, emphasizing the importance of combining face-to-face and online learning to enhance teacher training.	Teacher Training for Blended Learning	HL	DBR
[18]	They examined the use of algorithms in higher education decision-making processes, focusing on their impact on students and educators. It highlights the importance of designing these algorithms with a human-centered approach to ensure fairness, transparency, and ethical considerations.	Algorithms for decision making in higher learning	HL	SLR
[19]	The SLR research focused on learners' psychological and emotional states, along with their abilities and motivations, while also examining the role of personalized instruction in e-learning.	Learners' Psychological and Emotional State	HL	SLR
[20]	The SLR research examined the impact and application of Convolutional Neural Networks (CNNs) in education, particularly in areas such as face detection and student recognition.	CNN and for students' detection	HL	SLR
[21]	The SLR identified Special Educational Needs (SEN) as a form of Personalized Learning for primary and elementary students in the digital era, particularly those with attention challenges who depend on teacher support.	Learners with special educational needs	K-12	SLR
[22]	The SLR explored personalization strategies and gamification techniques in Virtual Reality (VR) to enhance educational outcomes in schools.	Gamification and virtual reality	HL/K-12	SLR
[23]	The scoping review analyzed ChatGPT's impact on higher education for researchers, teachers, and students globally, highlighting its benefits and concerns regarding accuracy, reliability, and academic integrity.	ChatGPT and academic integrity	HL	SSR
[24]	The SLR paper explored the role of AI chatbots as teaching assistants, aiding both educators and students, with a particular focus on their contribution to personalized education in higher learning.	Chatbots as teaching assistance	HL	SLR
[25]	The SLR examined how AI and Data Science integrate sustainable development into education, enhancing personalization, learning prediction, and dropout prevention.	Data Science and Sustainable Development	HL	SLR
[26]	The research explored computer-based methodologies for teaching English, advocating for personalized learning using AI, Text-to-Speech (TTS), and Natural Language Understanding (NLU).	Language Teaching	HL	SLR
[27]	This systematic literature review examines the personalization of e-learning models to adapt to students' learning styles and pace in Indonesian schools.	E-learning models	K-12	SLR
[28]	This SLR explores the essential components for constructing a learner model grounded in learning theories for adaptive e-learning systems.	E-learning systems	HL	SLR
[29]	The SLR assesses research on pedagogical agents using empirical data to enhance personalization in e-learning, emphasizing their flexibility, diversity, simulation capabilities, and overall impact.	Pedagogical agents	HL	SLR
[30]	The research explored the application of AI concepts in science education at the secondary school level.	AI and Science education	K-12	SLR
[31]	The SLR examines the lack of consensus on DPL (Digital Personalized Learning) technology in primary and secondary education by analyzing various empirical studies.	DPL technology	K-12	SLR
[32]	The SLR explored AI methods for identifying learner traits, structuring content, recommending learning paths, and evaluating their benefits and limitations.	AI for learner identification	HL	SLR
[33]	This SMS examined the incomplete implementation of adaptive learning systems and their failure to fully address students' diverse needs, proposing improved AI integration as a solution.	Adaptive learning and students' diversity	HL	SMS
[34]	It highlights the need for more empirical studies to validate adaptive learning models and suggests exploring new adaptive techniques and technologies to enhance personalized learning experiences.	Adaptive learning environment	HL	SMS

Note: SLR (Systematic Literature Review), SMS (Systematic Mapping Studies), SSR (Systematic Scoping Review), S/R (Survey Research), DBR (Design-Based Research), HL (Higher learning, Tertiary education), and K-12 (Primary and Secondary School, Kindergarten 12).

6) Learners' knowledge level. The system should identify and tackle learners' background, prior knowledge, feedback, and current knowledge level that involves adapting educational content and activities to address gaps and challenge the learner based on their understanding and proficiency [38].

7) Flexible pacing. This is achieved by designing the system to allow each learner to move through content at a pace that matches their individual learning needs, prior

knowledge, and personal goals. Sturgis and Patrick also emphasized that well-designed systems should enable tracking individual progress and suggest customized learning pathways.

8) Systems and tools. There are available tools that can be incorporated into PL systems for providing learning resources and other supports, like Khan Academy, Knewton, and ALEKS.

9) Smart learning environment. Effective smart learning environments should promote PL by fostering sensitivity, suggestions, self-reflection, assessment, constructive criticism, and enthusiasm. Examples include Intelligent Tutoring Systems, Virtual and Augmented Reality, SNAPP, Minecraft, etc.

10) Learning analytics (LA) and data mining (DM): Data Mining (DM) and Learning Analytics (LA) are powerful tools that support the learning process by collecting and analyzing data. They can help evaluate learning methods, predict the expected performance, and identify areas for improvement [39]. LA is particularly useful for analyzing data from diverse learning environments and can be used to create PL activities tailored to the needs and goals of each learner. As interest in LA continues to grow, it has the potential to transform learning by promoting personalized learning experiences.

Based on these twenty related works in Table II, it is evident that none of them focused primarily on reviewing or examining the existence of PL frameworks or models, and approaches that are subject-specific and focused on the peculiarities of the PU level. Most of the studies focused on advanced levels, and the use of AI in PL is less addressed in the studies. As a result, the next phase of this SLR will extensively review the AI-driven PL models, outcomes, and prospects. Additionally, it will explore further issues and gaps, as well as the components of PL that are appropriate for the pre-university level.

III. Review Process

This review adopts the SLR method outlined by Barbara Kitchenham [40]. These steps are illustrated in Fig. 2 and further explained in the sections ahead. The search process consisted of four stages. The first step was to formulate the research questions, where three research questions were formulated as presented in the sections ahead. The second step was the selection of articles using defined keywords on the identified resources (Scopus-indexed publications only). The search results were then filtered and recorded accordingly.



Fig. 2. Four stages of the review process based on an existing protocol [40].

The articles were then sorted systematically and organized by publication year and type (article, conference, review, conference review, book, or book chapter). The third step is on the inclusion and exclusion criteria: only articles reported in the English language that can be comprehended easily, and adherence to keywords is considered. Moreover, articles that are considered related to PL with AI techniques and centered on either components or factors of PL, model, framework, or approach are also selected. Only 69 articles were considered and selected for the final stage. Finally, the selected articles were used to answer the research questions outlined.

A. Research Questions

In framing the research questions, some requirements or criteria were considered, including population, intervention, comparison or justification, and outcomes, as presented in Table III. These requirements paved the boundaries upon which the direction of search and results were guided.

TABLE III. Criteria for Research Questions

Criteria (Component)	Scope (Coverage)
Population	Articles that propose a framework, model, or approach for PL at any level or consideration
Intervention	model, framework, or approach that addressed the application of AI techniques for PL at any level.
Comparison/ Justification	Strengths, weaknesses, and prospects for each approach.
Outcomes	Problems/issues/gaps addressed on the applicability of AI techniques suitable for PL and the PL components used, and at what level of learning?

Based on the criteria outlined in Table III, the research questions (RQs) were formulated as follows:

1) RQ1. Do the articles discuss the use of AI Techniques in PL, and what are the components of PL involved?

2) RQ2. Are these techniques clearly defined? Can they be replicated in pre-university computer science courses?

3) RQ3. Do the articles present gaps and points of interest attracting future research?

B. Search Process

The main purpose of the search was to obtain existing research on AI applications in the educational domain in the form of PL. Explicitly, PL models, frameworks, and approaches are the center of the review, to see which AI techniques were used and at what level were they applied, and to see the applicability of the techniques at the pre-university level. What are the PL components involved in this process?

1) The search process was carried out using the Scopus database. When papers are not available for download, they are then viewed and downloaded from the publisher's site, ResearchGate, or Google Scholar.

2) Organizing and categorizing articles based on keywords, type of publication, year of publication, and related information.

3) Recording search results according to keywords and publication type.

4) Refining the search involves the search process, and the choice of words is refined to ensure that all relevant papers are obtained.

Table IV presents the list of the seven keywords used in the search process, K1 to K7.

TABLE IV. SEARCH KEYWORDS

Keyword	Representation
K1	Personalized Learning
K2	Approach
K3	Model
K4	Framework
K5	Secondary School
K6	High School
K7	K-12

The search process from the repository considered the content and suitability of the articles by their abstract, publication title, and keywords, derived using the outlined search strings (SS1 to SS7) in Table V, which are partial combinations of the keywords in Table IV.

TABLE V. SEARCH STRINGS

Code	String
SS1	Personalized Learning Approach
SS2	Personalized Learning Framework
SS3	Personalized Learning Model
SS4	Personalized Learning Model OR Personalized Learning Framework
SS5	Personalized Learning Models AND High School OR Secondary School or K-12
SS6	Personalized Learning Approach AND High School OR Secondary School or K-12
SS7	Personalized Learning Framework AND High School OR Secondary School or K-12

Table VI presents the number and its percentage for each search string SS1 to SS7 versus the five document types. The highest found string is for SS5 that is 105 works in the scope of models and high school or secondary school or K-12.

TABLE VI. INITIAL SEARCH RESULTS FOR EACH DOCUMENT TYPE VERSUS SEARCH STRING

Type	SS1	SS2	SS3	SS4	SS5	SS6	SS7	Total 100 (%)
Article	10	3	6	23	43	5	13	103 (39.16)
Conference Paper	9	9	13	23	45	3	16	118 (44.87)
Conference Review	0	0	2	2	15	1	5	25 (9.51)
Book Chapter	0	1	3	4	1	2	1	12 (4.56)
Book	1	1	1	1	1	0	0	5 (1.90)
Total	20	14	25	53	105	11	35	263 (100)

Fig. 3 visualizes the percentage of the publication type, with conference papers as the highest in the list with 118 (44.8%) publications, followed by journal articles with 103 (39.16%) publications, 25 (9.51%) conference reviews, 12 (4.56%) book chapters, and five books (1.9%).

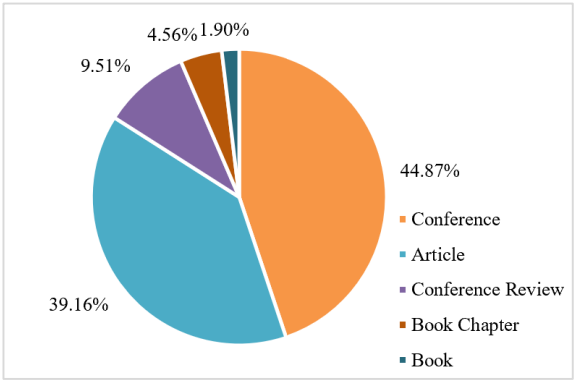


Fig. 3. Percentage of articles by publication type.

C. Inclusion and Exclusion Criteria

At this stage of the SLR, many criteria were considered for inclusion or exclusion. First, the articles were sorted by document type (regardless of the publishers, if the paper is found in the Scopus database), as displayed in Table VI. Second, only articles that were written or reported in English were selected. Third, the selected articles had to be from computer science, information technology, sciences, technology, or topics related to learning, AI, and education. Finally, the selected articles must not negate the search keywords and strings, as enumerated in Tables IV and V, respectively.

Articles that did not meet the inclusion criteria were excluded; articles that did not appear in a relevant field but in conformity with the search criteria were excluded; and articles that did not apply AI techniques in the framework, model, or approach were excluded, except for a few that included other technologies closely related to AI.

D. Quality Evaluation

To ensure the quality and conformity of the relevant articles, Table VII serves as a guide for choosing relevant studies based on the research questions outlined earlier. For each research question, the article should answer yes or no. Those who were neither "Yes" nor "No" were excluded from the study. Those who partially answered were further evaluated for inclusion or exclusion.

TABLE VII. EVALUATION BASED ON RESEARCH QUESTIONS

Research Question (RQ)	Possible Answer
RQ1: Do the articles discuss the use of AI Techniques in personalized learning, and what are the components of PL involved	Yes/No/Partially
RQ2: Are the techniques clearly defined? And can they be replicated at pre-university level computer science courses?	Yes/No/Partially
RQ3: Do the articles discuss gaps and points of interest attracting future research?	Yes/No/Partially

Fig. 4 shows the flow of the evaluation process and how the articles were evaluated to ensure conformance. The classification was adapted to evaluate the articles further, as follows:

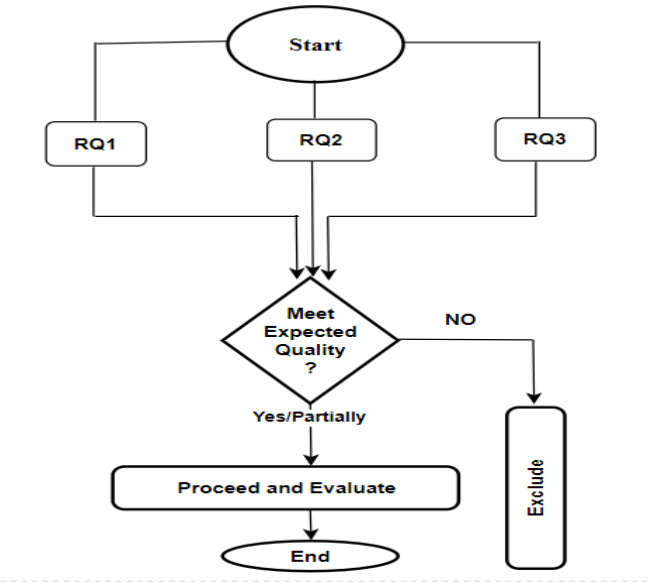


Fig. 4. Evaluation process flow.

1) The article's title or abstract mentioned the framework, model, or approach within a specific field of learning.

2) AI techniques or technology used for personalization were clearly stated.

3) The paper must equally discuss issues and gaps for further research.

IV. Results

The SLR reviews AI techniques and other associated technologies applied in developing PL frameworks, models, and approaches in different educational settings and levels. The SLR considers the frequencies of the techniques and the most considered elements (factors and components) of PL in the models, framework, or approach. In the initial stage of the search process, many results were obtained using the search strings listed in Table V but were narrowed by field and publication dates. For example, the search keyword S5 (personalized learning models AND high school OR secondary school or K-12) yielded 207 documents; only 107 were relevant and further screened to obtain the final selections.

The 69 publications selected are listed in the Appendix, capturing: focus of the article (model, framework, or approach), the PL elements (components, factors, approach) involved, the technique and technology applied, and the educational level (higher learning, K-12, or just general) considered. The results are further discussed in the next section; however, some graphs are presented here.

The articles selected according to the search strings are listed in Table VIII, with S5 (14 out of 69) having the highest number of publications, followed by S3 and S7 with 12 publications respectively, S1 with 10, S4 with 9, S2 with 8, and S6 with at least four publications.

TABLE VIII. Selected Articles by Search Strings

Search String	Selected Articles	Total
S1	[41][42][43][44][45][46][47][48][49][50]	10
S2	[51][52][53][54][55][56][57][58]	8
S3	[59][60][61][62][63][64][65][66][67][68][69][70]	12
S4	[71][72][73][74][75][76][77][78][79]	9
S5	[80][81][82][83][84][85][86][87][88][89][90][91][92][93]	14
S6	[94][95][96][97]	4
S7	[98][99][100][101][102][9][103][104][105][106][107][108]	12
Grand Total		69

Fig. 5 provides a graphical representation of the number of publications selected for the review, along with their respective publication years, 2013 to 2023. In total, 69 publications were included, with 2013 and 2014 having the lowest number of publications, and 2021 having the highest. The curve shows the trend in the number of publications over time.

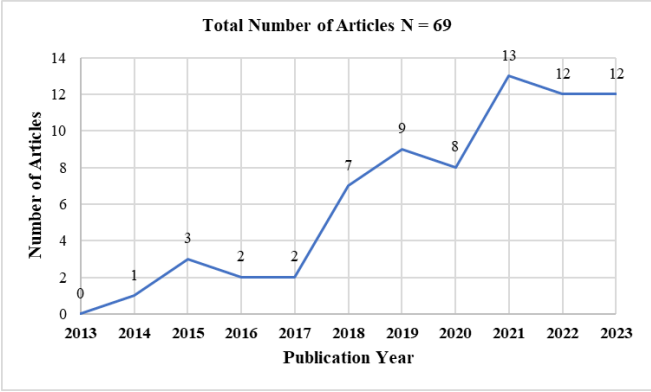


Fig. 5. Number of publications versus year of publication for the 69 studies.

Fig. 6 presents the chart for learning levels for the selected publications in the form of K-12 (kindergarten or primary and secondary school level), HL (higher learning: colleges, polytechnics, and university levels), and GL (general level with no emphasis on the study level or capacity of the learners).

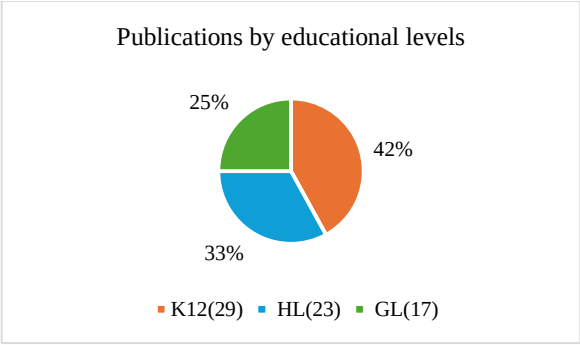


Fig. 6. Number of publications by level of education in the selected studies.

Fig. 7 shows the top five countries based on publisher locations, with Switzerland having 48.40%, the USA (22.60%), and the UK (16.00%), are the leading contributors to research in personalized learning, education, artificial intelligence approaches, models, frameworks, and their applications in education. China and India each have 6.50%, indicating their emerging roles in this scope of study.

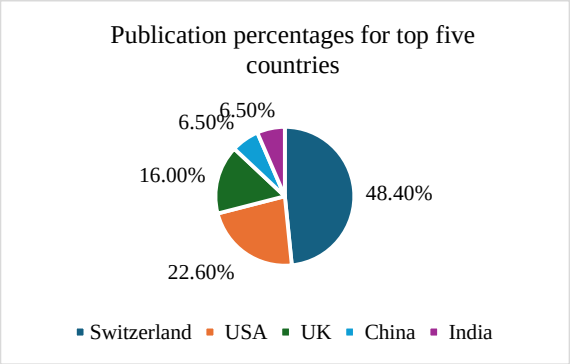


Fig. 7. Top countries with publications on personalized learning education with concepts of artificial intelligence.

This distribution highlights the significant investment and focuses these countries have on advancing educational technologies and AI-driven PL systems, while also underscoring the need for increased research contributions from developing countries to enrich the global research landscape.

Fig. 8 presents the top 10 publishers by the number of publications within the scope and data analyzed. Sustainability (Switzerland) dominates with 11 publications, demonstrating its prominent role and high level of scholarly activity in the area. ACM International Conference Proceedings Series and IEEE Access follow, each contributing 6 publications, indicating their strong engagement in disseminating research in the area. Other publishers like Artificial Intelligence in Medicine and Applied Sciences (Switzerland) have each possessed four publications, while several others, including Technology, Knowledge and Learning, Computers and Education, and Computers in Education: Artificial Intelligence, each have around 3 publications.

Finally, Computers in Biology and Medicine rounds off the list with two publications. This distribution highlights the significant role of multidisciplinary and technology-focused journals and proceedings in advancing PL and AI technology, with Swiss publishers particularly standing out in their contribution.

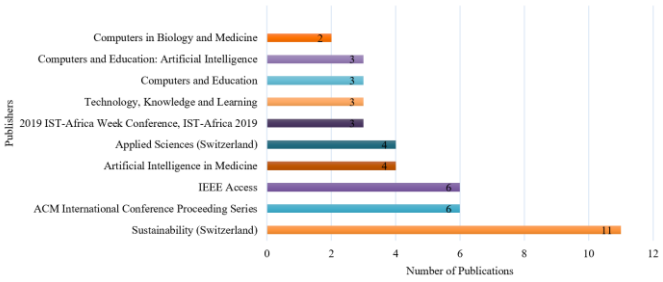


Fig. 8. Top 10 publishers by number of publications (two and above).

V. Discussion

As outlined in the earlier sections, the selected articles were intended to answer the three research questions guiding the SLR's findings. The percentages of articles concerning each research question are RQ1 (39%), RQ2 (42%), and RQ3 (19%), as shown in Table IX.

TABLE IX. Total Number of Articles Answering each Research Question

Research Question	Total	%
RQ1: Do the articles discuss the use of AI techniques in PL, and what components of PL are involved?	27	39
RQ2: Are the techniques clearly defined? And can they be replicated at pre-university level computer science courses?	29	42
RQ3: Do the articles discuss gaps and points of interest attracting future research?	13	19

A. Research Question 1 (RQ1)

RQ1 is stated as: "Do the articles discuss the use of AI techniques in PL, and what components of PL are involved?"

To answer this question, the Appendix presents all the articles in this review, the techniques, and elements of PL involved in the framework, model, or proposed approach. Fig. 9 is a graphical representation of the AI techniques and related technologies used in developing PL frameworks and models captured in the publications selected for this study.

The ten most frequent techniques were considered, in which machine learning tends to be the highest, followed by deep learning. Although other technologies (OT) are seen to be at the peak of the curve, it is a collection of various technologies, including big data, smart technology, digital, and mobile technology, as used in the context of the selected articles, which we deemed are not solely AI techniques, but contains elements of AI and its approach.

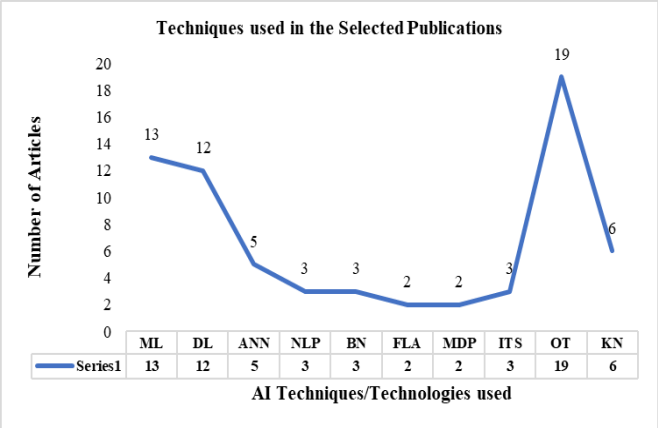


Fig. 9. AI techniques and other related technologies for building PL models and frameworks.

Key: ML (Machine Learning), DL (Deep Learning), ANN (Analytical Neural Network), NLP (Natural Language Processing), BN (Bayesian Network), FLA (Fuzzy Logic/Agent), MDP (Markov Decision Process), ITS (Intelligent Tutoring System), OT (Other Technologies), and KN (Knowledge Representation).

Although technology serves as the primary element in enabling and enhancing PL, there is a lack of consensus regarding the essential components of a dynamic PL approach. Furthermore, technology enables customization of learning experiences, but still no consensus on the essential elements required to create a distinct and effective learning experience tailored to individual learners [46].

Fig. 10 shows the PL components and their frequencies. The existing study [56] presented an additional ten attributes and components of PL as noted under the figure.

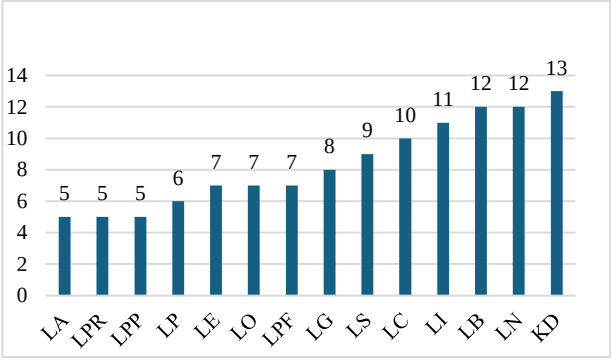


Fig. 10. Frequencies of PL components extracted from the selected studies.

Note: LPP (Learning Pace/Pacing), LP (Learning Path), LB (Learner/Learning Behavior), LS (Learning Style), LC (Learning Content), LPF (Learner Profile), LI (Learning Interest), LPR (Learning/Learners Preference), LN (Learners Need), LG (Learning/Learners Goal), LE (Learner/Learning Experience), KD (Knowledge Delivery), LO (Learning Outcome), LA (Learners/Learning Attitude).

Culture, emotional or mental state, socialization, motivation, learning preferences, prior knowledge, educational background, learning and cognitive style, navigation, and learning goals. Table X presents a brief description of fourteen of the components alongside reference articles for each. These components adapt dynamically based on the needs and preferences of the target user, educational level, and other considerations. They can be selected at random to suit the issues and purpose of development (framework, model, or approach for a particular educational level). The terms and names vary from one author to the other (since research on AIED is still new and dynamic). Knowledge delivery had the highest frequency, then learners' behavior, and needs towards learning, learners' interest, until LA, LPR, and LPP, which had the lowest frequencies, see Fig. 10.

Table X and Fig. 10 indicate that nearly all reviewed publications highlight a specific component related to personalized learning. The study concludes that knowledge delivery, learner behavior, and learner needs are the most emphasized aspects, whereas learning pace, attitude, path, and preferences receive less attention. This suggests that prioritizing the former may inadvertently diminish the significance of the latter.

B. Research Question 2 (RQ2)

RQ2 is stated as "Are the techniques clearly defined? Can they be replicated in pre-university level Computer Science courses?" The answer is "yes". Some of the articles clearly defined the techniques and justified them in the context of why such techniques were used in their work and specifically mentioned the level of study considered. However, some works have stated it as general, without targeting any educational level. Although in Fig. 6, K-12 level (primary, secondary, or pre-university), possesses the largest sector of the three, not all articles centered around AI techniques for PL, as they also consider other technologies.

Furthermore, the studies only point to the need for research on the K-12 educational level as a gap left unfilled; most of the publications are more about the feasibility of AI at that level, considering their peculiarities. Some of the works include the BPRMF model based on deep learning, neural networks, and data mining, proposed by Peng [93] to be applied to the problem of recommendation for the Civics and Political Science course, while another work proposes a system by hybridizing Visual/Aural/Read, Write/Kinesthetic (VARK) presentation or gamification and exercises difficulty scaffolding through skipping or hiding, or reattempting using Deep Q-Network Reinforcement Learning (DQN-RL) [81]. Another study proposed a novel approach for evaluating the co-learning performance of human intelligence (HI) and machine intelligence (MI) using a Knowledge Graph-based genetic fuzzy agent technique and a genetic algorithm [60].

Maddalora [77] introduced a PL model that recommends the Shortest Learning Sequence (SLS) to remediate students with learning difficulties, using Item Response Theory (IRT). On the other hand, a study provides an in-depth analysis of English course recommendation techniques through a combination of the bee colony algorithm and neural network algorithm, a deep learning model combined with collaborative filtering, to recommend suitable courses for users [88]. Other studies have used deep learning [62], [93], [43], [42], [72], neural networks [65], [105], [88], intelligent tutoring systems [46], [44], [92], natural language processing [53], [92], [87], knowledge representation [43], [89], [91], [66], and other techniques that are captured in Fig. 9.

TABLE X. DESCRIPTION OF FACTORS USED IN THE SELECTED STUDIES

Comp.	Description	Related Works
LPP	The speed at which a learner progresses through the materials is based on their capabilities and understanding.	[73][46][101] [108] [41]
LP	A strategically designed sequence of courses or modules aimed at guiding learners through a specific subject or program, ensuring a structured and progressive learning experience.	[44][73][46][65] [64] [47]
LB	Actions and responses demonstrated by a learner in a learning environment aim to cultivate positive and productive engagement.	[91][61][44][67] [65][42][86][85] [50][92] [47][81]
LS	An individual's preferred approach to acquiring and processing information is often categorized as visual, auditory, or kinesthetic.	[52][95][102][105][66][81][48] [56][47]

Comp.	Description	Related Works
LC	The material or resources that are used for educational purposes, including text, multimedia, and interactive elements.	[78][75][87][84] [27][68][70][79] [49][81]
LPF	Learner's characteristics, preferences, strengths, and weaknesses serve as a foundation for personalizing instruction.	[95][70][61][94] [46] [64][56]
LI	Specific topics or subjects that capture a learner's curiosity and engagement influence PL experiences.	[97][43][76][46] [90][45][58][85] [50][47] [48]
LPR	Individual choices and inclinations of a learner in terms of instructional methods, content formats, and assessment styles.	[44][41][93][50] [56]
LN	The individual needs and requirements of a learner, focusing on bridging gaps in understanding and skill development.	[43][95][79][46] [90][41][103][67][86][50][47] [107]
LG	The desired outcomes or achievements that a learner aims to attain through the learning process.	[61][9][44][46] [41] [96][85][56]
LE	The overall encounter a learner has during the educational journey encompasses interactions, challenges, and outcomes.	[78][84][43][27] [95] [70][83]
KD	The strategies and tools employed to deliver information and enhance learning experiences, according to individual preferences.	[78][87][52][27] [70][83][61][9] [76][49][94][54] [55]
LO	It refers to quantifiable outcomes and accomplishments that reflect a learner's proficiency in specific knowledge or skills.	[78][87][52][84] [68] [95][70]
LA	Addresses learners' mindset, approach, and attitude toward the learning process, which significantly impact engagement and success.	[95][61][44][46] [86]

Note: Comp stands for components (see also the note below Fig. 10)

C. Research Question 3 (RQ3)

Addressing RQ3, which is "Do the articles discuss gaps and points of interest attracting future research?", the answer is "yes". The selected publications identified existing research gaps, highlighted their contributions, and proposed future directions for further studies, as outlined in the abstract, conclusion, and future research sections. Table XI summarizes ten publications, detailing their contributions and recommended areas for future research.

Additionally, other studies have presented varied perspectives on the methods and requirements for personalizing education at the K-12 level. An existing study suggests gamification of learning content to drive better outcomes [81], while another study suggests further research to examine the implementation and impact of PL strategies in different areas and educational levels, as they only focused on higher learning in Saudi Arabia [41].

Xia and Cheng [70] highlighted four issues: lack of holistic learning data, lack of an acceptable learning profile, equipping educational robots is a big challenge, and there is no function for choosing proper strategies. In addition, some studies have suggested the need to explore key technologies, such as big data, data mining, and recommendation algorithms, for supporting PL [67], and the need for researchers to explore the learning outcomes of PL models based on specific practices and strategies [78].

VI. Conclusion and Future Direction

PL involves tailoring teaching approaches at both macro and micro levels to align with the unique needs of individual learners and is a focal point in global educational policy initiatives. Despite its widespread adoption, the methods of personalization vary, accommodating diverse priorities and influencing transformations at both individual and societal scales. The underlying philosophy emphasizes placing the learner at the core of the education system. However, upon closer examination, it becomes apparent that while these policies aim to diminish educational disparities, they occasionally contribute inadvertently to their exacerbation. This calls for a deeper understanding of what PL entails, its components, factors, target level (learners' level and need), policies, mission, vision, and goals of a particular country in terms of education, economy, and industrialization.

This study focused on the PL elements (components, factors, and technologies) involved in building frameworks and models at various educational levels. This study also presents the most frequent techniques and components used in 69 studies. It is evident that most studies concentrate on higher learning and proposals for different approaches to PL, regardless of educational policies, goals, and targets of stakeholders, but on mere assumptions. There is a lack of research that considers industrial demand, marketability of courses, tolerance in competitiveness, and employment chances upon one decision in choosing what to study or what to personalize. Personalization enhances one's technical skills, and hence, high chances of employability and adaptability, promoting learning autonomy and readiness for the workforce. Overall, PL promotes greater academic gain by narrowing learners' tasks to a specific target and promoting expertise. The scope of the study used mainly the Scopus database and other supporting primary research in the area to avoid voluminous and duplicate information from other sources, since Scopus has larger coverage. It can then be one of the limitations of this study as other databases may contain additional information regarding the area, and other researchers can explore further.

Advancements in AI are scaring nontechnical minds about job displacement, as many employees forecast that machines built on AI technology may render many people jobless. In contrast, AI has been proven to promote economic growth, boost productivity, enhance skills, and create new jobs. Today, information and communication technology and ICT-related industries have the highest prospects in Computer Science-related fields and subfields, especially those specializing in AI, Software Engineering, web development, and data analytics. Unfortunately, from the 69 selected studies, almost none focused on Computer Science and related fields for preparing

students at the pre-university level to study Computer Science when they enter the tertiary level.

Our future work will focus on designing a framework for PL at the pre-university level, to prepare students for various fields of Computer Science that the world needs today. The framework will prioritize and expose students to various opportunities and industrial needs, awaiting those who have studied Computer Science and have catalyzed solving issues using computer-based technologies.

TABLE XI. TEN SELECTED STUDIES: PROBLEM, CONTRIBUTION, TECHNIQUE, AND PROSPECT

Ref.	Identified problem	Paper Contribution	Applied Technique	Potential gap and prospects
[9]	The absence of psychological ownership of learning content among students and teachers.	Perception and assessment framework for the I-Learn portal.	Non-AI (Educational Technologies and Conjecture Mapping)	The necessity of creating tools for assessing and validating psychological ownership.
[43]	Standardized teaching approaches and ineffective skill development strategies in colleges and universities.	Proposed a personalized learning model utilizing IoT and deep learning algorithms.	AI-IoT integration with clustering algorithms, hierarchical clustering, collaborative filtering, and AI-based recommendations	The need for a flexible database for recommendations and content adjustments, enabling the integration of diverse materials.
[55]	The absence of study ownership and the impact of PL implementation in universities.	Proposed a framework on students' study ownership and learning outcomes.	Knowledge management through thematic analysis.	A necessity for research on the impact of personalization within authentic educational environments.
[57]	The ineffectiveness of traditional systems in tracking student needs during emergency remote education.	Proposed a framework for tertiary learning during and beyond the pandemic era.	Computer Technology	The necessity for research to validate the framework through longitudinal studies, utilizing technology for PL across all disciplines.
[62]	The absence of collaborative learning and limited facilitator-learner engagement in addressing students' needs.	Proposed a model centered on learner foundations, resources, community involvement, and diverse learning approaches.	Deep learning algorithms and learning analytics.	Future research should employ an alternative classifier to assess and validate the proposed approaches.
[73]	The absence of customized experiences that enhance students' learning journeys.	The research proposes a Personalized model to standardize entry-level requirements for health schools, emphasizing.	Knowledge Management.	Future research should incorporate evaluation and assessment components into the model and explore design enhancements to improve learning
[80]	Traditional one-size-fits-all systems are characterized by inefficient scheduling methods and a lack of individualized learning approaches.	An adaptive large neighborhood metaheuristic search integrated with an integer linear programming model.	Integer linear programming approach and Metaheuristic approach	Future research should explore modifications in school structure and staffing, along with extensions to accommodate longer timeframes.
[88]	Current research is insufficient for practical implementation, lacking sensitivity and collaborative strategies.	The research article introduces a novel recommendation technique that integrates bee colony optimization with neural network algorithms.	Bee colony, deep learning algorithm, and neural network techniques.	Future research should investigate user behavior analysis components, content evaluation mechanisms, and the integration of NLP for semantic analysis.
[106]	Lack of preparedness and absence of effective strategies for technology-enhanced learning.	A framework for Digital Learning Implementation.	Delphi Method	Future researchers should explore the practical applications of the proposed framework while prioritizing learners' needs.
[109]	Highlights the lack of learner data tracking, the dynamic nature of learners, and the insufficient development of computational thinking skills in existing studies.	A model that is supported by human-computer cooperation.	Big data technology and Machine Intelligence.	Future studies should expand on the proposed work to address the diverse learning needs of students in rural areas, both during pandemics and in post-pandemic settings.

ACKNOWLEDGMENT

This study is partly supported by the Centre for Advancement in Rural Education Information (iCARE) UTM-KEJORA Cost Centre No. 4B905 and its matching grant from UTM, Cost Centre No. 05M01.

REFERENCES

[1] Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence Unleashed: An Argument for AI in Education. Pearson.

[2] S. Paek and N. Kim, "Analysis of worldwide research trends on the impact of artificial intelligence in education," *Sustain.*, vol. 13, no. 14, 2021, doi: 10.3390/su13147941.

[3] F. Miao, W. Holmes, R. Huang, H. Zhang, and Unesco., *AI and education : Guidance for policymakers.*, vol. 3. 2200.

[4] A. Bhutoria, "Personalized education and Artificial Intelligence in the United States, China, and India: A systematic review using a Human-In-The-Loop model," *Comput. Educ. Artif. Intell.*, vol. 3, no. January, p. 100068, 2022, doi: 10.1016/j.caeai.2022.100068.

[5] R. I. Fariani, K. Junus, and H. B. Santoso, "A Systematic Literature Review on Personalised Learning in the Higher Education Context," *Technol. Knowl. Learn.*, vol. 28, no. 2, pp. 449–476, 2023, doi: 10.1007/s10758-022-09628-4.

[6] M. L. A. P. Liyanage *et al.*, "AI Solution to Assist Online Education Productivity via Personalizing Learning Strategies and Analyzing the Student Performance," *2022 IEEE 13th Annu. Ubiquitous Comput. Electron. Mob. Commun. Conf. UEMCON 2022*, pp. 609–615, 2022, doi: 10.1109/UEMCON54665.2022.9965735.

[7] M. L. Bernacki, M. J. Greene, and N. G. Lobczowski, *A Systematic*

Review of Research on Personalized Learning: Personalized by Whom, to What, How, and for What Purpose(s)?, vol. 33, no. 4. Educational Psychology Review, 2021. doi: 10.1007/s10648-021-09615-8.

[8] S. Hashim, M. K. Omar, H. Ab Jalil, and N. Mohd Sharef, "Trends on Technologies and Artificial Intelligence in Education for Personalized Learning: Systematic Literature Review," *Int. J. Acad. Res. Progress. Educ. Dev.*, vol. 11, no. 1, pp. 884–903, 2022, doi: 10.6007/ijarped/v11-i1/12230.

[9] K. Rajagopal, R. Van Schoors, S. Vanbecelaere, L. de Bie, and F. Depaepe, "Designing personalised learning support for K12 education: learner control, motivation and psychological ownership," *Interact. Des. Archit.*, vol. 45, pp. 155–176, 2020.

[10] N. Kucirkova, L. Gerard, and M. C. Linn, "Designing personalised instruction: A research and design framework," *Br. J. Educ. Technol.*, vol. 52, no. 5, pp. 1839–1861, 2021, doi: 10.1111/bjet.13119.

[11] G. Lampropoulos, E. Keramopoulos, K. Diamantaras, and G. Evangelidis, "Augmented Reality and Gamification in Education: A Systematic Literature Review of Research, Applications, and Empirical Studies," *Appl. Sci.*, vol. 12, no. 13, 2022, doi: 10.3390/app12136809.

[12] C. Walkington and M. L. Bernacki, "Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions," *J. Res. Technol. Educ.*, vol. 52, no. 3, pp. 235–252, 2020, doi: 10.1080/15391523.2020.1747757.

[13] E. B. G. d. Sousa, B. Alexandre, R. Ferreira Mello, T. Pontual Falcão, B. Vesin, and D. Gašević, "Applications of Learning Analytics in High Schools: A Systematic Literature Review," *Front. Artif. Intell.*, vol. 4, no. September, pp. 1–14, 2021, doi: 10.3389/frai.2021.737891.

[14] K. Aguar, S. Safaei, H. R. Arabnia, J. B. Gutierrez, W. D. Potter, and T. R. Taha, "Reviving Computer Science Education through Adaptive, Interest-Based Learning," *Proc. - 2017 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2017*, pp. 1161–1166, 2018, doi: 10.1109/CSCI.2017.202.

[15] K. Aguar, H. R. Arabnia, J. B. Gutierrez, W. D. Potter, and T. R. Taha, "Making CS inclusive: An overview of efforts to expand and diversify CS education," *Proc. - 2016 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2016*, pp. 321–326, 2017, doi: 10.1109/CSCI.2016.0067.

[16] R. Fariani, K. Junus, and H. Santoso, "Personalised Learning in Vocational Higher Education : A Model for Diverse Learners," *Res. Sq.*, pp. 1–31, 2023.

[17] M. Casper and A. Van Meegen, ""' Blending 4 Futures ": A teacher training course in blended learning for vocational teachers , based on design-research „ Blending 4 Futures ": A Teacher Training Course in Blended Learning for Vocational Teachers , Based on Design-Research," no. January, 2023, doi: 10.5281/zenodo.8208359.

[18] K. McConvey, S. Guha, and A. Kuzminykh, "A Human-Centered Review of Algorithms in Decision-Making in Higher Education," *Conf. Hum. Factors Comput. Syst. - Proc.*, 2023, doi: 10.1145/3544548.3580658.

[19] L. A. Alharbi, "A Systematic Literature Review on AI Algorithms and Techniques Adopted by e-Learning Platforms for Psychological and Emotional States," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, pp. 453–464, 2023, doi: 10.14569/IJACSA.2023.0140254.

[20] L. C. E. Silva *et al.*, "Applications of convolutional neural networks in education: A systematic literature review," *Expert Syst. Appl.*, vol. 231, no. June, p. 120621, 2023, doi: 10.1016/j.eswa.2023.120621.

[21] F. Gunnars, "A Systematic Review of Special Educational Interventions for Student Attention: Executive Function and Digital Technology in Primary School," *J. Spec. Educ. Technol.*, vol. 0, no. 0, pp. 1–13, 2023, doi: 10.1177/01626434231198226.

[22] A. Marougkas, C. Troussas, A. Krouska, and C. Sgouropoulou, *How personalized and effective is immersive virtual reality in education? A systematic literature review for the last decade*, no. 0123456789. Springer US, 2023. doi: 10.1007/s11042-023-15986-7.

[23] A. N. Ansari, S. Ahmad, and S. M. Bhutta, *Mapping the global evidence around the use of ChatGPT in higher education: A systematic scoping review*, no. 0123456789. Springer US, 2023. doi: 10.1007/s10639-023-12223-4.

[24] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," *Int. J. Educ. Technol. High. Educ.*, vol. 20, no. 1, pp. 1–17, 2023, doi: 10.1186/s41239-023-00426-1.

[25] V. Aguilar-Esteva, A. Acosta-Banda, R. Carreño Aguilera, and M. Patiño Ortiz, "Sustainable Social Development through the Use of Artificial Intelligence and Data Science in Education during the COVID Emergency: A Systematic Review Using PRISMA," *Sustain.*, vol. 15, no. 8, 2023, doi: 10.3390/su15086498.

[26] A. Valledor, A. Olmedo, C. J. Hellín, A. Tayebi, S. Otón-Tortosa, and J. Gómez, "The Eclectic Approach in English Language Teaching Applications: A Qualitative Synthesis of the Literature," *Sustain.*, vol. 15, no. 15, 2023, doi: 10.3390/su151511978.

[27] E. Jando, "Personalized E-learning Model : A Systematic Literature Review," no. November, pp. 238–243, 2017.

[28] S. A. Ochukut, R. O. Oboko, E. Miriti, and E. Maina, "Research Trends in Adaptive Online Learning: Systematic Literature Review (2011–2020)," *Technol. Knowl. Learn.*, vol. 28, no. 2, pp. 431–448, 2023, doi: 10.1007/s10758-022-09615-9.

[29] U. C. Apoki, A. M. Ali Hussein, H. K. M. Al-Chalabi, C. Badica, and M. L. Mocanu, "The Role of Pedagogical Agents in Personalised Adaptive Learning: A Review," *Sustain.*, vol. 14, no. 11, pp. 1–17, 2022, doi: 10.3390/su14116442.

[30] D. M. Heeg and L. Avraamidou, "The use of Artificial intelligence in school science: a systematic literature review," *EMI. Educ. Media Int.*, vol. 60, no. 2, pp. 125–150, 2023, doi: 10.1080/09523987.2023.2264990.

[31] R. Van Schoors, J. Elen, A. Raes, and F. Depaepe, "An overview of 25 years of research on digital personalised learning in primary and secondary education: A systematic review of conceptual and methodological trends," *Br. J. Educ. Technol.*, vol. 52, no. 5, pp. 1798–1822, 2021, doi: 10.1111/bjet.13148.

[32] V. K. Nadimpalli, F. Hauser, D. Bittner, L. Grabinger, S. Staufer, and J. Mottok, "Systematic Literature Review for the Use of AI Based Techniques in Adaptive Learning Management Systems," *ACM Int. Conf. Proceeding Ser.*, pp. 83–92, 2023, doi: 10.1145/3593663.3593681.

[33] T. Kabudi, I. Pappas, and D. H. Olsen, "AI-enabled adaptive learning systems: A systematic mapping of the literature," *Comput. Educ. Artif. Intell.*, vol. 2, no. December 2020, p. 100017, 2021, doi: 10.1016/j.caeai.2021.100017.

[34] N. N. Jamal, D. N. A. Jawawi, R. Hassan, and R. M. Irsyad Kamil, "Adaptive Learning in Computing Education: A Systematic Mapping Study," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 864, no. 1, pp. 0–5, 2020, doi: 10.1088/1757-899X/864/1/012069.

[35] Bhutoria, A., Aljabri, N. & Bose, S. Responsive early childhood parenting, preschool enrollment, and eventual student learning outcomes: a cross-country analysis using hierarchical linear modeling with TIMSS 2019 data. ICEP 19, 3 (2025). https://doi.org/10.1186/s40723-025-00141-6

[36] Y. Wang, C. Mu, X. Li, and Y. Yang, "An applied research on big data analysis and mining technology in education," *Proc. - 2019 Int. Conf. Sensing, Diagnostics, Progn. Control. SDPC 2019*, pp. 116–120, 2019, doi: 10.1109/SDPC.2019.00029.

[37] C. Chin and J. Osborne, "Students' questions: A potential resource for teaching and learning science," *Stud. Sci. Educ.*, vol. 44, no. 1, pp. 1–39, 2008, doi: 10.1080/03057260701828101.

[38] A. Shemshack and J. M. Spector, "A systematic literature review of personalized learning terms," *Smart Learn. Environ.*, vol. 7, no. 1, 2020, doi: 10.1186/s40561-020-00140-9.

[39] I. Yuyun and D. Suherdi, *Components and Strategies for Personalized Learning in Higher Education: A Systematic Review*, vol. 2. Atlantis Press SARL, 2023. doi: 10.2991/978-2-38476-054-1_23.

[40] Barbara Kitchenham, "Procedures for Performing Systematic Reviews," *Keele Univ. Tech. Rep.*, vol. 33, no. 2004, pp. 1–26, 2014.

[41] A. Alenezi, "Personalized Learning Strategies in Higher Education in Saudi Arabia: Identifying Common Approaches and Conditions for Effective Implementation," *TEM J.*, vol. 12, no. 4, pp. 2023–2037, 2023, doi: 10.18421/TEM124-13.

[42] C. Xu *et al.*, "SCEI: A Smart-Contract Driven Edge Intelligence Framework for IoT Systems," *IEEE Trans. Mob. Comput.*, vol. PP, no. X, pp. 1–14, 2023, doi: 10.1109/TMC.2023.3290925.

[43] M. Wang and Z. Lv, "Construction of personalized learning and knowledge system of chemistry specialty via the internet of things and

clustering algorithm," *J. Supercomput.*, vol. 78, no. 8, pp. 10997–11014, 2022, doi: 10.1007/s11227-022-04315-8.

[44] T. Ingkavara, P. Panjaburee, N. Srisawasdi, and S. Sajjapanroj, "The use of a personalized learning approach to implementing self-regulated online learning," *Comput. Educ. Artif. Intell.*, vol. 3, no. June, p. 100086, 2022, doi: 10.1016/j.caeai.2022.100086.

[45] H. J. Corbin, D. Chandran, C. Van Wingerden, and J. Baker-Sennett, "The learning camera: A personalized learning model for online pedagogy in human services education," *J. Technol. Hum. Serv.*, vol. 37, no. 4, pp. 334–346, 2019, doi: 10.1080/15228835.2019.1620151.

[46] A. Shemshack, Kinshuk, and J. M. Spector, "A comprehensive analysis of personalized learning components," *J. Comput. Educ.*, vol. 8, no. 4, pp. 485–503, 2021, doi: 10.1007/s40692-021-00188-7.

[47] V. P. Desai, K. S. Oza, and R. K. Kamat, "Adaptive Neuro Fuzzy Approach for Assessment of Learner's Domain Knowledge," *Lect. Notes Networks Syst.*, vol. 290, no. August, pp. 50–57, 2021, doi: 10.1007/978-981-16-4486-3_5.

[48] A. S. M. Ghazali, S. F. M. Noor, and S. Saad, "Review of personalized learning approaches and methods in e-learning environment," *Proc. - 5th Int. Conf. Electr. Eng. Informatics Bridg. Knowl. between Acad. Ind. Community, ICEEI 2015*, pp. 624–627, 2015, doi: 10.1109/ICEEI.2015.7352574.

[49] J. Lindstedt, R. Kauppinen, and A. Lagstedt, "Personalizing the Learning Process with Wihi," in *Proceedings of the 33rd Bled eConference: Enabling Technology for a Sustainable Society*, Bled, Slovenia, 2020, pp. 306-318. doi: 10.18690/978-961-286-362-3.21.

[50] H. Aberbach, A. Jeghal, A. Sabri, H. Tairi, and L. Laaouina, "A Personalized Learning Approach based on Learning Speed," *J. Comput. Sci.*, vol. 17, no. 3, pp. 242–250, 2021, doi: 10.3844/JCSSP.2021.242.250.

[51] C. T. Lam, T. N. Hoang, B. Kian, H. Low, and P. Jaillet, "Model Fusion for Personalized Learning," *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, pp. 5948–5958, 2021, [Online]. Available: https://proceedings.mlr.press/v139/lam21a.html

[52] P. Melzer, *A conceptual framework for personalised learning: Influence factors, design, and support potentials*. 2018. doi: 10.1007/978-3-658-23095-1.

[53] M. Taeb and H. Chi, "A Personalized Learning Framework for Software Vulnerability Detection and Education," *Proc. - 2021 Int. Symp. Comput. Sci. Intell. Control. ISCSIC 2021*, pp. 119–126, 2021, doi: 10.1109/ISCSIC54682.2021.00032.

[54] P. Melzer and M. Schoop, "A conceptual framework for task and tool personalisation in IS education," *2015 Int. Conf. Inf. Syst. Explor. Inf. Front. ICIS 2015*, no. Zimmerman 1989, pp. 1–17, 2015, doi: 10.1007/978-3-658-23095-1_3.

[55] S. Kontrimiene *et al.*, "Implementing the Personalised Learning Framework in University Studies: What Is It That Works?," *Acta Paedagog. Vilnensia*, vol. 47, pp. 122–142, 2021, doi: 10.15388/ActPaed.2021.47.9.

[56] M. Maravanyika and N. Dlodlo, "An Adaptive Framework for Recommender-Based Learning Management Systems," *2018 Open Innov. Conf. OI 2018*, pp. 203–212, 2018, doi: 10.1109/OI.2018.8535816.

[57] L. S. Iyer, S. Bharadwaj, S. H. Shetty, V. Verma, and M. Devanathan, "Advancing equity in digital classrooms: A personalized learning framework for higher education institutions," *Socioecon. Incl. Dur. an Era Online Educ.*, no. January 2024, pp. 225–245, 2022, doi: 10.4018/978-1-6684-4364-4.ch011.

[58] W. Jiang, Q. Li, L. Su, C. Miao, Q. Gu, and W. Xu, "Towards personalized learning in mobile sensing systems," *Proc. - Int. Conf. Distrib. Comput. Syst.*, vol. 2018-July, pp. 321–333, 2018, doi: 10.1109/ICDCS.2018.00040.

[59] B. Huynh, J. Orlosky, and T. Hollerer, "In-situ labeling for augmented reality language learning," *26th IEEE Conf. Virtual Real. 3D User Interfaces, VR 2019 - Proc.*, pp. 1606–1611, 2019, doi: 10.1109/VR.2019.8798358.

[60] C. S. Lee, M. H. Wang, C. Y. Chen, M. Reformat, Y. Nojima, and N. Kubota, "Knowledge Graph-Based Genetic Fuzzy Agent for Human Intelligence and Machine Co-Learning," *IEEE Int. Conf. Fuzzy Syst.*, pp. 1–6, 2023, doi: 10.1109/FUZZ52849.2023.10309699.

[61] M. Z. Islam, K. Mehmood, and H. S. Kim, *Reinforcement Learning Based Interactive Agent for Personalized Mathematical Skill Enhancement*, vol. 11989 LNCS, no. March. Springer International Publishing, 2020. doi: 10.1007/978-3-030-43120-4_31.

[62] A. B. Firdausiah Mansur, N. Yusof, and A. H. Basori, "Personalized Learning Model based on Deep Learning Algorithm for Student Behaviour Analytic," *Procedia Comput. Sci.*, vol. 163, pp. 125–133, 2019, doi: 10.1016/j.procs.2019.12.094.

[63] E. Jando, Meyliana, A. N. Hidayanto, H. Prabowo, H. L. H. S. Warnars, and Sasmoko, "Personalized E-learning Model: A systematic literature review," *Proc. 2017 Int. Conf. Inf. Manag. Technol. ICIMTech 2017*, vol. 2018-Janua, no. November 2020, pp. 238–243, 2017, doi: 10.1109/ICIMTech.2017.8273544.

[64] E. M. McCarthy, Y. Liu, and K. L. Schauer, "Strengths-based blended personalized learning: An impact study using virtual comparison group," *J. Res. Technol. Educ.*, vol. 52, no. 3, pp. 353–370, 2020, doi: 10.1080/15391523.2020.1716202.

[65] Z. Chen, "Research on the Teaching Reform of Physical Education Combined with Civic and Political Education in the New Era under the Background of Big Data," *Appl. Math. Nonlinear Sci.*, 2023, doi: 10.2478/amns.2023.2.01383.

[66] B. Yi *et al.*, "Research on personalized learning model under informatization environment," *Proc. - 2017 Int. Symp. Educ. Technol. ISET 2017*, pp. 48–52, 2017, doi: 10.1109/ISET.2017.19.

[67] L. Zhou, F. Zhang, S. Zhang, and M. Xu, "Study on the personalized learning model of learner-learning resource matching," *Int. J. Inf. Educ. Technol.*, vol. 11, no. 3, pp. 143–147, 2021, doi: 10.18178/ijiet.2021.11.3.1503.

[68] K. Aoki, *Teaching and Learning Methods at the Open University of Japan: Transition From the Broadcasting Model to a More Personalized Learning Model*. Elsevier Inc., 2016. doi: 10.1016/B978-0-12-803637-2.00006-3.

[69] J. Chang and X. Lu, "The study on students' participation in personalized learning under the background of artificial intelligence," *Proc. - 10th Int. Conf. Inf. Technol. Med. Educ. ITME 2019*, pp. 555–558, 2019, doi: 10.1109/ITME.2019.00131.

[70] X. Yu, J. Xia, and W. Cheng, "Prospects and Challenges of Equipping Mathematics Tutoring Systems with Personalized Learning Strategies," *IEIR 2022 - IEEE Int. Conf. Intell. Educ. Intell. Res.*, pp. 42–45, 2022, doi: 10.1109/IEIR56323.2022.10050082.

[71] Z. Wang and S. Wang, "Construction of Personalized Learning Model Supported by Human-computer Cooperation," *ACM Int. Conf. Proceeding Ser.*, pp. 39–44, 2022, doi: 10.1145/3582580.3582588.

[72] M. Ashok, A. Chinnasamy, K. Ramasamy, Y. H. Gokul, and J. B. Douglas, "A Systematic Survey on Personalized Learning Framework based Recommendation System," *2022 Int. Conf. Data Sci. Agents Artif. Intell. ICDSAAI 2022*, vol. 01, pp. 1–5, 2022, doi: 10.1109/ICDSAAI55433.2022.10028809.

[73] L. S. H. Cardenas, L. Castano, C. C. Guzman, and J. P. N. Alvarez, "Personalised learning model for academic leveling and improvement in higher education," *Australas. J. Educ. Technol.*, vol. 38, no. 2, pp. 70–82, 2022, doi: 10.14742/ajet.7084.

[74] E. V. Soboleva, K. K. Zhumakulov, K. P. Umurkulov, G. I. Ibragimov, L. V. Kochneva, and M. O. Timofeeva, "Developing a Personalised Learning Model Based on Interactive Novels to Improve the Quality of Mathematics Education," *Eurasia J. Math. Sci. Technol. Educ.*, vol. 18, no. 2, 2022, doi: 10.29333/EJMSTE/11590.

[75] E. V. Soboleva, T. N. Suvorova, M. I. Bocharov, and T. I. Bocharova, "Development of the Personalized Model of Teaching Mathematics by Means of Interactive Short Stories to Improve the Quality of Educational Results 0f Schoolchildren," *Eur. J. Contemp. Educ.*, vol. 11, no. 1, pp. 241–257, 2022, doi: 10.13187/ejced.2022.1.241.

[76] Y. Peng, "Design and Practice of Cloud Platform for Online Music Courses Based on the Mobile Internet Era," *J. Phys. Conf. Ser.*, vol. 1881, no. 2, 2021, doi: 10.1088/1742-6596/1881/2/022096.

[77] A. L. M. Maddalora, "Personalized learning model using item response theory," *Int. J. Recent Technol. Eng.*, vol. 8, no. 1 Special Issue 4, pp. 811–818, 2019.

[78] A. J. Bingham, J. F. Pane, E. D. Steiner, and L. S. Hamilton, "Ahead of the Curve: Implementation Challenges in Personalized Learning School Models," *Educ. Policy*, vol. 32, no. 3, pp. 454–489, 2018, doi: 10.1177/0895904816637688.

[79] B. Waldrip, J. J. Yu, and V. Prain, "Validation of a model of personalised learning," *Learn. Environ. Res.*, vol. 19, no. 2, pp. 169–180, 2016, doi: 10.1007/s10984-016-9204-y.

[80] N. A. Wouda, A. Aslan, and I. F. A. Vis, "An adaptive large neighbourhood search metaheuristic for hourly learning activity planning in personalised learning," *Comput. Oper. Res.*, vol. 151, no. September 2022, p. 106089, 2023, doi: 10.1016/j.cor.2022.106089.

[81] W. S. Sayed, A. M. Noeman, A. Abdellatif, M. Abdelrazek, M. G. Badawy, A. Hamed, and S. El-Tantawy, "AI-based Adaptive Personalized Content Presentation and Exercises Navigation for an Effective and Engaging E-learning Platform," *Multimedia Tools and Applications*, vol. 82, pp. 3303–3333, 2023, doi: 10.1007/s11042-022-13076-8.

[82] C. Xiao, S. X. Xu, K. Zhang, Y. Wang, and L. Xia, "Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, no. Bea, pp. 610–625, 2023, doi: 10.18653/v1/2023.bea-1.52.

[83] Y. Yang, H. Shang, N. Jamal, and F. Ullah, "Video Resources Recommendation for Online Tourism Teaching in Interactive Network," *Mob. Networks Appl.*, no. 0123456789, 2023, doi: 10.1007/s11036-023-02193-9.

[84] Y. Yu, L. Han, X. Du, and J. Yu, "An Oral English Evaluation Model Using Artificial Intelligence Method," *Mob. Inf. Syst.*, vol. 2022, 2022, doi: 10.1155/2022/3998886.

[85] U. Maier and C. Klotz, "Personalized feedback in digital learning environments: Classification framework and literature review," *Comput. Educ. Artif. Intell.*, vol. 3, no. March, p. 100080, 2022, doi: 10.1016/j.caeai.2022.100080.

[86] M. Bustamante-León *et al.*, "The Personalized and Inclusive MOOC: Using Learning Characteristics and Quality Principles in Instructional Design," *Sustain.*, vol. 14, no. 22, pp. 1–17, 2022, doi: 10.3390/su142215121.

[87] X. Zhang, A. Tlili, K. Shubeck, X. Hu, R. Huang, and L. Zhu, "Teachers' adoption of an open and interactive e-book for teaching K-12 students Artificial Intelligence: a mixed methods inquiry," *Smart Learn. Environ.*, vol. 8, no. 1, 2021, doi: 10.1186/s40561-021-00176-5.

[88] Y. Fang and J. Li, "Application of the Deep Learning Algorithm and Similarity Calculation Model in Optimization of Personalized Online Teaching System of English Course," *Comput. Intell. Neurosci.*, vol. 2021, 2021, doi: 10.1155/2021/8249625.

[89] J. Gao, L. Wang, and F. Xu, "Research on the Construction of Course Knowledge Graph of High School Information Technology," *Proc. - 2020 Int. Conf. Artif. Intell. Educ. ICAIE 2020*, pp. 211–215, 2020, doi: 10.1109/ICAIE50891.2020.00056.

[90] Y. Rosmansyah and A. F. Alim, "Personalized Blended Mobile Learning Framework for the Koran Memorization Process," *Proc. Int. Conf. Electr. Eng. Informatics*, vol. 2019-July, no. July 2019, pp. 278–283, 2019, doi: 10.1109/ICEEI47359.2019.8988837.

[91] N. Liu, "Improving online ideological and political precision education in universities based on knowledge mapping," *Appl. Math. Nonlinear Sci.*, vol. 3, no. 2, pp. 571–582, 2023.

[92] T. J. Gong, X. X. Yao, and W. Ma, "GRE: An adaptive and personalized exercise model for K-12 online education," *ACM Int. Conf. Proceeding Ser.*, pp. 48–54, 2018, doi: 10.1145/3291078.3291118.

[93] D. Peng, "Civic education reform based on deep reinforcement learning model.," *Appl. Math. Nonlinear Sci.*, vol. 3, no. 2, pp. 571–582, 2023.

[94] K. Takami, B. Flanagan, Y. Dai, and H. Ogata, "Personality-based tailored explainable recommendation for trustworthy smart learning system in the age of artificial intelligence," *Smart Learn. Environ.*, vol. 10, no. 1, 2023, doi: 10.1186/s40561-023-00282-6.

[95] M. N. K. Singh, Q. Sun, and C. Weber, "Evaluating the use of a personalized learning management system to increase student enrollment in high school physics (evaluation, diversity)," *ASEE Annu. Conf. Expo. Conf. Proc.*, 2019, doi: 10.18260/1-2--32761.

[96] D. Lee, Y. Huh, C. Y. Lin, C. M. Reigeluth, and E. Lee, "Differences in personalized learning practice and technology use in high- and low-performing learner-centered schools in the United States," *Educ. Technol. Res. Dev.*, vol. 69, no. 2, pp. 1221–1245, 2021, doi: 10.1007/s11423-021-09937-y.

[97] C. Walkington and M. L. Bernacki, "Motivating students by 'personalizing' learning around individual interests: A consideration of theory, design, and implementation issues," *Adv. Motiv. Achiev.*, vol. 18, pp. 139–176, 2014, doi: 10.1108/S0749-742320140000018004.

[98] M. A. N. Elmaadaway and Y. A. M. Abouelenein, "In-service teachers' TPACK development through an adaptive e-learning environment (ALE)," *Educ. Inf. Technol.*, vol. 28, no. 7, pp. 8273–8298, 2023, doi: 10.1007/s10639-022-11477-8.

[99] N. Chakraborty, S. Roy, W. L. Leite, M. K. S. Faradonbeh, and G. Michailidis, "The effects of a personalized recommendation system on students' high-stakes achievement scores: A field experiment," *Proc. 14th Int. Conf. Educ. Data Mining, EDM 2021*, no. Edm, pp. 588–594, 2021.

[100] H. Fake and N. Dabbagh, "The Personalized Learning Interaction Framework: Expert Perspectives on How to Apply Dimensions of Personalized Learning to Workforce Training and Development Programs," *ACM Int. Conf. Proceeding Ser.*, pp. 501–509, 2021, doi: 10.1145/3486011.3486503.

[101] A. Aslan, I. Bakir, and I. F. A. Vis, "A dynamic thompson sampling hyper-heuristic framework for learning activity planning in personalized learning," *Eur. J. Oper. Res.*, vol. 286, no. 2, pp. 673–688, 2020, doi: 10.1016/j.ejor.2020.03.038.

[102] T. Kajonmanee, P. Chaipidech, N. Srisawasdi, and K. Chaipah, "A personalised mobile learning system for promoting STEM discipline teachers' TPACK development," *Int. J. Mob. Learn. Organ.*, vol. 14, no. 2, pp. 215–235, 2020, doi: 10.1504/IJMLO.2020.106186.

[103] J. Guan, Z. Ma, Q. Liu, H. Chen, and S. WanG, "Design of personal learning management platform based on information technology," *ACM Int. Conf. Proceeding Ser.*, pp. 79–86, 2019, doi: 10.1145/3352740.3352754.

[104] Š. Karolčík, I. Zilinskiene, A. Slotkiene, and E. Čipková, "Analysis of e-learning environment for geography: Opportunities for personalized active learning," *Balt. J. Mod. Comput.*, vol. 7, no. 3, pp. 405–418, 2019, doi: 10.22364/bjmc.2019.7.3.07.

[105] M. Momenzad, B. Majidi, and M. Eshghi, "Deep Summarization of Academic Textbooks for Adaptive Gamified Virtual Learning Environments," *2018 2nd Natl. 1st Int. Digit. Games Res. Conf. Trends, Technol. Appl. DGRC 2018*, pp. 88–94, 2018, doi: 10.1109/DGRC.2018.8712065.

[106] C. A. Haynes and K. Shelton, "Beyond the Classroom: A Framework for Growing School Capacity in a Digital Age," *J. Res. Technol. Educ.*, vol. 50, no. 4, pp. 271–281, 2018, doi: 10.1080/15391523.2018.1451791.

[107] N. Mallios and M. G. Vassilakopoulos, "Evaluating students' programming skill behaviour and personalizing their computer learning environment using 'the Hour of Code' paradigm," *Proc. Int. Conf. e-Learning 2015, E-LEARNING 2015 - Part Multi Conf. Comput. Sci. Inf. Syst. 2015*, pp. 131–135, 2015.

[108] E. S. B. Abenga, E. O. Okono, M. Awuor, and S. Otanga, "Framework for technology-enriched active class learning of physics in secondary schools in Kenya," *Digit. Solut. Case Africa's Sustain. Dev.*, pp. 131–151, 2020, doi: 10.4018/978-1-7998-2967-6.ch009.

[109] Z. Wang and S. Wang, "Construction of Personalized Learning Model Supported by Human-computer Cooperation," in *ACM International Conference Proceeding Series*, 2022, pp. 39–44. doi: 10.1145/3582580.3582588.

APPENDIX: SELECTED PUBLICATIONS BASED ON SEARCH STRINGS

Research Paper	Search Strings	Approach/ Strategy for PL	Framework	Model	Component, Attribute, and other Consideration for Effective Implementations of PL Framework, Model, Approach, and System	Technique/Technology Applied	Learning Level		
							HL	K-12	GL
[41]	SS1	✓			Flexible pacing, adaptive assessment, learner agency, feedback, and student-centered learning.	Knowledge-based	✓		
[42]	SS1	✓	✓	✓	Data privacy, distribution, behavior, time, and location.	Deep Learning, Blockchain, IoT	✓		
[43]	SS1	✓		✓	Previous knowledge, Learning: interest, effectiveness, data, content, and satisfaction	Deep learning, IoT	✓		
[44]	SS1	✓			Feedback, autonomy, learning: path, pace, behavior, suggestions, and goal.	Machine Learning-Self regulated approach		✓	
[45]	SS1	✓		✓	Learning interest, previous knowledge, skills, and pedagogy.	Knowledge Management	✓		
[46]	SS1	✓		✓	Learners: profile, interest, goal, experience, knowledge and belief, attitude, and paths/pace.	Intelligent Tutoring System, Learning Analytics, Machine Learning, Data mining			✓
[47]	SS1	✓		✓	Students: interest, domain, skills, behavior, Learning: style, outcome, satisfaction, recommendation.	Analytical Neural Network, Fuzzy Logic			✓
[48]	SS1	✓			Learning: style, approach, interest, preferences. Profile, content, recommendation, pedagogy.	Data-driven (Bayesian Network)			✓
[49]	SS1	✓			Learner's needs, motivation, feedback, paths, data content, and tracking.	Machine Learning, Data Analytics	✓		
[50]	SS1	✓			Learning: speed, preference, needs, interest, experience, choices, behavior, skills.	Markov Decision Process, Data Mining			✓
[51]	SS2			✓	Encoding alignment, variational approximation algorithm, Meta-learning.	Deep generative model, ML algorithm			✓
[52]	SS2		✓		Learning content, technology, skills, methods, and outcomes.	Machine learning, learning analytics	✓		
[53]	SS2		✓		Learning skills, habits, and awareness.	Natural language processing, Deep Learning	✓		
[54]	SS2	✓	✓		Learning approach, task, and tools.	Theory of cognitive fit, Deep learning	✓		
[55]	SS2		✓		Flexible content, learning environment, and support.	Thematic Analysis, Data-driven Decision making	✓		
[56]	SS2		✓		Learner profile, style, behavior, pedagogy, and previous knowledge.	Intelligent Tutoring System	✓		
[57]	SS2		✓		Learning capability, competency level, skills, and passion.	Intelligent tutoring system, Technology-supported system.	✓		
[58]	SS2		✓		Individual data, level, uniqueness, sharing and feedback.	Machine learning	✓		
[59]	SS3		✓		Learning outcomes, contents, and students' growth.	Neural Network, Machine learning	✓		
[60]	SS3	✓		✓	Learner profile, knowledge domain, content, and outcome.	Genetic Algorithm, Fuzzy Agent	✓		
[61]	SS3	✓	✓	✓	Learning progress, profile, attitudes, skills, and target goals.	Exp3 algorithm, Q-Matrix approach			✓
[62]	SS3		✓	✓	Learner foundation, facilities, community engagement, learning approach.	Deep learning algorithm		✓	
[63]	SS3	✓	✓	✓	Personality, Knowledge, Behavior, interest, preferences.	Datamining, Bayesian Network			✓
[64]	SS3			✓	Learning: strength, support, path, flexibility, measurement, achievement.	Strength-Based learning		✓	
[65]	SS3			✓	Attention mechanism, learning paths, learning: awareness, supervision, cognition, behavior, and outcome.	Bigdata Technology, Recurrent Neural Network	✓	✓	
[66]	SS3			✓	Learning methods, style, ability, stages, prediction, and achievements.	Knowledge Visualization	✓		
[67]	SS3			✓	Learner information, behavior, learning resources, and matching degree	Support Vector Machine, Collaborative Filtering			✓
[68]	SS3			✓	Learning speed, performance, and pedagogical strategies	Formative Assessment	✓		

Research Paper	Search Strings	Approach/ Strategy for PL	Framework	Model	Component, Attribute, and other Consideration for Effective Implementations of PL Framework, Model, Approach, and System	Technique/Technology Applied	Learning Level		
							HL	K-12	GL
[69]	SS3			✓	Learners' participation, technology reliability, learners' autonomy, feedback and assessment, experience, methods, and beliefs.	Bigdata, AI Technology	✓		
[70]	SS3	✓		✓	Learning strategies, learners' profile, pedagogy, and Mistake diagnosis.	Bigdata, Datamining			✓
[71]	SS4			✓	Learning dynamics, teaching ideas, evaluation, and learner support.	Big Data, Machine Learning		✓	
[72]	SS4		✓		Learning process and relevant knowledge.	Deep learning, content filtering	✓		
[73]	SS4			✓	Mastery level, knowledge level, requirement, performance, learning flexibility, and student autonomy	Knowledge leveling	✓		
[74]	SS4			✓	Competency and pedagogical experience.	AXMA Story Maker, Interactive Novels		✓	
[75]	SS4			✓	Test learning effectiveness, economic requirements in mathematics.	Digital Technologies AXMA Story Maker		✓	
[76]	SS4			✓	Learning evaluation, curriculum, resources, cooperative, and independent.	Datamining, Deep Mining		✓	
[77]	SS4			✓	Lerner's ability, learning process, and diversity.	Item Response Theory	✓		
[78]	SS4		✓	✓	Learner Profile, Mastery level, learning path, and flexible environment.	Technology enabled		✓	
[79]	SS4			✓	Assessment, academic achievement, efficacy, and student wellbeing.	Technology enabled		✓	
[80]	SS5	✓			Course module, course planning, feasible instruction, and self-regulation.	Integer linear programming, Metaheuristic Algorithm		✓	
[81]	SS5	✓			Cognitive level, learning style, reinforcement, and gamification.	Deep Q-Network Reinforcement Learning		✓	
[82]	SS5	✓			Learning style and learner profile	Generative AI (Chat GPT)		✓	
[83]	SS5			✓	User behavior, interest, teaching strategy, notification, recommender component, and learning process tracking.	Intelligent Technology, Recommendation Algorithm			✓
[84]	SS5			✓	Oral English evaluation, tracking, recording, and evaluation	Data Mining, Machine Learning		✓	
[85]	SS5		✓		Feedback, progress, learning goals, and emotional state.	AI-Machine learning			✓
[86]	SS5			✓	Learner profile, characteristics, needs, and attitude.	Data mining			✓
[87]	SS5		✓		Learner's attitude, curriculum design, and feedback.	Machine learning, NLP		✓	
[88]	SS5			✓	Learning resources, data, and recommendations, learner characteristics.	Deep learning algorithm, Bee colony, Neural Network		✓	
[89]	SS5	✓			Learning content, curriculum, and learners' characteristics.	AI-Knowledge management			✓
[90]	SS5		✓		Learning flexibility, assessment, style, and learners' needs.	M-Learning Technology			✓
[91]	SS5			✓	Cognitive level, collaboration, and path recommendation	Bigdata Technology	✓		
[92]	SS5			✓	Learner recognition and exercise generation, learner behavior.	Natural Language Processing		✓	
[93]	SS5			✓	Interpretability of recommendation results and learning preferences.	Deep learning		✓	
[94]	SS6				Cognitive perception, personality, and student profile.	AI Systems, Bayesian Knowledge Tracing		✓	✓
[95]	SS6			✓	Learning style, needs, and attitude.	Technology Driven		✓	
[96]	SS6	✓			Planning, learning, record keeping, and assessment.	Technology Driven		✓	
[97]	SS6	✓			Prior Knowledge, motivation, and achievement.	Smart Technology		✓	

Research Paper	Search Strings	Approach/ Strategy for PL	Framework	Model	Component, Attribute, and other Consideration for Effective Implementations of PL Framework, Model, Approach, and System	Technique/Technology Applied	Learning Level		
							HL	K-12	GL
[98]	SS7	✓			Motivation, interest profile, and level.	Technology Driven			✓
[99]	SS7	✓			Interest, self-regulation, and learning speed.	Markov Decision Process		✓	
[100]	SS7		✓		Skill, flexibility, interface, choice, and feedback.	Delphi Technique			✓
[101]	SS7	✓	✓		Learning pace, goal, method, and flexibility.	Hyper-heuristic		✓	
[102]	SS7	✓	✓		Learning style, content, and behavior.	Mobile Technology		✓	
[9]	SS7	✓			Learner motivation and ownership.	Technology Driven		✓	
[103]	SS7		✓		Course plan, selection, assessment, ability, and achievement.	Information Technology	✓		
[104]	SS7		✓		Learning content, goals, a flexible environment, and a learner profile.	Datamining		✓	
[105]	SS7		✓		Text summarizer, gamification, video lessons, learner profile, learning path, style, and experience.	Machine learning (Deep Neural Network)		✓	
[106]	SS7		✓		Evaluation and assessment, school capacity, and digital learning.	Delphi Method		✓	
[107]	SS7		✓		Learners' needs, interests, behavior, and learning evaluation.	Hour of Code		✓	
[108]	SS7		✓		Learning ability, profile, level, and pace.	Technology enriched		✓	

Note: HL is Higher Learning, K-12 is Kindergarten 12 (Primary to Secondary School/Pre-university Level), SS (Search Strings 1 to 7)

Mobile Application Using Convolutional Neural Networks for Preliminary Diagnosis of Rosacea

Angie Fiorella Sapaico-Alberto, Rosalynn Ornella Flores-Castañeda
Facultad de Ingeniería y Arquitectura, Universidad César Vallejo, Lima, Perú

***Abstract*—Rosacea is a chronic skin disease affecting millions of people worldwide, characterized by redness and inflammatory lesions on the face. Given the need to improve early detection, this research aims to develop a mobile application using convolutional neural networks to improve the preliminary diagnosis of rosacea. For this purpose, increases in sensitivity, specificity and accuracy percentages were evaluated. The study was applied, with a quantitative approach and an experimental design, specifically pre-experimental. The study variable was the preliminary diagnosis of rosacea, and the sample consisted of 100 images: 50 from rosacea patients and 50 from healthy people. The technique used for data collection was observation. The results of the implementation of the mobile application showed an increase in sensitivity of 2.7%, specificity of 1.97% and accuracy of 0.10%. In conclusion, the use of the mobile application with convolutional neural networks improves the preliminary diagnosis of rosacea by optimizing the indicators evaluated.**

***Keywords*—Convolutional neural networks; mobile application; rosacea; preliminary diagnosis; sensitivity; specificity**

I. Introduction

In dermatology, the initial and correct detection of rosacea is considered a challenge. This may be because the symptoms are very different and can even be confused with other skin diseases. Thus, the development of a mobile application using convolutional neural networks is proposed to reach a better preliminary detection and thus increase the rates of sensitivity, specificity and accuracy in the diagnosis.

Rosacea is a chronic, non-cancerous dermatological disease affecting the patient's face [1]. The symptoms that appear can be mild, such as spontaneous facial flushing, to more severe cases, such as persistent inflammatory lesions, visibly dilated blood vessels and thickening of the nasal and ocular skin [2].

Rosacea is a skin problem; although the exact cause is unknown, it results from a combination of environmental factors, heredity and specific triggers [3]. Genetic susceptibility, constant intense exposure to ultraviolet (UV) radiation, climate and stress are also some of the main factors contributing to this condition [4]. In addition, certain foods and drinks, such as spicy foods, alcohol and dairy products, can cause rosacea flare-ups in people with sensitive skin [5].

This disease impacts approximately 5.5% of people in the world [6]. Although it affects both genders, it is more common in women and occurs between the ages of 20 and 50 in women but is usually more severe in men and can occur at any age, including childhood [7]. Individuals with rosacea often suffer from anxiety, depression and embarrassment, which affect their quality of life and disrupt their relationships [8]. In terms of physical discomfort, it can be debilitating, but the effect on self-perception and social interaction can be equally stifling [9].

Diagnosing rosacea often presents challenges due to the variability of its symptoms, because it can often be confused with other skin diseases such as seborrheic dermatitis or acne [10]. In addition, rosacea can vary from person to person, and its symptoms can evolve over time, especially in the early stages, complicating the accurate identification of rosacea.

Although dermatological medicine has progressed, there are still major challenges in the early and correct diagnosis of rosacea [11]. Furthermore, the subjectivity of clinical assessment, the diversity in the clinical symptoms of rosacea and the absence of standardized diagnostic standards are just some of the challenges faced by practitioners in their efforts to diagnose and treat rosacea appropriately [12]. Therefore, it is crucial that emerging technologies are used to support the diagnosis of dermatological diseases [13].

For this reason, a mobile application using convolutional neural networks was proposed as a solution to enable more effective picture-based diagnosis of rosacea. This application was developed using convolutional neural networks, which identify characteristics of rosacea. The aim was to increase sensitivity to detect positive cases, increase specificity to minimize false positives and improve accuracy. In this way, this tool provides a technological solution to patients and dermatologists for the effective detection of the disease.

This study is organized as follows: Section II presents the related works. In Section III, the methodology is applied. Section IV presents the results. Section V, the discussion of results and in Section VI, the conclusions of the research is presented.

II. Related Works

In [14], the authors studied deep learning frameworks to establish skin lesions from dermatological images. The study tested the use of convolutional neural networks to detect and categorize diseases at different levels. Some of the findings included efficacy in determining the presence of diseases and their type, with potential breakthroughs that would help individualize treatments for each patient. These results highlighted that the tool is potentially powerful and efficient for dermatological diagnosis in clinical settings.

In [15], the authors conducted a study to optimize the effectiveness of deep learning model training in detecting rosacea lesions from facial images. They performed further tests and observed that their technique not only trained the model

faster (up to 25.6 times faster) but also maintained equivalent accuracy. Finally, they indicated that this method is a fairly accurate and efficient method for detecting rosacea, which could prove to be of great value in the future for clinical uses.

Also in [16], a study to classify skin lesions using images through a customized deep learning framework designed with fewer convolution layers was explored in India for comparison with the previously trained VGG16, ResNet50 and InceptionV3 models. The study focused on collecting dermatoscopic images of patients with various skin conditions using the new customized framework. The main tools included the use of advanced image processing techniques and optimized algorithms to handle complicated visual data. Results were reduced by 95%, 85% and 84% against VGG16-TrL, ResNet50-TrL and InceptionV3-TrL, respectively, highlighting the accuracy of skin lesion classification against conventional methods.

On the other hand, [17] the current applications and results of implementing deep learning algorithms in dermatology were extensively examined. The study population covered a wide range of clinical and research dermatological applications using deep learning algorithms for multiple skin conditions. The tools examined were mainly convolutional neural networks, predicting in the image category. The main features highlighted in this case indicated that, although deep learning provides a noticeable increase in diagnostic accuracy compared to conventional methods, it also involves risks such as the need for large amounts of recorded data and model interpretation.

In addition, [18] investigated the effectiveness of collaboration between dermatologists and artificial intelligence in skin cancer identification, including a large set of dermatological images related to the disease. The techniques used for image processing were multiclass regression and content-based image retrieval (CBIR) to identify patterns in skin cancer. Primary results showed that the fusion of human- and computer-generated analysis not only produces more accurate diagnostic results than conventional methods but also reduces assessment time, suggesting a potential model for future use in routine dermatology.

Also in [19], the authors detected brain tumors by using deep learning and transfer learning. A database of MRI scans from a medical center with a random sample of patients was used. Advanced convolutional neural networks, including VGG, ResNet, MobileNet and DenseNet, were used for this dataset of MRI images that were trained and provided. The research showed that the implementation of the transfer learning technique in the proposed models improved brain tumor detection with an accuracy result of 99.75% for MobileNetv3.

On the other hand, [20] implemented an artificial intelligence model to detect and discern between actinic keratosis and seborrheic keratosis. They used a dataset of 2000 images, with training, validation and test subsets. The AI model was developed using deep learning techniques and implemented using the Python programming language and frameworks such as TensorFlow and PyTorch. Furthermore, taking advantage of the Google Collaboration Platform, the model was efficiently trained in two hours and thirteen minutes, using the computing resources provided by the platform. The model study showed a high ability to distinguish between images of both diseases, with high levels of accuracy, precision and specificity.

III. Methodology

This section describes the methodology used in this research, which was applied because it focused on the use of existing artificial intelligence techniques with the aim of developing a mobile application for the preliminary diagnosis of rosacea. The approach was quantitative and a pre-experimental design because it sought to demonstrate that the independent variable: mobile application using neural networks can provide a preliminary diagnosis of rosacea, this being the dependent variable. The population consisted of images obtained from the free repositories Dermatology Atlas, Fitzpatrick 17k and Flickr-Faces-HQ Dataset (FFHQ). For the sampling, images of people with and without a diagnosis of rosacea were considered for further evaluation. Observation was used to collect data. For data analysis, formulae were applied to calculate sensitivity, specificity and accuracy, respectively. Additionally, the Receiver Operating Characteristic (ROC) curve was used to evaluate the test performance. The ROC curve illustrates the connection between the true positive rate and the false positive rate across different thresholds.

A. Case Development

The methodology that was implemented is Mobile-D, which has an agile approach adapted specifically for mobile application development, combining agile and software engineering practices to address the needs of this type of project [23]. Regarding the population, a total of 600 images were considered, and for the distribution of images in the dataset, 450 images were used for training, a total of 30 images for validation and 20 images as a test set. Finally, for the post-test, 100 images were used. Thus, the sample consisted of 100 images selected from the described population, of which 50 were images of people diagnosed with rosacea and the other 50 were images of healthy people. The phases of the Mobile-D methodology are specified below.

B. Phase 1: Exploration

The product is defined, and customer requirements are understood, an initial plan is drawn up, and potential risks are identified and mitigated [21]. In this initial phase, priority is given to understanding the problem and clearly defining the essential functionalities of the rosacea detection application.

1) Outreach. Develop a mobile application using convolutional neural networks for the preliminary detection of rosacea.

2) Set of functional requirements. Key functional requirements are identified that will define the core capabilities of the application (see Table I).

TABLE I FUNCTIONAL REQUIREMENTS

Requirement	Description of the Requirement
RF01	• The system should allow users to register and authenticate themselves.
RF02	• The app should provide reliable information about rosacea and provide links to additional resources.

RF03	• The app should offer clear instructions to ensure that photos are taken correctly.
RF04	• The user should be able to select images from the gallery or take a picture from the camera to be analyzed.
RF05	• The results of the analysis should be clearly displayed, including the probability of each diagnosis.
RF06	• The system should generate reports summarizing the results of the analyzed images.
RF07	• The system should display the results and generate reports on metrics and indicators.
RF08	• The system should include an 'about the mobile application' section, mentioning that it is a university project, detailing its purpose and context.

C. Phase 2: Initialization

The development environment is established, the architecture and initial design are defined, and teams and their roles are formed [21]. The objective is to have a basic functional version of the mobile application that allows validating the first hypotheses and adjusting it to the user's expectations.

1) Configuration of the environment. Hardware: HP AMD Ryzen 3 laptop and mobile device (mobile phone).

Software: Android Studio IDE Koala 2024, Java, Firebase, Google Colab and Tensor Flow.

2) CNN-based application workflow diagram. Fig. 1 presents the workflow diagram of the convolutional neural network (CNN)-based application for rosacea detection that starts with the acquisition and augmentation of images resized to 224x224 pixels, which are divided into training, validation and test sets. The deep learning model is trained to identify rosacea features, tuned with the validation set, and the Application architecture diagram is evaluated with the test set. Once optimized, the model is integrated into the mobile app, which classifies the images and saves the results, providing a preliminary diagnosis.

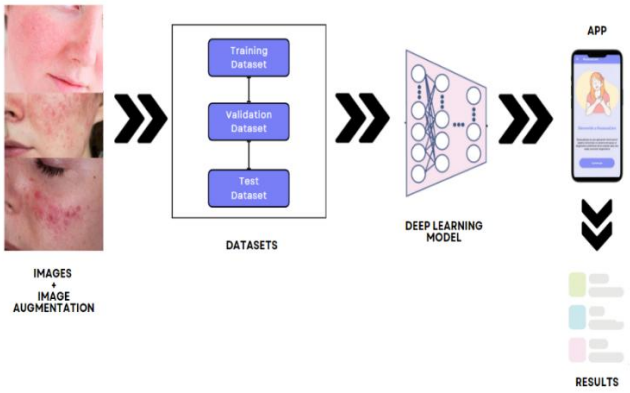


Fig. 1. Workflow diagram of the CNN-based application

a) Application architecture diagram: The mobile application was developed using Android Studio, with integration of TensorFlow for image classification and Firebase as a real-time database. Python and Google Colab were used for training and tuning the classification model, while Java was used for implementing the business logic in the mobile application (see Fig. 2).

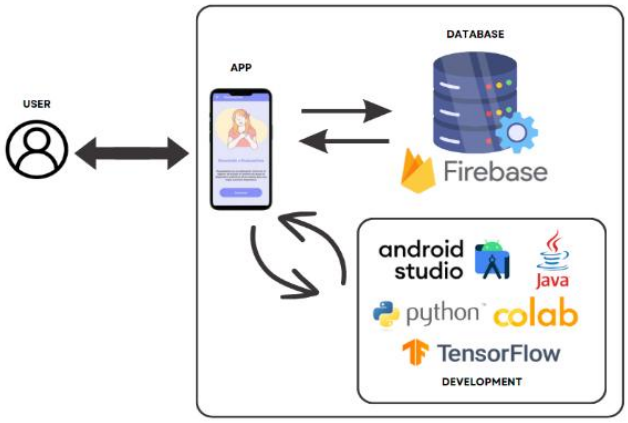


Fig. 2. Application architecture diagram

3) Android project structure. The project structure in Android Studio is organized hierarchically for easy code management and maintenance. The project includes a manifest.xml file that contains application settings and metadata. The source code (Java) directory contains Java files that implement the application logic. The res folder is where the resources of the application are stored. Within res, there are subfolders such as drawable for images, layout for XML files that define the user interface, mipmap for application icons and values for resources such as text strings and colors. The menu folder contains XML files that define the application's menus, such as nav_menu.xml, which specifies the navigation menu items. Finally, the model.tflite file in the assets folder is the deep learning model in TensorFlow Lite format.

D. Phase 3: Production

Work is underway to deliver incremental versions of the application, adding features such as detection, visualization of results, and integration of intuitive interfaces (see Table II).

TABLE II MODULE PLAN

Module	Code	Process	Requirement
User Authentication Module	M01	User authentication process, allowing the user to log in and save their diagnostic data.	RF01
Rosacea Information Module	M02	Process of presenting information about the disease rosacea, including symptoms, causes and recommended treatments.	RF02
Instructions Module	M03	Process of presenting instructions and recommendations on how to take pictures of your face properly.	RF03
Image Upload Module	M04	Process of classifying images using a machine learning model, which determines whether rosacea is present or not.	RF04
Results Viewing Module	M05	Process of visualization of the classification results. Displays diagnostic probabilities and conclusion (positive/negative).	RF05

Report Generation Module	M06	Report generation and visualization process, showing diagnostic results in tabular form, including probabilities and diagnosis.	RF06
Performance Metrics Module	M07	Application performance metrics visualization and report generation process.	RF07
About the Application Module	M08	Process of presenting information about the application, including its purpose, functionalities.	RF08

E. Stabilization

Extensive testing and bug fixes, optimizing performance and ensuring product quality prior to release [21]. The objective is to ensure that the application is ready to be used by end-users without critical problems.

1) Recommendations for mobile equipment. It is recommended to use a mobile device with sufficient storage space, and it is suggested that the device has access to a stable internet connection to perform the analysis in the cloud and synchronize the results with the database (see Table III).

TABLE III RECOMMENDATION FOR THE MOBILE PHONE

Hardware	Software
• RAM: 4GB or more. • Display: HD resolution (720x1280).	• Operating system: Android 8.0 or higher. • Network: 4G

F. Testing

The testing phase involves the final validation of the application under real conditions of use. A detailed review is carried out to ensure that all implemented functionalities meet the requirements established during the exploration phase. Continuous support is provided, and the product is updated with enhancements and new functionalities based on user feedback, version management and post-release problem solving.

1) Unit test 05. Results Display Module: The unit test on the results display module is shown (see Table IV).

TABLE IV TEST RESULT DISPLAY TEST

Code	Name
M05	• Results Display Module
Target	• The application displays the results of the image classification to the user.
Steps	• Access the results module after classification. • Review probabilities and diagnosis. • View analysis details for each loaded image.
Results achieved	• Classification results are clearly displayed to the user. • Probabilities and diagnosis are displayed correctly. • Information is presented without errors or interruptions.

IV. RESULTS

The aim of this research is to determine whether a mobile application using convolutional neural networks improves the preliminary diagnosis of rosacea.

A. Calculation of Indicators

Table V shows a confusion matrix showing the performance of the application with 100 images. Of 50 images of people with rosacea, 47 were correctly classified as positive, while 3 were incorrectly classified as negative. Of 50 images of people without rosacea, 49 were correctly identified as negative, and 1 was incorrectly classified as positive. With a sensitivity of 94% and a specificity of 98%, the model shows an accuracy of 96%. The matrix was developed using SPSS software to analyze the results of the application and evaluate the ability of the classification model on a controlled set of images, this analysis allowed the effectiveness of the model to be determined.

TABLE V CONFUSION MATRIX

	Prediction					
Current		Sensitivity	Specificity	Total	**Function test**	
		Person with rosacea	Person without rosacea			
	Positive	**47**	**1**	48	0.9791666667	Rosacea
	Negative	**3**	**49**	52	0.9423076923	Without rosacea
	Total	50	50	**100**		
	Laboratory test	0.94	0.98		0.96	

From the confusion matrix analysis, the following results were obtained:

% Sensitivity = 94%

% Specificity = 98%

% Accuracy = 96%

1) Analysis of the ROC curve and the area under the curve (AUC). The ROC curve, represented in blue, is shown in Fig. 3, which shows outstanding performance, characterized by high sensitivity and a low false positive rate. In contrast, the red diagonal line represents a random classifier with an AUC of 0.5, equivalent to making random predictions. The blue curve, located above this line, confirms significantly superior performance, highlighting the model's ability to make accurate classifications during testing.

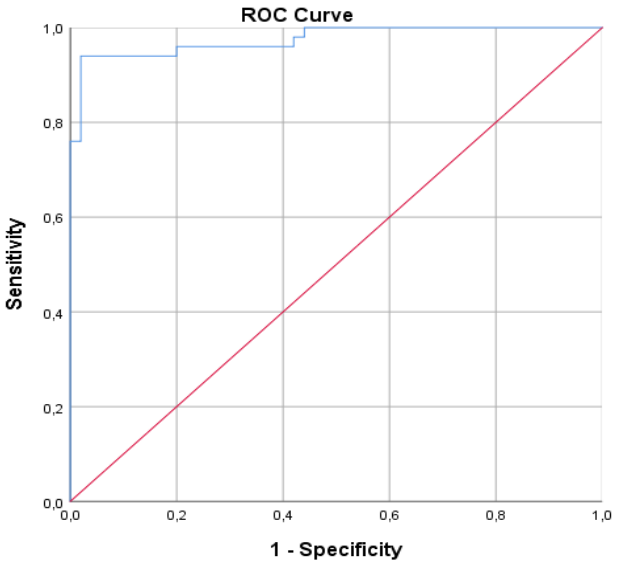


Fig. 3. ROV curve.

The AUC value is 0.975, indicating that the app has a 97.5% ability to correctly classify between positive and negative cases of rosacea (see Table VI).

TABLE VI AREA UNDER THE CURVE

Area	Std. Error	Asymptotic significance	95% asymptotic confidence interval	
			Lower limit	Upper limit
,975	,014	,000	,948	1,000

B. Analysis of the Pre-Experimental Design

a) Pre-test: Table VII shows the accuracy, sensitivity and specificity values of the EfficientNet-b4 model for different dermatological diseases; to verify the increase in percentages in the indicators, it was compared with the rosacea values of the model with the proposed mobile application [22].

TABLE VII PRE-TEST VALUES RESULTS

Disease category	Accuracy	Sensitivity	Specificity
Overall	0.948	0.934	0.950
Lichen planus	0.969	0.873	0.975
Rosacea	0.959	0.920	0.961
Viral warts	0.932	0.944	0.930
Acne vulgaris	0.935	0.974	0.927
Keloid and hypertrophic scar	0.969	0.934	0.970
Eczema/dermatitis	0.877	0.924	0.866
Dermatofibroma	0.987	1.000	0.987
Seborrheic dermatitis	0.926	0.811	0.933
Seborrheic keratosis	0.953	0.858	0.957
Melanocytic nevus	0.960	0.961	0.960
Hemangioma	0.984	0.975	0.984
Psoriasis	0.886	0.920	0.882
Port wine stain	0.963	0.989	0.961
Basal cell carcinoma	0.979	0.971	0.979

b) Post-test: The post-test was performed with 100 images from the dataset, the results of which were obtained with the report generated from the mobile application. Table VIII shows the metrics and percentages of the indicators, indicating sensitivity of 94%, specificity of 98% and accuracy of 96%.

TABLE VIII RESULTS OF THE INDICATORS FOR THE POST-TEST

Quantity	VP	VN	FN	FP	Sensitivity	Specificity	Accuracy
100	47	49	3	1	94%	98%	96%

Table IX shows the comparison between the percentages of sensitivity, specificity and accuracy indicators between the EfficientNet-b4 model and the mobile application.

TABLE IX COMPARISON OF RESULTS

Indicator	Efficientnet-B4	Mobile application
Sensitivity	92%	94%
Specificity	96.1%	98%
Accuracy	95.9%	96%

C. Hypothesis Testing

a) Specific scenario HE1: HE10: A mobile application using convolutional neural networks does not increase sensitivity in the preliminary diagnosis of rosacea.

HE11: A mobile app using convolutional neural networks increases sensitivity in the preliminary diagnosis of rosacea.

% Sensitivity variation = (94-92) / 92 X 100 = 2.17%

An increase in sensitivity was found in the percentage variation, obtaining a positive result of 2.17%; therefore, the alternative hypothesis is accepted, and the null hypothesis is rejected.

b) Specific scenario HE2: HE20: A mobile app using convolutional neural networks does not increase specificity in the preliminary diagnosis of rosacea.

HE21: A mobile app using convolutional neural networks increases specificity in the preliminary diagnosis of rosacea.

% Change in specificity = (98-96.1) / 96.1 X 100 = 1.97%

An increase in specificity was found in the percentage variation, giving a positive result of 1.97%; therefore, the alternative hypothesis is accepted, and the null hypothesis is rejected.

c) Specific scenario HE3: HE30: A mobile app using convolutional neural networks does not increase accuracy in the preliminary diagnosis of rosacea.

HE31: A mobile app using convolutional neural networks increases the accuracy in the preliminary diagnosis of rosacea.

% Variation in accuracy = (96-95.9) / 95.9 X 100 = 0.10%

An increase in accuracy was found in the percentage variation, giving a positive result of 0.10%; therefore, the alternative hypothesis is accepted, and the null hypothesis is rejected.

D. Results of Indicator Increments

Table X shows the percentage increments for each indicator, a positive number is the reason for acceptance of the hypothesis and a negative number is the reason for the increase for each indicator.

TABLE X RESULTS OF THE INCREMENTS FOR EACH INDICATOR

Increased sensitivity	Increased specificity	Increased accuracy
2.17%	1.97%	0.10%

E. General Hypothesis

HG0: A mobile application using convolutional neural networks does not improve the preliminary diagnosis of rosacea.

HG1: A mobile application using convolutional neural networks improves the preliminary diagnosis of rosacea.

By verifying the specific hypotheses HE1, HE2, and HE3, the general hypothesis is accepted.

V. DISCUSSION

The results showed a sensitivity of 94%, resulting in an increase of 2.17%, due to the use of advanced convolutional neural network architectures, which allowed accurate extraction of rosacea-related features. This result is lower than that obtained by [18], who achieved a sensitivity of 95.2% in skin cancer recognition, possibly because they used the ResNet34 architecture previously trained with ImageNet, together with the cross-entropy loss function and the Adam optimizer. Also, the results of this study are lower than those obtained by [20], which was 99.01% in the detection of actinic and seborrheic keratosis; this difference was possibly due to the use of a larger dataset consisting of 2000 high-resolution images. This shows that, although an increase in sensitivity for rosacea diagnosis was achieved, performance depends on factors such as network architecture, dataset size and preprocessing techniques.

In addition, 98% specificity and an increase of 1.97% were obtained because of image preprocessing and data enhancement techniques, which improved the model's ability to differentiate patterns, contributing significantly to the reduction of false positives in diagnosis. This outperformed the result of [18], which only scored 59.2% when detecting skin cancer. This could be attributed to class imbalance in the HAM10000 dataset, with some classes underrepresented. When compared to the study of [20], which obtained 100% specificity and with a higher value than this research, is an outstanding result due to factors such as a balanced and high-quality dataset, as well as distinctive visual patterns that facilitate the differentiation between actinic and seborrheic keratosis in high quality dermatological images. Based on the above, it is evident that image preprocessing and data augmentation contribute to the creation of more robust models with greater specificity.

On the other hand, the results revealed an accuracy of 96%, resulting in an increase of 0.10%, attributed to the training of the model with a carefully labeled high-resolution dataset, which shows its ability to identify relevant features of the dermatological condition. This result is superior to that obtained by [23], where the accuracy was 88-90% in detecting rosacea, this may be because the original images have a high resolution (4608 × 3072 pixels), but were divided into patches of different sizes (64 × 64, 128 × 128, etc.), which may have limited the ability of the model to capture global features. However, the accuracy achieved in this work is lower than that reported by [19], who achieved an accuracy of 99.75% in the detection of brain tumors, as a possible result of the use of fine-tuning and the large number of images used. Taken together, these findings demonstrate that accuracy is strongly influenced by factors such as the quality and resolution of the images used in the model. Skin diseases pose a major global health challenge, and emerging digital tools show promise in enhancing their diagnosis and treatment [24]. Finally, the results confirmed that the application significantly increased the key diagnostic indicators on the basis that in the post-test, there was a marked improvement in the sensitivity, specificity and accuracy of the results.

VI. CONCLUSION

The developed mobile application, based on convolutional neural networks, represents a significant advance in the preliminary diagnosis of rosacea, by considerably improving the sensitivity, specificity and accuracy of the model. A sensitivity of 94% with a variance of 2.7% was achieved, indicating a high ability to minimize false negatives and ensure that affected patients are correctly identified. This is attributed to the innovative implementation of the network architecture, advanced data augmentation techniques and the diversity of the training set, which contributed to better generalization of the model. Likewise, specificity reached 98% with a variation of 1.97%, reflecting a higher accuracy in the identification of negative cases, a result positively influenced by a rigorous image preprocessing and data enhancement process. Finally, the overall accuracy of the model was 96% with a variation of 0.10%, evidencing a substantial improvement in predictive capacity, is supported by using a high-quality data set and rigorous labeling. This study not only provides scientific value by demonstrating the effectiveness of convolutional neural networks applied on mobile devices for dermatological diagnosis but also offers an accessible and practical approach that can facilitate early diagnosis and improve clinical management of rosacea, especially in settings with limited access to specialists.

Among the main limitations of the system is the need to improve diagnostic accuracy, especially in complex cases or with low-quality images, which could affect reliability in diverse clinical settings. In addition, the model currently does not include additional evaluation parameters, such as F1-score or processing time, which would allow a more comprehensive and objective comparison with other existing methodologies. Another relevant limitation is the absence of user feedback mechanisms, which hinders the early identification of errors and the continuous incorporation of improvements based on real experience.

To overcome these limitations, future research should consider implementing ensemble models that combine different architectures to reduce individual errors and increase diagnostic robustness. It is also advisable to extend the evaluation parameters to include metrics such as F1-score, accuracy, recall, and runtime analysis to facilitate comprehensive comparisons with other approaches. It is also suggested to develop a direct feedback module for users to report bugs or suggestions, thus driving continuous improvement of the application. Finally, the incorporation of multilingual support will favor the accessibility and global adoption of the system, adapting it to different cultural and linguistic contexts.

REFERENCES

[1] D. Piccolo, I. Fusco, T. Zingoni, and C. Conforti, "Effective Treatment of Rosacea and Other Vascular Lesions Using Intense Pulsed Light System Emitting Vascular Chromophore-Specific Wavelengths: A Clinical and Dermoscopical Analysis," J Clin Med, vol. 13, no. 6, p. 1646, Mar. 2024, doi: 10.3390/jcm13061646.

[2] G. Galluccio et al., "Advances in the Pathogenesis and Treatment of Rosacea: A Phenotype-Based Therapeutic Approach," Cosmetics, vol. 11, no. 1, p. 11, Feb. 2024, doi: 10.3390/cosmetics11010011.

[3] P. Abdi, Z. Haq, M. J. Diaz, and H. I. Maibach, "Psychiatric comorbidities associated with rosacea: a propensity score-matched case–control study in the All of Us database," Clin Exp Dermatol, vol. 49, no. 4, pp. 400–403, Apr. 2024, doi: 10.1093/ced/llad417.

[4] Z. Yang et al., "Quality of life, sleep and anxiety status among patients with rosacea in the Yunnan plateau region: A 2-year retrospective study," Skin Research and Technology, vol. 30, no. 3, p. e13616, Mar. 2024, doi: 10.1111/srt.13616.

[5] K. Zujko-Kowalska, J. Masłowska, M. Knaś-Dawidziuk, J. Hamulka, and M. E. Zujko, "Dietary Antioxidants May Support Cosmetic Treatment in Patients with Rosacea," Antioxidants, vol. 13, no. 3, p. 381, Mar. 2024, doi: 10.3390/antiox13030381.

[6] Y. Pan, K. Jia, S. Yan, and X. Jiang, "Effectiveness of VISIA system in evaluating the severity of rosacea," Skin Research and Technology, vol. 28, no. 5, p. 740, Sep. 2022, doi: 10.1111/SRT.13194.

[7] Z. Qi, F. Wang, Y. Huang, and P. Wang, "Analysis of the correlation between skin barrier function and age in rosacea patients in Qinghai region," J Cosmet Dermatol, vol. 23, no. 3, pp. 999–1003, Mar. 2024, doi: 10.1111/jocd.16041.

[8] P. V. Chernyshov et al., "Bullying in persons with skin diseases," Journal of the European Academy of Dermatology and Venereology, vol. 38, no. 4, pp. 752–760, Apr. 2024, doi: 10.1111/jdv.19683.

[9] L. Montali and S. Garnieri, "Feeling uncomfortable in your own skin: a qualitative study of problematic skin picking in Italian women," Current Psychology, vol. 43, no. 14, pp. 12870–12881, Apr. 2024, doi: 10.1007/s12144-023-05377-4.

[10] L. Ji and X. Hongfu, "How to avoid misdiagnosis of rosacea[如何避免玫瑰痤疮的误诊]," Chinese Journal of Dermatology, vol. 57, no. 2, pp. 119–122, Feb. 2024, doi: 10.35541/cjd.20230419.

[11] Y. Peng et al., "Clinical characteristics of the well-defined upper eyelid vascular network pattern in patients with rosacea," Int J Dermatol, vol. 63, no. 3, pp. 337–344, Mar. 2024, doi: 10.1111/ijd.16946.

[12] Y. R. Woo and H. S. Kim, "Deciphering Childhood Rosacea: A Comprehensive Review," J Clin Med, vol. 13, no. 4, p. 1126, Feb. 2024, doi: 10.3390/jcm13041126.

[13] R. O. Flores-Castañeda, S. Olaya-Cotera, M. López-Porras, E. Tarmeño-Juscamaita, and O. Iparraguirre-Villanueva, "Technological advances and trends in the mining industry: a systematic review," Mineral Economics, pp. 1–16, Jul. 2024, doi: 10.1007/S13563-024-00455-W/METRICS.

[14] S. S. Noronha, M. A. Mehta, D. Garg, K. Kotecha, and A. Abraham, "Deep Learning-Based Dermatological Condition Detection: A Systematic Review With Recent Methods, Datasets, Challenges, and Future Directions," IEEE Access, vol. 11, pp. 140348–140381, 2023, doi: 10.1109/ACCESS.2023.3339635.

[15] H. Binol, M. K. K. Niazi, A. Plotner, J. Sopkovich, B. H. Kaffenberger, and M. N. Gurcan, "A multidimensional scaling and sample clustering to obtain a representative subset of training data for transfer learning-based rosacea lesion identification," Progress in Biomedical Optics and Imaging - Proceedings of SPIE, vol. 11314, p. 1131415, Jan. 2020, doi: 10.1117/12.2549392.

[16] S. R. Sahoo, R. Dash, and R. K. Mohapatra, "A customized deep learning framework for skin lesion classification using dermoscopic images," Comput Animat Virtual Worlds, vol. 34, no. 5, p. e2132, Oct. 2023, doi: 10.1002/cav.2132.

[17] H. K. Jeong, C. Park, R. Henao, and M. Kheterpal, "Deep Learning in Dermatology: A Systematic Review of Current Approaches, Outcomes, and Limitations," JID Innov, vol. 3, no. 1, p. 100150, Jan. 2023, doi: 10.1016/J.XJIDI.2022.100150.

[18] P. Tschandl et al., "Human–computer collaboration for skin cancer recognition," Nat Med, vol. 26, no. 8, pp. 1229–1234, Aug. 2020, doi: 10.1038/s41591-020-0942-0.

[19] S. K. Mathivanan, S. Sonaimuthu, S. Murugesan, H. Rajadurai, B. D. Shivahare, and M. A. Shah, "Employing deep learning and transfer learning for accurate brain tumor detection," Sci Rep, vol. 14, no. 1, p. 7232, Dec. 2024, doi: 10.1038/s41598-024-57970-7.

[20] S. Reddy, D. Giri, and R. Patel, "Artificial Intelligence-Based Distinction of Actinic Keratosis and Seborrheic Keratosis," Cureus, vol. 16, no. 4, p. e58692, Apr. 2024, doi: 10.7759/CUREUS.58692.

[21] Y. D. Balaguera Amaya, "Metodologías ágiles en el desarrollo de aplicaciones para dispositivos móviles. Estado actual," Revista de Tecnología, ISSN 1692-1399, Vol. 12, No. 2, 2013 (Ejemplar dedicado a: Transportes sustentables), págs. 111-123, vol. 12, no. 2, pp. 111–123, 2013, Accessed: Jun. 26, 2024. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=6041502&info=resumen&idioma=SPA

[22] C. Y. Zhu et al., "A Deep Learning Based Framework for Diagnosing Multiple Skin Diseases in a Clinical Environment," Front Med (Lausanne), vol. 8, p. 626369, Apr. 2021, doi: 10.3389/FMED.2021.626369/BIBTEX.

[23] H. Binol, A. Plotner, J. Sopkovich, B. Kaffenberger, M. K. K. Niazi, and M. N. Gurcan, "Ros-NET: A deep convolutional neural network for automatic identification of rosacea lesions," Skin Research and Technology, vol. 26, no. 3, pp. 413–421, 2020, doi: 10.1111/srt.12817.

[24] A. F. Sapaico-Alberto, S. Olaya-Cotera, and R. O. Flores-Castañeda, "Analysis of the use of digital technologies in the preliminary diagnosis of dermatological diseases: a systematic review," Arch Dermatol Res, vol. 317, no. 1, pp. 1–11, Dec. 2025, doi: 10.1007/S00403-024-03650-5/METRICS.

Hybrid PSO-ACO Optimization for Rice Leaf Disease Classification Using Random Forest and Support Vector Machines

Avip Kurniawan[1], Tri Retnaningsih Soeprobowati[2], Budi Warsito[3]
Department of Information System, School of Postgraduate Studies, Universitas Diponegoro, Semarang, Indonesia[1]
Department of Informatics, Faculty of Engineering Universitas Krisnadwipayana, Jakarta, Indonesia[1]
Department of Biology, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia[2]
Cluster for Paleolimnology Universitas Diponegoro, Semarang, Indonesia[2]
Department of Statistics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia[3]

***Abstract*—This study proposes a hybrid machine learning framework for rice leaf disease detection by combining handcrafted feature extraction with metaheuristic optimization and classical classifiers. Using a dataset of 6,000 rice leaf images across seven classes, features including color, texture, shape, and edge were extracted and optimized using Spider Monkey Optimization (SMO), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO). Classification was conducted using Random Forest Classifier (RFC) and Support Vector Classifier (SVC), both with and without hyperparameter tuning. Experimental results revealed that PSO consistently outperformed other optimizers, achieving 91.00% accuracy with RFC and 94.64% with SVC when all features and optimal parameters were used. While SMO also showed strong performance, ACO yielded less consistent results. These findings highlight the importance of combining comprehensive feature engineering with adaptive optimization strategies to improve classification accuracy. Compared to previous SMO-based approaches, the proposed PSO-ACO framework demonstrated improved stability and scalability. The proposed framework is interpretable, efficient, and scalable, making it suitable for practical deployment in precision agriculture. Future research directions include integrating deep learning with handcrafted features, developing adaptive metaheuristics, and implementing real-time mobile detection systems.**

***Keywords*—Rice leaf disease; particle swarm optimization (PSO); support vector machine (SVM); feature extraction; precision agriculture**

I. INTRODUCTION

Rice (*Oryza sativa*) remains a staple food for more than half of the global population, playing a pivotal role in ensuring food security particularly in Asia and Africa [1]. With rising global demand, sustainable rice production has become a critical priority. However, rice cultivation faces persistent threats from a variety of leaf diseases such as Bacterial Leaf Blight, Leaf Blast, and Tungro, which can lead to yield losses of up to 70% in severely infected areas [2],[3]. Early and accurate identification of rice leaf diseases is therefore essential to mitigate such risks and enable effective precision agriculture. Traditional diagnostic techniques such as manual inspection and laboratory testing are labor-intensive, time-consuming, and often prone to human error, limiting their scalability [4]. To overcome these challenges, recent advances in machine learning and computer vision have been harnessed to develop automated disease detection systems with improved accuracy and efficiency [5], [6].

In our previous work, we introduced a hybrid detection model named SMOREF-SVM, which integrates Spider Monkey Optimization (SMO), Random Forest (RF), and Support Vector Machine (SVM) [7]. This model achieved a remarkable classification accuracy of up to 98% for rice leaf diseases. However, despite its success, the SMOREF-SVM approach posed challenges related to optimization efficiency and computational complexity. As with many swarm-based algorithms, SMO often suffers from unstable convergence behavior in high-dimensional feature spaces [8]. In response, alternative metaheuristic strategies such as Particle Swarm Optimization (PSO) [9] and Ant Colony Optimization (ACO) [10] have been considered promising due to their robust search capabilities and efficiency in feature selection tasks. This extended study applies PSO and ACO to optimize handcrafted features extracted from rice leaf images. Two machine learning classifiers, Random Forest and Support Vector Machine, are employed to evaluate classification performance, both with and without hyperparameter tuning. The proposed framework introduces a PSO-ACO-based hybrid optimization pipeline designed to enhance accuracy and efficiency in rice leaf disease classification. This study includes a comparative analysis between PSO and ACO against the previously established SMO-based approach. Furthermore, it examines how different feature types color, texture, shape, and edge, impact classification accuracy across the two classifiers. The role of hyperparameter tuning is also analyzed to assess its influence when combined with metaheuristic-based optimization. The findings support the hypothesis that PSO and ACO are viable and scalable alternatives to SMO in plant disease detection frameworks.

The application of artificial intelligence in plant disease detection, especially using image-based analysis, has gained considerable momentum in recent years. Panchal et al. [11] demonstrated the capability of convolutional neural networks (CNNs) in identifying plant diseases across fourteen different crop species, including rice, using a large publicly available

dataset. Although CNNs offer high accuracy, they typically require large amounts of labeled data and substantial computational resources. Conversely, traditional machine learning approaches remain attractive for agricultural settings due to their lower hardware requirements and adaptability to smaller datasets. Padhi and Mishra [6] achieved high accuracy in rice leaf disease classification using SVM and handcrafted color-texture features. Similarly, Sunil et al. [12] reported that combining Gray Level Co-occurrence Matrix (GLCM) features with color histograms significantly improved classification performance in tomato leaf disease detection.

Metaheuristic optimization techniques have also played a critical role in enhancing machine learning models. Nguyen et al. [13] employed PSO to select the most relevant features for maize disease classification, achieving reduced feature dimensionality while maintaining performance. ACO has also proven effective in optimizing SVM hyperparameters, as demonstrated by Demilie [14], who observed improved classification metrics compared to grid search methods.

Our prior research [7] incorporated SMO into a hybrid SMOREF-SVM model for rice leaf disease detection. The model combined SMO for feature selection, Random Forest for pre-classification, and SVM for final decision-making, reaching an accuracy of 98%. Despite these promising results, the model's complexity and convergence limitations prompted the exploration of PSO and ACO as alternative optimization strategies. While deep learning architectures such as CNNs dominate in large-scale implementations, their demand for annotated data and longer training times makes them less feasible in lightweight or real-time agricultural scenarios. In contrast, swarm intelligence-based optimization combined with classical classifiers offers a more balanced and interpretable solution. Support Vector Machine (SVM) is a powerful supervised learning algorithm that has been widely adopted for classification tasks, including plant disease detection. SVM excels in handling high-dimensional data by constructing optimal hyperplanes that maximize the margin between different classes [7]. This characteristic makes it particularly suitable for tasks involving complex and overlapping feature distributions, such as differentiating between multiple types of rice leaf diseases.

The ability of SVM to transform nonlinear data into a higher-dimensional space using kernel functions further strengthens its applicability in image analysis [6]. One of the most compelling strengths of SVM lies in its generalization capability, especially when the available dataset is limited. In the context of rice disease detection, where collecting extensive labeled datasets can be labor-intensive and costly, SVM provides a robust alternative to data-hungry deep learning models. Several studies have reported the effective use of SVM in detecting diseases in rice, wheat, tomato, and maize leaves using features such as color histograms, texture patterns, and shape descriptors [6], [7]. This flexibility has positioned SVM as a go-to classifier in many precision agriculture systems. Moreover, SVM has shown great compatibility with metaheuristic algorithms such as Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Genetic Algorithms. These combinations aim to optimize SVM hyperparameters like the penalty term (C), kernel type, and gamma values, which are crucial for achieving high classification accuracy. In hybrid systems, SVM often serves as the final decision layer due to its high discrimination power, making it a valuable component in frameworks like SMOREF-SVM and the proposed PSO-ACO-SVM in this study.

Random Forest (RF) is an ensemble learning algorithm that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [8]. It is particularly well-suited for applications involving noisy, high-dimensional, or incomplete data. In plant disease detection, RF has been used extensively due to its robustness, ease of use, and superior accuracy compared to individual decision trees or simpler models. Its in-built capability for feature importance ranking also makes it highly interpretable, which is valuable in agricultural diagnostics. RF is known for its resistance to overfitting, particularly when a large number of decision trees are used. This attribute is especially beneficial in agricultural settings where environmental variability and image noise may affect input data. Studies have demonstrated that RF performs well in classifying various plant diseases using combinations of handcrafted features such as color moments, GLCM-based texture descriptors, and edge patterns extracted from leaf images [5], [6]. Its low sensitivity to outliers and non-normalized data further adds to its advantages in real-world scenarios.

Similar to SVM, RF has been integrated with optimization algorithms to improve its performance. Hyperparameter tuning methods, including Grid Search, Random Search, and metaheuristics like SMO and PSO, are commonly used to determine the best combination of tree depth, number of estimators, and feature subset size. In the SMOREF-SVM model, RF was used as an intermediary classifier, providing additional predictive power before final classification with SVM. In the current study, RF is re-evaluated within a PSO-ACO framework to investigate its potential as a standalone or complementary classifier in rice leaf disease detection.

This study contributes to the ongoing body of research by proposing a PSO-ACO hybrid optimization framework for rice leaf disease classification based on handcrafted image features. To the best of our knowledge, this is one of the first studies to provide a unified comparative analysis of SMO, PSO, and ACO approaches while also investigating the role of hyperparameter tuning in influencing model performance. Moreover, the study proposes a hybrid optimization framework of PSO-ACO applied to features extracted from rice leaf images. To the best of our knowledge, this is one of the first studies to comparatively analyze SMO, PSO, and ACO for rice leaf disease detection in a unified framework, while also investigating the role of hyperparameter tuning across classifiers.

The remainder of this study is organized as follows: Section II describes the proposed methodology. Section III presents the experimental setup and results. Section IV discusses findings and comparative evaluations. Finally, Section V concludes the study and outlines directions for future work.

A. Literature Review

Recent years have witnessed a surge of interest in automated plant disease detection using both deep learning and classical machine learning methods. Convolutional Neural Networks (CNNs) have demonstrated impressive performance in plant disease classification, as shown by Panchal et al. [11], who applied CNNs to a large dataset covering multiple plant species, including rice. Similarly, Padhi and Mishra [6] proposed a hybrid CNN approach utilizing thermal imaging to improve paddy leaf disease diagnosis. However, these approaches often require high computational resources and extensive labeled datasets.

Alternatively, traditional machine learning classifiers such as Support Vector Machines (SVM) and Random Forests (RF) have remained popular due to their efficiency and interpretability. Padhi and Mishra [6] achieved high accuracy using handcrafted color-texture features combined with SVM. Sunil et al. [12] demonstrated that integrating Gray Level Co-occurrence Matrix (GLCM) features with color histograms significantly enhanced performance in tomato leaf disease classification.

Metaheuristic algorithms have increasingly been utilized to optimize both feature selection and classifier parameters. Nguyen et al. [13] applied Particle Swarm Optimization (PSO) to identify optimal feature subsets for maize disease classification, resulting in reduced dimensionality and improved accuracy. Ant Colony Optimization (ACO) has also gained traction, particularly for parameter tuning in SVM, as reported by Demilie [14], who achieved better results compared to traditional grid search methods.

Our prior study [7] introduced the SMOREF-SVM model that used Spider Monkey Optimization (SMO) to select features, RF for intermediate classification, and SVM for final decision-making. Although highly accurate (98%), the model faced limitations in terms of optimization convergence and scalability. This current work explores PSO and ACO as more robust alternatives.

Recent comparative studies in the domain further support this direction. For instance, the study by Lakshmanaprabu et al. [15] reported a 93.40% accuracy using DWT, SIFT, and GLCM for rice disease classification. Tyralis et al. [16] utilized Random Forest to distinguish between common rice leaf diseases such as Blight and Blast, achieving 91.47% accuracy. Sheykhmousa et al. [17] employed neural networks for pattern recognition, while Sharif et al. [18] proposed an optimized segmentation approach for citrus diseases.

Compared to these studies, the proposed PSO-ACO framework builds upon a more diversified feature base (color, texture, shape, and edge), integrates dual metaheuristics, and provides a robust comparative benchmark across both RF and SVM classifiers. Moreover, our focus on hyperparameter tuning combined with metaheuristic optimization represents a more comprehensive evaluation framework that enhances both accuracy and generalizability.

II. Research Method

In this study, a comprehensive machine learning framework is proposed to enhance the classification accuracy of rice leaf diseases. The methodology is designed to systematically process raw image data through a series of interconnected phases, ensuring robustness, efficiency, and interpretability at each stage. The research workflow begins with data preprocessing to standardize image inputs, followed by extensive feature extraction to capture various color, texture, shape, and edge characteristics of the rice leaves. Several metaheuristic algorithms, including Spider Monkey Optimization (SMO), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO), are employed to optimize the extracted features and reduce redundancy. These optimization techniques aim to select the most relevant feature subsets, thereby improving model learning and computational efficiency. Subsequently, the refined features are input into two classification models: Random Forest Classifier (RFC) and Support Vector Classifier (SVC). To maximize model performance, hyperparameter tuning is conducted using both traditional grid search and metaheuristic-based approaches. Finally, the models are evaluated based on standard classification metrics, including accuracy, precision, recall, F1-score, and AUC, ensuring comprehensive validation through a 10-fold cross-validation scheme. The overall research methodology is illustrated in Fig. 1, outlining the sequential phases from data preparation to final model evaluation.

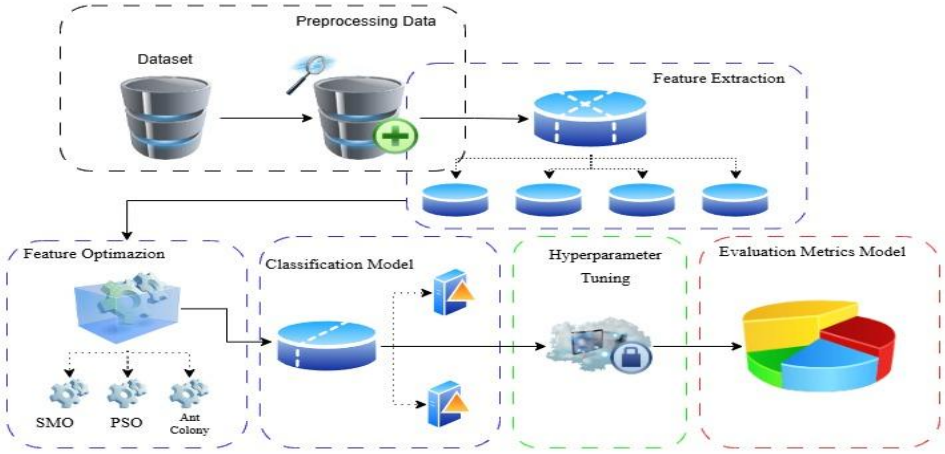


Fig. 1. Workflow of the proposed rice leaf disease classification framework, illustrating the data preprocessing, feature extraction, feature optimization, classification, hyperparameter tuning, and evaluation phases.

B. Dataset Description

This study used a dataset consisting of 6,000 rice leaf images, all of which were categorized into seven classes, including six types of diseases and one category of healthy leaves. Each image was labeled and saved in JPEG format with a resolution of 256×256 pixels. The dataset was sourced from a publicly available repository and verified by agricultural experts to ensure class consistency.

C. Data Preprocessing

Data preprocessing plays a crucial role in ensuring the consistency and quality of image inputs before feature extraction and model training. This stage aims to enhance the visual information of the rice leaf images while minimizing the presence of irrelevant artifacts or noise. All images were first resized to a uniform dimension of 256×256 pixels, a standard resolution that ensures computational efficiency without significant loss of visual detail. Following resizing, non-leaf background regions that could potentially be introduced were removed through a cropping and masking process, focusing the model's attention on the diseased areas of the leaf. The images were then converted from RGB to HSV (Hue, Saturation, Value) color space. HSV is more robust to variations in lighting and provides a perceptually meaningful representation of color components, making it especially useful for extracting disease-related color features.

Histogram equalization was applied to enhance contrast and detail. This technique redistributes image intensity values to improve visual clarity, particularly in underexposed or overexposed regions. Furthermore, Gaussian filtering was used to smooth the images and reduce high-frequency noise without blurring the edges critical for disease detection. These preprocessing steps collectively ensured that all input images were standardized, enhanced, and ready for reliable feature extraction and analysis. Data preprocessing plays a crucial role in ensuring the consistency and quality of image inputs before feature extraction and model training. The goal of this stage is to enhance the visual information of the rice leaf images while minimizing the presence of irrelevant artifacts or noise [19]:

$$I_{resized}(x,y) = \sum_{i=0}^{1} \sum_{n=1}^{\infty} W_{ij} \,.\, I(i + \lfloor x \rfloor, j + \lfloor y \rfloor) \quad (1)$$

Background pixels were removed using binary masking [20], where the binary mask $M \in \{0,1\}$ isolates the leaf region:

$$I_{masked}(x,y) = I_{resized}(x,y) \,.\, M\,(x,y) \quad (2)$$

The images were then transformed into the HSV color space [20]:

$$\mathrm{H} = \begin{cases} 60^0 \,.\, \frac{G-B}{\Delta}, & if\ C_{max} = R \\ 60^0 \,.\, (2 + \frac{B-R}{\Delta}), & if\ C_{max} = G \\ 60^0 \,.\, (4 + \frac{R-G}{\Delta}), & if\ C_{max} = B \end{cases}, \; S = \frac{\Delta}{C_{max}}, V = C_{max} \quad (3)$$

where, $\Delta = C_{max} - C_{min}$

To improve contrast, histogram equalization [19] was applied:

$$P_{equalized}(i) = \frac{1}{MN} \sum_{j=0}^{i} h\,(j) \quad (4)$$

Lastly, Gaussian filtering [21] was used to reduce noise

$$G(x,y) = \frac{1}{2\pi\sigma^2}\, exp\, \left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (5)$$

D. Feature Extraction

Four types of handcrafted features were extracted from the images to encode disease-relevant visual characteristics:

1) Color features. HSV histograms were calculated for each image channel to capture color distribution and intensity. Let $h_c(i)$ denote the histogram value for bin i in channel bin in channel c $\in \{H, S, V\}$.

2) Texture features. Texture features were computed using the Gray Level Co-occurrence Matrix (GLCM) [22], including contrast C, correlation ρ, and energy E:

$$\mathrm{C} = \sum_{i,j} (i-j)^2 \mathrm{P}(i,j), p = \frac{\sum_{i,j}(i-\mu i)(j-\mu j)P(i,j)}{\sigma_i \sigma_j}$$

$$\mathrm{E} = \sum_{i,j} P(i,j)^2 \quad (6)$$

3) Shape features. Aspect ratio $AR = \frac{width}{height}$, solidity, and circularity $C = \frac{4\,\pi.Area}{Perimeter^2}$ were computed for leaf boundary analysis [23]

4) Edge features. Sobel operators were applied to compute edge gradients:

$$G_x = \frac{\partial I}{\partial x}, G_y = \frac{\partial I}{\partial y}, \mathrm{G} = \sqrt{G_x^2 + G_y^2} \quad (7)$$

All features were concatenated into a final feature vector:

$$\mathrm{f} = \left[F_{color}, F_{texture}, F_{shape}, F_{edge}\right] \quad (8)$$

E. Feature Optimazion

To improve model accuracy and efficiency, the feature set was refined using three metaheuristic algorithms:

1) Spider monkey optimization (SMO). A population-based search that simulates foraging behavior.

2) Particle swarm optimization (PSO). Particles update positions and velocities to find optimal feature subsets [24]:

$$\mathrm{v}_i(t+1) = \omega \mathrm{v}_i(t) + c_1 r_1 \left(\mathrm{p}_i - x_i(t)\right) + c_2 r_2 \left(\mathrm{g} - \mathrm{x}(t)\right) \quad (9)$$

$$x_i(t+1) = x_i(t) + \mathrm{v}_i(t+1)$$

Ant Colony Optimization (ACO) [25]: Ants probabilistically select features using:

$$P_i^k = \frac{[T_i]^{\propto}[\eta_i]^{\beta}}{\sum_{j \in allowed_k} [T_j]^{\propto}[\eta_j]^{\beta}} \quad (10)$$

F. Classification Model

Optimized features were classified using:

Support Vector Machine (SVM) [26]: Solves:

$$min_{w,b,\varepsilon}\, \frac{1}{2}\|w\|^2 + C \sum_i \varepsilon_i \;\mathrm{s.t.}\; y_i(w^T \phi(x_i) + b) \geq 1 - \varepsilon_i \quad (11)$$

Random Forest (RF) [27]: output majority vote from T decision trees:

$$\hat{y} = \text{mode}\{h_1(x), h_2x, \ldots, h_T(x)\} \quad (12)$$

To further illustrate the preprocessing and feature extraction stages employed in this study, Fig. 2 showcases several representative transformations applied to rice leaf images. Fig. 2(a) to Fig. 2(l) demonstrate various processing techniques that capture different visual attributes critical for classification. Fig. 2(a) presents the original RGB image, while Fig. 2(b) illustrates a pseudo-color mapping used to enhance color contrast. Fig. 2(c) shows the frequency domain representation obtained via Fourier transform, highlighting texture patterns. In Fig. 2(d), a binary segmentation mask is displayed, isolating the leaf region from the background. Fig. 2(e) applies bounding box detection to identify the target area. Fig. 2(f) to Fig. 2(l) exhibit various edge detection and enhancement methods: Fig. 2(f) shows Sobel edge detection, Fig. 2(g) and 2(h) present different gradient edge visualizations, Fig. 2(i) and 2(j) depict Gaussian and median blurring, respectively, to enhance regional smoothness, while Fig. 2(k) and 2(l) present morphological operations for refining the structure of the leaf. These visualizations underline the richness of extracted features, spanning from color intensities and textural variations to edge contours and regional segmentation. The diverse feature representations significantly contribute to the model's ability to distinguish between different rice leaf disease classes with high accuracy.

G. Hyperparameter Tuning

Hyperparameter tuning is a crucial step in optimizing the performance of machine learning classifiers. Unlike model parameters learned during training (e.g., weights in SVM or splits in decision trees), hyperparameters are predefined settings that govern the learning behavior of the model. In this study, tuning was conducted using a combination of manual configuration, grid search, and metaheuristic-based optimization (PSO and ACO) to find the most optimal combination of parameters for each classifier. For the Support Vector Machine (SVM) classifier, three key hyperparameters were considered:

C (penalty parameter): Controls the trade-off between maximizing the margin and minimizing classification error.

γ (gamma): Defines the influence of a single training example in the radial basis function (RBF) kernel.

Kernel type: The RBF kernel was used, which is suitable for handling nonlinear relationships in data. The search space was defined as:

$$C \in \{0.1,1,10\}, \gamma \in \{0.01,0.1,1\} \quad (13)$$

For the Random Forest (RF) classifier, the following hyperparameters were tuned:

n_estimators: The number of decision trees in the ensemble.

max_depth: The maximum depth of the trees.

criterion: The function to measure the quality of a split (Gini or Entropy).

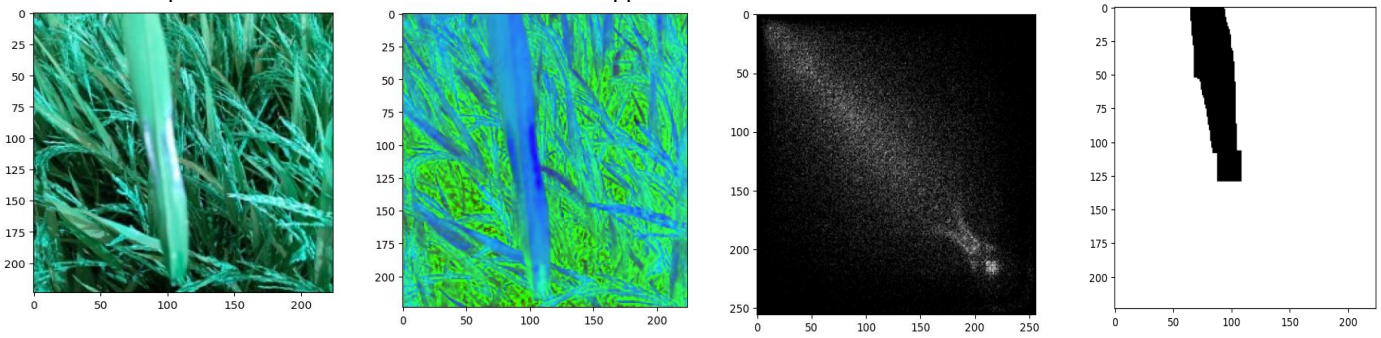
max_features: The number of features considered when looking for the best split.

Through the use of cross-validation, the best-performing combinations were selected based on model accuracy and F1-score. Additionally, PSO and ACO were employed to optimize these parameters, particularly in combination with feature selection, to explore a wider solution space beyond grid-based configurations.

H. Evaluation Metrics Model

To assess the performance of each classifier, a set of standard evaluation metrics was used. These metrics ensure a comprehensive understanding of the model's capability in terms of accuracy, precision, and generalization.

Accuracy measures the overall correctness of the model:



(a) (b) (c) (d)

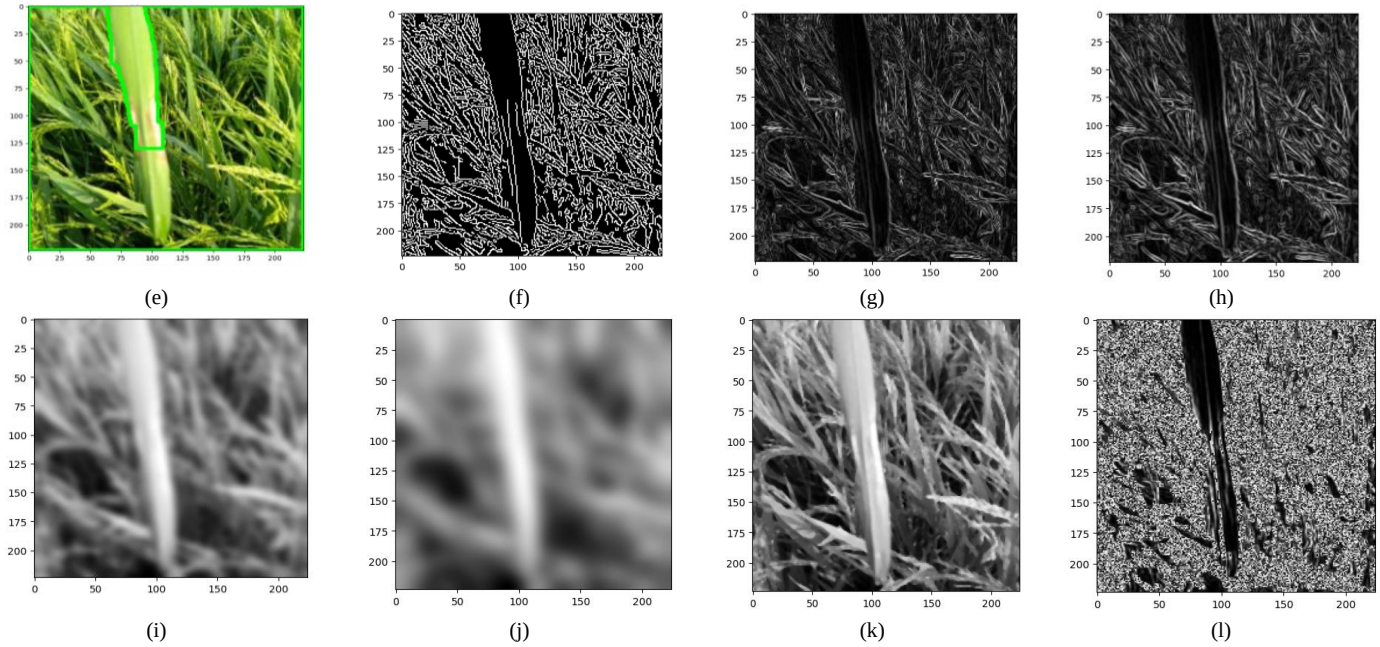


Fig. 2. Visualization of feature extraction and preprocessing steps on rice leaf images: (a) Original RGB image, (b) Pseudo-color enhancement, (c) Frequency domain transformation, (d) Binary mask segmentation, (e) Bounding box detection, (f) Sobel edge detection, (g–h) Gradient edge visualizations, (i–j) Gaussian and median smoothing, (k–l) Morphological enhancement and noise reduction.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

where, TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

Precision quantifies how many of the predicted positives are actually positive:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (15)$$

Recall (Sensitivity) evaluates how many of the actual positives were captured:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (16)$$

F1-Score is the harmonic mean of precision and recall:

$$\text{F1} - \text{Score} = 2.\frac{Precision\ .Recall}{Precesion+Recall} \quad (17)$$

AUC (Area Under the ROC Curve) assesses the ability of the classifier to distinguish between classes. A higher AUC indicates better separability.

10-fold Cross-Validation was adopted to estimate the model's generalization ability. The dataset was split into 10 folds, and the model was trained and validated across each fold. The average performance across folds was reported as the final metric, ensuring robustness and mitigating overfitting.

III. Result

This study evaluated the performance of Random Forest Classifier (RFC) and Support Vector Classifier (SVC) in detecting rice leaf diseases, using a combination of handcrafted features (color, texture, shape, and edge) and optimization methods including Spider Monkey Optimization (SMO), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO). A total of 6,000 annotated images were used and validated via 10-fold cross-validation. Both classifiers were tested with and without hyperparameter tuning. Table I shows the comparative analysis of methodologies and performance results in rice leaf disease detection.

Table II and Fig. 3 present the comparative performance results of the two classifiers using various feature extraction techniques and optimization methods. Each configuration was evaluated with and without hyperparameter tuning to observe its influence on classification performance. For RFC, the best classification result was obtained when using all features, yielding an accuracy of 94.36% without any tuning. With PSO-based feature optimization and tuning, the model achieved 91.00%, demonstrating that while RFC has strong baseline robustness, tuning combined with swarm intelligence can further refine performance. SMO also produced solid results (87.86%), while ACO achieved lower consistency, with accuracy dropping to 77.71% post-tuning.

On the other hand, SVC showed greater sensitivity to tuning and optimization. Without tuning, the model performed poorly on individual features (e.g., 27.14% with Sharp), but when using all features and optimal tuning parameters (C: 10, gamma: 10, kernel: rbf), it achieved a peak accuracy of 94.64%, the highest recorded in the study. As shown in Fig. 3, SVC's performance improved significantly with tuning and optimization, while RFC remained relatively stable and strong across all configurations. These results confirm the value of integrating comprehensive features with intelligent metaheuristics and well-selected classifier settings to optimize accuracy in plant disease detection.

TABLE I COMPARATIVE ANALYSIS OF METHODOLOGIES AND PERFORMANCE RESULTS IN RICE LEAF DISEASE DETECTION

Year	Methodology Summary	Research Focus / Similarity	Performance Result
[24] 2019	- Image Acquisition - Preprocessing, Segmentation - DWT, SIFT, GLCM Feature Extraction	Focus on rice leaf disease classification using hybrid image processing and ML techniques	93,40%
[25] 2019	- Classification using KNN, BPNN, Naïve Bayes, Multiclass SVM - Image Acquisition and Preprocessing - Feature Extraction with Intensity Moments - Random Forest Classifier	Use of Random Forest for classifying Blight, Blast, and Spot diseases	91.47%
[26] 2020	- Image Acquisition - Grayscale Conversion, Segmentation - Pattern Recognition using Neural Networks	Detection of rice leaf diseases using artificial neural networks	92.5%
[27]2021	- Image Acquisition, Preprocessing - Thresholding, Edge Detection, Color Slicing - RGB Feature Calculation for Classification	Emphasis on color-based techniques for Brown Spot and Narrow Brown Spot detection	89.00%
[7] 2024	- SMO for Feature Optimization - Random Forest and SVM Hybrid Model - ROC, Precision, Recall, F1-Score Analysis	Focus on accuracy improvement through SMO + RF + SVM integration	98.00% (AUC = 0.98, Precision = 94%, Recall = 92%, F1 Score = 93%)
Proposed method 2025	- PSO and ACO for Feature Optimization - Random Forest and SVM Evaluation - Performance Benchmarking and Tuning	Enhancement of rice disease classification using PSO-ACO hybrid optimization with classical classifiers	94.64% (SVC + All Features + Tuning) 91.00% (RFC + PSO + Tuning)

Note: This study integrates dual-metaheuristic optimization (PSO and ACO) with multi-classifier evaluation, providing a scalable and generalizable framework compared to prior studies.

TABLE II PERFORMANCE COMPARISON OF RANDOM FOREST AND SUPPORT VECTOR CLASSIFIERS WITH VARIOUS FEATURE EXTRACTION AND OPTIMIZATION TECHNIQUES

Classifier	Feature Type / Method	Hyperparameter Tuning	Best Parameters	Accuracy (%)
Random Forest Classifier	Color	No	–	85.79%
	Texture	No	–	87.43%
	Sharp	No	–	55.14%
	Edge	No	–	87.00%
	All	No	–	94.36%
	All	Yes	criterion=entropy, max_depth=10, max_features=log2, n_estimators=200	82.00%
	SMO	No	–	87.43%
	SMO	Yes	criterion=entropy, max_depth=10, max_features=log2, n_estimators=200	87.86%
	PSO	No	–	87.36%
	PSO	Yes	criterion=entropy, max_depth=10, max_features=log2, n_estimators=200	91.00%
	Ant Colony	No	–	89.50%
	Ant Colony	Yes	criterion=entropy, max_depth=10, max_features=log2, n_estimators=200	77.71%
Support Vector Classifier	Color	No	–	51.57%
	Texture	No	–	58.92%
	Sharp	No	–	27.14%
	Edge	No	–	42.00%
	All	No	–	71.14%
	All	Yes	C=10, gamma=10, kernel=rbf	94.64%
	SMO	No	–	74.86%
	SMO	Yes	C=10, gamma=10, kernel=rbf	90.00%
	PSO	No	–	76.79%

	PSO	Yes	C=10, gamma=10, kernel=rbf	90.57%
	Ant Colony	No	–	59.57%
	Ant Colony	Yes	C=10, gamma=10, kernel=rbf	64.29%

Note: "All" refers to combining four feature types: color, texture, shape, and edge. Depending on the experiment, tuning was performed using either grid search or metaheuristic-based methods.

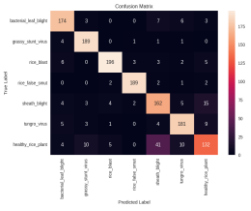
IV. Discussion

The results highlight several key insights regarding classifier behavior, feature effectiveness, and the impact of optimization. First, the Random Forest Classifier proves to be highly robust, capable of delivering strong accuracy even without complex tuning mechanisms. Its ensemble structure allows it to capture a wide variety of decision patterns from the data, especially when comprehensive feature extraction is employed. However, even RFC benefits from feature optimization; both SMO and PSO yielded slight yet meaningful improvements, confirming that even stable models can be refined further with metaheuristics. Support Vector Classifier, in contrast, displayed a different dynamic. Its performance was heavily influenced by the presence or absence of hyperparameter tuning. Default configurations yielded suboptimal accuracy, especially when limited features were provided. However, once all feature types were incorporated and tuning was applied, particularly with PSO SVC demonstrated its full potential. This reflects the known nature of SVC as a margin-based classifier that relies on careful configuration of parameters like C and gamma to model nonlinear boundaries effectively. Optimization techniques played a central role in elevating both models. Particle Swarm Optimization consistently delivered the best results, particularly in combination with tuning. Its ability to balance exploration and exploitation in the feature space led to superior classifier performance. Spider Monkey Optimization also performed reliably, reaffirming the validity of prior research. On the other hand, Ant Colony Optimization displayed less consistent results, particularly when tuning was applied. This suggests that ACO's performance may be highly sensitive to its internal settings, and could benefit from future refinement or hybridization.

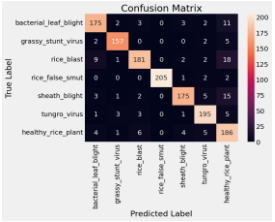
Compared to SMO-only approaches, PSO provided more consistent accuracy with less variance across folds, while ACO showed instability during tuning. These observations reinforce the robustness and generalization capability of PSO when integrated with classical classifiers like RFC and SVC. Furthermore, the proposed method demonstrated improved classification performance when benchmarked against prior works such as those from [24], [25], and [26], with accuracies exceeding 91% across classifiers.

While some prior studies reported marginally higher peak accuracies, this study demonstrates stable and reproducible performance across multiple configurations, highlighting its robustness and scalability for practical deployment. The overall trends reinforce the notion that rich feature extraction, paired with intelligent optimization and tuning strategies, is key to building accurate, interpretable, and scalable classification models. Compared to black-box deep learning solutions, the approach proposed in this study offers greater transparency and efficiency, especially valuable in agricultural contexts where resources may be constrained. Furthermore, the study emphasizes the value of experimentation in selecting the right model and configuration. Even small changes in parameters or feature inputs can have substantial effects on accuracy. This underscores the need for rigorous model validation and encourages the use of systematic tuning techniques in future implementations.

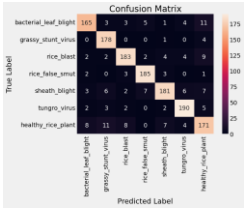
Random Forest Classifier



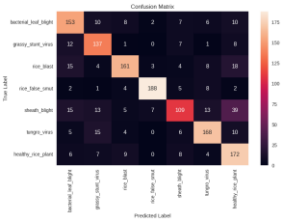
Cofusion Matrix Non- Hypertuning with PSO



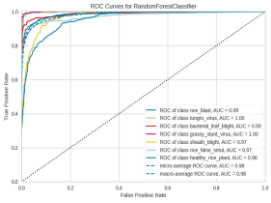
Cofusion Matrix Hypertuning Parametric with PSO



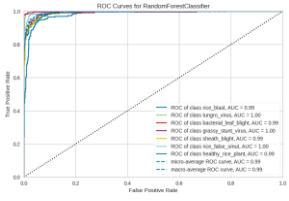
Cofusion Matrix Non- Hypertuning with ACO



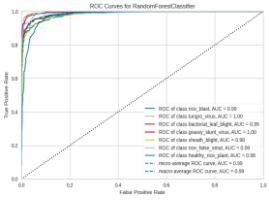
Cofusion Matrix Hypertuning Parametric with ACO



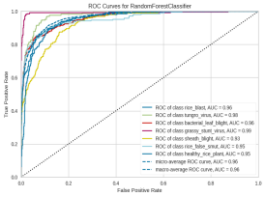
ROC AUC Non- Hypertuning with PSO



ROC AUC Hypertuning Parametric with PSO

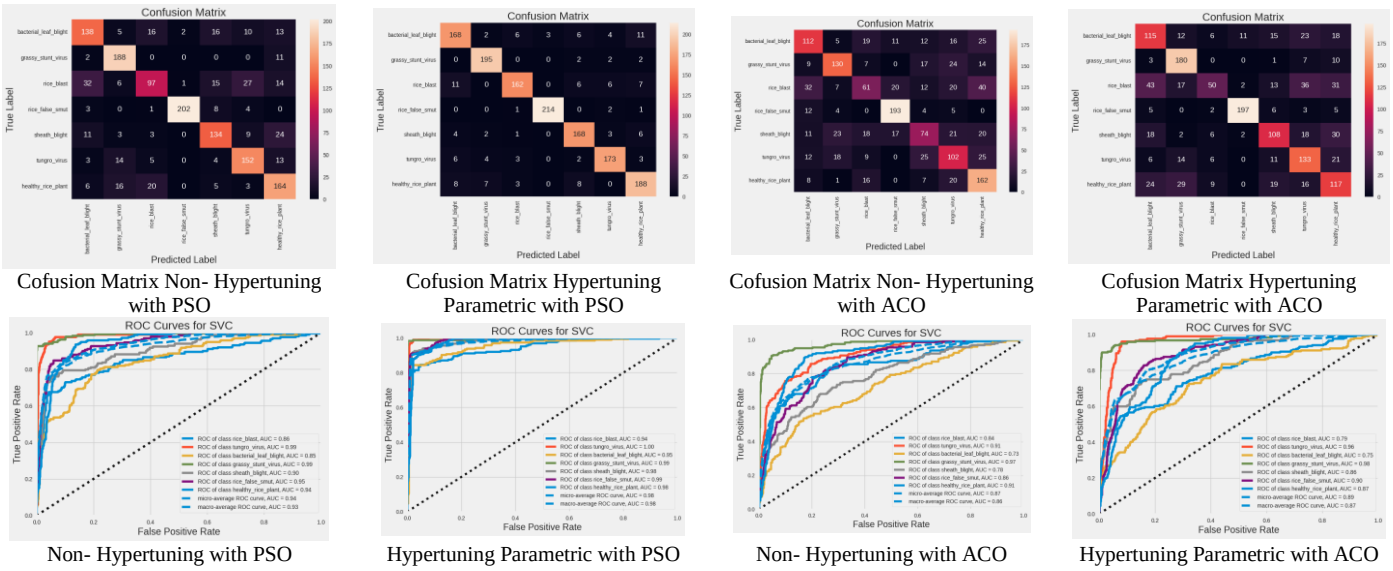


ROC AUC Non- Hypertuning with ACO



ROC AUC Hypertuning Parametric with ACO

Support Vector Classifier



Cofusion Matrix Non- Hypertuning with PSO | Cofusion Matrix Hypertuning Parametric with PSO | Cofusion Matrix Non- Hypertuning with ACO | Cofusion Matrix Hypertuning Parametric with ACO

Non- Hypertuning with PSO | Hypertuning Parametric with PSO | Non- Hypertuning with ACO | Hypertuning Parametric with ACO

Fig. 3. Performance comparison of RFC and SVC models with PSO and ACO optimization.

V. Conclusion

This study has conducted an in-depth evaluation of rice leaf disease detection using two classical machine learning classifiers Random Forest Classifier (RFC) and Support Vector Classifier (SVC) combined with handcrafted feature extraction and metaheuristic optimization techniques, including Spider Monkey Optimization (SMO), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO). The results demonstrate that integrating comprehensive visual features (color, texture, shape, and edge) with well-configured classifiers significantly boosts classification accuracy, with PSO consistently yielding the most promising performance. Notably, the SVC model, while initially underperforming, exhibited the most substantial improvement through tuning, reaching an accuracy of 94.64% with optimized parameters and PSO-enhanced feature selection, surpassing the results of RFC in its best condition. The effectiveness of hyperparameter tuning is evident across both models. Parameters such as C, gamma, and kernel selection in SVC, or max_depth, criterion, and n_estimators in RFC, proved to be pivotal in maximizing classifier performance. Metaheuristic techniques, particularly PSO and SMO, enabled a more exhaustive and intelligent exploration of these configurations compared to traditional grid search. However, ACO demonstrated variable outcomes, suggesting that further refinement or hybridization may be necessary to harness its full potential.

The implications of these findings extend beyond technical metrics. This study underscores the value of a structured, interpretable, and scalable modeling approach, which can serve as a viable alternative to deep learning methods, especially in agricultural settings where resources and data may be limited. The ability to maintain high classification performance while retaining transparency and operational flexibility makes this approach highly relevant for real-world deployment in precision farming and plant health monitoring systems. Looking ahead, there are several promising avenues for future research. First, integrating deep feature representations such as those extracted from convolutional neural networks with handcrafted features could create hybrid models that leverage both abstraction and interpretability. Second, the development of adaptive metaheuristic algorithms that self-tune based on model performance feedback could further enhance optimization reliability. Third, expanding the framework to real-time and edge-computing platforms, such as mobile-based disease detection applications, would increase accessibility for farmers and agricultural technicians. Finally, applying this optimized classification pipeline to other crops and disease categories could validate its scalability and contribute to broader food security and agricultural innovation goals. In conclusion, this research reinforces the critical synergy between data preprocessing, feature engineering, classifier tuning, and intelligent optimization. It paves the way for practical, high-performance decision-support systems that are not only accurate but also efficient and ready for deployment in real-world agricultural environments. Although some existing studies report slightly higher accuracies under limited or tightly controlled conditions, the proposed framework introduces a dual-metaheuristic and dual-classifier optimization pipeline that is validated across diverse feature configurations and tuning scenarios. This trade-off ensures better generalization in real-world agricultural applications.

Despite its promising results, this study has certain limitations. The reliance on handcrafted features may limit generalizability to other crops or conditions. Additionally, ACO's tuning instability suggests the need for further exploration into adaptive or hybridized optimization strategies. Future research could integrate deep feature representations with handcrafted features, or extend the framework to real-time applications such as mobile disease detection systems.

References

[1] Food and Agriculture Organization of the United Nations, "FAO RICE MARKET MONITOR," https://www.fao.org/markets-and-trade/commodities/rice/rmm/en.

[2] C. G. Simhadri, H. K. Kondaveeti, V. K. Vatsavayi, A. Mitra, and P. Ananthachari, "Deep learning for rice leaf disease detection: A systematic literature review on emerging trends, methodologies and

techniques," 2024, China Agricultural University. doi: 10.1016/j.inpa.2024.04.006.

[3] M. Kannan, M. M. Saad, N. Talip, S. N. Baharum, and H. Bunawan, " Complete Genome Sequence of Rice Tungro Bacilliform Virus Infecting Asian Rice (Oryza sativa) in Malaysia ," Microbiol Resour Announc, vol. 8, no. 20, May 2019, doi: 10.1128/mra.00262-19.

[4] M. Arsenovic, M. Karanovic, S. Sladojevic, A. Anderla, and D. Stefanovic, "Solving current limitations of deep learning based approaches for plant disease detection," Symmetry (Basel), vol. 11, no. 7, Jul. 2019, doi: 10.3390/sym11070939.

[5] B. Nagachandrika, R. Prasath, and I. R. Praveen Joe, "An automatic classification framework for identifying type of plant leaf diseases using multi-scale feature fusion-based adaptive deep network," Biomed Signal Process Control, vol. 95, Sep. 2024, doi: 10.1016/j.bspc.2024.106316.

[6] J. Padhi, K. Mishra, A. K. Ratha, S. K. Behera, P. K. Sethy, and A. Nanthaamornphong, "Enhancing paddy leaf disease diagnosis -a hybrid CNN model using simulated thermal imaging," Smart Agricultural Technology, vol. 10, Mar. 2025, doi: 10.1016/j.atech.2025.100814.

[7] A. Kurniawan, T. R. Soeprobowati, and B. Warsito, "Revolutionizing Rice Leaf Disease Detection: Next-Generation SMOREF-SVM Integrating Spider Monkey Optimization and Advanced Machine Learning Techniques," 2024. [Online]. Available: www.ijacsa.thesai.org

[8] G. P. Guru, N. K. Naveen, A. K. Jain, J. Syed, and R. U. Baig, "Convergence of improved particle swarm optimization based ensemble model and explainable AI for the accurate detection of food adulteration in red chilli powder," Journal of Food Composition and Analysis, vol. 143, Jul. 2025, doi: 10.1016/j.jfca.2025.107577.

[9] A. Kumar, V. Patidar, D. Khazanchi, and P. Saini, "Optimizing Feature Selection Using Particle Swarm Optimization and Utilizing Ventral Sides of Leaves for Plant Leaf Classification," in Procedia Computer Science, Elsevier B.V., 2016, pp. 324–332. doi: 10.1016/j.procs.2016.06.079.

[10] Y. Xiaofeng and M. Yuzhuo, "Research on ant colony algorithm based on motion optimization and adaptive threshold for pantograph slider edge detection," Measurement (Lond), vol. 245, Mar. 2025, doi: 10.1016/j.measurement.2024.116598.

[11] A. V. Panchal, S. C. Patel, K. Bagyalakshmi, P. Kumar, I. R. Khan, and M. Soni, "Image-based Plant Diseases Detection using Deep Learning," Mater Today Proc, vol. 80, pp. 3500–3506, Jan. 2023, doi: 10.1016/j.matpr.2021.07.281.

[12] S. C.K., J. C.D., and N. Patil, "Tomato plant disease classification using Multilevel Feature Fusion with adaptive channel spatial and pixel attention mechanism," Expert Syst Appl, vol. 228, Oct. 2023, doi: 10.1016/j.eswa.2023.120381.

[13] B. Nguyen, B. Xue, and P. Andreae, A particle swarm optimization based feature selection approach to transfer learning in classification. 2018. doi: 10.1145/3205455.3205540.

[14] W. B. Demilie, "Plant disease detection and classification techniques: a comparative study of the performances," J Big Data, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-023-00863-9.

[15] S. K. Lakshmanaprabu, K. Shankar, M. Ilayaraja, A. W. Nasir, V. Vijayakumar, and N. Chilamkurti, "Random forest for big data classification in the internet of things using optimal features," International Journal of Machine Learning and Cybernetics, vol. 10, no. 10, pp. 2609–2618, 2019, doi: 10.1007/s13042-018-00916-z.

[16] H. Tyralis, G. Papacharalampous, and A. Langousis, "A brief review of random forests for water scientists and practitioners and their recent history in water resources," May 01, 2019, MDPI AG. doi: 10.3390/w11050910.

[17] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," 2020, Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/JSTARS.2020.3026724.

[18] M. Sharif, M. A. Khan, Z. Iqbal, M. F. Azam, M. I. U. Lali, and M. Y. Javed, "Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection," Comput Electron Agric, vol. 150, pp. 220–234, 2018, doi: https://doi.org/10.1016/j.compag.2018.04.023.

[19] R. C. . Gonzalez and R. E. . Woods, Digital image processing. Pearson, 2018.

[20] C. Solomon and T. Breckon, Fundamentals of Digital Image Processing A Practical Approach with Examples in Matlab, 7th Edition. Boca Raton: CRC Press, 2018. doi: https://doi.org/10.1201/b18983.

[21] "Image Processing, Analysis, and Machine Vision."

[22] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," IEEE Trans Syst Man Cybern, vol. SMC-3, no. 6, pp. 610–621, 1973, doi: 10.1109/TSMC.1973.4309314.

[23] D. Zhang and G. Lu, "Shape-based image retrieval using generic Fourier descriptor," Signal Process Image Commun, vol. 17, no. 10, pp. 825–848, 2002, doi: https://doi.org/10.1016/S0923-5965(02)00084-X.

[24] A. G. Gad, "Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review," Archives of Computational Methods in Engineering, vol. 29, no. 5, pp. 2531–2561, Aug. 2022, doi: 10.1007/s11831-021-09694-4.

[25] M. Dorigo and T. Stützle, Ant Colony Optimization. The MIT Press, 2004. doi: 10.7551/mitpress/1290.001.0001.

[26] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press, 2000. doi: DOI: 10.1017/CBO9780511801389.

[27] L. Breiman, "Random Forests," 2001.

Reinforcement Learning for Real-Time Scheduling in Dynamic Reconfigurable Manufacturing Systems

Salah Hammedi[1], Abdallah Namoun[2], Mohamed Shili[3]
Networked Objects, Control, and Communication Systems (NOCCS), ENISo, University of Sousse, Tunisia[1]
AI Center, Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia[2]
Innov'COM Laboratory, National Engineering School of Cartahage, Tunisia[3]

***Abstract*—This study presents a novel application of Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) for scheduling optimization in Reconfigurable Manufacturing Systems (RMFS). The performance of these approaches is quantitatively evaluated and compared with traditional scheduling methods, specifically Shortest Processing Time (SPT) and Earliest Due Date (EDD), across several key metrics, including makespan, tardiness, resource utilization, and adaptability to disturbances. Our results show a significant reduction in makespan, with RL achieving a 20% improvement and DRL a 28.57% improvement over SPT. Moreover, RL and DRL outperform classical methods in minimizing tardiness and improving resource utilization. DRL also demonstrates superior adaptability under dynamic disruptions such as machine breakdowns, with only a 5% deviation in makespan compared to 16.67% for SPT. These findings confirm the benefits of RL and DRL for real-time decision-making in dynamic manufacturing environments. The study discusses the robustness and scalability of RL and DRL approaches, as well as the challenges related to their computational cost. The novelty lies in integrating RL and DRL into RMFS scheduling to offer a scalable, adaptive solution that improves production efficiency.**

***Keywords*—Adaptability, deep reinforcement learning (DRL); makespan; manufacturing systems; reinforcement learning (RL); resource utilization; scheduling optimization; shortest processing time (SPT); tardiness; traditional scheduling methods**

I. Introduction

Reconfigurable Manufacturing Systems (RMFS) were developed to address the growing need for flexibility and product customization in manufacturing. Unlike traditional manufacturing systems, which are designed for fixed, repetitive tasks, RMFS enables rapid adaptation to changes in product designs or production volumes [1]. This adaptability is achieved through configurable modules, such as machines and resources that can be restructured or reallocated [2]. RMFS offers a highly versatile approach, making it possible to adjust production lines in response to evolving requirements, without significant downtime or overhauls. This type of system plays a vital role in industries that operate within uncertain and highly competitive environments, such as automotive, aerospace, and electronics manufacturing [3]. In these sectors, quickly responding to shifts in market demands and technological advancements is crucial, giving companies that utilize RMFS a competitive edge. Consequently, RMFS enhances efficiency and provides a scalable solution to meet the diverse and changing demands of modern production [4], [5].

A. Complexity of the Scheduling Process

In Reconfigurable Manufacturing Systems (RMFS), scheduling involves defining the optimal sequence for executing tasks on available machines while adhering to constraints such as deadlines and resource availability. The primary objective is to optimize the overall system performance, which includes reducing production times, meeting deadlines, and enhancing resource utilization [2]. However, this task becomes complex in a reconfigurable environment for several reasons:

1) System dynamics. Machine configurations can change in real-time, necessitating adjustments to the scheduling algorithm [1].

2) Frequent introduction of new orders. Customized products often require adjustments in production sequences to accommodate various specifications and customer demands [3].

3) Breakdowns and interruptions. Machines may experience unexpected failures or require unscheduled maintenance, which disrupts the planned sequence of tasks and the scheduling process [4].

B. Specific Challenges of Scheduling in RMFS

The challenges associated with scheduling in RMFS include:

1) Responsiveness to unforeseen events. The system must quickly modify the sequence during disruptive occurrences, such as a machine breakdown [5].

2) Resource conflicts. Multiple tasks may require the same machines or resources, creating bottlenecks that hinder efficient operation [2].

3) Multi-criteria optimization. It is not just about minimizing production time; balancing various objectives, such as cost, deadlines, and quality, is also crucial [3].

C. Necessity for an Intelligent Approach to Scheduling

Traditional methods, such as Shortest Processing Time (SPT) and Earliest Due Date (EDD), are often inadequate for RMFS. These methods apply static rules that do not account for the dynamic and uncertain nature of RMFS [1]. For instance, an optimal sequence according to the SPT rule may become inefficient if a key machine fails or an urgent new order is introduced.

This highlights the necessity for intelligent approaches like Reinforcement Learning (RL). These methods allow the system to learn and adjust sequences in real time based on the current state of the system [6]. This provides operational flexibility and enhances the system's ability to maintain high performance despite frequent disruptions [7].

D. Impact of Optimized Scheduling on RMFS Performance

Effective scheduling in RMFS offers several strategic advantages:

1) Reduction of downtime. Machines are utilized more effectively by optimizing the sequence, thereby minimizing idle periods [5].

2) Improved customer satisfaction. Meeting delivery deadlines becomes easier, even when dealing with customized or unexpected orders [2].

3) Reduction in operational costs. Operational costs are lowered with less resource wastage and shortened production lead times [3].

4) Increased adaptability. The system can respond swiftly to new market demands, ensuring a competitive advantage [1].

Scheduling plays a central role in the effective management of RMFS. Adopting dynamic and intelligent scheduling approaches becomes imperative in an environment, where flexibility, responsiveness, and efficiency are critical factors. Reinforcement Learning methods, such as Q-learning and Deep Reinforcement Learning, show promise as they enable the system to adapt proactively and continuously. Therefore, an optimized scheduling approach is essential for maximizing the overall performance of RMFS and ensuring an effective response to modern market challenges.

Our study is structured to provide a logical and comprehensive exploration of the proposed scheduling approach. Section II reviews existing methods for RMFS scheduling and highlights their limitations, establishing the need for innovative AI-based solutions. Section III details the design and implementation of our Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) frameworks for intelligent scheduling. Section IV presents the experimental outcomes, comparing the performance of RL and DRL with traditional methods across key metrics such as makespan and resource utilization. Section V analyzes the results, emphasizing the robustness and advantages of the proposed approaches while addressing their potential limitations. Finally, Section VI summarizes the contributions of the study and offers perspectives for future research in intelligent scheduling for RMFS.

II. Related Work

Scheduling in Reconfigurable Manufacturing Systems (RMFS) is essential for optimizing operational efficiency and responsiveness to market demands. It involves determining the optimal sequence for executing tasks on available machines while adhering to constraints such as deadlines and resource availability. Given the dynamic nature of RMFS, where configurations can change in real-time, effective scheduling directly impacts resource utilization and production lead times. Reviewing existing literature on scheduling techniques is crucial for understanding the evolution from classical methods, like Shortest Processing Time (SPT) and Earliest Due Date (EDD), to more advanced artificial intelligence-driven approaches. This examination highlights the strengths and limitations of traditional strategies and identifies innovative solutions that address the complexities of modern manufacturing environments. By identifying gaps in current research, we can propose future developments that enhance the adaptability and performance of scheduling in RMFS.

A. Classical Scheduling Approaches

Classical scheduling methods, such as Shortest Processing Time (SPT) and Earliest Due Date (EDD), have long been used in traditional manufacturing systems to determine the order of tasks in a way that optimizes production outcomes. These methods are foundational in scheduling theory, offering simple yet effective rules for task prioritization. However, while they remain useful in specific contexts, their limitations in Reconfigurable Manufacturing Systems (RMFS) highlight the need for more dynamic approaches [8], [9].

1) Shortest Processing Time (SPT). The Shortest Processing Time (SPT) rule prioritizes tasks based on the time required to complete them, scheduling shorter tasks first to minimize the total production time, or "makespan" [10]. Historically, SPT has been widely applied in conventional manufacturing environments to reduce the average waiting time of tasks, making it an effective strategy for repetitive and stable production systems [11].

a) Advantages of SPT: The primary advantage of SPT is its effectiveness in reducing makespan and minimizing in-process inventory by accelerating task completion [12]. This can be particularly valuable in batch production systems, where rapid completion of smaller tasks allows for quick turnover [13].

b) Limitations in RMFS: While SPT can improve efficiency in traditional settings, its static prioritization approach does not align well with the dynamic demands of RMFS. The SPT rule is not inherently adaptable to sudden changes in task requirements, equipment reconfiguration, or machine breakdowns. As a result, it may struggle to keep up with the flexibility and responsiveness required in RMFS environments where tasks and resources are frequently reallocated [14], [15].

2) Earliest Due Date (EDD). The Earliest Due Date (EDD) rule is another classical approach that prioritizes tasks based on their deadlines, scheduling the task with the earliest due date first [16]. In traditional manufacturing, EDD has been effectively used to reduce tardiness and to ensure that tasks meet deadlines, which is particularly important in environments where adherence to delivery schedules is crucial [17].

a) Advantages of EDD: EDD helps to minimize the number of late tasks and is beneficial in scenarios, where customer satisfaction or regulatory compliance depends on timely delivery. This approach is instrumental in make-to-order environments where each task has a distinct due date [18].

b) Limitations in RMFS: In an RMFS, the EDD rule's rigid focus on due dates may lead to inefficiencies, especially when production requirements change frequently. If a task with a nearer due date is delayed due to a reconfiguration or machine breakdown, EDD may not respond flexibly, potentially resulting in longer delays and resource underutilization [19], [20].

3) Comparison of SPT and EDD. SPT and EDD have distinct performance metrics that influence their effectiveness in different scenarios:

a) Performance metrics: SPT tends to excel in minimizing makespan and average waiting time, while EDD is more effective for reducing tardiness and ensuring timely task completion. These metrics are particularly relevant when comparing their situational effectiveness within traditional and reconfigurable manufacturing settings [8], [21].

b) Situational appropriateness: SPT is best suited to scenarios with high volumes of short, repetitive tasks where reducing total completion time is key. In contrast, EDD is more appropriate in environments where on-time delivery is critical, and tasks have varied due dates. However, the dynamic nature of RMFS often requires real-time adjustments, which neither SPT nor EDD can manage effectively without additional support [22], [23].

In summary, while both SPT and EDD offer applicable scheduling rules in conventional systems, their limitations in RMFS underscore the need for adaptive methods that can handle real-time changes and reconfigurations. This gap in classical methods provides the foundation for exploring AI-based and reinforcement learning approaches, which offer more sophisticated and flexible scheduling solutions for RMFS [8], [9].

B. AI-based Scheduling Approaches

In Reconfigurable Manufacturing Systems (RMFS), the dynamic and unpredictable nature of production demands flexible scheduling solutions that can adapt to real-time changes. Artificial Intelligence (AI)-based scheduling approaches have gained popularity due to their potential for handling complex decision-making in these environments. Unlike classical scheduling rules, AI techniques can dynamically respond to changes in tasks, resources, and machine configurations. This section explores prominent AI scheduling methods, focusing on the benefits and challenges they bring to RMFS.

1) Overview of AI techniques in scheduling. AI techniques such as Genetic Algorithms (GA), Neural Networks (NN), and other machine learning methods have been widely applied in scheduling. Genetic Algorithms, inspired by natural selection, are known for their ability to search large solution spaces and optimize complex scheduling problems by iteratively improving task sequences [8], [24]. They are beneficial for multi-objective scheduling problems, where objectives like minimizing makespan, reducing lateness, or maximizing resource utilization must be balanced [9].

Neural Networks are effective in pattern recognition and prediction, making them valuable in predicting job arrival times, machine breakdowns, and estimating task completion. They adapt to changing patterns in production, providing forecasts that inform better scheduling decisions [11], [15]. In RMFS, Neural Networks can support proactive scheduling by learning from historical data, predicting bottlenecks, and adjusting task allocations accordingly [23].

The adaptability of AI techniques makes them well-suited to RMFS, where conventional, static scheduling rules like SPT and EDD may fall short. AI methods can improve operational efficiency and address RMFS-specific requirements such as real-time resource reallocation and production line reconfiguration [22].

2) Reinforcement learning for scheduling. Reinforcement Learning (RL) is an advanced AI technique suitable for dynamic environments like RMFS. In RL, an agent learns to make sequential decisions by interacting with the environment, receiving feedback in the form of rewards, and continuously improving its policy to maximize cumulative reward over time [19]. RL techniques such as Q-Learning and Deep Reinforcement Learning (DRL) have shown promising results in complex scheduling tasks, where decisions need to adapt based on real-time feedback [20].

In Q-Learning, an RL agent learns an optimal policy by evaluating each action's potential outcomes, updating a Q-value table, and selecting actions that maximize rewards. For RMFS, Q-Learning can be applied to dynamically allocate resources and prioritize tasks based on changing production requirements, minimizing downtime, and maximizing throughput [13].

Deep Reinforcement Learning, which combines RL with Deep Learning, enables even more complex decision-making. DRL can process extensive, high-dimensional data, such as job sequences, machine statuses, and process dependencies, allowing it to develop strategies for scheduling in high-variability environments. Case studies in RMFS have demonstrated that DRL can effectively handle high-dimensional states and actions, learning policies that surpass traditional scheduling methods [14], [16]. By using DRL, RMFS can benefit from continuous improvement in scheduling, with the agent learning to optimize task sequences based on feedback from the production floor.

3) Comparison with classical approaches. Compared to classical methods like SPT and EDD, AI-based scheduling approaches exhibit higher adaptability, efficiency, and responsiveness in RMFS. Classical methods are limited by their static nature and cannot effectively handle unexpected events like machine breakdowns or sudden changes in task priority. AI methods, however, are equipped to re-evaluate schedules in real-time, adjusting for new conditions and optimizing for multiple objectives simultaneously [8], [18].

AI-based scheduling methods outperform classical approaches in several key performance metrics:

a) Adaptability: AI techniques, especially RL, can adjust schedules dynamically based on real-time feedback, maintaining optimal performance under changing conditions [21].

b) Efficiency: Genetic Algorithms and Neural Networks have been shown to reduce makespan and tardiness in complex scheduling environments, ensuring that production schedules remain efficient even in high-variability contexts [11], [12].

c) Responsiveness: Unlike static scheduling rules, AI-based methods can quickly adapt to disruptions, minimizing the impact of delays and enabling faster recovery times [23].

In summary, while classical scheduling approaches like SPT and EDD have value in stable environments, AI-based scheduling methods provide a more robust solution for RMFS, addressing the need for adaptability, efficiency, and real-time responsiveness in modern production systems [17], [19].

C. Hybrid Approaches

1) Combining classical and AI techniques. Hybrid scheduling approaches integrate traditional rules like Shortest Processing Time (SPT) and Earliest Due Date (EDD) with advanced AI techniques, such as Genetic Algorithms (GA) and Reinforcement Learning (RL). These approaches aim to capitalize on the stability and interpretability of classical methods while enhancing adaptability and efficiency through AI. Studies show that combining classical heuristics with AI algorithms enhances scheduling performance in dynamic environments like RMFS, especially under fluctuating demand and resource availability. For instance, recent research in sustainable edge computing applied hybrid methods to optimize task distribution and found that this integration reduced makespan and improved scalability [19], [25].

2) Case studies and results. Various implementations of hybrid approaches have demonstrated their potential in real-world RMFS environments. For example, a 2023 study applied a hybrid GA-SPT algorithm in cloud-based systems and observed a significant improvement in task completion rates and resource utilization [26]. These studies underscore the efficiency of hybrid methods in handling the specific demands of RMFS, such as reconfigurability and task prioritization, which are challenging for standalone classical or AI methods alone.

D. Gaps in Current Research

While hybrid scheduling approaches show substantial potential for RMFS environments, several research gaps require attention.

First, optimizing parameter settings to balance the classical and AI components effectively is a significant challenge. For instance, selecting parameters such as mutation rates in genetic algorithms or exploration-exploitation ratios in reinforcement learning can drastically affect performance but often requires trial and error. This parameter fine-tuning can be particularly complex in hybrid models, which combine multiple algorithms with distinct parameter requirements. Studies suggest that integrating automated hyperparameter tuning techniques, such as Bayesian optimization, may enhance hybrid model performance without extensive manual adjustment [27], [28].

Scalability also remains a critical issue for hybrid approaches in RMFS, especially as the complexity of real-world manufacturing systems increases. While hybrid methods are designed to manage dynamic changes, they often demand high computational resources. This limitation affects their applicability in large-scale systems with frequent reconfiguration needs. Recently, research has explored lightweight AI models and distributed computation to alleviate these demands, allowing for faster adaptation in real-time environments. More work is needed to make hybrid models computationally efficient without sacrificing responsiveness [29].

Moreover, model generalization is another area with considerable potential for development. Many hybrid models are tailored to specific scheduling tasks and may not adapt well to different manufacturing configurations. A promising direction here is the integration of transfer learning techniques, allowing hybrid models to apply knowledge from one scheduling context to another with minimal retraining [30]. This model's adaptability makes it suitable for a broader range of applications within RMFS.

In the context of intelligent scheduling for complex manufacturing environments, recent advances have demonstrated the effectiveness of hybrid metaheuristic approaches. Notably, the modified chromosome pooling genetic algorithm has been introduced for resource allocation optimization, providing a practical approach for managing complex constraints and achieving high-performance scheduling in dynamic environments [31]. Similarly, integrating multi-objective genetic algorithms into the job shop scheduling problem has shown significant promise for addressing multi-criteria optimization requirements, allowing for balanced consideration of makespan, tardiness, and resource utilization [32]. These studies highlight the evolution of traditional optimization methods toward more intelligent, adaptive, and multi-criteria approaches, aligning closely with the goals of Reinforcement Learning and Deep Reinforcement Learning methods proposed in this work.

Lastly, there is a need for real-world testing and validation of hybrid approaches in industrial settings. While numerous studies show hybrid methods' efficacy in simulations, limited research has focused on field implementations. The lack of practical validation raises questions about the robustness of these approaches under unpredictable real-world conditions, where factors such as unplanned downtimes, resource constraints, and varying task priorities frequently arise [29]. Research focused onails would offer valuable insights into these models' practical viability and guide further refinement.

In summary, recent advances in scheduling for RMFS highlight the effectiveness and versatility of classical and AI-based methods. Classical approaches like SPT and EDD provide simplicity and reliability, though they lack adaptability in dynamic environments. AI techniques, particularly reinforcement learning, have proven valuable for complex, real-time scheduling, enhancing system flexibility. Hybrid

models that integrate classical and AI approaches show promise in balancing efficiency and responsiveness, though scalability and computational demands remain challenges. Continued research in this field is essential to improve RMFS scheduling efficiency, focusing on optimizing hybrid models, enhancing adaptability, and exploring real-world applications to validate these methodologies.

III. Proposed Methodology

A. Introduction to Reinforcement Learning for Scheduling

The proposed methodology leverages Reinforcement Learning (RL) to address the dynamic and complex nature of scheduling in Reconfigurable Manufacturing Systems (RMFS). Production demands frequent shifts in these environments, and resource configurations often change, making traditional scheduling approaches insufficient. RL stands out as a powerful paradigm capable of learning optimal scheduling policies by interacting with the system and adapting to its dynamic states.

1) Objective. The primary goal is to utilize RL to optimize task allocation and resource utilization in real-time, minimizing production delays and improving system responsiveness. RL frameworks can dynamically adjust scheduling strategies as they continuously learn from the system's feedback.

2) Approach selection. The choice of RL technique depends on the complexities and requirements of RMFS environments.

Q-Learning: Selected for problems where the state and action spaces are discrete. Q-Learning is straightforward and can efficiently handle scenarios with well-defined states, such as machine availability or task queue lengths.

Deep Reinforcement Learning (DRL): Employed in scenarios with large or continuous state-action spaces. Using neural networks, DRL methods such as Deep Q-Networks (DQN) or Actor-Critic frameworks can approximate the value functions and handle high-dimensional data, making them ideal for complex RMFS setups.

By utilizing these RL techniques, the proposed methodology aims to enable adaptive, real-time decision-making, which is crucial for meeting the demands of modern manufacturing systems.

B. System Architecture

1) Overview. The architecture of the proposed system integrates Reinforcement Learning (RL) with Reconfigurable Manufacturing Systems (RMFS) for intelligent scheduling. It included the following key components:

a) Task queues: Dynamic queues representing pending tasks, each characterized by parameters like processing time, priority, and deadlines.

b) Machine configurations: A representation of the current state of resources, including machine availability, operational capabilities, and current workloads.

c) Feedback mechanisms: A loop that captures system performance metrics (e.g., task completion times, resource utilization) to refine the learning process and adapt real-time scheduling decisions.

d) RL module: The core decision-making unit, which interacts with the environment (RMFS), learns from feedback, and outputs optimized scheduling actions.

2) State definition. The state represents the current environment snapshot, encompassing:

a) Machine configurations: Status of each machine (idle, busy, or under maintenance).

b) Task queue: The tasks awaiting processing, including task-specific details like priority, duration, and dependencies.

c) System load: Current workload distribution across the system.

Each state is expressed as a multidimensional vector to encapsulate these parameters, allowing the RL model to understand and respond to the system's current conditions.

3) Action space. The action space defines the possible decisions that the RL agent can make, including:

a) Task assignment: Assigning tasks to specific machines based on their capabilities and current state.

b) Machine allocation: Activating or deactivating machines dynamically to optimize energy use and resource utilization.

c) Scheduling adjustments: Re-sequencing tasks in response to unexpected events or delays.

Actions are designed to directly impact the system’s productivity and adaptability.

4) Reward structure. The reward function guides the RL agent by quantifying the effectiveness of its actions. It is designed to achieve:

a) Minimizing production time: Encouraging decisions that reduce the makespan of all tasks.

b) Minimizing delays: Penalizing actions that increase task tardiness.

c) Balancing resource utilization: Promoting even workload distribution across all machines to avoid bottlenecks.

d) Energy efficiency: Incentivizing reduced machine downtime and energy consumption.

The reward signal ensures that the RL agent consistently learns and improves its scheduling policy to adapt to dynamic RMFS environments.

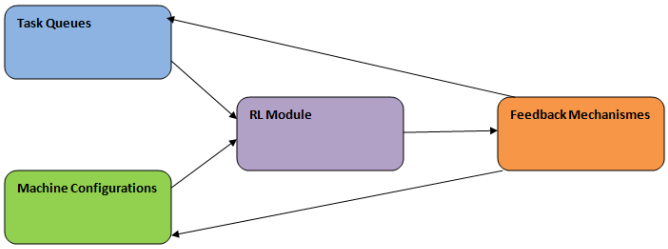


Fig. 1. System architecture diagram for reinforcement learning-based scheduling in RMFS.

The architecture diagram (Fig. 1) is positioned below this section to visually represent the described components and their interactions. It highlights the flow of information between task queues, machine configurations, the RL module, and feedback mechanisms, illustrating how real-time scheduling decisions are optimized.

C. Q-Learning-based Scheduling Approach

1) Algorithm explanation. Q-Learning is an RL algorithm that finds the best action-selection policy through trial and error, updating Q-values as it interacts with the environment. As shown in Fig. 2, the agent initializes a Q-table, observes states, chooses actions via an ε-greedy policy, receives rewards, and updates its Q-values until the end of an episode, gradually learning an optimized scheduling policy for RMFS.

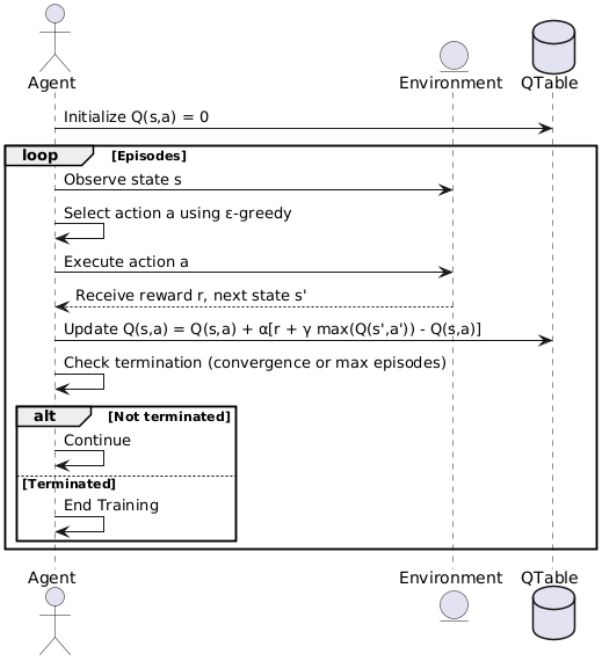


Fig. 2. Q-Learning algorithm flow diagram for RMFS scheduling.

2) Q-Learning-based scheduling model. This subsection describes the Q-Learning algorithm used for dynamic scheduling optimization in RMFS, detailing its initialization, learning process, and training procedure.

a) Q-Table Initialization:

- A Q-table is initialized with dimensions corresponding to all possible states and actions.
- Initially, all Q-values are set to zero or small random values. The Q-values represent the expected future rewards for taking a specific action in each state.

b) State-Action Pair Selection:

- The agent selects an action a in the current state s using a policy, such as ε-greedy.
- ε-greedy policy: The agent explores randomly with probability ϵ and exploits the current knowledge (selecting the action with the highest Q-value) with probability $1-\epsilon$.

c) Update Rule:

- After taking action a, the agent observes the reward r and the next state s'.
- The Q-value for the state-action pair $Q(s,a)$ is updated using the formula:

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma max_{a'} Q(s',a') - Q(s,a)] \quad (1)$$

where,

α: Learning rate (controls how much new information overrides old knowledge).

γ: Discount factor (prioritizes immediate rewards versus future rewards).

r: Immediate reward for taking action$a$ in state a.

$max_{a'} Q(s',a')$: The maximum Q-value of the next states$'$ over all possible actionsa'.

3) Exploration versus exploitation. Balancing exploration and exploitation is crucial for effective learning:

a) Exploration: Trying untested state-action pairs helps discover new and potentially better actions.

b) Exploitation: Focuses on leveraging the agent's current knowledge to maximize rewards.

The ε-greedy policy dynamically adjusts ϵ. At the start of training, ϵ is high to encourage exploration. As training progresses, ϵ decreases, favoring exploitation.

4) Learning process. The learning process is iterative and involves:

a) State observation: The agent observes the current state s.

b) Action selection: An action a is selected using the ε-greedy policy.

c) Action execution and feedback: The agent executes aaa, receives immediate reward r, and observes the next state s'.

d) Q-Value update: The Q-value for the state-action pair Q(s, a) is updated using the Q-Learning formula.

e) Repeat: Steps 1 to 4 are repeated for each episode until convergence or the termination condition is met.

5) Training Procedure.

a) Setup:

- Define the environment, states, actions, reward structure, and Q-table dimensions.
- Initialize learning parameters (α,γ,ϵ).

b) Episodes:

- Train the agent over a predefined number of episodes.
- Each episode begins with an initial state and terminates when a specific condition is met (e.g., all tasks are scheduled or a time limit is reached).

c) Termination Conditions:

- Convergence: No significant changes in Q-values.
- Fixed maximum number of iterations or episodes.

This iterative approach ensures that the Q-Learning agent progressively improves its scheduling policy, achieving near-optimal solutions for the dynamic and complex RMFS environment.

D. Deep Reinforcement Learning (DRL) for Real-Time Adaptation

Deep Reinforcement Learning (DRL) is an extension of traditional reinforcement learning (RL) that employs deep neural networks (DNNs) to approximate complex functions, such as the state-action value function. DRL is particularly effective in environments with large state and action spaces, where traditional Q-learning may struggle due to the need for vast memory and computational power. Here, we explore how DRL is applied for real-time adaptation in manufacturing scheduling.

1) Neural network design. In DRL, the neural network architecture approximates the state-action value function Q(s, a), which defines the expected future rewards for taking an action "a" in a state "s". The neural network acts as a function approximator, mapping states to a vector of action values.

a) Input features: The input layer of the neural network receives the state representation, which may include variables such as:

- Machine configurations (e.g., status of machines, idle or busy)
- Task queue information (e.g., tasks to be processed, task priorities, deadlines)
- System load (e.g., the current distribution of tasks and resource utilization)
- Environmental factors (e.g., external disturbances or machine failures)

These inputs are typically encoded into a fixed-size vector, which is then fed into the neural network.

b) Layer Configurations: The network typically consists of multiple hidden layers, such as:

- Convolutional layers (if spatial relationships between components of the state need to be captured)
- Fully connected layers (to process high-dimensional inputs, such as task queues)
- Activation functions (e.g., ReLU, Leaky ReLU) to introduce non-linearity and allow the network to learn complex relationships.

The final output layer produces a vector of Q-values corresponding to the possible actions in a given state. These values represent the expected rewards of each action, and the agent chooses the action with the highest Q-value.

c) Training DRL network: The neural network is trained using backpropagation and an optimization algorithm like Adam to minimize the error between the predicted Q-values and the target Q-values. The target Q-values are calculated using the Bellman equation, similar to Q-learning:

$$Q(s,a) = r + \gamma max_{a'} Q(s',a') \quad (2)$$

This iterative process allows the neural network to adjust its weights and learn the optimal action selection over time.

2) Continuous learning. Real-time adaptation in manufacturing systems requires the agent to continuously update its knowledge to respond to environmental changes. DRL facilitates this by enabling online learning where the agent updates its policy based on the most recent experiences.

a) Online Learning:

- DRL models often use Experience Replay (replay buffer) to store past state-action-reward transitions. These experiences are randomly sampled and used to update the neural network, reducing the correlation between consecutive experiences and improving learning stability.
- Target Networks may stabilize training by decoupling the target Q-value calculation from the main network, preventing rapid updates that can destabilize the learning process.

b) Real-Time Adaptation:

- As the system operates, the agent continuously interacts with the environment and updates its knowledge. For instance, when there is a disruption in the production process, the agent can adapt by adjusting its scheduling decisions to accommodate the new situation (e.g., reallocating resources, adjusting task priorities).
- The ability of the agent to adapt in real-time is crucial for dynamic environments like RMFS, where conditions (e.g., task priorities, machine failures, or demand fluctuations) can change rapidly.

3) Reward optimization. In DRL, the reward function is critical in guiding the agent's learning. For manufacturing scheduling, the reward function must reflect the key objectives, such as minimizing production time, balancing resource utilization, and ensuring high system throughput. Here's how reward optimization works:

a) Reward signal design: The reward function is designed to drive behavior that improves the scheduling

process. It may be composed of several terms that reward or penalize specific actions:

- Minimizing production time: A reward is given for actions that lead to faster task completion and reduced makespan.
- Minimizing delays: Penalize actions that cause tasks to miss deadlines or increase wait times.
- Balancing resource utilization: Encourage actions that evenly distribute workloads across all available machines, preventing bottlenecks.
- Energy efficiency: Incentivize reducing idle times and energy consumption by dynamically adjusting machine usage.

b) Optimizing scheduling decisions: The agent is trained to maximize the cumulative reward over time. By learning from the rewards it receives, the DRL agent adapts its scheduling decisions to improve production efficiency. For example, the agent optimizes overall system performance by allocating resources to critical tasks first or adjusting the schedule to respond to unexpected machine breakdowns.

c) Dynamic adaptation: As the agent interacts with the environment and gathers feedback from its actions, it continuously adjusts its policies, improving its scheduling decisions based on real-time data. This adaptability is crucial in real-world RMFS where system conditions change constantly.

4) Summary. Deep Reinforcement Learning provides a robust framework for adaptive, real-time decision-making in manufacturing scheduling. By leveraging neural networks to approximate complex value functions and employing continuous learning processes, DRL can dynamically adjust to evolving production demands and optimize scheduling performance. The reward structure is key to guiding the agent towards optimal decisions, balancing competing objectives such as efficiency, resource utilization, and energy consumption.

E. Experimental Configuration

We designed a comprehensive simulation environment representing a dynamic and reconfigurable job shop to evaluate the performance and robustness of the proposed Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) scheduling approaches. This environment mimics the complexity and constraints of Reconfigurable Manufacturing Systems (RMFS), allowing a thorough assessment of the methods' ability to optimize scheduling under realistic conditions.

1) Environmental setup. The experimental environment consists of five machines and twenty independent tasks, each with a processing time ranging from three to fifteen time units. Tasks are assigned random priority levels and due dates within a 30 to 60 time unit window to reflect realistic delivery constraints [33]. To test the resilience and adaptability of the proposed approaches, the environment incorporates disturbances such as machine breakdowns and the sudden arrival of urgent, high-priority jobs. These dynamics enable the evaluation of the proposed methods in a context inspired by recent DRL-based scheduling studies, as illustrated in Fig. 3.

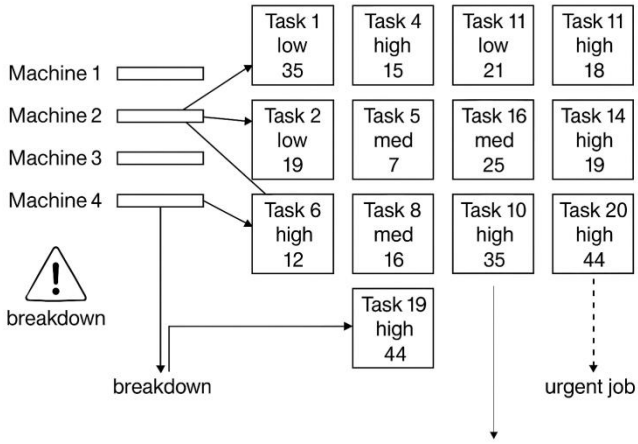


Fig. 3. Experimental scheduling environment: machines, tasks, and disturbance events.

2) Data specifications. The data used in the experiments comprises both simulated and, when available, historical inputs:

a) Historical scheduling data: Includes task priorities, processing times, due dates, and machine availability patterns sourced from actual RMFS environments, providing a realistic basis for model evaluation.

b) Simulated inputs: When historical data is unavailable, synthetic data are generated to mimic realistic operational characteristics, including random task arrivals, varied processing times, and machine downtime events. These inputs enable robust testing across a range of scenarios and system configurations.

3) Performance metrics. The effectiveness of the Q-Learning and DRL methods is measured across several performance indicators:

a) Makespan: Total time required to complete all tasks, indicating the overall efficiency of the scheduling method.

b) Machine utilization: Percentage of active processing time relative to total available machine time, assessing the workload distribution across resources.

c) Tardiness rate: Deviating actual task completion times from their due dates, evaluating the ability to adhere to deadlines and maintain service quality.

d) Energy efficiency: Evaluation of idle times and energy usage across machines, measuring the system's ability to reduce waste.

e) Adaptability: The ability of the approach to maintain performance levels despite unexpected disturbances, such as equipment failures or urgent task arrivals.

4) Reward function. The learning process for the RL and DRL agents is guided by a multi-criteria reward function:

$$Reward = -(Makespan) - \lambda \cdot Tardinesse + \mu \cdot Utilization \quad (3)$$

where,

- λ=0.5: weight for penalties associated with tardiness,
- μ=0.3: weight for promoting higher resource utilization.

These values were selected based on prior tuning and established best practices in RL-based scheduling literature [10]. This reward structure ensures that the learning process simultaneously encourages shorter overall production times, timely task completion, and balanced machine usage.

5) Summary. The experimental configuration provides a realistic, data-rich, and challenging testbed for assessing the Q-Learning and DRL approaches in RMFS environments. By combining dynamic disturbances, multi-criteria performance metrics, and a tailored reward structure, this setup allows for a thorough examination of the effectiveness, efficiency, and resilience of intelligent scheduling methods in modern manufacturing.

F. Comparison with Classical Approaches

This section compares traditional scheduling methods and the proposed Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) techniques. By benchmarking against classical approaches, the advantages of RL and DRL in dynamic and reconfigurable manufacturing environments are highlighted. Fig. 4, Fig. 5, Fig. 6 and Table I summarize the findings.

1) Benchmarking against traditional methods. To evaluate the effectiveness of RL and DRL, their performance was compared with classical scheduling techniques:

a) Shortest processing time (SPT): SPT prioritizes tasks with the shortest processing times, offering computational efficiency. However, as shown in Table 1, SPT exhibits high tardiness rates and limited adaptability due to its inability to account for task deadlines or machine workloads.

b) Earliest due date (EDD): EDD focuses on scheduling tasks based on deadlines, ensuring that tasks with earlier due dates are prioritized. While it demonstrates moderate tardiness performance (Table 1), it fails to optimize resource utilization or adapt to system disturbances effectively.

c) First-come, first-served (FCFS): FCFS processes tasks in the order of arrival, without accounting for task priorities or processing times. This simplicity leads to inefficiencies, as reflected by its poor makespan and utilization scores in Table I.

d) Hybrid approaches: These combine elements of the above methods with basic heuristics to address specific scheduling challenges. While slightly better than standalone traditional methods, they still lack the flexibility and responsiveness required in dynamic environments.

2) Comparison metrics. Performance was evaluated using the following metrics, summarized in Table I:

a) Flexibility: Ability to adapt to changes such as task re-prioritization or machine breakdowns.

b) Response time: Speed of updating scheduling decisions in response to new data.

c) Computational efficiency: Resources required to compute optimized schedules.

3) Case study results. A simulated Reconfigurable Manufacturing System (RMFS) environment benchmarked RL and DRL against traditional methods under identical conditions.

a) Performance Gains with RL and DRL:

- Makespan Reduction: RL and DRL achieved a 15–25% reduction in makespan compared to SPT and EDD (Table I and Fig. 4).

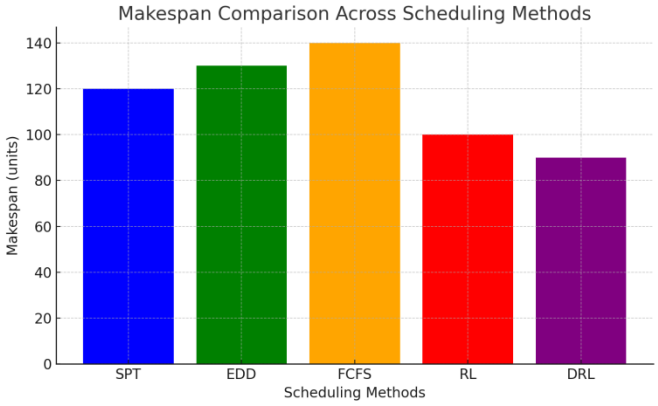


Fig. 4. Makespan comparison across scheduling methods.

Fig. 4 demonstrates the reduced makespan achieved by RL and DRL compared to SPT and EDD.

- Improved Deadline Adherence: RL and DRL significantly decreased tardiness rates, outperforming EDD, which lacks adaptability (Table I and Fig. 5).

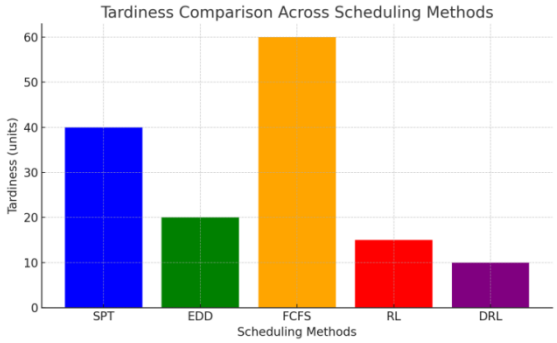


Fig. 5. Tardiness comparison across scheduling methods.

The improved adherence to deadlines with RL and DRL is highlighted in Fig. 5.

- Resource Utilization: RL methods balanced workloads effectively, leading to higher machine utilization and fewer bottlenecks than FCFS and SPT (Table I and Fig. 6).

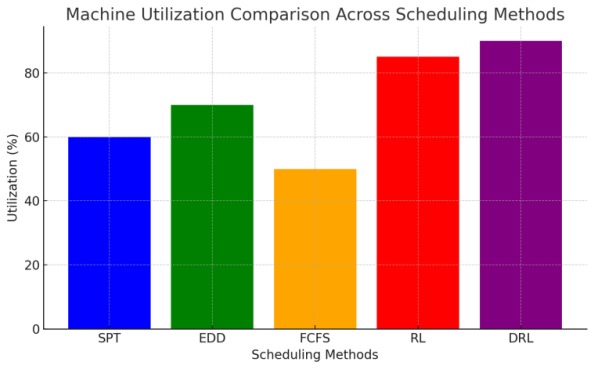


Fig. 6. Machine utilization comparison.

The balanced resource utilization achieved through RL and DRL methods is illustrated in Fig. 6.

- Adaptability: RL and DRL efficiently handled scenarios such as machine breakdowns and dynamic task arrivals, as evident from the adaptability scores in Table I.

b) Computational Trade-offs:

- Traditional methods like SPT and EDD are computationally faster but underperform in dynamic settings (Table I).
- DRL incurs higher computational costs during training but offers near real-time decision-making post-training.

c) Specific Scenarios:

- Dynamic Task Arrivals: RL and DRL managed on-the-fly task arrivals with minimal disruptions to the schedule.
- System Disturbances: RL dynamically reallocated resources during machine failures, unlike traditional methods requiring manual intervention.

4) Visualization of Results

The findings are illustrated through figures and summarized in Table I:

TABLE I. PERFORMANCE METRICS FOR SCHEDULING METHODS

Method	Makespan (min)	Tardiness (%)	Utilization (%)	Adaptability Score
SPT	250	30	70	Low
EDD	230	25	65	Low
FCFS	270	35	60	Low
RL	200	10	85	High
DRL	190	08	90	High

5) Summary of improvements. The comparison confirms that RL and DRL outperform traditional methods in dynamic RMFS environments. While classical approaches excel in simplicity and computational speed, their lack of flexibility and adaptability limits their applicability. RL and DRL provide an optimal balance between decision-making speed and performance, as demonstrated in Table I. The accompanying figures validate these findings further.

G. Summary of Methodology

Our proposed methodology leverages Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) to develop an intelligent scheduler tailored for Reconfigurable Manufacturing Systems (RMFS). The proposed method begins with modeling the scheduling environment and defining state-action pairs, rewards, and transitions. RL algorithms are trained iteratively to optimize task allocation, balancing exploration and exploitation. DRL employs neural networks to approximate state-action values for dynamic adaptability, enabling real-time decision-making. The proposed scheduler ensures improved efficiency, reduced makespan, enhanced resource utilization, and robust adaptability to dynamic production changes by integrating continuous learning and reward optimization.

IV. RESULTS

This section presents the experimental outcomes of our proposed Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) approaches in Reconfigurable Manufacturing Systems (RMFS). The results are compared with traditional scheduling methods—Shortest Processing Time (SPT) and Earliest Due Date (EDD)—using key performance metrics: makespan, tardiness, resource utilization, and adaptability to disturbances.

A. Simulation Objectives

We conducted simulation experiments to evaluate the following:

1) Reduction in makespan. Minimize the total production time.

2) Deadline adherence. Ensure tasks meet deadlines, minimizing tardiness.

3) Resource utilization. Optimize workload distribution and prevent bottlenecks.

4) Adaptability. Assess the system's ability to handle disruptions, such as machine breakdowns.

B. Performance Metrics

1) Makespan reduction. Our approach significantly reduced the total production time compared to traditional scheduling methods such as SPT and EDD. The simulation results show that RL reduced the makespan by 20%, while DRL achieved a 28.57% reduction relative to SPT. Among all methods tested, DRL consistently recorded the lowest average makespan. Table II presents the detailed comparison of average makespan values and percentage improvements across all methods.

TABLE II. AVERAGE MAKESPAN COMPARISON

Method	Average Makespan (seconds)	Improvement (%)
SPT	420	--
EDD	390	7.14
RL	336	20
DRL	300	28.57

2) Deadline adherence. RL and DRL achieved better deadline adherence than EDD, with lower tardiness across simulations. Unlike EDD, DRL adapts to dynamic changes while maintaining compliance. Fig. 7 highlights the reduced tardiness of RL and DRL compared to traditional methods.

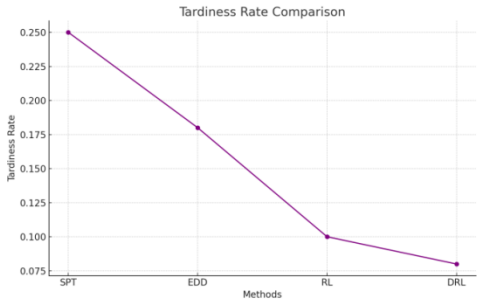


Fig. 7. Line graph of tardiness rates.

Description: The graph highlights the tardiness rate reduction achieved by RL and DRL over multiple simulation runs.

3) Resource utilization. RL and DRL improved workload balance and machine utilization by allocating tasks based on real-time states, reducing bottlenecks and idle times compared to static rules like SPT and EDD. Fig. 8 shows their more balanced resource distribution and enhanced production flow.

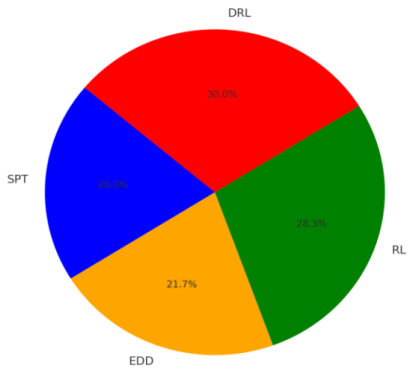


Fig. 8. Resource utilization efficiency (in percentage)

Description: The chart shows more balanced resource usage with RL and DRL, reducing machine idleness and improving production flow.

4) Adaptability to disturbances. Dynamic disruptions such as machine breakdowns were simulated to evaluate the adaptability of each scheduling approach. The results show that RL and DRL effectively minimized performance deviations during disruptions. DRL, in particular, demonstrated the highest level of resilience, with only a 5% increase in makespan compared to normal conditions. In contrast, traditional methods like SPT and EDD experienced more significant deviations. Table III provides a detailed comparison of the makespan values under normal and disrupted conditions and the percentage deviations for each method.

TABLE III. MAKESPAN COMPARISON UNDER DISRUPTIONS

Method	Normal Makespan (seconds)	Disrupted Makespan (seconds)	Deviation (%)
SPT	420	490	16.67
EDD	390	450	15.38
RL	336	360	7.14
DRL	300	315	5.00

The scatter plot in Fig. 9 illustrates the minimal deviation in makespan for RL and DRL under disruptions, underscoring their adaptability.

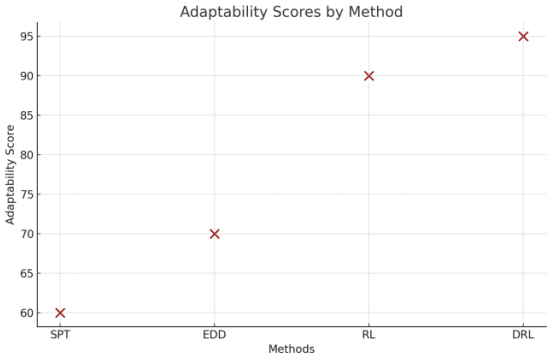


Fig. 9. Scatter plot of adaptability scores.

C. Visualization of Results

1) Gantt charts for task sequences. To visually illustrate the effectiveness of the different scheduling methods, Fig. 10 presents Gantt charts comparing the task sequences generated by RL (Reinforcement Learning), DRL (Deep Reinforcement Learning), and traditional methods such as SPT (Shortest Processing Time) and EDD (Earliest Due Date). This visualization highlights differences in task start and end times, particularly showcasing the reduction of idle times and optimizing transitions achieved through learning-based approaches.

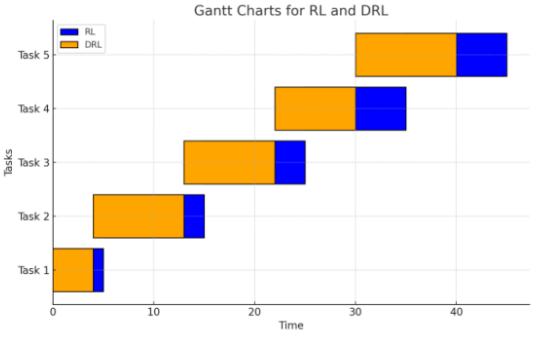


Fig. 10. Gantt charts for task sequences.

2) Comparative performance metrics. Fig. 11 compares key performance indicators—makespan, tardiness rate, and resource utilization—showing clear advantages of RL and DRL over traditional SPT and EDD methods.

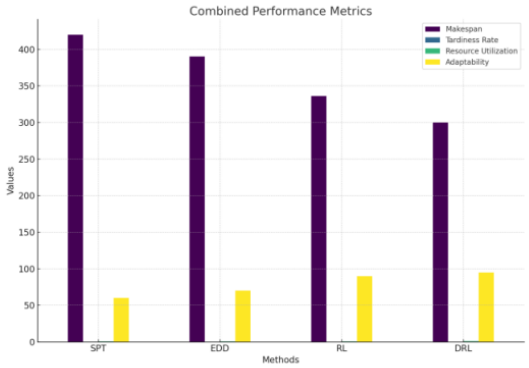


Fig. 11. Combined performance graph.

D. Implementation Results

To validate our method using a Python implementation. Below are details of the implementation outputs:

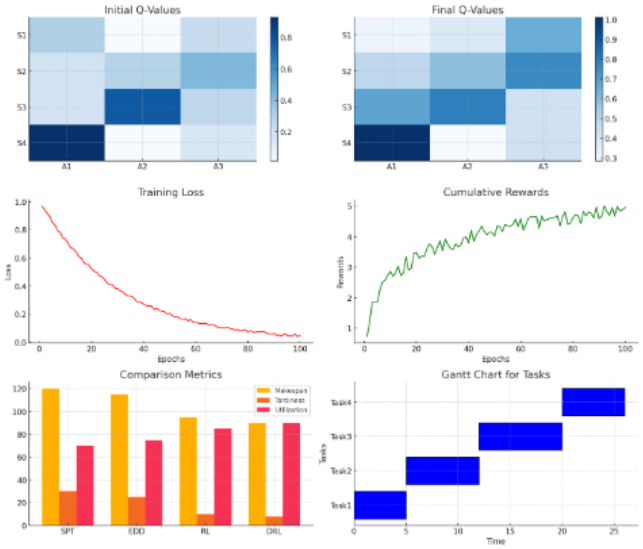


Fig. 12. Result of implementation outputs.

Fig. 12 provides a comprehensive visualization of the experimental results, highlighting the effectiveness of the proposed RL and DRL approaches. Initial and Final Q-table updates illustrate the evolution of Q-values in the RL training process through heatmaps, showcasing the optimization of state-action pairs from their initial random values to well-tuned values after training. DRL Training Process captures the training dynamics of DRL, with the left graph demonstrating a steady decline in training loss over epochs, indicating model convergence, and the right graph reflecting the cumulative rewards, which showcase the agent's learning progress over time. Comparison Metrics compares key performance metrics—makespan, tardiness, and resource utilization—through a bar chart, revealing significant improvements achieved by RL and DRL over traditional methods (SPT and EDD). Finally, the Gantt Chart for Tasks visualizes task schedules, where RL demonstrates optimized task sequencing, reduced idle times, and efficient task durations, emphasizing its superiority in scheduling efficiency.

The experimental results underscore the significant advantages of Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) approaches over traditional scheduling methods, such as Shortest Processing Time (SPT) and Earliest Due Date (EDD), in Reconfigurable Manufacturing Systems (RMFS). Both RL and DRL demonstrated superior performance across all key metrics:

1) Efficiency. DRL achieved the best results in minimizing makespan, reducing it by up to 28.57% compared to SPT, while maintaining low tardiness rates and balanced resource utilization.

2) Adaptability. RL and DRL effectively responded to disruptions such as machine breakdowns, with DRL showing exceptional resilience and minimal performance deviations.

3) Scalability. DRL's ability to handle complex scheduling scenarios highlights its potential for dynamic and reconfigurable production environments.

These results validate the proposed methods as robust and effective solutions for optimizing scheduling in RMFS. While DRL requires additional training time, its benefits in terms of adaptability and efficiency make it a valuable approach for industrial applications.

V. Discussion

This section provides an in-depth analysis of the experimental results, exploring the reasons behind the superior performance of the Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) approaches compared to traditional scheduling methods. Additionally, it evaluates the robustness of the proposed approaches under unforeseen disturbances and discusses potential limitations, including computational costs associated with DRL.

A. Analysis of RL and DRL Performance

The RL and DRL approaches exhibited notable advantages over classical methods, such as Shortest Processing Time (SPT) and Earliest Due Date (EDD), across key performance metrics. This performance can be attributed to several factors:

1) Dynamic decision-making. Unlike SPT and EDD, which rely on static rules, RL and DRL dynamically adapt their scheduling decisions based on the evolving state of the system. This capability allows them to consider multiple factors, such as task priorities, machine availability, and future system states, leading to optimized scheduling solutions.

2) Learning from experience. RL and DRL learn to identify patterns and optimize decision-making through iterative training processes. The DRL approach, leveraging neural networks, can capture complex, non-linear relationships within the system, further enhancing its ability to make efficient scheduling decisions.

3) Makespan optimization. The results show that RL reduced the makespan by 20% and DRL by 28.57% compared to SPT. These improvements stem from the algorithms' ability to minimize idle times and optimize task transitions, ensuring efficient resource utilization.

4) Tardiness reduction. While EDD prioritizes deadlines, its lack of adaptability limits its effectiveness in dynamic environments. In contrast, RL and DRL demonstrated superior deadline adherence by balancing task prioritization with system constraints.

5) Resource utilization. RL and DRL achieved more balanced workload distributions, avoiding bottlenecks and reducing machine idleness. This balance enhances overall system efficiency and throughput.

B. Robustness to Disturbances

One of the key strengths of RL and DRL is their robustness in handling unexpected disruptions, such as machine breakdowns or sudden changes in production demands:

1) Adaptability. Both methods responded effectively to disturbances, maintaining minimal deviations in performance metrics. DRL, in particular, demonstrated exceptional resilience, with only a 5% increase in makespan under disruptions compared to 16.67% for SPT and 15.38% for EDD.

2) Real-time decision-making. The ability of RL and DRL to make real-time decisions based on updated system states enabled quick recovery from disruptions, minimizing their impact on production flow.

3) Scalability. The scalability of DRL makes it particularly suitable for complex and dynamic RMFS environments, where traditional methods often struggle.

C. Potential Limitations

Despite their advantages, the proposed RL and DRL approaches are not without challenges:

1) Computational costs. The training phase of DRL requires significant computational resources and time, particularly for complex systems with large state and action spaces. This limitation may pose practical challenges in deploying DRL in resource-constrained environments.

2) Data requirements. Effective training of RL and DRL models depends on sufficient and representative data availability. In cases, where historical data is limited or unbalanced, the training process may be less effective.

3) Complexity of implementation. Implementing DRL in industrial settings requires machine learning and system modeling expertise, which may not always be readily available.

4) Overfitting risk. In some cases, DRL models may overfit to the training environment, potentially reducing their effectiveness when applied to new or unseen scenarios.

D. Practical Implications

The results and discussion highlight the potential of RL and DRL to revolutionize scheduling in RMFS by providing efficient, adaptable, and scalable solutions. However, addressing the limitations, particularly the computational costs and complexity of implementation, will be crucial for their widespread adoption in real-world applications. Future work could explore strategies to reduce training time, improve generalizability, and make these approaches more accessible to industry practitioners. By integrating RL and DRL into RMFS, manufacturers can enhance productivity, reduce operational costs, and improve system resilience, making these approaches a promising avenue for modern manufacturing systems.

VI. Conclusion

This study presented a comprehensive evaluation of Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) approaches for scheduling optimization in Reconfigurable Manufacturing Systems (RMFS). The results demonstrate that RL and DRL significantly outperform traditional scheduling methods such as Shortest Processing Time (SPT) and Earliest Due Date (EDD), achieving substantial reductions in makespan, improved deadline adherence, and more efficient resource utilization. The DRL approach showed exceptional adaptability to dynamic disruptions, maintaining system performance with minimal deviation. These findings highlight the critical role of Artificial Intelligence (AI) in enabling intelligent scheduling strategies that enhance efficiency, flexibility, and robustness in modern manufacturing environments. Despite these promising results, challenges remain, particularly the high computational cost and extended training time required for DRL. Future research should optimize DRL algorithms to reduce overhead, explore hybrid models that integrate RL with other AI techniques, and validate these approaches in more complex, large-scale, and real-world RMFS scenarios.

References

[1] Tang, J., Haddad, Y., & Salonitis, K. (2022). Reconfigurable manufacturing system scheduling: a deep reinforcement learning approach. Procedia CIRP, 107, 1198-1203.

[2] Yang, S., & Xu, Z. (2022). Intelligent scheduling and reconfiguration via deep reinforcement learning in smart manufacturing. International Journal of Production Research, 60(16), 4936-4953.

[3] Bezoui, M., Kermali, A., Bounceur, A., Qaisar, S. M., & Almaktoom, A. T. (2023, November). Deep Reinforcement Learning for Multiobjective Scheduling in Industry 5.0 Reconfigurable Manufacturing Systems. In International Conference on Machine Learning for Networking (pp. 90-107). Cham: Springer Nature Switzerland.

[4] Li, H., Zhang, H., He, Z., Jia, Y., Jiang, B., Huang, X., & Ge, D. (2024). Solving the integrated process planning and scheduling problem via graph neural network-based deep reinforcement learning. arXiv preprint arXiv:2409.00968.

[5] Jang, J., Klabjan, D., Liu, H., Patel, N. S., Li, X., Ananthanarayanan, B., ... & Juang, T. H. (2024). Scalable Multi-agent Reinforcement Learning for Factory-wide Dynamic Scheduling. arXiv preprint arXiv:2409.13571.

[6] Hammedi, S., & Chniti, H. Combining Petri Nets and AI Techniques to Improve Dynamic Production Scheduling Optimization.

[7] Zhou, L., Zhang, L., & Horn, B. K. (2020). Deep reinforcement learning-based dynamic scheduling in smart manufacturing. Procedia Cirp, 93, 383-388.

[8] Rossit, D. A., Tohmé, F., & Frutos, M. (2019). A data-driven scheduling approach to smart manufacturing. Journal of Industrial Information Integration, 15, 69-79.

[9] Pinedo, M. L., & Pinedo, M. L. (2016). Design and implementation of scheduling systems: More advanced concepts. Scheduling: Theory, Algorithms, and Systems, 485-508.

[10] Hofer, A., Brandl, F., Bauer, H., Haghi, S., & Reinhart, G. (2020). A framework for managing innovation cycles in manufacturing systems. Procedia CIRP, 93, 771-776.

[11] Pulansari, F., & Nugraha, I. (2024). Single machine production scheduling analysis using fcfs, spt, lpt, and edd methods to increase work productivity in furniture companies. Nusantara Science and Technology Proceedings, 15-19.

[12] Hammedi, S., Elmeliani, J., & Nabli, L. (2024). Optimization of Scheduling in Reconfigurable Production Systems: An Approach Based on Intelligent Petri Nets. Saudi J Eng Technol, 9(9), 433-441.

[13] Galán, R. (2008). Hybrid heuristic approaches for scheduling in reconfigurable manufacturing systems. Metaheuristics for scheduling in industrial and manufacturing applications, 211-253.

[14] Li, L., & Mao, C. (2020). Big data supported PSS evaluation decision in service-oriented manufacturing, IEEE Access, vol. 8.

[15] Kusiak, A. (2017). Smart manufacturing must embrace big data. Nature, 544(7648), 23-25.

[16] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. nature, 518(7540), 529-533.

[17] Almeida, F. S. D., & Nagano, M. S. (2023). Heuristics to optimize total completion time subject to makespan in no-wait flow shops with sequence-dependent setup times. Journal of the Operational Research Society, 74(1), 362-373.

[18] Cox, I. I. I., & Spencer, M. S. (2024). The constraints management handbook. CRC Press.

[19] Hammedi, S., Elmeliani, J., & Nabli, L. (2024, November). Embracing the Future: The Evolution of Flexible Reconfigurable Manufacturing Systems. In 2024 International Symposium of Systems, Advanced Technologies and Knowledge (ISSATK) (pp. 1-6). IEEE.

[20] IX, S. (2012). ting S ystem.

[21] Khadivi, M., Charter, T., Yaghoubi, M., Jalayer, M., Ahang, M., Shojaeinasab, A., & Najjaran, H. (2025). Deep reinforcement learning for machine scheduling: Methodology, the state-of-the-art, and future directions. Computers & Industrial Engineering, 110856.

[22] Kibira, D., Morris, K. C., & Kumaraguru, S. (2016). Methods and tools for performance assurance of smart manufacturing systems. Journal of Research of the National Institute of Standards and Technology, 121, 282.

[23] Grumbach, F., Badr, N. E. A., Reusch, P., & Trojahn, S. (2023). A memetic algorithm with reinforcement learning for sociotechnical production scheduling. IEEE Access, 11, 68760-68775.

[24] Hammedi, Salah & Jalloul, el méliani & Nabli, Lotfi. (2025). Optimizing resource allocation in job shop production systems with seasonal demand patterns. International Journal of Reconfigurable and Embedded Systems (IJRES). 14. 12. 10.11591/ijres.v14.i1.pp12-25. http://doi.org/10.11591/ijres.v14.i1.pp12-25

[25] Abd Elaziz, M., Attiya, I., Abualigah, L., Iqbal, M., Ali, A., Al-Fuqaha, A., & El-Sappagh, S. (2023). Hybrid enhanced optimization-based intelligent task scheduling for sustainable edge computing. IEEE Transactions on Consumer Electronics.

[26] Abbas, B., & Shah, M. A. (2023, August). An Empirical Study on Hybrid Scheduling Algorithms in Cloud Computing. In 2023 28th International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE.

[27] Hammedi, S., Elmelliani, J., Nabli, L., Namoun, A., Alanazi, M. H., Aljohani, N., ... & Alshmrany, S. (2024). Optimizing Production in Reconfigurable Manufacturing Systems with Artificial Intelligence and Petri Nets. International Journal of Advanced Computer Science & Applications, 15(10).

[28] Danishvar, M., Danishvar, S., Katsou, E., Mansouri, S. A., & Mousavi, A. (2021). Energy-aware flowshop scheduling: A case for AI-driven sustainable manufacturing. IEEE Access, 9, 141678-141692.

[29] Napoleone, A., Andersen, A. L., Brunoe, T. D., & Nielsen, K. (2023). Towards human-centric reconfigurable manufacturing systems: Literature review of reconfigurability enablers for reduced reconfiguration effort and classification frameworks. Journal of Manufacturing Systems, 67, 23-34.

[30] Serrano-Ruiz, J. C., Mula, J., & Poler, R. (2022). Development of a multidimensional conceptual model for job shop smart manufacturing scheduling from the Industry 4.0 perspective. Journal of Manufacturing Systems, 63, 185-202.

[31] Mateev, V., & Marinova, I. (2023, December). Modified chromosome pooling genetic algorithm for resource allocation optimization. In AIP Conference Proceedings (Vol. 2939, No. 1). AIP Publishing.

[32] Xi, S., Chen, Q., MacGregor Smith, J., Mao, N., Yu, A., & Zhang, H. (2020). A new method for solving buffer allocation problem in large unbalanced production lines. International Journal of Production Research, 58(22), 6846-6867.

[33] Bai, J., Fang, S., Xu, X., & Tang, R. (2022). LMPF: A novel method for bill of standard manufacturing services construction in cloud manufacturing. Journal of Manufacturing Systems, 62, 402-416.

A Proposed Framework for Loan Default Prediction Using Machine Learning Techniques

Mona Aly SharafEldin[1], Amira M. Idrees[2], Shimaa Ouf[3]
Business Information Systems, Helwan University, Egypt[1]
College of Business, King Khalid University, Saudi Arabia[2]
Information Systems Department, Helwan University, Egypt[3]

***Abstract*—The accurate prediction of loan defaults is critical for the risk management strategies of financial institutions. Traditional credit assessment approaches have often relied on subjective judgment, leading to inconsistent decisions and heightened financial risk. This study investigates the application of machine learning techniques—namely Random Forest, Decision Tree, and Gradient Boosting—to predict loan defaults using customer data from the Agricultural Bank of Egypt. The research emphasizes the role of feature selection in enhancing model performance, utilizing both embedded and recursive methods to isolate key predictive attributes. Among the evaluated features, loan balance, due amount, and delinquency history emerged as the most influential, while demographic variables like gender and employment status were found to be less significant. The Decision Tree model demonstrated superior performance with an overall accuracy of 88%, a recall of 53%, and a specificity of 89%, making it the most effective among the tested classifiers. The findings highlight the importance of combining robust feature selection with interpretable models to support informed decision-making in banking.**

***Keywords*—Random forest; decision trees; gradient boosting machines; feature selection; feature importance; loan default**

I. Introduction

In recent decades, the financial sector has increasingly adopted credit lending as a central business activity. However, many of these institutions are still facing serious problems with respect to the proper assessment of credit risk, mainly because there is a lack in traditional evaluating models that do not count on the real behavior of the borrower. Accurate estimation of credit risk is a key factor in maintaining financial stability and improving the performance of institutions in the data-driven economy of today [1].

shimaaouf@commerce.helwan.edu.egWith the proliferation of data and computing capabilities, machine learning has become available for mainstream risk management tasks. The latest technologies provide enhanced prediction on loan repayment and default identification. One of the biggest problems in creating accurate predictive models focusing on relevant features is the process of eliminating noise and irrelevant signals and zeroing in on what really matters: input variables to those most impactful on prediction results. Removing irrelevant or repetitive features not only reduces the risk of overfitting but also simplifies the model and enhances interpretability [2].

Mona.Aly21@commerce.helwan.edu.eg[1]
aidrees@kku.edu.sa[2]
shimaaouf@commerce.helwan.edu.eg[3]

To address these challenges, various feature selection approaches, such as filtering, wrapping, and embedded methods, are employed to refine datasets and ensure that learning algorithms can generalize effectively. When assessing the likelihood of loan default, classification algorithms, like decision trees, support vector machines, and random forests, are commonly used. Furthermore, ensemble learning methods, such as AdaBoost [3], have proven to improve performance by combining multiple predictive models.

These algorithms analyze borrower-specific financial data, including income levels, employment details, credit scores, and historical payment patterns. The integration of feature selection with these advanced models enables financial institutions to build more reliable predictive systems that not only improve forecast accuracy but also provide clear insights into the variables that influence repayment behavior [4].

Essentially, a loan is a financial arrangement in which a borrower receives funds from a lender under specific repayment terms. This agreement includes the repayment of the principal amount, interest, and other applicable charges. To approve a loan, lenders typically examine a borrower's financial history and current obligations to assess their capacity to fulfill the repayment terms. If the borrower fails to meet payment deadlines beyond a predetermined grace period, the loan may be categorized as default [5].

As the financial sector continues to evolve, early identification of potential loan defaults has become increasingly important for minimizing institutional risk. By leveraging machine learning methods and carefully selected data inputs, lenders can make more informed, accurate, and strategic credit decisions.

The primary objective of this study is to identify the key factors that influence decision-making in the context of loan defaults. By applying machine learning methods, the study aims to explore which attributes significantly affect default risk and to evaluate the effectiveness of different machine learning methods in this domain. Additionally, the study seeks to determine the most suitable technique for accurately predicting loan default outcomes, thereby supporting more informed and data-driven financial decision-making.

The contributions were made to the field of financial risk modelling by focusing on loan default prediction in a specific context: livestock-based microfinance within Egypt's agricultural sector. The study applies machine learning

techniques tailored to the particular characteristics and challenges of this domain. The contributions can be summarized as follows:

A. Loan Default Prediction in Livestock-Based Lending

The study addresses a relatively under-researched area: livestock-based lending using real-world data obtained from the Agricultural Bank of Egypt, which serves rural communities across more than 1,190 branches. Unlike commonly studied domains such as credit cards or personal loans, this dataset reflects borrowing behavior in a specialized agricultural finance context.

B. Application of Machine Learning Techniques

Various supervised machine learning algorithms, including Random Forest, Decision Tree, and Gradient Boosting, were utilized to evaluate feature significance and create predictive models. Important variables, such as balance (BAL), due amount (DUE), and delinquency score (DELI), emerged as strong predictors of loan defaults. Conversely, attributes like gender and occupation showed minimal predictive power and were omitted to streamline the models and enhance clarity.

C. Decision Tree Analysis for Interpretability

The Decision Tree model, which recorded the highest accuracy (88%) among the models tested, was chosen for deeper examination due to its interpretability. Analysis of the root node indicated that a balance threshold of ≤ 2140 was a key risk indicator. Additional branches included factors like region (ZONE_NAME), DELI, and DUE, allowing the model to recognize interactions within borrower attributes.

D. Regional Influence on Risk Classification

Borrowers from the "East Delta" region exhibited more uncertain classification outcomes, suggesting that regional factors may influence loan repayment behavior. This highlights the value of incorporating geographic context in risk assessment models, especially in rural or sector-specific lending scenarios.

E. Gini Index as a Measure of Classification Certainty

The Gini index was employed to assess the certainty of classification outcomes across various decision tree paths. Lower Gini values indicated greater confidence in predictions, whereas moderate values signaled more variability in borrower characteristics. This measure facilitated the identification of segments where the model's forecasts were more or less dependable, aiding in practical decision-making.

F. Comprehensive Performance Assessment

In addition to accuracy, the study included further evaluation metrics like precision, recall, and F1-score to offer a more thorough insight into model performance. This methodology is particularly important in financial contexts, where both false positives and false negatives have significant operational ramifications.

The research is structured into eight sections: Introduction is given in Section I. Related Work in Section II. Proposed Framework in Section III. Data Integration in Section IV. Data preprocessing in Section V. Feature Selection in Section VI. Model Building in Section VII. Model Evaluation in Section VIII. Results in Section IX. Discussion in Section X. Finally, Conclusion and Future Work in Section XI and XII.

II. Related Work

This section provides an overview of existing studies related to the prediction of default loans using machine learning methods. In recent years, this area has been the focus of numerous studies.

Puli (2024) [6] utilized various machine learning methods, including random forests, naïve Bayes, gradient boosting, support vector machines, neural networks, k-nearest neighbors, and decision trees, along with statistical approaches, such as logistic regression. The findings indicate that neural networks and random forest models demonstrated notable efficacy in predicting banking crises in India. Additionally, the research evaluated the performance of multiple algorithms, including neural networks, naïve Bayes, and ensemble techniques, such as random forest bagging, boosted decision trees, stacking of logistic regression, support vector machines, and neural networks, for forecasting loan defaults. The ensemble models consistently outperformed individual classifiers. Specifically, boosted decision trees achieved the highest accuracy (84.9 %), followed by random forest (83.1 %). Neural networks recorded an accuracy of 80.3%, surpassing other individual classifiers, whereas naïve Bayes exhibited significantly lower performance at 46.5%. These results underscore the superior predictive capability of the ensemble methods in this domain.

Saini (2023) [7] explored the application of various machine learning algorithms to predict loan approval decisions. The study implemented Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression (LR) to evaluate their predictive capabilities. The comparative analysis revealed that the Random Forest model delivered superior performance, achieving an accuracy of 98.04%, thus indicating its robustness and effectiveness in loan approval prediction tasks.

Jovanne (2023) [8] employed various supervised machine learning algorithms, k-nearest neighbors (k-NN), decision trees (J48), naïve Bayes, and logistic regression to predict loan default using a dataset of 1,000 instances. Preprocessing steps included missing value imputation, normalization, and SMOTE for class balancing. Feature selection was conducted using correlation, information gain, and wrapper methods. The models were evaluated using classification accuracy, F-measure, and kappa statistics. Results showed that k-NN (k=3) and logistic regression achieved the highest accuracy rates of 78.38% and 77.31%, respectively. These findings highlight the potential of data mining techniques in enhancing credit risk assessment in financial institutions.

Alaradi and Hilal (2020) [9] utilized a predictive model for assessing loan approval using a variety of decision tree-based algorithms, ranging from simple decision trees to more complex ensemble methods like random forests. Their findings indicated that basic decision trees struggled to deliver strong performance, likely due to their limited capacity to capture the intricate and highly correlated relationships among key loan-related features. Despite this, the decision tree (DT) model demonstrated a strong balance between accuracy, interpretability, and practical

relevance. It achieved a test accuracy of 97.25%, making it a viable option for automating and accelerating the evaluation of loan applications based on applicant characteristics. The study recommended that using the DT-based prediction model can support more transparent and efficient decision-making in loan processing systems.

Güder and Köse (2024) [10] conducted a study to develop predictive models for home loan approval decisions using machine learning algorithms. The primary goal was to enhance the efficiency and accuracy of loan processing in the banking sector, minimizing human error and improving decision-making speed. The researchers employed four popular supervised learning techniques: k-nearest neighbors (KNN), random forest (RF), support vector machines (SVM), and logistic regression (LR). Two separate datasets were used to assess the performance of these models across varying data conditions. Each dataset was split into 90% for training and 10% for testing. The performance metrics included accuracy, precision, recall, F1-score, specificity, and others. The results indicated that SVM achieved the best performance on the first dataset with an accuracy of 88.7%, while Random Forest performed best on the second dataset, reaching an accuracy of 98.8%. Additionally, Random Forest achieved the highest precision (98.9%) and recall (99.3%) in the second dataset, highlighting its strong predictive capability in more homogeneous data environments. In contrast, the performance of algorithms on the smaller, more variable dataset was slightly lower, illustrating the impact of data quality and size on model accuracy.

Sampurna and Vidya (2023) [11] discuss how the XGBoost and Random Forest algorithms outperformed decision trees, logistic regression, and regularized logistic regression in terms of classification performance. The research highlights the importance of evaluating multiple algorithms to identify the most effective one.

Tabassum, Namita, and Prachi (2025) [12] conducted a study to evaluate the effectiveness of various machine learning models in forecasting financial distress among publicly listed companies in Vietnam. The results indicate that the Extreme Gradient Boosting (XGBoost) model achieved the highest predictive accuracy at 95.66%, while the Artificial Neural Network (ANN) yielded the lowest accuracy at 91.68%. To enhance the interpretability of the models, SHAP (Shapley Additive Explanations) values were employed to identify the most influential financial indicators. Key variables, such as the long-term debt-to-equity ratio, enterprise value-to-sales ratio, accounts payable-to-equity ratio, and diluted earnings per share (EPS), were found to play a critical role in predicting financial instability. The study not only offers a novel analytical tool for credit rating agencies to assess default risks but also contributes to the broader effort to make complex machine learning models more transparent and explainable.

Hussain (2024) [13] conducted a study employing supervised machine learning algorithms, including Decision Tree, Naive Bayes, Multilayer Perceptron, and a Stacking Ensemble Model, to predict loan approval outcomes in the banking sector. The dataset used comprised borrower-related features, such as income, credit history, loan amount, and employment status. Data preprocessing was enhanced using K-Nearest Neighbors (KNN) for missing value imputation, alongside clustering techniques (K-Means) to segment applicants based on income and dependent attributes.

The study applied classification algorithms to both general and cluster-specific data. The Stacking Model demonstrated the highest testing accuracy (83.24%), outperforming Naive Bayes (82.16%), Decision Tree (80%), and Multilayer Perceptron (73.51%). Additionally, clustering models based on applicant income and marital/dependent status yielded varied accuracies across clusters, ranging from 33.33% to 81%, indicating the effectiveness of group-specific modeling.

Chen (2023) [14] proposed a framework for detecting financial statement fraud in publicly listed companies by utilizing several machine learning models, including LightGBM, XGBoost, Gradient Boosting Decision Trees (GBDT), and Random Forest. To enhance model performance, the study also introduced an integrated feature selection approach. Furthermore, the use of the Synthetic Minority Over-sampling Technique (SMOTE) effectively addressed the issue of class imbalance, leading to a notable improvement in fraud detection accuracy. Among the evaluated models, GBDT demonstrated superior performance, achieving the highest Area Under the Curve (AUC) and sensitivity metrics.

Ali (2023) [15] introduced a fraud detection model based on the XGBoost algorithm, aimed at identifying fraudulent financial activities within companies across the Middle East and North Africa (MENA) region. To address the issue of class imbalance in the dataset, the study employed the Synthetic Minority Over-sampling Technique (SMOTE). Various machine learning algorithms were implemented using Python to predict instances of financial statement fraud. The experimental results demonstrated that XGBoost achieved the highest accuracy, reaching 96.05%, outperforming other models such as Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM).

Gupta and Mehta (2021) [16] conducted a study to examine the application of statistical and machine learning methodologies in detecting financial statement fraud. The findings indicate that the fuzzy technique, neural network, decision tree, SVM, logistic regression, and probit regression achieved the highest precision rates of 89.5%, 71.7%, 73.6%, 90.4%, and 86.8%, respectively. Consequently, machine learning techniques demonstrated superior performance compared with statistical methods in predicting the likelihood of fraud in businesses, particularly when the sample had limited data access, achieving a precision rate of 96.05%.

Venkatesan, A. Kumar and S. Sabni, (2020) [17] used a machine learning algorithm to detect credit card fraud based on the customer's transaction, and the accuracy of the system was predicted using machine classification algorithms, such as logistic regression and KNN, based on credit card data. The KNN algorithm delivered the best results in terms of statistical measures, such as precision 0.95), recall (0.72), and f1-score (0.82) for fraud.

Vahid (2024) [18] conducted research employing machine learning models, Logistic Regression, KNN, SVM, Decision Tree, and Random Forest, to predict loan approval. Using a

Kaggle dataset, the study tested the impact of two-feature selection methods: K-Best and Recursive Feature Elimination (RFE). The Random Forest model achieved the highest accuracy (97.71%) when paired with RFE and cross-validation. Results showed that selecting key features significantly improves model performance, and Random Forest is especially effective in loan prediction tasks.

Kozinaa, (2023) [19] proposed automated models for predicting defaults in leasing companies using machine-learning techniques. These models use a combination of internal data from company systems and external sources from financial supervisory institutions, incorporating features related to leasing contracts and asset specifications. The experimental results showed that the Random Forest algorithm achieved the highest precision for default prediction, reaching up to 75%, whereas deep neural networks recorded the highest recall at 81.6%. For non-default predictions, the deep learning methods demonstrated the highest precision, whereas AdaBoost achieved the best recall. Although these models were developed in the context of leasing firms, the authors suggest that their flexibility allows them to be applied across various financial institutions to support credit risk management. However, a notable limitation is the substantial amount of detailed input data required, which could present practical challenges for real-world deployment. The authors recommend that future studies explore statistical forecasting techniques to benchmark and possibly enhance current models.

Jumaa (2023) [20] introduced a deep-learning model to predict loan default risks among UAE bank customers. Using data from 1000 respondents, the model translated demographic and financial factors, such as income, debt ratio, and credit card use, into predictive inputs. After pre-processing, a neural network was developed with 25 input features and tested using Keras and TensorFlow. The model achieved 97.6% training accuracy and 95.2% training accuracy on the test data, showing a strong potential in supporting lending decisions. The authors emphasized the model's practical value and called for further research using regression and variance analyses.

Viswanatha (2023) [21] addressed the ongoing difficulties faced by banks in accurately identifying suitable loan applicants amid rising demand. To enhance the selection process, the study suggested integrating various machine learning techniques with ensemble learning methods. The research applied algorithms, such as Random Forest, Naive Bayes, Decision Tree, and K-Nearest Neighbors (KNN), achieving a maximum accuracy of 83.73%, with Naive Bayes outperforming the others. This approach not only improves selection precision but also significantly reduces loan processing time, benefiting both applicants and banking staff by streamlining the approval procedure.

Btoush (2025) [22] proposed a hybrid ensemble framework combining classical machine learning techniques—such as decision tree (DT), random forest (RF), support vector machine (SVM), XGBoost, CatBoost, and logistic regression (LR)—with deep learning models, including convolutional neural networks (CNN) and bidirectional LSTM (BiLSTM) enhanced with attention mechanisms. The model employed a stacking approach to integrate predictions from all classifiers, using Random Forest as the final estimator. To address the challenge of data imbalance in detecting credit card fraud, the authors applied robust pre-processing and evaluation strategies. The hybrid model achieved outstanding performance, with an accuracy of 99.97%, precision of 97.62%, recall of 83.67%, and an F1-score of 90.11%. These results demonstrated its effectiveness in minimizing both false positives and undetected fraudulent transactions. The study highlighted the model's scalability, interpretability, and suitability for real-time fraud detection.

Several previous studies have explored the use of machine learning methods for classification tasks. Nevertheless, many of these studies have primarily relied on publicly accessible or benchmark datasets and have not thoroughly tackled significant challenges such as class imbalance or the relevance of findings in real-world scenarios. Although models such as XGBoost, Random Forest, and SVM have shown promising outcomes in certain contexts, the degree to which these results translate to operational environments remains unclear.

This study aims to fill two identified gaps in the existing literature. First, regarding the choice of algorithms, it employs established machine learning techniques—specifically, Decision Tree, Random Forest, and XGBoost—selected for their effective combination of predictive accuracy, interpretability, and resilience across various data distributions. Second, and more importantly, the research utilizes actual operational banking data from the Agricultural Bank of Egypt, as opposed to depending on conventional or synthetic datasets. This choice offers a more realistic basis for developing and assessing models.

In Table I, the results of the machine learning techniques utilized in the previous studies have been presented.

TABLE I. RESULTS OF THE EMPLOYED TECHNIQUES IN PREVIOUS STUDIES

Reference	Techniques	Accuracy	Recall	Precision	F1-Score
Puli (2024) [6]	DT	84.9%9	98.98%	100%	99.49%
	RF	83.10%	-	-	-
	NN	80.30%	-	-	-
	Naïve Bayes	46.50%	-	-	-
Saini (2023) [7]	RF	98.04%	-	-	-
Jovanne (2023) [8]	DT	78.38%	-	-	-
	k-NN	78.58%			
	Naïve Bayes	76.54%			

	Logistic	77.31%			
Alaradi &Hilal(2020) [9]	LR	79.50%	-	-	-
	KNN	78.41%	-	-	-
	SVM	93.78%	-	-	-
	DT	87%	-	-	-
	RF	97.68%	-	-	-
Güder a& Köse (2024) [10]	Data set1 KNN	87%	95.70%	88.20%	91.80%
	Data set1RF	87.00%	95.70%	88.20%	91.80%
	Data set1LR	87.00%	97.90%	95.70%	92%
	Data set1SVM	88.70%	97.90%	87%	93.10%
	Data set2 KNN	92.90%	92.50%	96.10%	94.30%
	Data set2RF	98.80%	92.50%	98.90%	99.10%
	Data set2LR,	90.80%	94%	92.90%	92.70%
	Data set2SVM	93.90%	94%%	96.20%	95.10%
Sampurna & Vidya (2023) [12]	XGBoost, RF, LR				
Tabassum (2025) [13]	XGBoost	95.66%,	-	-	-
	ANN	91.68%.	-	-	-
Hussain (2024) [14]	DT	80%	-	-	-
	Naïve Bayes	82.16%	-	-	-
	Multilayer Perceptron	73.51%	-	-	-
	Stacking Model	83.24%	-	-	-
Chen (2023) [15]	RF+SMOTE	60.55%	-	-	-
	GBDT+SMOTE	72.24%	-	-	-
	XGBoost + SMOTE	66.46%	-	-	-
	LGB+SMOTE	69.47%	-	-	-
Ali (2023) [16]	LR	73.88%	83.44%	80.55%	81.96%
	DT	82.22%	50%	41.11%	45.13%
	SVM	88.88%	84.11%	80.34%	82.18%
	RF	80.55%	50%	40.27%	0.4461%
	AdaBoost	83.33%	50%	41.66%	45.45%
	XGBoost	93.66%	86.37%	79.38%	82.72%
Gupta & Mehta (2021) [17]	SVM	90.4%	-	-	-
	LR	86.8%	-	-	-
	DT	73.6%	-	-	-
Vengatesan (2020) [18]	LR, KNN		72%	95%	82%
Vahid (2024) [19]	LR	79.50%	-	-	-
	KNN	78.41%	-	-	-
	SVM	93.78%	-	-	-
	DT	87.02%	-	-	-
	RF	97.68%	-	-	-
Kozina & Michaland (2023) [20]	RF	-	81.6 %	75.0 %	-
Jumaa (2023) [21]	NN	95.2%	-	-	-
Viswanatha (2023) [22]	RF,	77.23%			
	NB	83.73%,	-	-	-
	DT	63.41%	-	-	-
	KNN	77.23%	-	-	-
Btoush (2025) [23]	DT	99.93%	81.63%	89.89%	-
	RF	99.96%	76.53%	97.40%	-
	SVM	99.94%	66.33%	97.02%	-
	XGBoost,	99.96%	77.55%	95.00%	-
	LR	99.92%	60.20%	88.06%	-

III. Proposed Framework

Fig. 1 illustrates the proposed framework for predicting loan defaults using machine-learning techniques. The model began with data integration and selection, combining historical customer records and geographic information from the Egyptian Agricultural Bank. Following this, a comprehensive data preprocessing phase ensures data quality through steps such as handling missing values, normalization, encoding, and outlier detection. Feature selection is then conducted to extract the most informative attributes and uncover significant variable interactions. The refined dataset was subsequently used to train predictive models using machine learning algorithms. Finally, the different models are evaluated using robust performance metrics and visualized through interpretive tools to facilitate informed decision-making in credit risk assessment.

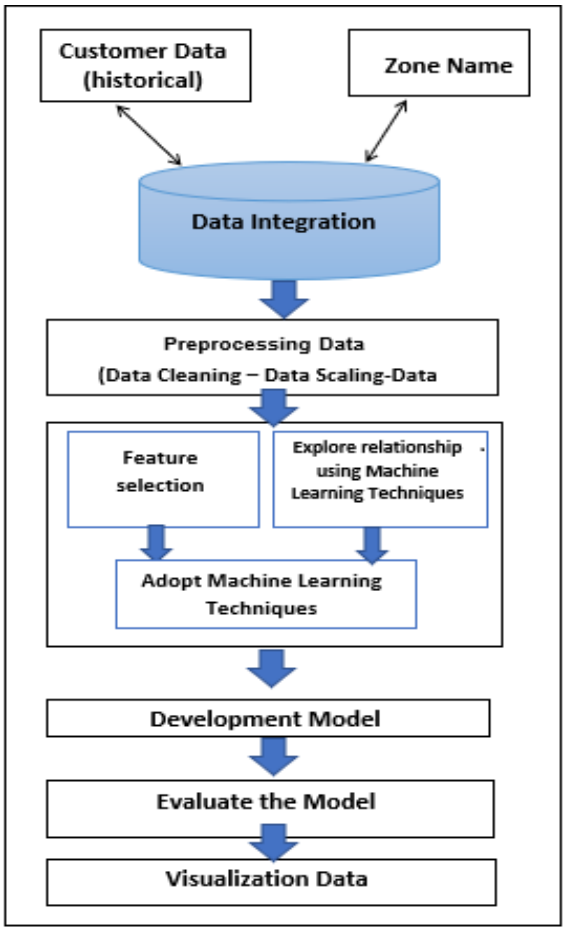


Fig. 1. Proposed framework.

IV. Data Integration

The first stage of the proposed framework centers on merging two key data sources: historical customer information and geographic zone data sourced from the Egyptian Agricultural Bank. The combined dataset consists of 168,95100 entries, each featuring attributes that represent the demographic, financial, and geographic characteristics of individual borrowers. This integration process guarantees the data's consistency, completeness, and relevance, thus setting it up for subsequent analytical tasks. As illustrated in Table II, the chosen features were directly sourced from customer records. These attributes were selected based on their potential to predict loan defaults and include variables such as gender, age category, employment status, loan amount, history of delinquency, loan due date, and the corresponding geographic zone. The final compilation of selected features is as follows:

TABLE II. The List of Attributes of ABE Data

Feature	Data	Column name
Name of zone	Each zone name included the branch it follows.	ZONE_NAME
Gender	Gender and Sex of borrowers.	GENDER
Age of the borrower	age range	AGE_RANGE
job of the borrower	yes\no	JOB
Balance of loan	Loan amount requested by the borrower	BAL
Delinquency	The number of times the borrower had been delinquent for 30+ days in the past 3 years	DELI
Due date for loan	DATE of due loan	LOAN_DUE_DATE
Classification	target for prediction (loan status according to table (0\1)	CLASSIFICATION

To get ready for model training, the target variable (CLASSIFICATION) was established by transforming the original delinquency status into a binary format (Yes/No), based on the days past due (DPD). As explained in the corresponding delinquency classification in Table III, records with DPD values of 91 days or more (i.e., buckets 4 to 7) were categorized as "defaulted (Yes=1)", while those with values below this threshold (Current to Bucket-3) were designated as "non-defaulted (No=0)".

TABLE III. Show Retail Loans Classification

Delinquency buckets	DPD as per Due Date	Retail Classification ORR	Non-accrual status	Classification based on past due days count
Current	0	Performing	no	
bucket-1	Jan-30		no	
bucket-2	31-60	Sub-standard	no	
bucket-3	61-90		no	
bucket-4	91-120	Doubtful	yes	
bucket-5	121-150	bad/non-performing	yes	
bucket-6	151-180		yes	
bucket-7	181+		yes	

A. Descriptive Analysis

1) Sample distribution by geographic zone. Table IV and associated graph presented in Fig. 2 below offer a descriptive summary of the dataset, emphasizing the geographic distribution of loan applicants. This analysis indicates that the West Delta region is the primary source of observations, representing the largest share of the sample. Conversely, the Branches Sector makes up the smallest portion.

This distribution highlights the regional concentration of applicants, suggesting a potential geographic influence on lending patterns that may be explored further in subsequent analyses.

TABLE IV. BRIEF RESULT OF THE DISTRIBUTION OF THE SAMPLE ACCORDING TO ZONE

Category	Frequency	Percentage
"West Delta"	1729335	35.93%
"East Delta"	1079627	22.43%
"South Upper Egypt"	1044724	21.70%
"North Upper Egypt"	932144	19.37%
"Branches sector"	27686	0.58%

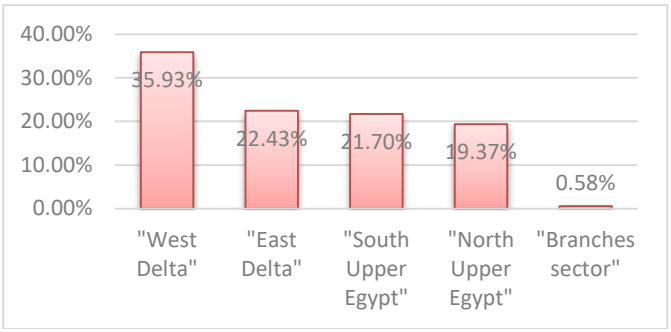


Fig. 2. Distribution of the sample according to zone.

2) Loan default status. The following table summarizes the dataset concerning the loan default status of applicants. It is clear that a significant majority of the instances are classified in the "No Fault" category, signifying that most borrowers have not defaulted on their loans. A mere fraction of the sample is linked to default (see Table V).

TABLE V. BRIEF RESULT OF THE DISTRIBUTION OF THE SAMPLE ACCORDING TO LOAN DEFAULT

Fault	Frequency	Percentage
No	4638395	4%
Yes	173299	96%

3) Age distribution. Table VI illustrates the age distribution of the sample. It is clear that the highest percentage of individuals is in the 56 to 60 age bracket, comprising 37.9% of the overall sample. In contrast, the 21 to 25 age cohort constitutes the smallest portion, with merely 0.8% of the cases. This indicates that the dataset is predominantly made up of older individuals, whereas younger age groups are significantly underrepresented. Fig. 3 shows the distribution of the sample according to age range.

4) Gender distribution. The distribution of the dataset by gender indicates a significant imbalance, with male clients comprising the majority at approximately 79%, while female clients represent only 21%, as shown in Table VII. Disparity may reflect broader demographic or socioeconomic trends within the bank's customer base, and highlights the importance of examining whether gender has a substantial impact on loan default prediction

TABLE VI. BRIEF RESULT OF THE DISTRIBUTION OF THE SAMPLE ACCORDING TO AGE RANGE

AGE_RANGE	Frequency	Percentage
21-25	39602	0.8%
26-30	211996	4.4%
31-35	298947	6.2%
36-40	400692	8.3%
41-45	428711	8.9%
46-50	447490	9.3%
51-55	483336	10.0%
56-60	1825130	37.9%
61-65	417645	8.7%
66-70	259967	5.4%

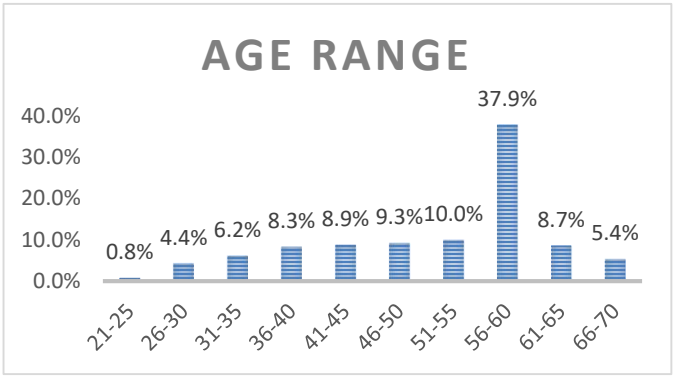


Fig. 3. Display of the distribution of the sample according to age range.

TABLE VII. BRIEF RESULT OF THE DISTRIBUTION OF THE SAMPLE ACCORDING TO GENDER

Gender	Frequency	Percentage
M	3799435	79%
F	1014081	21%

5) Job distribution. Table VIII illustrates the distribution of the sample according to employment status. It is clear that the overwhelming majority of individuals in the dataset are employed, representing 99.7% of the total sample, whereas a negligible portion (0.3%) are not currently employed.

TABLE VIII. BRIEF RESULT OF DISTRIBUTION OF THE SAMPLE ACCORDING TO JOB

Job	Frequency	Percentage
Work	4798416	99.7%
Not Work	15100	0.3%

After presenting how the data are integrated and presenting descriptive analysis of the data, now the second step is to present how the data is preprocessed.

V. Data Preprocessing

Before conducting the analysis, the integrated dataset underwent several preprocessing steps to ensure data quality and prepare it for modeling. These steps are essential, particularly in financial applications such as bank loan risk assessment, where the reliability of data directly impacts model performance.

A. Data Cleaning

Handling Missing Values: in this study, an available case method was used, which is that only complete records were retained for analysis. Although this approach may lead to information loss especially if the proportion of missing values is high it was considered appropriate given the large dataset size. Retaining complete cases helped maintain data consistency without significantly affecting the representativeness of the sample.

Outlier Detection: The dataset was reviewed for outliers using the Z-score test, but no significant extreme values were found. While outlier treatment is important for models sensitive to such values (e.g., linear regression), it is less critical for more robust methods like decision trees and ensemble algorithms.

B. Data Scaling

To prepare the dataset for machine learning models and maintain uniformity among various variable types, a range of data transformation methods was utilized. These preprocessing procedures were driven by the organization and meaning of the accessible data, which comprises a combination of categorical and numerical attributes, including ZONE_NAME, GENDER, AGE_RANGE, JOB, YEAR, MONTH, BAL, DUE, DELI, and LOAN_DUE_DATE.

1) Normalization of numerical features. Continuous variables, particularly balance (BAL), amount due (DUE), and delinquency (DELI), were standardized to align their scales and enhance computational efficiency. For example, the balance values, which frequently attained high magnitudes, were converted to thousands to facilitate interpretation and minimize computational demands during training. This normalization guaranteed that individual features would not unduly impact the learning process because of variations in unit scales.

2) Removing duplicates and noise. The dataset underwent a thorough examination to spot and remove duplicate entries, along with irrelevant or inconsistent data that might skew the analysis, which accounted for no more than 4 per cent of the total. This process was essential to uphold data integrity and avoid misleading trends from affecting model training. However infrequent, entries with unlikely balance or delinquency figures were assessed in relation to their corresponding loan due dates and financial behavior patterns prior to deciding whether to keep or discard them.

C. Data Splitting

In this study, the dataset was divided into two parts: a training set and a testing set, utilizing an unconventional split ratio. After trying out several partitioning methods, a 60% training and 40% testing split was chosen due to its superior performance in initial assessments. This selected ratio guaranteed that the training subset was adequate for effective model learning, while the somewhat larger test set facilitated a more thorough and reliable evaluation of the model's generalization abilities. This division was also reinforced by implementing K-fold cross-validation, which improved the robustness and credibility of the performance evaluation process.

VI. Feature Selection

Feature selection is a critical step in building effective machine learning models, particularly in domains such as bank loan risk assessment. Broadly, feature selection techniques can be categorized into three types: filter, wrapper, and embedded methods. Filter methods evaluate the statistical relevance of features independently of any learning algorithm. Examples include the correlation coefficient, which measures linear relationships between features and the target variable; the chi-square test, which assesses statistical independence in categorical data; and mutual information, which quantifies how much information a feature provides about the target. These methods are computationally efficient and help in quickly identifying potentially useful features [23, 24].

Wrapper methods, on the other hand, assess the performance of feature subsets by training models and selecting features based on predictive accuracy. Techniques like Recursive Feature Elimination (RFE) iteratively remove the least important features to find an optimal subset. Forward selection and backward elimination also follow similar principles but vary in their direction of feature inclusion or exclusion. Embedded methods integrate feature selection directly into model training. Notable examples include Lasso regression, which shrinks less relevant feature coefficients to zero through L1 regularization, and tree-based models like Decision Trees, Random Forests, and Gradient Boosting, which inherently perform feature selection by evaluating feature importance during the learning process [25]. These methods not only improve model performance but also enhance interpretability, a critical factor for institutions like banks.

When working with datasets that contain a mix of categorical and numerical features common in banking applications certain techniques stand out for their effectiveness. Both RFE and tree-based models are suitable, especially when categorical variables are appropriately encoded. However, tree-based models offer several advantages over RFE: they naturally handle mixed data types, capture non-linear relationships, are robust to multicollinearity, and provide computational efficiency [26]. Furthermore, tree-based models deliver built-in feature importance scores, simplifying the selection process, and advancements have been made to reduce biases in these scores. Their interpretability and versatility across classification and regression tasks make them particularly valuable in the context of loan risk modeling. Consequently, this research prioritizes tree-based feature selection methods to enhance both model performance and practical applicability. Accordingly, in this research, tree-based models were adopted.

In this section, we will highlight the most important features using tree-based methods: Decision Trees, Random Forests, and Gradient Boosting. These methods are applied to both dependent variables. The first one is to use classification in its original

format, and the other one is to categorize the dependent variable into two categories only. The first one is not default, and the second category is loan default. The following are the results of the three methods: Random Forest, decision tree, and gradient boosting.

A. Classification (Four Categories)

In this section, feature importance was evaluated using tree-based ensemble techniques—specifically, Decision Trees, Random Forest, and Gradient Boosting—applied to the original multi-class classification target variable. This classification represents the internal credit grading standards followed by the bank. The objective was to identify the most influential predictors affecting loan status as classified into four distinct risk categories as outlined in the preceding credit rating table.

1) Random forest feature importance. The Random Forest algorithm was employed to assess the input features based on their significance in contributing to classification accuracy. As shown in Table IX and depicted in Fig. 4, the findings reveal that loan balance (BAL) emerged as the most significant feature by a considerable margin, trailed by delinquency history (DELI) and amount due (DUE). In contrast, gender and job status showed little importance, indicating limited predictive capability in the realm of loan classification.

TABLE IX. BRIEF RESULT OF THE RANDOM FOREST FEATURE IMPORTANCE

Variable name	Importance	Variable name	Importance
ZONE_NAME	0.009437	GENDER	0.002049
AGE_RANGE	0.015941	JOB	0.000363
YEAR	0.030530	BAL	0.672653
DUE	0.130977	DELI	0.138049

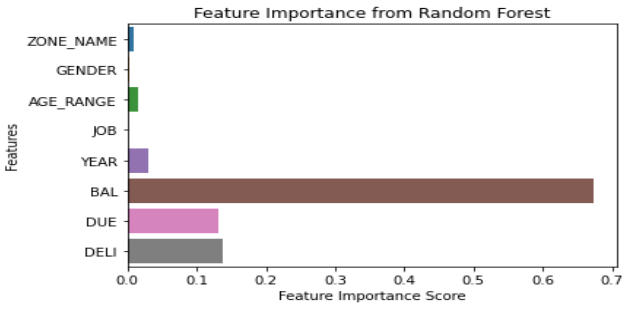


Fig. 4. Feature importance from Random Forest.

2) Decision tree feature importance. The Decision Tree was employed to compute feature importance scores. This method provides a straightforward interpretation by evaluating how each feature contributes to reducing impurity at decision nodes. As shown in Table X and visualized in Fig. 5, the most dominant variable was the loan balance (BAL), followed by delinquency history (DELI) and year of the record (YEAR).

3) Gradient boosting feature importance. From the following Table XI and Fig. 6, the least two important factors are gender and job.

TABLE X. BRIEF RESULT OF THE DECISION TREE FEATURE IMPORTANCE

Variable name	Importance	Variable name	Importance
ZONE_NAME	0.012394	GENDER	0.006294
AGE_RANGE	0.028994	JOB	0.000518
YEAR	0.048181	BAL	0.644940
DUE	0.042377	DELI	0.216300

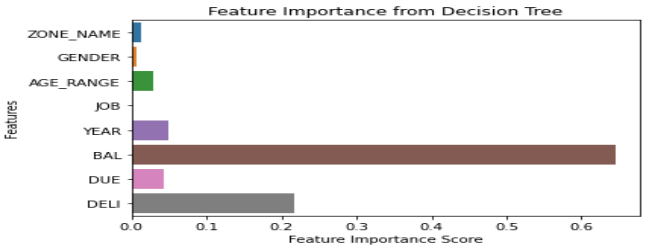


Fig. 5. Feature importance from Decision Tree.

TABLE XI. BRIEF RESULT OF THE GRADIENT BOOSTING FEATURE IMPORTANCE

Variable name	Importance	Variable name	Importance
ZONE_NAME	0.017195	GENDER	0.000568
AGE_RANGE	0.003375	JOB	0.000873
YEAR	0.151289	BAL	0.074997
DUE	0.461394	DELI	0.290309

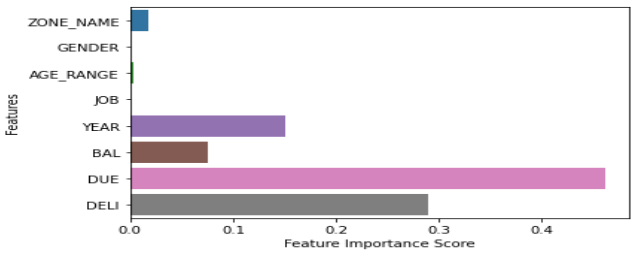


Fig. 6. Feature importance score from Gradient Boosting.

Three tree-based machine learning methods—Decision Tree, Random Forest, and Gradient Boosting (Table XI)—were utilized to evaluate the importance of various features in predicting loan defaults. The analysis revealed a consistent trend in the ranking of variables, although some differences in their significance were noted among the models. The most critical features identified across all models were Loan Balance (BAL), Amount Due (DUE), Delinquency (DELI), and Year (YEAR). Conversely, the least significant features in all three models were, distinctly, Gender (GENDER) Job status (JOB). These two variables consistently received the lowest importance ratings, reflecting their limited influence on default prediction. Consequently, gender and job status were removed from the final modeling stage to improve efficiency and decrease noise, without affecting the framework performance. Fig. 6 shows feature importance score from Gradient boosting.

B. Loan Default (Binary Classification)

During this stage, the target variable was transformed into a binary format (either defaulted or non-defaulted). The findings

reveal that most of the loan records within the sample are categorized as non-defaulted.

1) Random forest feature importance. As shown in the Table XII, the Random Forest model was used to identify the most influential features in predicting loan defaults. The analysis reveals that gender and job status had the lowest importance scores, suggesting they contributed minimally to the predictive performance of the model.

TABLE XII. BRIEF RESULT OF THE RANDOM FOREST FEATURE IMPORTANCE IN LOAN DEFAULT

Variable name	Importance	Variable name	Importance
ZONE_NAME	0.006080	GENDER	0.001318
AGE_RANGE	0.011007	JOB	0.000294
YEAR	0.033104	BAL	0.655241
DUE	0.168674	DELI	0.124281

2) Decision tree feature importance. In the analysis of binary classification utilizing the Decision Tree algorithm, the model assessed the significance of each variable in forecasting loan default results. As shown in Table XIII, gender and employment status displayed the least importance scores compared to all other features.

TABLE XIII. BRIEF RESULT OF THE DECISION TREE FEATURE IMPORTANCE IN LOAN DEFAULT

Variable name	Importance	Variable name	Importance
ZONE_NAME	0.011449	GENDER	0.004770
AGE_RANGE	0.025480	JOB	0.000544
YEAR	0.048742	BAL	0.621828
DUE	0.266804	DELI	0.020383

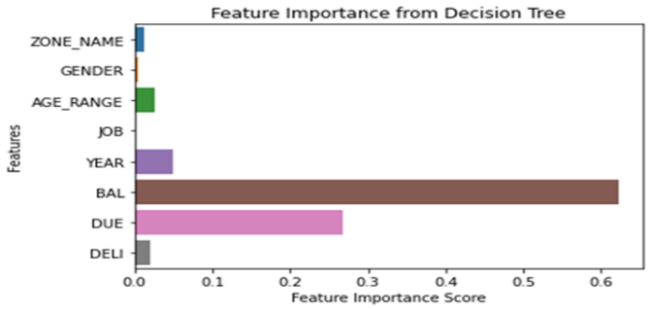


Fig. 7. Feature importance score from Decision Tree (loan default results).

In the analysis of binary classification utilizing the Decision Tree algorithm, the model assessed the significance of each variable in forecasting loan default results. As shown in Table XIV, gender and employment status displayed the least importance scores compared to all other features (see Fig. 7).

3) Gradient boosting feature importance. In the context of the binary classification task, the Gradient Boosting algorithm was utilized to assess the significance of each input feature in predicting loan default. The findings are presented in Table XIV.

It reveals that gender and job status possess the least importance scores, indicating a limited effect on the results of the model (Fig. 8).

TABLE XIV. BRIEF RESULT OF THE GRADIENT BOOSTING FEATURE IMPORTANCE IN LOAN DEFAULT

Variable Name	Importance	Variable Name	Importance
ZONE_NAME	0.004796	GENDER	0.000187
AGE_RANGE	0.001554	JOB	0.000184
YEAR	0.133705	BAL	0.042997
DUE	0.680054	DELI	0.136524

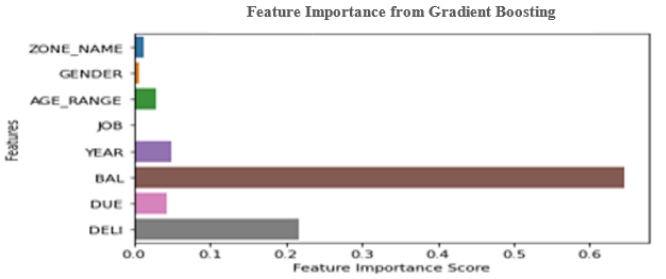


Fig. 8. Feature importance score from Gradient Boosting (loan default results).

VII. MODEL BUILDING

Following the selection of relevant features and the identification of key relationships among variables, suitable Machine learning algorithms are applied to develop predictive models. Techniques such as Random Forest, Decision Trees, and Gradient Boosting Machines are employed due to their proven effectiveness in handling complex, high-dimensional financial data. These algorithms are chosen for their ability to capture non-linear patterns and interactions within the dataset, thereby enhancing the accuracy and reliability of loan default predictions.

Before deciding, we have run more than one model: KNN, DT, RF, and Gradient Boost, and the following table shows that the best one is decision tree (Table XV).

The performance of the developed predictive model is assessed using a set of standard evaluation Metrics:

1) Accuracy. This metric measures the overall correctness of the model by calculating the proportion of correctly predicted instances out of the total number of cases.

2) Precision and recall. Precision reflects the proportion of correctly identified positive cases among all cases predicted as positive, while recall (also known as sensitivity) measures the model's ability to identify actual positive cases. Together, these metrics help evaluate the trade-off between false positives and false negatives [27].

3) Specificity and sensitivity. Sensitivity indicates the model's effectiveness in detecting loan defaults (positive cases), whereas specificity measures its ability to correctly identify non-defaults (negative cases). These metrics provide a comprehensive understanding of the model's capability to distinguish between default and non-default scenarios [27].

So, we are going to run a decision tree model for the second target variable that is categorized into two categories. A decision tree is a popular machine learning algorithm used for classification and regression tasks. It is a flowchart-like structure where each internal node represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents a class label or a continuous value.

TABLE XV. THE EVALUATION METRICS OF VARIOUS ML MODELS

Model	F1-score	Accuracy	Recall	Precision	ROC curve
DT	66%	88%	53%	16%	80%
KNN	61%	82%	49%	11%	68%
RF	63%	84%	51%	12%	74%
Gradient boost	53%	70%	42%	10%	72%

A decision tree is a structured model used for both classification and regression tasks, consisting of several key components. The root node sits at the top and represents the entire dataset, initiating the first decision split. Internal nodes follow, each representing a decision based on a particular feature. Branches connect the nodes and illustrate the outcomes of those decisions, guiding the data down various paths. Finally, leaf nodes represent the terminal points, where a specific outcome or class label is assigned. This hierarchical structure enables the decision tree to map inputs to outputs through a series of straightforward rules.

One of the main advantages of decision trees is their interpretability. They are easy to understand and visualize, making them accessible to non-technical stakeholders. Additionally, they work well with both numerical and categorical data and are capable of modeling non-linear relationships. Unlike many other algorithms, decision trees do not require data normalization or scaling. They also inherently provide feature importance scores, offering insights into which variables most significantly impact predictions. However, a potential drawback is their tendency to overfit the training data, especially in the presence of noise, which can reduce their generalization performance. Despite this, their simplicity and explanatory power make decision trees a valuable tool in many machine learning applications.

The selected algorithms are implemented in Python. The model is trained on the cleaned dataset, with optimization techniques such as cross-validation applied to improve accuracy and prevent overfitting.

VIII. MODEL EVALUATION

The performance of the developed predictive model is assessed using a set of standard evaluation metrics:

Accuracy: This metric measures the overall correctness of the model by calculating the proportion of correctly predicted instances out of the total number of cases [28] [29].

Confusion Matrix: This provides a detailed view of the model's classification performance by showing the counts of true positives, true negatives, false positives, and false negatives [30]. Fig. 9 shows the confusion matrix of the model.

The overall accuracy is 88%, which means that the model could predict 88% of the non-fault and Fault loans correctly predicted, and this is a very high percentage. The sensitivity of the model 53%, which means that 53% of the Fault loans correctly specified. While specificity is 89%, which means that 89% of the non-Fault loans correctly specified.

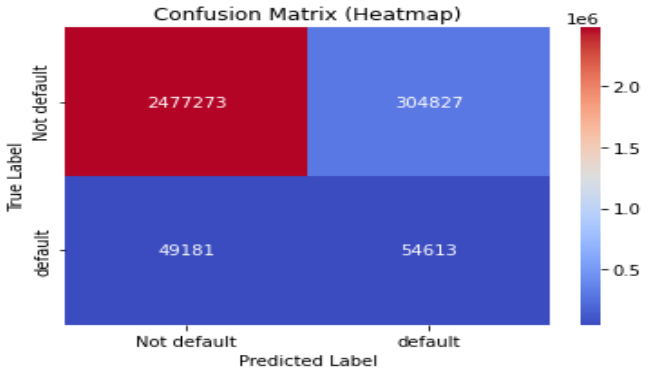


Fig. 9. Confusion matrix for model.

Another measure for the Goodness of fit is the ROC curve. The ROC Curve (Receiver Operating Characteristic Curve) is a graphical representation used to evaluate the performance of a binary classification model. It shows the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate) [29] at different classification thresholds. From Table XV, it is clear that the area under the curve is 80%, which indicates that the model is a good fit [30]. Fig. 10 shows the ROC curve.

$$TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$FPR = \frac{False\ Positives}{False\ Positives + TrueNegatives}$$

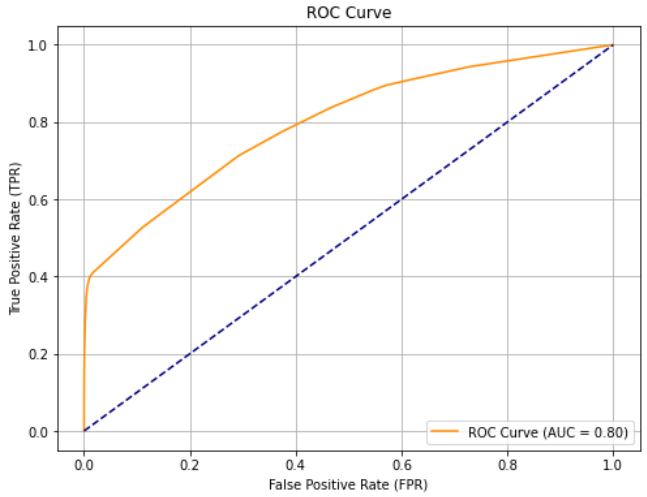


Fig. 10. ROC Curve.

A. Visualization and Analysis

In the final stage, model results are visualized to enhance interpretability and support effective communication with stakeholders. These visual representations enhance transparency, aid in validating the model's behavior, and ensure

that findings can be clearly communicated to both technical and non-technical audiences. Fig. 11 illustrates the structure of a decision tree.

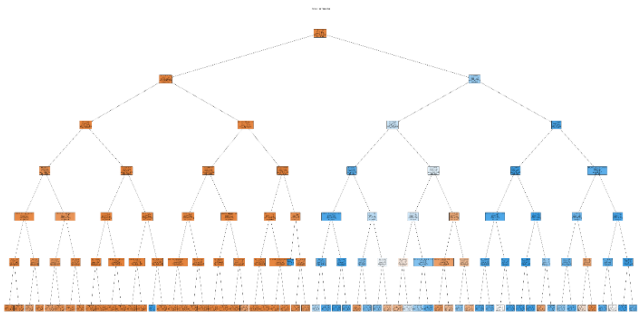


Fig. 11. Illustrates the structure of a decision tree.

The root node analysis of the decision tree highlights how the feature BAL (balance) serves as the most critical determinant in the classification of loan default risk. The initial split occurs at BAL ≤ 2140.0, making this threshold the most influential decision point. Lower values typically indicate a "Not Fault" classification, while higher balances often correspond with an increased likelihood of "Fault." This foundational split sets the tone for the downstream branches, guiding how further conditions refine the classification outcome.

- Path 1 reveals a scenario of high confidence for "Not Fault" classification. When BAL is further reduced to ≤ 2097.6 and the borrower's zone is not "East Delta," the Gini index is extremely low, signaling strong consensus in the data. This indicates that such a combination of low BAL and non-"East Delta" region consistently leads to non-default outcomes. The model is highly reliable in this path, showing minimal misclassification and suggesting that borrowers in this group present low risk.
- Path 2 follows a similar route but introduces moderate uncertainty. Here, while BAL remains ≤ 2097.6, the zone is identified as "East Delta." This regional factor slightly increases the Gini index, indicating some ambiguity. Although the majority of predictions still point to "Not Fault," the presence of the "East Delta" zone introduces variability, suggesting that geographical factors may contribute to nuanced risk assessments, especially in marginal financial conditions.
- In Path 3, the model addresses a more complex situation leading to a "Fault" classification. This occurs when BAL exceeds 2140.0, DUE is also high (above 0.68), but DELI (delinquency) remains low (≤ 0.15). Despite the low DELI, the combination of high BAL and DUE tips the classification toward "Fault." The Gini value here is moderate, suggesting a less decisive split. It reflects a group where most instances are defaults, but a few are not, highlighting the complexity of high-risk profiles, where even a favorable feature like low DELI cannot fully offset the effect of high balances and dues.
- Path 4 introduces a highly uncertain classification scenario. When BAL is within a borderline range (≤ 2140.0 but > 2097.6), DELI is above 0.05, and DUE varies, the model shows significant fluctuation in its output. With moderate DUE (≤ 11364.07), the classification remains "Not Fault," but once DUE exceeds this threshold, the likelihood of a "Fault" classification rises and the Gini index drops, indicating greater certainty. This path illustrates the dominance of the DUE variable in edge cases and how it can override other indicators in determining default risk.
- Finally, Path 5 emphasizes the strong influence of DELI in reducing risk. When BAL is low and the borrower is not in the "East Delta" zone, a DELI ≤ 0.05 leads to a highly confident "Not Fault" prediction, despite other variables suggesting moderate risk. The Gini index is low in this scenario, showing that low delinquency plays a powerful protective role. This finding underscores the importance of monitoring delinquency closely, as it can significantly alter the risk profile even when other conditions are less favorable.

IX. Results

The predictive model showcases a thoughtful and structured decision-making process shaped by the relative importance of selected customer attributes. Rather than relying on a single indicator, the model navigates through combinations of factors to reach decisions that are both data-driven and sensitive to nuanced behavioral patterns.

A. Balance (BAL): Entry Point for Risk Assessment

At the initial stage, the customer's balance acts as a key determinant: When the balance is less than or equal to 2140.0, the likelihood of loan default is low, and the model typically moves toward a "no fault" decision.

Balances exceeding this threshold trigger additional evaluation steps, as they may signal heightened financial pressure or exposure, leading to a higher probability of default.

This distinction reflects the model's cautious approach: it does not generalize but instead digs deeper when early indicators suggest potential risk.

B. Geographic Factor: ZONE_NAME

Regional context also plays an essential role: customers outside the "East Delta" region generally present lower risk, and the model tends to assign them to the "non-fault" category.

However, being from "East Delta" introduces a degree of uncertainty. This does not imply automatic risk but encourages the model to be more attentive, possibly reflecting local economic or demographic conditions.

Such consideration reflects how risk may be influenced, not just by individuals but also by broader regional dynamics.

C. Delinquency Score (DELI): Strong Behavioral Signal

This variable carries significant weight in the model's decisions: scores at or below 0.05 offer strong reassurance, often guiding the model to a "not fault" outcome, even in the presence of moderate financial concerns.

In contrast, scores above 0.15 often signal concern, sometimes shifting the model's outcome toward a "fault" decision—even if other variables seem relatively stable.

This behavior highlights the model's ability to recognize patterns in payment discipline as a reliable reflection of future risk.

D. DUE (Outstanding Amount): Key in Borderline Decisions

In more complex or uncertain cases, the outstanding amount becomes a tipping factor: values below 11,364.07 often help stabilize decisions in favor of a non-default classification.

Higher amounts, however, tend to increase the model's confidence that default risk may be present.

Here, the model demonstrates sensitivity to financial load, adjusting its judgment when individuals may be reaching unsustainable repayment levels.

E. Gini Index: Evaluating Predictive Fairness

The model's ability to separate risky from safe cases is validated through the Gini coefficient, a well-regarded metric in credit scoring. A higher Gini score indicates that the model can reliably distinguish between borrowers who are likely to default and those who are not.

This ensures that the decisions made are not only technically sound but also aligned with ethical considerations of fairness and accountability.

X. Discussion

This study offers a unique and novel contribution by concentrating on a very specific and little-studied domain: livestock-based loans within the Agricultural Bank of Egypt. This is in contrast to previous studies that have extensively examined the use of machine learning algorithms to predict loan approval and default risks, with a primary focus on general loan products or credit scoring mechanisms in various banking sectors. Just as Puli (2024), Saini (2023), which employed algorithms like Random Forest, SVM, and Neural Networks on public or generalized datasets, the present research leverages actual, real-world customer data from one of Egypt's largest rural-focused banking institutions, comprising 1,190 branches nationwide. The dataset reflects true client behaviors in repaying livestock-oriented loans—an area that remains largely unexplored in prior academic research.

XI. Conclusion

Machine learning has become an increasingly valuable tool in assessing loan risk within the banking sector. By leveraging these techniques, it is possible to identify patterns associated with loan default and to determine the most relevant predictive features. The present study emphasizes the role of appropriate feature selection in improving both model accuracy and interpretability. Approaches such as filter methods, wrapper strategies, and embedded techniques were considered to identify key variables.

Among the models implemented, algorithms such as Random Forest, Decision Tree, and Gradient Boosting demonstrated favorable performance, with Random Forest achieving an accuracy of approximately 88%. These models also offer a reasonable balance between predictive power and interpretability, making them suitable for practical use.

The feature importance analysis revealed that financial indicators had the greatest influence on prediction outcomes. Notably:

Balance (BAL) showed a strong association with default risk.

Do Amounts (DUE) were identified as critical in flagging potential non-repayment.

Delinquency Scores (DELI) provide insight into historical repayment behavior.

On the other hand, demographic attributes, such as gender and occupation, were found to have limited impact on model performance, suggesting that behavioral and financial factors may play a more central role in loan risk evaluation.

In conclusion, machine learning, when applied with thoughtful feature selection, can contribute meaningfully to improving credit risk assessment. While no single model can guarantee perfect accuracy, decision trees and ensemble methods have shown promise in balancing accuracy, transparency, and real-world applicability.

XII. Future Work

While the current study demonstrates the potential of traditional machine learning models such as Decision Tree and Random Forest in predicting loan defaults, future research can benefit from exploring more advanced algorithms and broader datasets. Incorporating modern ensemble and deep learning techniques, such as Extreme Gradient Boosting (XGBoost) Artificial Neural Networks (ANNs), may offer improved predictive accuracy, especially in handling complex, nonlinear relationships within financial data.

In addition, expanding the dataset to include behavioral patterns, real-time transactional data, and external credit factors may also lead to deeper insights and more generalized models.

Acknowledgment

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through the General Research Project under grant number GRP/6/46.

References

[1] S. Mestiri and S. Mestiri, "Credit scoring using machine learning and deep Learning-Based models," Data Science in Finance and Economics 2024 2:236, vol. 4, no. 2, pp. 236–248, 2024, doi: 10.3934/DSFE.2024009.

[2] M. Ranjan, K. Barot, V. Khairnar, V. Rawal, A. Pimpalgaonkar, S. Saxena and A. Sattar, "Python: Empowering Data Science Applications and Research," Journal of Operating Systems Development & Trends, 2023, Volume 10, Issue 1, 2023,pp27-33

[3] J. Quan and X. Sun, "Credit risk assessment using the factorization machine model with feature interactions," Humanities and Social Sciences Communications 2024 11:1, vol. 11, no. 1, pp. 1–10, Feb. 2024, doi: 10.1057/s41599-024-02700-7.

[4] N. Gulo, E. Nurninawati, R. A. R. S, and D. P. Kristiadi, "Decision Support System for Submitting Credit using Analytical Hierarchy Process (AHP) Method Based on Android on Save and Loan Cooperatives Cubg Pasar Kemis Tangerang," IJISTECH (International Journal of

Information System and Technology), vol. 6, no. 4, pp. 441–448, Dec. 2022, Accessed: Feb. 20, 2025. [Online]. Available: https://ijistech.org/ijistech/index.php/ijistech/article/view/259

[5] Tran, Phuong ,and , Nga Phan " Digital Transformation of the Banking Industry in Developing Countries " International Journal of Professional Business Review (2023), Vol. 8 , 5 pp. 8-0.

[6] Vahid sinap "AComparative Study of Loan Approval Predicayion Using Machine Learning Methods",Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Ankara, Türkiye,Apirl 2024,pp644-663

[7] P. S. Saini, A. Bhatnagar, and L. Rani, "Loan approval prediction using machine learning: A comparative analysis of classification algorithms," in 2023 3rd Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE), May 2 023, pp. 1821-1826.

[8] Jovanne C. Alejandrino1 , Jovito Jr. P. Bolacoy1 , John Vianne B. Murcia, " Supervised and unsupervised data mining approaches in loan default prediction," International Journal of Electrical and Computer Engineering (IJECE) , April 2023, Vol. 13, pp. 1837-1847

[9] M. Alaradi and S. Hilal, "Tree-based methods for loan approval," in 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), Oct. 2020, pp. 1-6.

[10] Güder, G., & Köse, U. (2024). Prediction of home loan approval with machine learning. *Advances in Artificial Intelligence Research (AAIR), 4*(2), pp.87–95.

[11] Sampurna , Sahil and Vidya " Predicting Credit Risk in European P2P Lending: A Case Study of "Bondora" Using Supervised Machine Learning Techniques" December 2023, 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2023, pp. 1-6.

[12] Tabassum, Namita, Vaishnavi and Prachi ,Aditya and" A Machine Learning Based Framework For Bankruptcy Prediction In Corporate Finances Using Explainable AI Techniques" Vol 49, No 15 (2025) ,pp 15–26.

[13] Hussain, M. Z., Ejaz, S., Batool, E., Hasan, M. Z., Mustafa, M., Khalid, A., Hussain, U., Khan, Z., Javaid, A., Ashraf, M. F., Awan, R., & Yaqub, M. A" Bank Loan Prediction System Using Machine Learning Models" 2024 IEEE 9th International Conference for Convergence in Technology No I2 ,pp. 1-7

[14] chen, Y. (2023). Financial Statement Fraud Detection based on Integrated Feature Selection and Imbalance Learning. Frontiers in Business, Economics and Management,VOL 8(3), pp46-48.

[15] Ali, A. A., Khedr, A. M., El-Bannany, M., & Kanakkayil, S. (2023). A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique. Applied Sciences, VOL 13(4), 2272

[16] A. Gupta, V. Pant, S. Kumar, and P. K. Bansal, "Bank Loan Prediction System using Machine Learning," in 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), Dec. 2020, pp. 423-426

[17] .K. Vengatesan , A. Kumar and S. Sabni,. Credit Card Fraud Detection Using Data Analytic Techniques. Advances in Mathematics: Scientific Journal 9, (2020)Vol3, pp1185–1196.

[18] Vahid SİNAP."A Comparative Study of Loan Approval Prediction Using Machine Learning Methods" Gazi University Journal of Science, Part C: Design and Technology, (2024),VOL 12(2), P.P644– 663.

[19] Agata Kozinaa , Łukasz Kuźmińskia , Michał Nadolnya , Karolina Miałkowskaa , Piotr Tutaka , Jakub Janusa , Filip Płotnickia , Ewa Walaszczy (2023)"The default of leasing contracts prediction using machine learning", 27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systemsp,p424–433

[20] Muhamad, Muhamad JUMA and Arif (2023) Improving Credit Risk Assessment through Deep Learning-based Consumer Loan Default Prediction Model, INTERNATIONAL JOURNAL OF FINANCE & BANKING STUDIES VOL 12,p.p85-92.

[21] Viswanatha, V. (2023)." Prediction of loan approval in banks using machine learning approach", International, ournal of Engineering and Management Research, Vol 13(4), P.P7–19

[22] Btoush, W., Alsharari, N. M., & Al-Dhamari, R. A. (2025). A hybrid deep learning and machine learning stacking model for credit card fraud detection. *Applied Sciences, 15*(6), 1–20. https://doi.org/10.3390/app15063320

[23] Hiba, A. (2024). A review of feature selection methods in big data. *International Journal of Intelligent Systems and Applications in Engineering*, VOL*12*(4), P.P4347–4366.

[24] Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data, 11*(44),P.P1-16.

[25] Pödör, Z., & Hekfusz, M. (2024). Comparing feature selection methods on metagenomic data using r andom forest classifier. *Transactions on Engineering and Computing Sciences, 12*(1), P.P175–187

[26] S. Rajora, D. Li, C. Jha, N. Bharill, O. Patel, S. Joshi, and M. Prasad(2018), "A comparative study of machine learning techniques for credit card fraud detection based on time variance," In 2018 IEEE symposium series on computational intelligence (SSCI), IEEE, pp. 1958-1963.

[27] Jasmina, Goran and Drago (2020)New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers, Advanced Engineering Informatics, Vol 45, p.p

[28] Silvia ,Federico and Emanuele Giovannini (2017),Solvency prediction for small and medium enterprises in banking, Decision Support Systems, Vo 102, P.P 91-97

[29] Cabot, J. H., & Ross, E. G. (2023). Evaluating Prediction Model Performance. *Surgery*, Elsevier VOL174,NO(3), P.P 723–726.

[30] Tariq Saeed (2020)The Application of Data Mining Techniques for Financial Risk Management: A classification framework, IJCSNS International Journal of Computer Science and Network Security, VOL.20 No.8,P.P84-93.

AUTHORS' PROFILE

Mona Sharafeldin received her master degree of computer science from the Sadat Academy for Management Sciences in 2014. She is currently PhD student in Business Information System, Helwan University.

Prof. Amira M. Idrees: I'm currently a professor in College of Business, King Khalid University. I have been the head of scientific departments and the vice dean of the community services and environmental development, Faculty of Computers and Information, Fayoum University. I have also been a professor in the Faculty of Computers and Information Technology in Future University, the head of IS department and the head of University Requirements Unit. My research interests include Knowledge Discovery, Text Mining, Opinion Mining, Cloud Computing, E-Learning, Software Engineering, Data Science, and Data warehousing.

Assoc.Prof/Shimaa Ouf: I am currently an Associate Professor at the Information Systems Department, Faculty of Commerce and Business Administration, Helwan University. I was born in Cairo, Egypt. She received an M.Sc. degree in Information Systems from the Faculty of Computers and Information, Helwan University, Egypt. Ph.D. degree in Information Systems from the Faculty of Computers and Information, Helwan University, Egypt. I have published many Scientific articles in International journals and conferences in the area of E-learning Ecosystems, Web 2.0 Technologies, Cloud Computing, Intelligent learning environments, Personalized E-learning Ecosystem using Ontology and Semantic Web Rule Language, Business Intelligence, Big Data, Artificial Intelligence, and Blockchain. I have deep experience in creating intelligent systems using the semantic web and Artificial Intelligence technologies. I am an active reviewer for numerous international journals.

Advanced AI-Driven Safety Compliance Monitoring in Dynamic Construction Environment

Aisha Hassan[1], Ali H. Hassan, Ph.D[2], Yasmin Christensen[3], Hussain Alsadiq Ph.D[4]
Undergraduate Student, Department of Architectural Engineering, University of Prince Mugrin, Medinah, Alaqool, Saudi Arabia[1]
Assistant Professor, Department of Computer Science, University of Prince Mugrin, Medinah, Alaqool, Saudi Arabia[2]
Independent Researcher, Medinah, Saudi Arabia[3]
Assistant Professor, Department of Mechanical Engineering, University of Prince Mugrin, Medinah, Alaqool, Saudi Arabia[4]

***Abstract*—Construction safety is a critical global concern due to the high-risk environment faced by workers, with accidents often leading to serious injuries and fatalities. To enhance construction management, this study proposes a scalable deep-learning model for real-time compliance monitoring of safety regulations. The research gap addressed is the lack of real-time, scalable AI solutions for safety compliance monitoring in dynamic construction environments. The YOLOv11n model was trained and evaluated to identify and track the use of safety helmets and vests in extreme dynamic environments, ensuring timely detection of non-compliance. It is hypothesized that the YOLOv11n model will outperform baseline models in accuracy and real-time monitoring speed. The YOLOv11n model outperformed other baseline models, with precision, recall, and mean average precision scores of 89.5%, 85%, and 91.6%, respectively, and a real-time processing speed of 71.68 FPS. Its lightweight size and performance make it suitable for deployment. Integrated with a person-detection framework, the system provides real-time desktop alerts for safety violations, enhancing safety compliance. These findings contribute to construction automation by advancing scalable AI-driven solutions for proactive safety compliance, reducing accidents, and improving operational efficiency on construction sites.**

Keywords*—YOLOv11n; personal protection equipment (PPE); construction safety; real-time object detection; deep learning; AI-driving compliance systems

I. Introduction

The construction industry is one of the most hazardous sectors globally, accounting for approximately 20% of workplace fatalities annually, according to [1]. Among these incidents, 38.4% result from trips, slips, and falls, highlighting the critical need for improved safety measures. Other more critical cases include being struck by objects and electrocutions, often due to a lack of compliance with personal protective equipment (PPE) protocols [2]**.** To address these risks, measures such as personal protective equipment (PPE) -including helmets, vests, gloves, and boots- are implemented to ensure workers' safety and reduce casualties. However, despite widespread adoption of safety measures, challenges remain in monitoring workers' compliance with PPE on construction sites due to several factors. First, the dynamic and complex conditions encountered in real-world construction sites with constantly changing weather and conditions can affect visual monitoring systems [3]. Second, manual monitoring is costly and difficult to scale for large-scale construction projects [4]. Third is environmental factors, as extreme dust, weather environments, and fluctuating lighting conditions may hinder accurate PPE detection. Moreover, studies [5] and [6] have highlighted that standard object detection models lack an accurate ability to detect small or overlapping objects and cluttered scenes, all of which are common on construction sites.

Recent advancements in artificial intelligence (AI) and deep learning have great potential for improving safety monitoring in construction sites. Similarly, in high-stakes domains like healthcare, the need for explainable AI models has been highlighted to encourage trust and aid in decision making [7]. Reinforcing the advances of applying AI models to dynamic construction settings. These technologies enable real-time data processing, object categorization, and predictive decision-making, allowing for the proactive identification of safety violations and potential risks [8]. It has been demonstrated that methods that combine deep learning with preprocessing techniques can improve recognition accuracy in areas like emotion recognition [9], [10]. Among these technologies are the You Only Look Once (YOLO) family of models, which are recognized for being fast and efficient deep learning algorithms for real-time object detection, identifying and classifying objects in a single step by predicting bounding boxes and class probabilities across an image [11].

Several studies have utilized YOLO in real-time detection of personal protection equipment (PPE) such as helmets, vests, and other safety gear, on construction sites. Studies by [6], [12], [13] emphasize the use of advanced YOLO versions, including YOLOv5, YOLOv8, and modified iterations, for real-time detection of PPE violations. The latter works highlight the models' strength in accurately identifying safety helmets, vests, and other protective equipment. In [6], the researchers focused on examining the feasibility of the addition of computer vision technology in construction sites through consistent training processes of YOLO models, to improve overall prediction accuracy. Likewise, the study conducted by [13] proposes an advanced PPE detection system that ensures safety compliance and enhanced productivity in high-risk environments.

In [12], the authors additionally suggests an improved YOLOv5 algorithm based on a Reverse-Unet framework (RUYOLO) to enhance small object detection and accuracy performance. Similarly, studies [14] and [15] have a shared emphasis on integrating advanced technologies (e.g. UAVs (Unmanned Aerial Vehicles) and IoTs (Internet of Things)) with deep learning algorithms to enhance safety monitoring devices. Both studies applied object detection models, such as YOLOv7

and MR-Net (which is a combination of MobileNet and Deep Residual Network components), to identify safety equipment gears. Bian et al. focus on the detection of safety helmets in specifically challenging areas such as construction power line energy fields. They proposed a YOLOv7 model integrated with an extended efficient layer aggregation network (E-ELAN), achieving high results in detection accuracy.

Despite significant advancements in object detection technologies, research on the latest YOLOv11 model for PPE detection remains limited. Most of the existing studies focus primarily on static environments and curated datasets, overlooking the dynamic and complex conditions encountered on real-world construction sites, such as low visibility and cluttered backgrounds. Furthermore, few studies compare YOLO versions under consistent training conditions, making model growth and trade-offs difficult to measure.

This study aims to bridge the gap by proposing and evaluating an optimized YOLOv11n, specifically focusing on helmet detection and PPE compliance to safety standards in construction work sites in dynamic construction settings. The model is trained using a domain-specific dataset augmented to improve detection accuracy under dynamic conditions, such as extreme weather conditions and background clutter. The performance is benchmarked against previous YOLO models, YOLOv3, YOLOv5, and YOLOv8, highlighting significant improvements in detection accuracy, inference speed, and deployment efficiency. Overall, this research aims to demonstrate how AI can be used to minimize safety violations and improve safety compliance in construction sites, with the specific use of improved deep learning models.

The key contributions of this study are as follows:

- Present a comprehensive analysis of multiple YOLO models to examine their potential in detecting PPE compliance under dynamic construction site conditions.
- Implementation of an optimized YOLOv11n model to achieve high object detection accuracy and real-time processing efficiency while reducing computational load and model size, making it suitable for real-time deployment.
- Integrated dual-model framework by merging two YOLO models to detect PPE compliance in construction sites, with real-time notification code that triggers desktop alerts when failure to comply with PPE standards is detected, demonstrating its practical use in live monitoring scenarios.

The rest of the study is structured as follows:

Section II explores previous studies of safety monitoring models. Section III details the training and optimization processes of the YOLOv11n model, including the data collection, preprocessing, model training, system integration, and performance evaluation. Section IV presents the results of the study. While Section V includes a discussion of the results highlighting the overall efficiency in real-time object detection. Section VI concludes the study, outlines the challenges of the study and suggests future research.

II. Literature Review

With the ever-growing construction industry, safety in construction is a great challenge faced worldwide [2]. For this reason, object detection for PPE equipment has become a priority to prevent accidents caused by safety violations [4]. Several studies have used deep learning models to combat this challenge, with the most popular choice being YOLO-based models, because of their ability for real-time detection and high effectiveness. A study made in the United Arab Emirates (UAE) developed a Convolutional Neural Network (CNN) model by using a model algorithm based on YOLOv3 for safety monitoring [16]. Their goal was to show that AI can be used to minimize accidents on construction sites by detecting workers' safety helmets, harnesses and lifelines. They were able to achieve a 94% accuracy rate in object detection, indicating the high success rate of the YOLO models.

This literature review investigates the effectiveness of deep learning models, particularly YOLO, in monitoring PPE compliance on construction sites. It compares different YOLO versions to identify which model offers the optimal safety compliance capabilities. The main research questions guiding this review are: How can AI deep learning models be used to enhance PPE compliance monitoring in construction settings, and what are the strengths and limitations of each of the YOLO models? Additionally, this review focuses on the gap by comparing these models' performances in detecting PPE violations.

A. YOLO Based Detection for Safety Compliance

In recent years, deep learning models, specifically YOLO models, have emerged as a popular choice for object detection because of their high effectiveness in the real-time detection of PPE equipment regarding speed and accuracy. Because of this, multiple studies have used this model, showing significant improvement in identifying PPE gear.

YOLO operates by dividing images into an S × S grid, where each grid cell detects objects within its boundaries as shown in Fig. 1. For each cell, the model predicts bounding boxes to determine the object's location and a confidence score indicating the likelihood of the object's presence [16]. These predictions include the object's coordinates, such as center, width, and height and its class probabilities, allowing YOLO to identify and localize multiple objects simultaneously [17]. This information is processed in a single forward pass which makes it efficient for real-time detection. Non-Maximum Suppression (NMS) is then applied to eliminate the overlapping prediction to ensure that most accurate bounding box is retained for each object.



Fig. 1. YOLO object detection workflow.

Since its first release in 2016, the YOLO models have undergone significant enhancements and improvements. YOLO was established as a single-stage detector, identifying and classifying objects through a single network. They are smaller in size and faster, which makes them faster to train and deploy [17]. With the development of YOLO versions such as YOLOv3, YOLOv5, and YOLOv8, the models are enhanced, and performance is faster and more accurate. The latest release of the YOLO model, YOLOv11, in September 2024, introduces advanced features further optimizing the model for object detection accuracy and precision.

B. Related Works on PPE Detection

The studies examined highlight the significance of YOLO-based object detection models in enhancing real-time PPE compliance monitoring systems. Notably, versions such as YOLOv5 and YOLOv8 have demonstrated high accuracy and speed, even in complex environments. As seen in a study conducted by [6], 4844 images were gathered by the authors and trained on both YOLO versions, YOLOv5 and YOLOv8, with the same dataset. The results showed that the validation accuracies achieved by both models were 94.7% and 95.4%, respectively, making the YOLOv8 model the superior of the two. Similarly, in a study by [18], the YOLOv8 was optimized to improve the overall accuracy and effectiveness in small object detection. The model used mosaic data augmentation, including scaling images and cropping, which improved its object detection performance.

The YOLOv3 and RCNN models to identify the highest performance in detecting safety helmets in a construction setting were compared in [19]. An overall of 5000 images were gathered and then processed using two techniques, resizing and data augmentation. The datasets were then split into a ratio of 8:2 based on training and testing, respectively. The study utilized mAP, precision and recall, comparing the two models. After training, the YOLOv3 and RCNN models achieved 97.12% and 96.05% mAPs, respectively. Subsequently, the YOLOv3 model acquired a precision of 96.53% and a recall of 97.87%, while the RCNN model showed a precision of 94.1% and a recall of 96.75%. Despite challenges the models faced, such as detecting small objects and image lighting effects, the YOLOv3 model outperformed the RCNN with greater precision in recognizing safety helmets. Researchers in [20] focused on detecting safety gear using the YOLOv3 model. In their study, they stressed the significance of not only focusing on safety helmets but also on other safety gear, such as safety vests. 300 images were gathered and divided into two classes: compliant and non-compliant workers. 80% of the image data was set for training, while the remaining was used for data verification. The data was then analyzed and labelled by using LabelImg, which is used to annotate images with a bounding box. The YOLOv3 model was then trained using the data sets, achieving 92.99% mAP during the second iteration of the study. The model was also tested on live capture and video data, likewise, presenting high accuracy results. The authors concluded that even with a limited amount of data, the model received high precision in detecting safety gear. They also suggest further development in detecting other safety equipment through additional recent versions, expanding datasets and training models.

The use of a Mobile Residual Network (MR-Net) for object detection in safety monitoring was explored in [15]. They proposed three consecutive systems for enhanced object detection, the first being a sensor unit connected to an IoT network. A cloud server was then deployed to issue a smart alert, followed by the processing of gathered image data for detection using an NLM filter. YOLOv3 was then used to detect objects and categorize them based on four core classes (person, machines, vehicle, and safety cones) and further classify objects using MR-Net. The study then compared previous models based on accuracy, precision, recognition, positive outcomes, and negative outcomes. The results achieved optimal values of 91.2%, 90.5%, 92.3%, 90.9%, and 91.9%, respectively.

Detecting safety helmets in construction settings is difficult due to two main challenges: the small size of helmets at a further distance and the sheer size of the background image in contrast to the helmet's proportion. For this reason, authors in [3] analyzed video data to improve helmet detection and prevent false detections from occurring. The test datasets were collected from construction sites and utilized with multi-scale feature detection and training to enhance the YOLOv3 model. The original YOLOv3 model was then compared with the new optimized model, resulting in 93.5% mAP for the optimized model and 91.3% mAP for the YOLOv3 model. The enhanced model shows greater accuracy in detecting safety helmet violations than the original YOLO model by 2.2%.

To address the challenges in detection models resulting from small object detection, high rate of overlapping, and frequent occlusions, the study in [21] introduced an improved safety helmet detection model based on YOLOv5, integrating EIOU-loss to enhance model improvement. 4,400 images were gathered, ranging from different environments and distances, then labelled into three categories: helmet, head, and uncertainty. The improved models, retitled as YOLO-ESC and YOLO-ESCA, were compared with YOLOv3 and standard versions of YOLOv5. While the previous YOLO versions achieved greater precision and recall than the improved models, the model sizes of the YOLO-ESC and YOLO-ESCA are smaller, making them better candidates for deployment. Despite the improved enhancements, the study recognizes the limitations of the model in small object detection, making it a goal for further research development.

The performance of advanced YOLOv5 model variants (n, s, m, l, x) for safety helmet detection has been examined in [22]. The study conducted a comparative analysis based on key performance indicators, including mean Average Precision (mAP), precision, and recall, across two target classes: “head” and “helmet”. The findings indicated that YOLOv5n demonstrated superior speed, achieving 62.5 FPS for image data and 70.4 FPS for video data. However, in terms of detection accuracy, the YOLOvX variant achieved the highest mAP score of 95.8%, making it the most precise among the evaluated models. Similarly, in [5], the authors proposed an improved lightweight YOLOv5s model to enhance safety helmet detection through image augmentation techniques, aiming to optimize model robustness in dynamic environments. A comparative analysis was conducted between the improved YOLOv5s model and other advanced models such as YOLOv3, YOLOv4, YOLOv7 and YOLOv8. The improved model attained a mAP

of 94.1%, exceeding that of the YOLOv7-Tiny and similar to that of the YOLOv8s model. The results indicate a notable improvement in the model while maintaining its lightweight characteristics. The study concludes with future research recommendations for further exploring systems, to diminish parametric defects and improve accurate detection of objects.

The lack of focus on other PPE elements such as safety vests, gloves, boots and glasses has been highlighted in previous research. To address this, a safety monitoring model was proposed that compares YOLOv5 and YOLOv7 [23]. A dataset of a total of 743 images was collected and categorized into 6 classes: boots, glasses, gloves, helmets, person, and safety vest. The data was then annotated using LabelImg and directed for training using both models individually. The metrics used for evaluation are mAP@0.5 of precision and recall. The final values of the models showed that the superior system in precision and recovery of image detection was YOLOv7, with a value of 87.3%. The value of the YOLOv5 was marginally lower at 79.6%, indicating a slight insufficiency of the process metric.

Due to the shortcomings of traditional safety compliance monitoring methods in construction sites, the need for integrating object detection systems has become increasingly important. Recent research has demonstrated that improved object detection algorithms, specifically YOLOv8, are efficient in detecting PPE items [24]. In this study, a total of 3,750 images were annotated and split into five classes: Vests, Safety shoes, No vests, No helmets, and No safety shoes. The dataset was then trained with YOLOv8 using a maximum of 100 epochs, where performance is evaluated periodically. The model achieved significant results in accuracy, scoring 100% and 98% in the detection of helmets and vests, respectively. The model faced a slight regression in detecting safety shoes with a value of 84%, despite this, the model's performance and effectiveness in PPE detection were prominent.

The limited focus on computer vision software and work safety and wellbeing has contributed to the study in [25], which explores two model frameworks based on YOLOv8 pretrained nano and small architectures. These models were trained for hard hat and head detection using datasets provided by Roboflow. The YOLOv8 nano and small models were evaluated based on precision, recall, mAP@50, and mAP@50-95. The YOLOv8 small model outperformed the YOLOv8 nano model, with a high precision value of 92%. The study discusses the restraints of exploring more enhanced models due to the computational constraints limiting the number of test models, suggesting further research on YOLOv8 structures with longer test phases.

C. Synthesis of Current Studies

The studies reviewed, demonstrate significant potential in enhancing PPE compliance detection using YOLO deep learning models. YOLOv5 and YOLOv8, in particular, presented substantial results in safety helmet and vest detection, reaching an accuracy value of 95% [6]. Nevertheless, the application of these models in real-world construction environments continues to face significant challenges, as highlighted by recent studies. These challenges include ensuring environmental robustness under various conditions, limitations in multi-PPE detection, and issues related to computational efficiency.

1) Environmental robustness. A major challenge in current detection models is their ability to provide accurate and comprehensive detection under various environmental conditions. As emphasized by [26], we need scalable energy-efficient solutions to balance model performance and resource constraints in real-world environments. The advancement of the YOLO model for safety equipment detection was explored in [3] and [14]. Their studies reported a significant drop when these models were exposed to complex site conditions, such as dust, low visibility and cluttered backgrounds- factors that are common in construction work sites. Despite the success in object detection in static backgrounds, the YOLO models still faced limitations in small object detection and compliance monitoring in dynamic environments- as current models often utilize clean, high-quality data, which limits their ability to comply with the ever-changing work environment.

2) Multi-PPE detection. Additionally, most existing research focuses primarily on helmet detection, overlooking other PPE items such as safety vests, gloves and shoes, which are equally essential for the safety of construction workers (Table I). According to studies by [23] and [20], most studies have focused on helmet detection while little attention has been on other critical safety gear such as gloves, safety vests, masks and goggles. The authors have further highlighted the importance of focusing on various safety gear to ensure safety compliance and minimize workplace accidents. Future research should focus on improving YOLO frameworks to further enhance small object detection and recognition of multiple types of PPE in dynamic environments. Additionally, further studies should evaluate these models using real-time data from live construction sites to assess their practical effectiveness.

3) Computational efficiency. The YOLO family of object detection models has evolved significantly over time, with each version showing improved capabilities in terms of speed, accuracy and computational efficiency as shown in Table II. Although the earlier models, such as YOLOv3 and YOLOv5s, demonstrated notable performance, they often required higher computational resources as they have relatively large model sizes, which are 117 MB and 7.06 MB, respectively and higher FLOPs of 39.2G and 7.70G, respectively. YOLOv7 improved upon previous versions, attaining a speed of 14.1ms despite its medium model size of 80.91 MB, indicating its efficiency in fast optimization. The YOLOv8n achieved a significantly smaller model size of 3.2 MB, reducing latency to 12.4 ms, resulting in an impressive FPS of 80.4.

While these high-preforming variants provide advanced accuracy, their practical deployment is limited by their computational demands. As [27] and [28] note, larger models require high computational resources such as large memory resources, which limit their scalability on real-life work sites. This challenge highlights the balance between deployability and model complexity in resource-constrained environments.

TABLE I. QUANTITATIVE COMPARISON OF OBJECT DETECTION MODELS FOR PPE COMPLIANCE MONITORING

Reference	Algorithm	Environment	Datasets	Class	Precision%	Recall%	mAP50%
[19]	YOLOv3	Static	5000	Helmet	96.53%	97.87%	97.12%
[21]	YOLO-ESC	Dynamic	44,000	All	96.5%	98.2%	98.7%
				Helmet	98.1%	99.6%	99.1%
				Head	97.5%	99.1%	99.1%
				Uncertainty	93.9%	95.8%	97.8
[22]	YOLOv5n	Static	7063	Helmet	89.3%	93.5%	94.2%
				Head	85.9%	93.5%	
[5]	Imp.YOLOv5s	Dynamic	36,918	Helmet	92.9%	91.6%	94.1%
[23]	YOLOv7	Static	743	All	95.7%	87.3%	92%
[25]	YOLOv8n	Dynamic	7035	Helmet	89.5%	44.9%	51.7%
				Head			
[24]	YOLOv8m	Static	2934	All	89%	92%	95.6%

TABLE II. PERFORMANCE COMPARISON BETWEEN PREVIOUS MODELS AND THE CURRENT IMPROVED YOLOV11N MODEL

Algorithm	Study	Model Size	Parameters (M)	Latency (ms)	FLOPs (G)	Frames per Second (FPS)
YOLOv3	[21]	medium	117	39.2	65.0	25.5
YOLOv5s	[27]	small	7.06	18	7.70	55.56
YOLOv7	[14]	medium	80.91	14.1	101.8	70.92
YOLOv8n	[29]	nano	3.2	12.4	8.7	80.4

III. METHODOLOGY

This research develops and evaluates the YOLOv11n model for PPE compliance monitoring in construction sites. High-quality datasets of 19,076 images were collected from various sources, annotated for safety helmets and vests, and preprocessed with resizing and data augmentation to improve detection accuracy under dynamic conditions such as extreme weather conditions and fluctuating lighting. The YOLOv11n model, selected for its computational efficiency, was trained on a T4 GPU using Google Colab, with key metrics such as precision, recall, and mAP used for performance evaluation. The model was then tested on prerecorded and live video feeds to validate its effectiveness in real-time safety monitoring.

A. Image Collection

High-quality images and videos of construction sites and PPE items were collected from multiple sources to ensure diverse and balanced datasets. Publicly available resources from sites like Unsplash and Pexels were gathered, increasing safety gear and construction site image variety. Additionally, two distinct datasets were forked from Roboflow [30], [31]. Each dataset specifically focused on safety helmets and safety vests, respectively, further enhancing the model with a wide range of PPE items and variations. The addition of these datasets was essential to develop a well-trained model with high accuracy in object detection. In total, 19,076 images were utilized for dataset preparation, providing an exemplary dataset for effective model training. The images collected represent a diverse collection of PPE imagery, considering factors such as lighting, dynamic environments, and movement, which ensured adaptation to real-time conditions.

B. Dataset Preparation

The datasets were merged, creating a single diverse PPE dataset, allowing for enhanced model training. The dataset was then annotated using Roboflow by implementing bounding boxes on PPE items based on two classes (Safety Helmet and Safety Vest) as shown in Fig. 2. This annotation process ensures precise labelling required for high-accuracy training. An additional dataset obtained from Roboflow was also labelled using a single class (Person) [32] to be integrated into a separate training for enhanced PPE compliance monitoring.

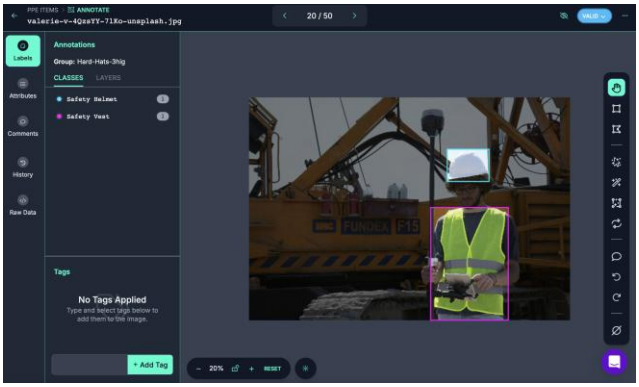


Fig. 2. Screenshot of Roboflow annotation interface based on two class labels: safety helmets and safety vests.

Each class was annotated with a count ratio of approximately 2:1 for safety helmets and safety vests, respectively. The ratio difference is due to the challenges AI models face in small object detection. Thus in order to enhance the detection of safety helmets in dynamic backgrounds, the annotation count was increased (Table III). For the PPE dataset, the images were split based on training, validation and testing. Each set was divided into 16,650, 1625, and 801 images, respectively, as shown in Table IV.

TABLE III. ANNOTATION COUNT FOR EACH CLASS

Class	Count
Safety Helmet	15,287
Safety Vest	6,590

TABLE IV. DATASET DISTRIBUTION FOR TRAINING, VALIDATION, AND TESTING

Training	Validation	Testing
16,650	1625	801

1) Preprocessing. To maintain a consistent processing size and improve model training, all images in the dataset were resized to a resolution of 640 × 640 pixels. This uniform size allows for standardized processing across the dataset and equally complies with the standard image size of YOLO models. Additionally, Auto-Orient was applied to correct the image orientation and ensure proper alignment for optimal object detection. This step ensures standardized preprocessing for consistent results across the dataset. This preprocessing stage is crucial for the success of data augmentation and model training processes while reducing computational overload.

2) Data augmentation. To further enhance model training, data augmentation is implemented in the dataset. Data augmentation is a critical process to enhance model recognition and generalization capabilities, as it utilizes the original dataset to generate a diverse dataset based on multiple variations. This process simulated dynamic conditions such as extreme weather and variable lighting, enhancing the model's robustness. The augmentation strategy, which included applying 25% grayscale alterations to images, assigning a blur with a 4.8-pixel radius, and addition of noise with a deviation of 1.8 pixels, were chosen for their ability to simulate real-world conditions (Fig. 3). For instance, the grayscale modifications imitate limited visibility due to stormy weather, while noise and blur alterations simulate dust, rain, or light fluctuations. By expanding the dataset diversity, this process plays a pivotal role in increasing model training accuracy.



Fig. 3. Example of data augmentation added to construction site image. Original image (Left) and augmented image (Right) with 25% Grayscale, blur 4.8px, and noise 1.8px.

C. Model Training

This research employs the YOLOv11n model obtained by Ultralytics and trains it for PPE detection using Google Colab. The YOLOv11n model is selected because of its lightweight architecture, optimized inference speed and low memory usage compared to the other YOLOv11 models (YOLOv11s, YOLOv11m, YOLOv11l, YOLOv11x), making it a significant choice where computational efficiency is a priority, such as construction sites [33]. The model is trained using a carefully curated dataset for PPE detection, ensuring high-speed training and performance. The training process was conducted on a T4 GPU via Google Colab, allowing for faster training time and model optimization. The coding environment was set up in both Google Colab and Spyder, Python 3.12. Table V further details the training environment.

TABLE V. MODEL TRAINING ENVIRONMENT

Component	Specifications
Hardware	GPU: T4 (Google Colab), CPU (MacOS CPU model)
Program	Google Colab, Anaconda (Spyder, Python 3.12)
Framework	PyTorch
Libraries	Ultralytics, OpenCV, NumPy, Roboflow, Matplotlib
Algorithm	YOLO11n

Hyperparameters are set in the training environment to ensure optimal accuracy and recall values in PPE detection (Table VI). The dataset used is defined as data.yaml, which contains data for classes and paths to images used for training. The model was trained for 20 epochs to establish sufficient exposure to the dataset and achieve optimal learning. A batch size of 16 and an image size of 640 pixels were selected for efficient model training and processing. The model weights were further optimized for real-time detection performance by implementing fine-tuning settings, including reducing precision to FP16, using a batch size of 8, and resizing input images to 320 × 320 pixels. These optimizations were crucial for enhanced speed and reduced computational load.

TABLE VI. MODEL TRAINING HYPERPARAMETERS

Hyperparameter	Value
Data	data.yaml
Epoch	20
Batch	16
Imgsz	640

1) Metric monitoring. After model training is complete, the model is evaluated using the following metrics:

Precision%: Percentage of accurately detected PPE as shown in Eq. (1):

$$Precision = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \quad (1)$$

Recall%: Percentage of actual PPE identified as computed according to Eq. (2):

$$Recall = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (2)$$

Mean average precision (mAP50): The average precision values are calculated at an intersection over union (IoU) threshold of 50%.

Mean average precision (mAP50-95): The average precision values are calculated at an IoU threshold of 50% to 95%.

Confusion matrix: The model's ability to identify objects correctly. Table VII defines each confusion matrix.

TABLE VII. CONFUSION MATRIX DEFINITIONS

Metric	Definition
True Positive (TP)	The number of instances predicted correctly as a positive class
True Negative (TN)	The number of instances predicted incorrectly as a positive class
False Positive (FP)	The number of instances predicted incorrectly as a negative class
False Negative (FN)	The number of instances predicted correctly as a negative class

Class Loss: Indicates how the model performs in classifying each PPE item.

Latency (*L*): As shown in Eq. (3), it is the time required to generate a prediction during the inference process.

$$L = \frac{\text{Total Inference Processing Time (s)}}{\text{Number of Images Processed}} \quad (3)$$

Frames per Second *(FPS*): The number of frames processed per second is calculated as in Eq. (4):

$$FPS = \frac{\text{Batch Size}}{\text{Latency (ms)}} \times 1000 \quad (4)$$

Floating Point Operations (FLOPs): Indicates the computational cost and efficiency of the model as measured in Eq. (5):

$$FLOPs = 2 \cdot K^2 \cdot C_{\text{in}} \cdot C_{out} \cdot H_{\text{out}} \cdot W_{\text{out}} \quad (5)$$

K: Kernel size

C_{in}: Number of input channels

C_{out}: Number of output channels

H_{out}: Output height

W_{out}: Output width

Complex O Notation (*O*): The computational complexity during inference can be estimated as shown in Eq. (6) for a single layer and Eq. (7) for the total complexity:

$$Single\ Layer = O(K^2 \cdot C_{\text{in}} \cdot C_{\text{out}} \cdot H_{\text{out}} \cdot W_{\text{out}}) \quad (6)$$

$$Total\ Complexity = O(\sum_{i=1}^{L} K_i^2 \cdot C_{\text{in},i} \cdot C_{\text{out},i} \cdot H_{\text{out},i} \cdot W_{\text{out},i}) \quad (7)$$

Monitoring these metrics is an important process in developing an efficient model to avoid overfitting or underfitting, and ensuring that the model performs well with real-time data.

D. System Integration

The PPE-trained YOLOv11n model is integrated with a pretrained YOLO model improved for person detection. This integration allows for enhanced PPE compliance detection and system notification, making the model effective for real-time safety compliance monitoring on construction sites. The two models are merged using Anaconda (Spyder 3.12), including a code that executes a desktop alert for failure to comply with PPE requirements, further enhancing the model for optimal construction site safety.

E. Performance Evaluation

To further evaluate the performance of the enhanced model, it is tested on prerecorded video data of construction sites and live video feed using an IP Webcam application. The evaluation focuses on the accuracy of the bounding boxes in detecting PPE objects, noting the model's ability to detect safety helmets, safety vests, and safety violations. Key performance metrics such as precision, recall, and mAP are recorded, assessing the model's capabilities in dynamic environments and its effectiveness in real-time compliance monitoring on construction sites.

IV. RESULTS

The YOLOv11n model received a training progression over 20 epochs. At the 20th epoch, the model achieved a precision of 89.48%, 85.01% recall, 91.62% mAP50, and 64% mAP50-95, indicating significant performance outcomes (Table VIII).

TABLE VIII. PERFORMANCE METRICS OF YOLOV11N BASED ON PRECISION, RECALL, AND MAP50

Algorithm	Precision%	Recall%	mAP50%
YOLOv11n	89.5%	85%	91.6%

The Precision-Recall curve shown in Fig. 4(a) illustrates the performance of each class and overall accuracy. Safety helmets received the highest confidence at 94.9%, revealing the model's efficiency in helmet detection. Likewise, safety vests achieved an optimal confidence of 88.3% resulting in an overall mAP50 of 91.6%. The model maintains exceptional results at an optimal threshold, achieving a high value of 97% as shown in Fig. 4(b) in the Recall-Confidence curve. This indicates the model's strong ability to detect minimal false negatives. Notably, as the confidence threshold increases, the recall value might slightly decrease to prevent false positives from showing, signifying the model's ability to balance between recall and precision.

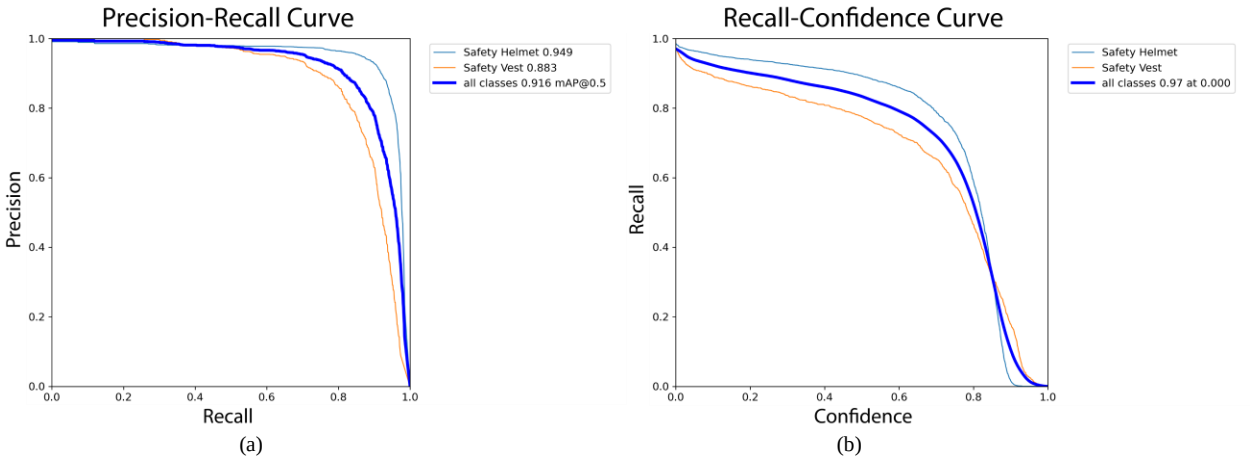


Fig. 4. YOLOv11n : (a) Precision-Recall Curve; (b) Recall-Confidence Curve.

The confusion matrix, which represents the model's capability in identifying objects correctly based on three classes (Safety Helmet, Safety Vest, and Background), indicates remarkable results. As shown in Fig. 5, the TP rate for safety helmets reached an optimal 94%, revealing the model's strength in detection accuracy and background separation. The model experienced false negatives primarily in poor lighting conditions and during object overlaps, where smaller or obscured PPE items were not detected. Conversely, false positives were observed when background elements, such as construction equipment, resembled PPE items. These findings indicate that environmental factors significantly impact detection accuracy. Despite the model experiencing slight confusion in distinguishing safety vests from background images due to similar features or overlap in color, the class nevertheless obtained a notable TP value of 86%. These results signify the robustness of the model's performance and effectiveness in identifying target classes.

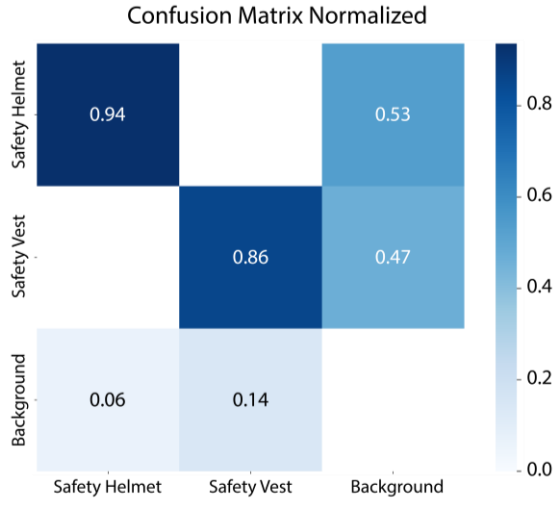


Fig. 5. Confusion matrix across three classes: safety helmets, safety vests, and background.

Instance distribution and bounding box visualization are equally analyzed, as shown in Fig. 6. The instance chart (Left) indicates a significant class count of safety helmets in contrast to safety vests, suggesting increased trained data contributing to higher accuracy in helmet detection. The bounding box visualization shown in Fig.6 (Right) illustrates the predefined shapes the model applies to determine object location and size. The anchor boxes are well distributed and vary in size, indicating the model's capability in detecting objects of various dimensions.

Following the optimization process, which included reducing computational load by implementing FP16 precision for faster inference times, resizing images to 320 pixels, and minimizing batch size from 16 to 8, the model is evaluated for latency, FPS and FLOPs using NVIDIA GPU. The model achieved a significant processing speed of 71.68 FPS, a paramount increase in contrast to the video playback rate of 25 FPS (Table IX). The average latency equally was 13.91ms indicating near real-time detection of objects. The computational complexity of the model is equally calculated achieving a value of 801.68 MMac and a compact model size of 2.59 MB, these values demonstrate the efficiency for deployment on edge devices with limited resources in dynamic construction environments.

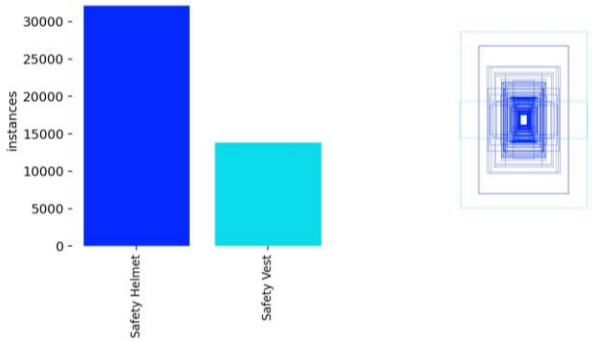


Fig. 6. Instance distribution and bounding box visualization analysis.

TABLE IX. PERFORMANCE METRICS OF OPTIMIZED YOLOV11N BASED ON FPS, LATENCY, MODEL SIZE, AND FLOPS

Algorithm	Video FPS	Processing FPS	Latency (ms)	Model Size (M)	FLOPs (G)
YOLO11n	25	71.68	13.951	2.59	0.801

The enhanced model is evaluated on the distributed dataset set for testing, prerecorded video data and live webcam feed, to further validate its effectiveness in real-time dynamic construction environments as shown in Fig. 7. The results displayed exceptional performance in both static and dynamic settings, with consistent detection accuracy despite fluctuations in lighting, worker movement and environmental conditions. These findings demonstrate the efficiency of the model's application in safety compliance monitoring on construction sites.

To ensure PPE compliance in construction sites, the improved YOLOv11n model is integrated with a pretrained YOLO model aimed specifically for person detection. The supplementary model is enhanced with additional training at 30 epochs receiving an overall precision value of 99.8% and recall of 100%. The newly merged dual model framework is additionally integrated with a desktop alert code to further implement PPE compliance on construction work sites and notify supervisors of non-compliance breaches. The model achieved optimal results displaying efficient performance in construction compliance monitoring. Fig. 8 shows a real-time PPE violation desktop alert sent on MacOS.



Fig. 7. Image evaluation: Training set dataset (left) and inference ran on prerecorded video data (right).



Fig. 8. Real-time detection of PPE compliance and safety violations.

V. DISCUSSION

The optimized Yolov11n model obtained higher accuracy in detecting PPE compliance, exceeding the precision value of both YOLOv8n and YOLOv5n in dynamic conditions [22], [25]. It demonstrated superior recall and precision, outperforming previous models and validating its suitability for real-time monitoring in construction sites. This aligns with research by [18] who emphasize the importance of real-time detection in

complex environments. While safety helmets obtained a significant TP value of 94%, the model experienced a slight regression in the TP value for safety vests, which could be due to the model's struggle in differentiating the class from complex backgrounds with similar color overlaps and feature distinction. This lower accuracy has equally been noted in previous research studies, such as that by [24]. Future iterations of the model should incorporate advanced background segmentation techniques to address this challenge.

Additionally, the model demonstrated significant ability in detection of objects in real-time environments, as evidenced by its increased FPS and reduced latency, making it highly efficient for fast paced construction sites. The model outperformed the inference speed of the earlier YOLOv3 model by 78.5% and the YOLOv5s model by 22.5% [21], [27]. The YOLOv11n model equally attained a lighter model size and reduced computational cost, in contrast to the heavy weight and high cost of the YOLOv7 architecture, signifying the ability to perform on edge devices in dynamic environments without compromising speed and accuracy [14]. While the YOLOv8n model demonstrated a slightly higher inference speed of 80.4 FPS, the improved YOLOv11n model displays a significant balance with lower FLOPs and model size, making it a superior choice for deployment on complex construction sites where computational efficiency and real-time performance are critical [29].

Despite the model's strengths, the research faced several challenges and limitations in achieving optimal accuracy results. This could be due to an imbalance in class counts between safety helmets and safety vests, which could have led to biased detection outcomes. Likewise, while the nano architecture of the YOLOv11 model enhanced detection speed and training efficiency, a larger model architecture might have potentially increased performance significantly. This limitation arises from the limited computational load of the MacOS CPU, and the restricted time limit of Google Colab used during training. Addressing these challenges, by using a balanced dataset and increasing GPU for longer training time, could enhance the model further for PPE detection.

Furthermore, due to the conversion to FP16 and resizing of the images to 320 pixels during the inference process in order to improve latency and reduce computational cost, the model could potentially face a slightly lowered detection performance. Finetuning the model further, with a larger batch size during training and exploring enhanced precision techniques, including merging FP32 and FP16 where necessary, could enhance overall performance in real-time applications. Expanding the dataset to include additional PPE items such as gloves, glasses, and boots is recommended for future research endeavors to develop a more comprehensive safety object detection model. Increasing the number of epochs and fine-tuning hyperparameters, while ensuring class balance, could further enhance the model's performance and detection accuracy. Moreover, deploying the model on edge devices, such as sensors and IoTs, could additionally advance the model in object detection making it crucial for real-time applications. Implementing these steps would not only potentially enhance detection capabilities but equally allow for widespread implementation in challenging conditions and high-risk environments ensuring worker safety and construction requirements compliance.

Overall, the results of the YOLOv11n model demonstrate exceptional performance in real-time detection within dynamic environments, with further optimization suggestions paving way for a more comprehensive model and its deployment in real-time critical settings.

VI. Conclusion

This research proposes a deep learning model using an optimized YOLOv11n algorithm to address safety challenges on construction sites. Crucial performance metrics such as precision, mAP50, Latency, and FLOPS, are evaluated to ensure real-time PPE compliance monitoring in resource-constrained dynamic environments. The YOLOv11n demonstrated significant results, achieving 89.5%, 85%, and 91.6% for precision, recall, and mAP50, respectively while maintaining a notable balance between speed and computational load. With a latency of 13.95 ms and computational cost of 0.801 GFLOPs, the model exhibits a lightweight and computationally efficient architecture capable of object detection on edge devices. Additionally, the study integrated a PPE compliance system with desktop notifications to alert supervisors of safety violations, demonstrating the model's robustness and practical application for real-time monitoring.

The research, however, encountered several challenges. The class imbalance led to a potential bias in detection results, and the limited computational resources prevented the exploration of larger, more complex architectures. Additionally, the need for increased speed and reduced model size in complex environments may slightly impact detection performance, practically for small object detection. These challenges underscore the importance for further enhancements and evaluation.

Future research recommendations include improving detection performance by further optimizing the model with additional hyperparameters, extended training epochs and balanced datasets. Expanding the model's scope to include additional PPE items, such as safety glasses, boots, and gloves, would further enhance the model's applicability. Furthermore, exploring the integration of the model on edge devices could significantly advance its capabilities for real-time detection in complex construction sites. These efforts aim to address current limitations and pave the way for safer, more efficient construction sites through advanced, scalable deep learning models and real-time compliance monitoring systems.

References

[1] U.S. Bureau of Labor Statistics, "A look at falls, slips, and trips in the construction industry : The Economics Daily: U.S. Bureau of Labor Statistics." Accessed: Nov. 09, 2024. [Online]. Available: https://www.bls.gov/opub/ted/2024/a-look-at-falls-slips-and-trips-in-the-construction-industry.htm

[2] A. Mohammadi, M. Tavakolan, and Y. Khosravi, "Factors influencing safety performance on construction projects: A review," Saf Sci, vol. 109, pp. 382–397, Nov. 2018, doi: 10.1016/J.SSCI.2018.06.017.

[3] J. Hu, X. Gao, H. Wu, and S. Gao, "Detection of Workers Without the Helments in Videos Based on YOLO V3," in 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, Oct. 2019, pp. 1–4. doi: 10.1109/CISP-BMEI48845.2019.8966045.

[4] S. Li and A. Zhu, "Recent Advancements with Human Behavior Recognition and AI in Construction Automation: A literature Review," in

ICARM 2024 - 2024 9th IEEE International Conference on Advanced Robotics and Mechatronics, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 759–764. doi: 10.1109/ICARM62033.2024.10715843.

[5] C. Du, X. Zhang, and C. Wu, "Method for Safety Helmet Detection Based on Improved YOLOv5," in Proceedings - 2023 International Conference on Computer, Internet of Things and Smart City, CIOTSC 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 43–48. doi: 10.1109/CIoTSC60428.2023.00013.

[6] K. Kim, K. Kim, and S. Jeong, "Application of YOLO v5 and v8 for Recognition of Safety Risk Factors at Construction Sites," Sustainability, vol. 15, no. 20, p. 15179, Oct. 2023, doi: 10.3390/su152015179.

[7] A. H. Hassan, R. bin Sulaiman, M. Abdulhak, H. Kahtan, and U. Kebangsaan Malaysia, "Bridging Data and Clinical Insight: Explainable AI for ICU Mortality Risk Prediction Institute of IR 4.0 (IIR4.0)," 2025. [Online]. Available: www.ijacsa.thesai.org

[8] M. Dol and A. Geetha, "A Learning Transition from Machine Learning to Deep Learning: A Survey," in Proceedings of the 2021 International Conference on Emerging Techniques in Computational Intelligence, ICETCI 2021, Institute of Electrical and Electronics Engineers Inc., Aug. 2021, pp. 89–94. doi: 10.1109/ICETCI51973.2021.9574066.

[9] Z. ullah, L. Qi, D. Binu, B. R. Rajakumar, and B. Mohammed Ismail, "2-D canonical correlation analysis based image super-resolution scheme for facial emotion recognition," Multimed Tools Appl, vol. 81, no. 10, 2022, doi: 10.1007/s11042-022-11922-3.

[10] M. A. A. M. Divyashree, "Machine Learning Technique for Image Classification System using Assembly," in International Journal of Engineering Research & Technology (IJERT), 2020. [Online]. Available: www.ijert.org

[11] K. Alshammari, "Enhancing Vision-Based Religious Tourism Systems in Makkah Using Fine-Tuned YOLOv11 for Landmark Detection," 2025. [Online]. Available: www.ijacsa.thesai.org

[12] L. Sun and L. Wang, "An improved YOLO V5-based algorithm of safety helmet wearing detection," in Proceedings of the 34th Chinese Control and Decision Conference, CCDC 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 2030–2035. doi: 10.1109/CCDC55256.2022.10033448.

[13] D. Thakur, P. Pal, A. Jadhav, N. Kable, V. Bhagyalakshmi, and S. Deshpande, "YOLOv8- Based Helmet and Vest Detection System for Safety Assessment," in 2023 International Conference on Network, Multimedia and Information Technology, NMITCON 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/NMITCON58196.2023.10275958.

[14] H. Bian et al., "Detection Method of Helmet Wearing Based on UAV Images and Yolov7," in ITNEC 2023 - IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1633–1640. doi: 10.1109/ITNEC56291.2023.10082536.

[15] L. R. Jadhav and M. Hudnurkar, "Mobilenet and Deep Residual Network for Object Detection and Classification of Objects in IoT Enabled Construction Safety Monitoring," in Proceedings of the 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems, ICSES 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICSES60034.2023.10465566.

[16] M. Z. Shanti et al., "A Novel Implementation of an AI-Based Smart Construction Safety Inspection Protocol in the UAE," IEEE Access, vol. 9, pp. 166603–166616, 2021, doi: 10.1109/ACCESS.2021.3135662.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Dec. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[18] B. Lin, "Safety Helmet Detection Based on Improved YOLOv8," IEEE Access, vol. 12, pp. 28260–28272, 2024, doi: 10.1109/ACCESS.2024.3368161.

[19] N. K. Anushkannan, V. R. Kumbhar, S. K. Maddila, C. S. Kolli, B. Vidhya, and R. G. Vidhya, "YOLO Algorithm for Helmet Detection in Industries for Safety Purpose," in 3rd International Conference on Smart Electronics and Communication, ICOSEC 2022 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 225–230. doi: 10.1109/ICOSEC54921.2022.9952154.

[20] R. C. Reyes, R. V. Sevilla, G. S. Zapanta, J. V. Merin, R. R. Maaliw, and A. Ferrer Santiago, "Safety Gear Compliance Detection Using Data Augmentation-Assisted Transfer Learning in Construction Work Environment," in 2022 IEEE International Conference on Electronics, Computing and Communication Technologies, CONECCT 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/CONECCT55679.2022.9865757.

[21] P. Jin, H. Li, W. Yan, and J. Xu, "YOLO-ESCA: A High-Performance Safety Helmet Standard Wearing Behavior Detection Model Based on Improved YOLOv5," IEEE Access, vol. 12, pp. 23854–23868, 2024, doi: 10.1109/ACCESS.2024.3365530.

[22] Kisaezehra, M. U. Farooq, M. A. Bhutto, and A. K. Kazi, "Real-Time Safety Helmet Detection Using Yolov5 at Construction Sites," Intelligent Automation and Soft Computing, vol. 36, no. 1, pp. 911–927, 2023, doi: 10.32604/iasc.2023.031359.

[23] M. R. Alvarez, C. Q. Vega, and L. Wong, "Model for Recognition of Personal Protective Equipment in Construction Applying YOLO-v5 and YOLO-v7," in International Conference on Electrical, Computer and Energy Technologies, ICECET 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICECET58911.2023.10389215.

[24] M. I. Al-Khiami and M. M. Elhadad, "ENHANCING CONSTRUCTION SITE SAFETY USING AI: THE DEVELOPMENT OF A CUSTOM YOLOV8 MODEL FOR PPE COMPLIANCE DETECTION," in Proceedings of the European Conference on Computing in Construction, European Council on Computing in Construction (EC3), 2024, pp. 577–584. doi: 10.35490/EC3.2024.307.

[25] P. Jankovic, M. Protic, L. Jovanovic, N. Bacanin, M. Zivkovic, and J. Kaljevic, "YOLOv8 Utilization in Occupational Health and Safety," in 2024 IEEE Zooming Innovation in Consumer Technologies Conference, ZINC 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 182–187. doi: 10.1109/ZINC61849.2024.10579310.

[26] A. H. Hassan, E. M. Ahmed, J. M. Hussien, R. bin Sulaiman, M. Abdulhak, and H. Kahtan, "A cyber physical sustainable smart city framework toward society 5.0: Explainable AI for enhanced SDGs monitoring," Research in Globalization, vol. 10, Jun. 2025, doi: 10.1016/j.resglo.2025.100275.

[27] L. Wang, X. Zhang, and H. Yang, "Safety Helmet Wearing Detection Model Based on Improved YOLO-M," IEEE Access, vol. 11, pp. 26247–26257, 2023, doi: 10.1109/ACCESS.2023.3257183.

[28] R. Ayachi, Y. Said, M. Afif, A. Alshammari, M. Hleili, and A. Ben Abdelali, "Assessing YOLO models for real-time object detection in urban environments for advanced driver-assistance systems (ADAS)," Jun. 01, 2025, Elsevier B.V. doi: 10.1016/j.aej.2025.03.077.

[29] Glenn Jocher, Ayush Chaurasia, and Jing Qiu, "YOLOv8 - Ultralytics YOLO Docs." Accessed: Dec. 12, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolov8/#key-features

[30] Universe Datasets, "Hard Hat Universe Dataset > Overview." Accessed: Dec. 02, 2024. [Online]. Available: https://universe.roboflow.com/universe-datasets/hard-hat-universe-0dy7t

[31] Roboflow Universe Projects, "Safety Vests Dataset > Overview." Accessed: Dec. 02, 2024. [Online]. Available: https://universe.roboflow.com/roboflow-universe-projects/safety-vests

[32] Yolov8, "Person Dataset." Accessed: Dec. 02, 2024. [Online]. Available: https://universe.roboflow.com/yolov8-mn29d/person-hecxr

[33] Ultralytics YOLO11, "YOLO11 🚀 NEW - Ultralytics YOLO Docs." Accessed: Dec. 04, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolo11/#citations-and-acknowledgements

Reinforcement Learning Improves SVM-Driven Algorithms for Classifying Multi-Sensor Data for Medical Monitoring

Zhiwei Xuan, Yajie Liu*
School of Information Engineering, Henan Vocational University of Science and Technology, Zhoukou, China

***Abstract*—Multi-sensor data in medical monitoring includes waveform changes in physiological signals and time-series characteristics of disease progression. These features typically exhibit high-dimensionality, large-scale, and time-varying characteristics. Nonlinear relationships exist between these features, increasing the difficulty of data processing and feature extraction, thereby reducing the classification capabilities of related algorithms. This study proposes a multi-sensor data classification processing method in medical monitoring based on reinforcement learning improved SVM. The algorithm employs the DBSCAN algorithm combined with Euclidean distance for clustering and data collection of multi-sensor data. Discrete wavelet transform is used to remove interference noise from the data, followed by convolutional neural networks for signal feature extraction from the denoised data. The Q-learning algorithm in reinforcement learning is used to improve the traditional SVM, with the extracted signal features input into the improved SVM. The classification results of medical monitoring multi-sensor data are output via a regression function. The experimental results show that the denoising results of medical monitoring data of the method are high, the signal-to-noise ratio is high, and the Kappa coefficient reaches up to 0.98. Therefore, it shows that the method can accurately classify medical monitoring multi-sensor data.**

Keywords*—Reinforcement learning; improved SVM; medical monitoring; multi-sensor; data; classification processing

I. INTRODUCTION

Medical monitoring systems [1] are increasingly being used to monitor patient health data. These include electrocardiograms (ECG), electroencephalograms (EEG), blood pressure, blood oxygen saturation, body temperature, and a variety of other physiological signals [2]. Through the processing of monitoring data, statistics on patient cases and health status [3] can be achieved to help doctors make more accurate clinical judgments and decisions; the abnormal classification of medical signals can also promote the development of medical intelligence, thus promoting the progress and innovation of medical technology [4]. The continuous improvement and application of medical signal processing technology are expected to bring more possibilities for clinical medicine and promote the development of the healthcare industry toward intelligence, personalization, and precision. Accurate classification of medical data can help improve the automation level of the medical monitoring system, reduce the workload of healthcare workers, and improve the efficiency and quality of medical services [5]. In addition, the relevant methods can also provide more objective and accurate data support for medical research and promote the progress of medical research and the innovation of clinical practice.

With the continuous development of medical monitoring, multi-sensor data analysis has become an important part of medical diagnosis. Traditional multi-sensor data classifications for medical monitoring often have problems such as low classification accuracy and slow processing speed [6]. Support vector machines (SVMs) are widely used as an effective classification algorithm in medical data classification. However, traditional SVM methods still face many challenges when processing complex and variable data, such as noise interference and inaccurate feature extraction.

Therefore, the research on multi-sensor data classification and processing methods for medical monitoring is not only of great significance for improving the function and performance of medical monitoring systems, but also provides technical support and theoretical guidance for improving medical services, promoting medical research, and safeguarding patients' health [7]. Relevant experts have been continuously exploring the field of data classification.

Singh and Khaiyum [8] proposed a data stream classification method based on the concept of drift, introducing an incremental semi-supervised learning model to regularize neural networks by incorporating auxiliary information such as label merging or pairwise constraints. However, due to the algorithm's efficiency and sensitivity to parameters, it is challenging to achieve better results in highly nonlinear environments with multimedia sensor data. Kenger and Ozceylan [9] proposed a data classification method based on mathematical modeling and improved online learning, using I0L_GFMM to generate initial cluster centers for the MILP model to enhance its efficiency, and combining fuzzy minimum and maximum neural networks with the MILP model. The proposed hybrid model demonstrated its applicability on both real and synthetic datasets. However, the aforementioned methods do not account for noise interference in medical data and are not suitable for high-noise environments in medical monitoring data. Zhai et al. [10] proposed a data classification method based on generative model diversity oversampling. When classifiers are faced with imbalanced multi-sensor-collected medical monitoring datasets, they typically favor the majority category, leading to poor classification performance. Liang et al. [11] proposed a data classification method based on flow regularization to overcome data quality issues and the high labeling cost of extracting data sample labels. However, while

This work was sponsored by Henan Provincial Higher Education Key Research Project Program (Project number of the fund：24B510006).

this classification algorithm can handle specific types of label noise, identifying the type of noise present in each medical data point remains a challenge.

Aiming at the problems of existing methods, this study proposes an improved SVM based on reinforcement learning for the classification of multi-sensor monitoring data. Compared to existing research, the reinforcement learning-enhanced SVM method proposed in this study aims to address the issue of insufficient classification accuracy in traditional algorithms under feature selection and imbalanced dataset scenarios. This is achieved through DWT denoising, CNN feature extraction, and Q-learning optimization. By combining the Discrete Wavelet Transform (DWT) with Multiresolution Analysis (MRA), precise denoising of medical monitoring data is realized, thereby improving the signal-to-noise ratio of the data. Furthermore, this method leverages the advantages of reinforcement learning to enhance the classification performance of medical data through interaction between the agent and the environment. By incorporating the Q-learning algorithm to improve the SVM, the method continuously learns from interactions with the environment to identify the optimal classification strategy, addressing the limitations of traditional support vector machines when processing multi-sensor data. The introduction of reinforcement learning algorithms enables decision automation in exploring optimal feature selection, optimizing model parameters, and balancing classification boundaries, thereby enhancing the model's performance and adaptability.

The subsequent sections of this study are arranged as follows: Section II introduces multi-sensor data acquisition and clustering methods. Section III elaborates on the data denoising process. Section IV discusses CNN feature extraction techniques. Section V provides a detailed explanation of the SVM classification model improved by Q-learning. Section VI validates the effectiveness of the method through experiments. Finally, Section VII summarize the research results and looks forward to future directions.

II. Multi-Sensor Data Acquisition and Clustering in Medical Monitoring

Multi-sensor data involved in medical monitoring are usually characterized by high dimensionality, large scale, and time-varying nature, such as waveform changes in physiological signals and time-series features of disease progression. The high dimensionality and complexity of such data lead to a reduced ability to categorize the data. The use of multi-sensors to collect data can obtain more dimensional information, which helps to improve the accuracy and stability of classification algorithms and enhance the effectiveness and reliability of classification. Therefore, in this study, multi-sensor data acquisition in medical monitoring is used to ensure data integrity and to support the accuracy of feature extraction through sufficiently acquired data, which can help to assess the patient's health status and process it for exhaustive classification more accurately.

The DBSCAN algorithm can discover the existence of potential clusters based on the density between data points, which helps to reveal hidden patterns and structures in the data. In medical monitoring, different types of medical surveillance sensing data may correspond to different health states or disease characteristics, which can be of help to better understand, by discovering clusters in the data. The DBSCAN algorithm identifies whether a data point is a core object or a boundary point, and from this, clusters with different levels of density are constructed. By identifying the core objects, the clusters to which the data points belong can be better distinguished, thus facilitating subsequent data classification tasks.

The DBSCAN algorithm divides the medical monitoring multi-sensor data into classes; those isolated few signal points will be considered as outliers to be eliminated, and the remaining data clusters as useful data. For the data object p, the ε neighboring object is denoted as $N_\varepsilon(p)$ and $N_\varepsilon(p)=\{q\in D \mid dist(p,q)\le \varepsilon\}$. For a given value $MinPts$, the object q is said to be directly density reachable from q if it is a neighboring object of the object ε of p. The point p is said to be directly density reachable from q.

For a given ε and $MinPts$, if there exists an object O in the set of data objects D such that the objects p and q are density reachable from O concerning the densities ε and $MinPts$,then the objects p and q are said to be density-connected about ε and $MinPts$.

For the set of data objects D, the cluster C is a non-empty subset of the set D that satisfies the following conditions:

1) $\forall p,q$, if $p\in C$, p and q are density reachable, then $q\in C$.

2) $\forall p,q\in C$, p and q are density-connected.

The DBSCAN algorithm requires two important parameters, a point p neighborhood ε radius, and the minimum number of points contained in the neighborhood $MinPts$. Firstly, the data set of a medical sensor monitoring object is specified, from which any data object p is selected, and all the objects are scanned to find out the density reachable set of p about ε and $MinPts$; if p is a core object, all the reachable objects of the set p form clusters; otherwise, p is not a core object, it is an outlier and should be eliminated.

The medical monitoring sensor data was expanded in the 3D numerical axes and the set of points obtained by the expansion was set to be G , and the distances of the points in the space were calculated using the Euclidean distances in Eq. (1) [12]:

$$d(P_i,Z_l)=\sqrt{\sum_{i=1}^{n}(P_{ij}-Z_{lj})^2} \tag{1}$$

where, $d(P_i,Z_l)$ represents the Euclidean distance between the coordinate value P_i of the i-th feature and the coordinate value Z_{lj} of the l-th cluster point, which is the basis for measuring the similarity between data points, P_{ij} denotes the

j-th dimensional coordinate value of the i-th feature point, and Z_{lj} denotes the j-th dimensional coordinate value of the l-th clustering point.

To cluster these points, the values of the parameters ε and $MinPts$ need to be developed first. The parameter values are influenced by the number of data points collected by the medical monitoring multi-sensory during the time and the positive proportionality that exists between them. The number of multi-sensor measurements, and hence the parameter values, are determined based on the needs of medical monitoring.

The execution flow of the DBSCAN algorithm for the medical monitoring sensor data acquisition process is as follows:

1) Arbitrarily select any point object P belonging to any data cluster in the point set G and create it as a new data cluster.

2) For all points P, if the points' number of P's neighborhood is no less than a minimum, P is served as the core point of the newly created data cluster and is compared with other data points to add the density reachable point to that cluster.

The determination of the density reachable is:

A point W is density reachable from a core point P, if there is a path from P to W, and each point on the path is a core point, and the distance between each adjacent two points on the path is within the neighborhood.

3) Loop step (2) until no points can be added to the cluster, then perform step (1) again.

4) Ends when all points in the point set belong to a particular data cluster.

The DBSCAN algorithm is shown in Fig. 1.

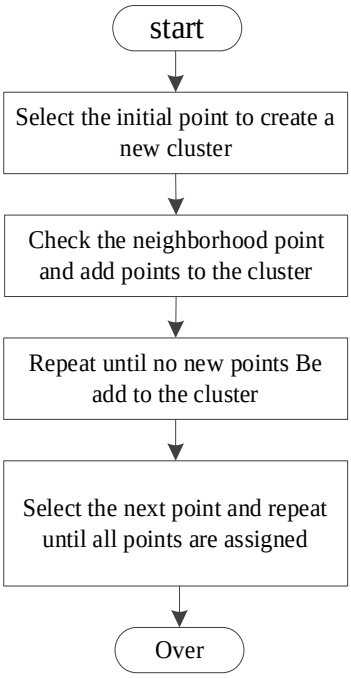


Fig. 1. The DBSCAN algorithm flowchart.

After the DBSCAN algorithm is processed, the medical surveillance multi-sensor data will exist in clusters and the number of points in the clusters a and $MinPts$ need to be compared to determine the validity of the clusters:

1) When $a < MinPts$ is present, the cluster will be considered as a cluster consisting of anomalies, which are anomalous medical monitoring sensor data, and will be excluded.

2) When $a \geq MinPts$, the cluster is composed of valid data, which will be retained to obtain the clustered collected medical monitoring multi-sensor data Q.

III. Sensor Data Denoising for Medical Monitoring

Medical monitoring data are subject to various interference during acquisition and transmission, such as electromagnetic interference, signal attenuation, and transmission failures. These factors can lead to noise in medical monitoring data, which can be mixed into the signal, masking or distorting useful information and reducing the quality and interpretability of the data, and the nonlinear relationships that exist in the noisy data can also make feature extraction difficult. Therefore, by denoising the acquired medical monitoring data, noise interference can be reduced, the accuracy and reliability of the data can be improved, it helps to highlight important features in the signal, and features related to the patient's condition can be extracted and analyzed more accurately, thus enhancing the discriminative power of the classification method.

In this study, after completing the medical monitoring data clustering, the Discrete Wavelet Transform (DWT) is used for denoising. The main advantages are: firstly, the discrete wavelet transform can effectively remove the noise in the signal and improve the quality of the data; secondly, by retaining the important features of the signal, the method can accurately denoise the signal while maintaining the detailed information of the signal as much as possible; in addition, the discrete wavelet transform can also analyse the signal at multiple scales and capture the features at different scales, which can help to understand the structure of the signal in a more comprehensive way and denoise it efficiently; Finally, the discrete wavelet transform has efficient real-time processing capability, which is suitable for scenarios that require rapid processing of medical monitoring data. The specific denoising process is shown in Fig. 2.

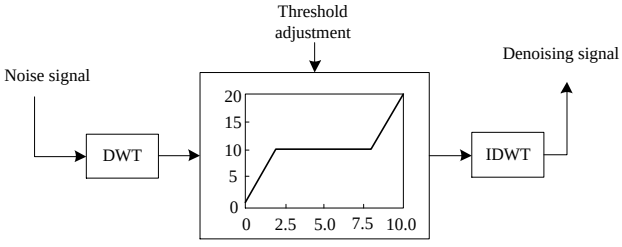


Fig. 2. Discrete wavelets transform denoising flowchart.

The algorithm flow is as follows:

Step 1. Decompose data $U(t)$ at t time into a set of wavelets with Eq. (2) [13]:

$$U(t)=\frac{1}{\sqrt{2^j}}\sum_k\sum_j DW_k^j\cdot\varphi(2^{-j}\cdot t-k) \tag{2}$$

where, φ and j denote the translation and expansion factors, respectively, and DW_k^j denotes the DWT coefficients in the wavelet basis.

Step 2. Use Multi-Resolution Analysis (MRA) data to extend analyses that provide different time and frequency resolutions at each level.

Multi-resolution analysis (MRA) provides different time and frequency resolutions at each analysis level by using the scale function $\phi(t)$ and the wavelet function $\varphi(t)$. $\phi(t)$ is the foundation of MRA, which describes the overall signal features on a coarser scale, similar to the low-frequency in the signal. It constructs the low-frequency approximation of the signal through its expansion and displacement. $\varphi(t)$ describes the local signal features on a finer scale, similar to the high frequency in the signal. It captures the signal details through expansion and displacement, that is, the high frequency in the signal at different scales.

The multi-sensor data $U(t)$ is calculated by Eq. (3), which decomposes the signal into a linear combination of the scale function and the wavelet function at different scales [14].

$$U(t)=\sum_k A_k^{j_0}\varphi_k^{j_0}(t)+\sum_{j=j_0}\sum_k D_k^j\phi_k^j(t) \tag{3}$$

where, $A_k^{j_0}$ and D_k^j denote the approximation coefficient and detail coefficient, respectively. According to the scale and wavelet bases, the equations for both can be expressed as Eq. (4) and Eq. (5) [15]:

$$A_k^{j_0}(t)=\sum_k f_k\cdot\varphi_k^{j_0}(t)=\sum_n h(n-2k)\cdot A_k^{j_0+1}(t) \tag{4}$$

$$D_k^j(t)=\sum_k f_k\cdot\phi_k^{j_0}(t)=\sum_n g(n-2k)\cdot A_k^{j+1}(t) \tag{5}$$

The wavelet decomposition tree constructed by MAR is shown in Fig. 3.

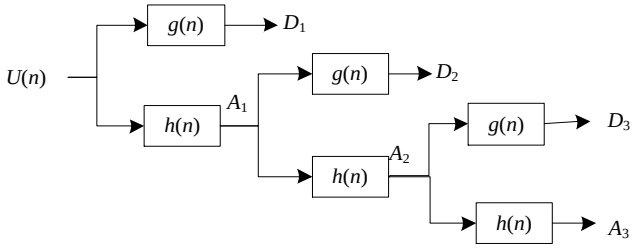


Fig. 3. Schematic diagram of MRA scale decomposition.

Step.3 Soft Threshold Calculation

In the process of quantizing the threshold λ, it is usually considered that anything less than λ is caused by noise, and anything greater than λ is caused by signals, and the soft threshold algorithm for searching towards zero according to a fixed amount is as follows in Eq. (6) [16]:

$$\hat{w}_{j,k}=\begin{cases}\mathrm{sgn}(w_{j,k})(w_{j,k}-\lambda), |w_{j,k}|\geq\lambda\\ 0, |w_{j,k}|\leq\lambda\end{cases} \tag{6}$$

In Eq. (6), λ and $\mathrm{sgn}(x)$ are defined as Eq. (7) and Eq. (8), respectively:

$$\lambda=\sigma\sqrt{2\log(n)} \tag{7}$$

$$\mathrm{sgn}(x)=\begin{cases}1, x>0\\ 0, x=0\\ -1, x<0\end{cases} \tag{8}$$

In the above equation, n denotes the sampling length of the noise signal, σ denotes the noise variance, and $w_{j,k}$ denotes the wavelet coefficients at different scales. The noise variance σ is calculated in Eq. (9) [17]:

$$\sigma=\left[\frac{median|\Lambda_{i,j}|}{0.6475}\right] \tag{9}$$

In Eq. (9), $\Lambda_{i,j}$ denotes the first layer of high-frequency coefficients of the wavelet transform of the noise-containing medical monitoring multisensory data, and $median|\Lambda_{i,j}|$ denotes the median.

Step.4 Positive and Negative Thresholding

The DWT exhibits low-frequency and high-frequency coefficients, respectively, and the data are characterized by two types of high-frequency oscillations: high-frequency vibrations with small amplitudes and low-frequency oscillations with large amplitudes. Both contain peaks and noise signals, respectively [18]. Therefore, the positive and negative thresholds are found through Eq. (10):

$$\lambda_i=\begin{cases}\lambda_+=\alpha*\max(d_i)\\ \lambda_-=\alpha*\min(d_i)\end{cases} \tag{10}$$

In Eq. (10), d_i denotes the high frequency detail coefficient of the n layer, and α denotes the empirical parameter.

The post-threshold low-frequency and high-frequency signals were reconstructed to obtain denoised medical monitoring multi-sensor data $X(t)$.

IV. Multi-Sensor Data Feature Extraction for Medical Monitoring

Due to the characteristics of high dimensionality and large-scale data, its direct use for classification processing will increase the complexity of the computational algorithm and reduce the accuracy of the algorithm. Feature extraction for the fused data after the denoising process can map the original data to a lower-dimensional feature space, reduce the data dimension, simplify the problem, improve the classification efficiency, help to extract the key information and features in the data, and strengthen the effective patterns and structures in the data. Therefore, in this study, CNN is used to extract the main features of the data to reduce the interference of unnecessary information and improve the classification ability of the data by finding meaningful features related to the classification task.

The convolutional layer is the core of the whole CNN, and it can effectively extract the sample features. The convolutional kernel traverses the whole sample by sliding to carry out feature extraction, and after the local features are extracted, the positional relationship between that part of the features and other features can be marked. CNN is non-transparent for feature extraction, and different from traditional network learning methods, the convolutional kernel operation has the feature of local perception, when a part of the features are learnt, the features will be inputted into the next mapping layer, and the mapping layer after that will continue to carry out classification learning. The samples after the convolutional layer use the spatial distribution characteristics of the surrounding samples, and the feature extraction accuracy can be obtained.

For the input medical monitoring multisensory data $X(t)$, a matrix W with a convolution kernel size $a \times b$ and bias B, the result after convolution is in Eq. (11) [19]:

$$h = g(X * W + B) \tag{11}$$

In Eq. (11), $*$ denotes the convolution operation and $g(.)$ denotes the activation function. The activation function used in this study is the ReLU function shown as Eq. (12):

$$g(x) = \max(0, x) \tag{12}$$

The sparse activation of the ReLU function can make certain neurons inactive, reduce the number of redundant connections and parameters, and improve the computational efficiency and generalization ability. The ReLU function solves the problem of gradient vanishing in traditional activation functions and avoids the performance degradation caused by the gradient being too small. In addition, the computation of the ReLU function is simple and efficient, without the need for complex exponential operations, which accelerates the training and inference process of the model. In addition, the ReLU function integrates linear and nonlinear properties, which is convenient to capture the nonlinear relationship of the data and enhances the expressive ability of the neural network. Finally, the ReLU function has no upper limit and can pass positive values to the next layer, further expanding the expressive ability of the network.

The pooling layer can effectively reduce the data dimensions, and the input of the pooling layer is the output of the convolutional layer. The output of the convolutional layer N feature maps is passed through the pooling layer to reduce the data size. In this research, the maximum pooling layer is introduced to sparsity the hidden layer data. The operation of maximization is to retain the maximum value of the pooled region, that is, to remove the non-extremely large values within the matrix, while extracting the extremely large value of the region as the representative value after pooling. The output of the pooling layer results in a substantial reduction of the network parameters, and the kernel size is generally 2×2 when maximum pooling is used, and the 2×2 size kernel enables the number of parameters to be effectively regulated, which helps to prevent the occurrence of overfitting situations and enhances the robustness of the convolutional neural network. Average pooling performs an averaging operation on the matrices in the region and uses the average value to sparse the non-overlapping target region.

The role of the Dropout layer is to drop some neurons at a certain rate during the network training process, which is a simple and effective regularization technique. However, choosing a suitable Dropout rate is critical to optimizing network performance. Higher Dropout rate causes the model to underfit because too many neurons are dropped, and difficult to learn enough information. However, a lower Dropout rate cannot effectively prevent overfitting. In addition, due to the randomness of Dropout during training, it helps the model learn more robust feature representations. Also, the difference between training and testing exists because all neurons are activated.

Therefore, it is often necessary to scale the output of all neurons to match the average level of activation at training, thus improving the generalization ability of the model. Choosing an appropriate Dropout rate is crucial to prevent model overfitting and improve generalization performance. In general, for larger networks or datasets, a lower Dropout rate (e.g. 0.2-0.5) can be chosen, as the network already has enough capacity to capture complex features, and too high a Dropout rate may limit its ability to learn. Conversely, for smaller networks or datasets, a higher Dropout rate (such as 0.5-0.8) can be selected to reduce the risk of overfitting.

The structure of the Dropout network is shown in Fig. 4.

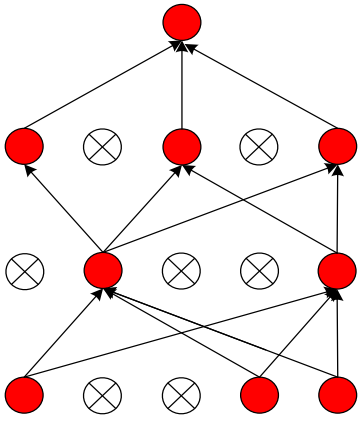


Fig. 4. Dropout network structure.

In the original network structure, all neurons need parameter training, while in the Dropout network structure, some neurons are not involved in the training, Dropout is a very effective method to suppress overfitting, and its use often brings a great improvement to the network generalization performance. The computational equation of the network using Dropout as in Eq. (13) [20]:

$$\left.\begin{aligned} r_i^l &\sim Bernoulli(p) \\ \tilde{y}^{(l)} &= r^{(l)} * y^{(l)} \\ z_i^{l+1} &= W_i^{(l+1)} \tilde{y}^{(l)} + B_i^{(l+1)} \\ y_i^{(l+1)} &= f(z_i^{(l+1)}) \end{aligned}\right\} \tag{13}$$

In Eq. (13), *Bernoulli* function to generate the probability r vector, that is, randomly generate a $[0,1]$ vector, $y_i^{(l+1)}$ is the output of the convolutional neural network for medical monitoring multi-sensor data feature extraction results.

V. Reinforcement Learning for Improved SVM Classification Models

Through feature extraction, we have extracted the most representative and distinguishing features from the original multi-sensor data. The classification algorithm can better distinguish different data types by enriching the feature space and improving classification accuracy. Many physiological signals in medical monitoring data are time-varying, and the trend of these time-varying features is very important for predicting the direction of disease development. Considering these characteristics of sensor monitoring signals, a method combining Q-learning with SVM is proposed to improve the effectiveness of SVM in multi-sensor data classification of medical monitoring.

Reinforcement learning can reduce the dimensionality of the feature space of sensor signals, improve the efficiency of the algorithm, can better capture the dynamic change law of time-varying features, and improve the classification method's ability to process time-ordered data. Therefore, in this study, the Q learning algorithm in reinforcement learning is used to improve SVM, to provide better classification effect of medical monitoring data by SVM.

SVM is a powerful supervised learning algorithm that is widely used in classification and regression tasks. In classification, the goal of SVM is to find a hyperplane that maximizes the spacing between different classes; in regression, the goal of SVM is to find a hyperplane that minimizes the distance of all data points to the hyperplane, while allowing for a certain error boundary.

In SVM regression, the objective function aims to minimize the sum of distances from all data points to the hyperplane while taking into account the error boundary. For a given training data set $\{(x_i, y_i)\}_{i=1}^n$, where x_i is the input vector and y_i is the corresponding output value, the goal of SVM regression is to minimize Eq. (14) objective function [21]:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} L_\varepsilon(y_i - (w \cdot \phi(x_i) + b)) \tag{14}$$

where, w and b are the normal vector and intercept of the hyperplane, respectively; $\phi(x_i)$ represents the mapping from input space to feature space; C represents a regularization parameter, which is used to control the balance between the complexity of the model and the training error; L is ε - insensitive loss.

In SVM, support vectors are those training samples on or within the error boundary, which have a decisive effect on the hyperplane position for their determination of the constraints on interval maximizing. The support vectors are equally important in regression because of both the definition of hyperplane location and affection of the generalization ability. By keeping the key sample points, SVM achieves better generalization performance while ensuring training errors.

Q-learning method is an important machine learning method in reinforcement learning. It is similar to other methods in reinforcement learning in that it learns the optimal policy for its dynamic system by perceiving the state information of the environment and interacting with the environment continuously, and it improves the behavior of the system by interacting with the environment through trial and error methods, and it is a kind of on-line learning method that can be applied to real-time environments. Therefore, it has been widely researched in the fields of intelligent control, machine learning, and so on.

The Q learning system views learning as a process of trial and error, and its basic model is shown in Fig. 5.

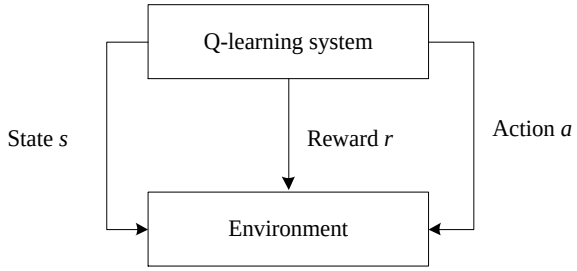


Fig. 5. Basic model of Q-learning.

In the learning process, the Q-learning system selects an action a to act on the environment, the environment receives the action and changes, and at the same time produces a reinforcement signal r which is fed back to the learning system, which then selects the next action according to the reinforcement signal and the current state of the environment, the principle of selection is to make the probability of receiving a positive reward increase. The chosen action affects not only the immediate reinforcement value, but also the next state of the environment and the final reinforcement value.

From the Q learning system, it is known that the set of states in the Q learning system is discrete and finite in general, but often in some practical situations, the set of state variables of the environment is large-scale or continuous, resulting in a difficult learning process and a slow response speed of the system. To solve this problem, this study introduces Q-learning to improve

SVM, where Q-learning is used to preprocess and extract features, and then SVM is used for classification. It not only captures the dynamic variation of time-varying features, but also reduces the complexity of SVM when processing high-dimensional data.

Q-learning improves the behavior of the system by constantly interacting with the environment through trial-and-error methods to find a strategy that maps the state of the environment into the corresponding action. This strategy of reinforcement learning inevitably increases the difficulty of learning the system, resulting in very much blind learning, and leading to longer learning times.

To overcome these problems, the features optimized by Q-learning are passed to the SVM as input to make efficient classification decisions, thus improving the overall system performance. For large-scale systems with reinforcement learning , the introduction of SVM not only reduces the complexity due to solving quadratic programming problems in standard support vector machines, but also for improving the convergence speed of large-scale reinforcement learning systems, Q-learning improved SVM as shown in Fig. 6.

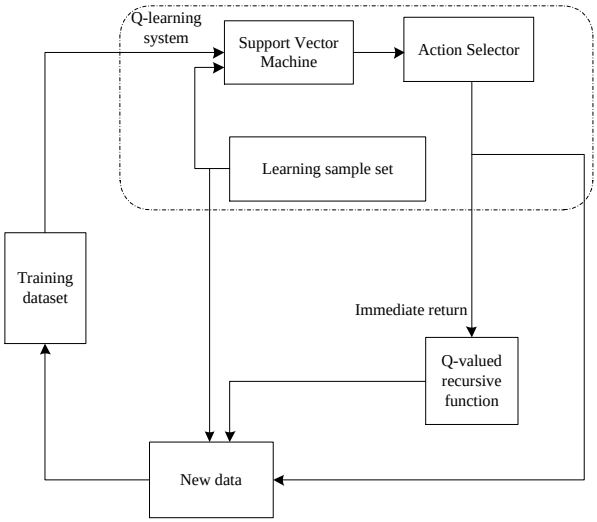


Fig. 6. Q-learning improved SVM structure.

The aim of the Q-learning algorithm shown as Eq. (15) and Eq. (16) is to find a policy π such that the value $Q^{\pi}(s)$ of each state s reaches the end at the same time, that is, it tries to find a policy $\pi : s \to a$ that is able to maximize the value of each state [22].

$$Q^{\pi}(s,a) = E\{r_1 + \gamma r_2 + \cdots + \gamma^{i+1} r_i + \cdots \mid s_0 = s, a_0 = a\} \tag{15}$$

$$Q^{\pi}(s,a) = \max_{\pi}(Q^{\pi}(s,a)) \tag{16}$$

where, r_i denotes the immediate reward at the moment t, $\gamma(\gamma \in [0,1])$ denotes the decay coefficient, and $Q^*(s,a)$ denotes the optimal value function. The corresponding optimal strategy as in Eq. (17):

$$\pi^* = \arg\max Q^*(s,a) \tag{17}$$

One of the simplest forms of Q-learning is single-step Q-learning with a modified consensus of Q-values as shown in Eq. (18) [23]:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \beta\left[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right] \tag{18}$$

where, β denotes the learning rate and γ denotes the discount rate.

Given a training set $S = (y_i, P_i)$, where y_i denotes the input vector and P_i is the output classification objective value. Determine an optimal function $F(x)$ as the objective of the regression problem such that $F(x)$ can correctly regress the unknown input vector with the highest possible probability. In SVM, the regression function $F(x)$ has the form as in Eq. (19) [24]:

$$F(x) = \upsilon^T \delta(x) + b \tag{19}$$

where, $\delta(.)$ denotes the mapping from the input space to the feature space, υ and b denote the coefficient vector and the deviation term respectively, which are the quantities to be sought.

The unknown quantity can be determined by Eq. (20) optimization problem [25]:

$$\begin{aligned} &\min_{\upsilon,b,\varepsilon} Q(\upsilon,b,e) = \frac{1}{2}\|\upsilon\|^2 + \frac{\gamma}{2}\sum_{i=l}^{l} e_i^2 \\ &s.t. y_i = \upsilon^T \delta(x_i) + b + e_i \end{aligned} \tag{20}$$

Its Lagrangian function as in Eq. (21):

$$L(\upsilon,b,e,\beta) = Q(\upsilon,b,e) - \sum_{i=1}^{l} \beta_t\left[\upsilon^T \delta(x_i) + b + e_i - y_i\right] \tag{21}$$

The KKT condition for the above equation is shown in Eq. (22) [26]:

$$\begin{cases} \dfrac{\partial L}{\partial \upsilon} = 0 \Rightarrow \upsilon - \sum_{i=1}^{l} \beta_i \delta(x_i) = 0 \\ \dfrac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=0}^{l} \beta_i = 0 \\ \dfrac{\partial L}{\partial e_i} = 0 \Rightarrow Ce_i - \beta_i = 0 \\ \dfrac{\partial L}{\partial \beta_i} = 0 \Rightarrow \upsilon^T \delta(x_i) + b + e_i - y_i = 0 \end{cases} \tag{22}$$

Writing Eq. (22) in matrix form and eliminating υ and e, Eq. (23) is obtained:

$$\begin{bmatrix} 0 & \vec{1}^T \\ \vec{1} & \Omega+\gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \beta \end{bmatrix}=\begin{bmatrix} 0 \\ y \end{bmatrix} \tag{23}$$

where, $\Omega_{ij}=k(x_i,x_j)$. A regression function completes the classification of multi-sensor data for medical monitoring by solving Eq. (24):

$$F(x)=\sum_{i=1}^{l}\beta_i K(x,x_i)+b \tag{24}$$

VI. Experimental Validation

Test experiments are conducted to validate the effectiveness of the proposed reinforcement learning based improved SVM for multi-sensor data classification processing in medical monitoring.

A. Experimental Data

A schematic diagram of the medical monitoring multi-sensor arrangement is shown in Fig. 7.



Fig. 7. Schematic diagram of multi-sensor arrangement for medical monitoring.

Here, suitable types of sensors, including heart rate sensors or blood pressure sensors, are selected based on monitoring needs and placed at specific locations on the patient's body to accurately collect physiological parameters and health data. Subsequently, the data collected by the sensors is transmitted to the monitoring device or data system for processing and storage.

The multi-sensor data included ECG data (heart rate, heart rate), blood pressure data (including systolic and diastolic blood pressure), oxygen data (e.g., oxygen saturation), temperature data, respiratory data (respiratory rate), exercise data (number of steps, intensity of activity), including blood glucose data, and pulse waveform data. Examples of selected samples are shown in Table I.

TABLE I. Examples of Selected Samples

Data Type	Numerical Value
heart rate	75 bpm
pulse rate	normalcy
systolic blood pressure	120 mmHg
diastolic blood pressure	80 mmHg
oxygen saturation	98 per cent
(body) temperature	36.5°C
respiratory rate	16 times/minute
number of steps	8000 steps
Activity intensity	medium
blood sugar	5.0 mmol/L

In this setting, the parameter configurations of Q- learning and SVM model are planned in detail, and the specific parameter settings are as follows:

1) The state space of Q-learning algorithm selects parameters in multi-sensor data as characteristics, such as heart rate and systolic blood pressure. The action space includes the selection of different SVM improvement methods.

2) Setting the learning rate of Q-learning is 0.1 to control the updating speed of the Q value; the discount rate is 0.9 to balance the importance of current and future rewards; the exploration rate gradually decreases from 0.9 to 0.1 to promote the algorithm's balance between exploration and utilization.

3) Setting the initial regularization parameter C of SVM is 1.0, and the kernel function is RBF (radial basis function). The space is reserved for adjusting other parameters through cross-validation.

The above parameter configuration aims to improve the repeatability of the experiment and the reliability of the results.

B. Experimental Program

Based on the above collected data, using the signal-to-noise ratio, data feature extraction effect, and Kappa coefficient as indicators, the method of this study is compared and validated with the data classification method based on the concept of drift, and the classification method based on mathematical modelling and improved online learning.

1) Signal-to-noise ratio. The signal-to-noise ratio is the ratio of the strength of the useful signal to the strength of the noise signal and is used to measure the relative strength of the signal in relation to the noise. A high signal-to-noise ratio indicates that the useful signal is stronger relative to the noise signal, indicating a higher quality signal and clearer information.

2) Data feature extraction effect. Data feature extraction effect refers to the meaningful features that can characterize the relevant information of the signal extracted from the original data by feature extraction methods.

3) Kappa coefficient. The Kappa coefficient measures the accuracy of the classifier while considering the random factor in classification. The closer the Kappa coefficient is to 1, the better the performance of the classifier.

C. Analysis of Experimental Results

1) Signal-to-noise ratio. The signal-to-noise ratio is a measure of the relative strength between signal and noise. By performing a comparative signal-to-noise ratio test, the quality of different data can be assessed, and it can be determined which data has a higher signal quality and is more reliable. Signal-to-noise ratio is one of the key factors in signal classification and processing. Lower signal-to-noise ratios may lead to interference and distortion of the signal by noise,

reducing the performance and accuracy of the classification algorithm. By performing a signal-to-noise ratio comparison test, data with lower signal-to-noise ratios can be identified, so that methods such as appropriate preprocessing or adjusting parameters can be taken to improve classification performance. The signal-to-noise ratio results of the three methods are shown in Fig. 8.

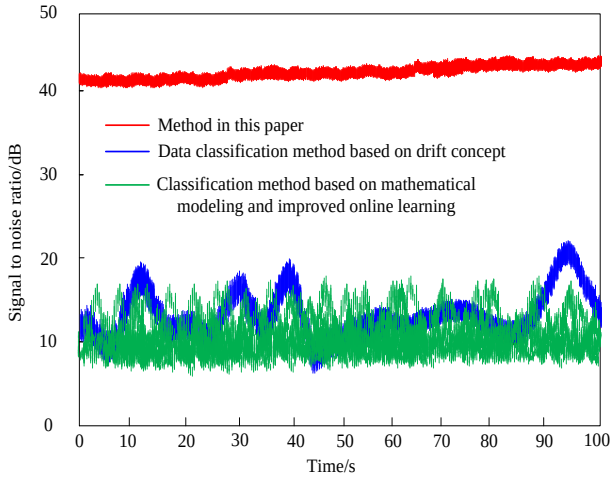


Fig. 8. Signal-to-noise ratio results.

Observing the results of the signal-to-noise ratio shown in Fig. 8, the signal-to-noise ratio of this study's method is always higher and smoother compared to the two literature comparison methods, and the signal-to-noise ratio of this study's method is always above 40 dB. On the other hand, the signal-to-noise ratio of the data classification method based on the concept of drift and the classification method based on mathematical modelling and improved online learning not only fluctuates more, but also has a lower value, with a maximum of only about 25 dB. Therefore, it shows that the method in this study can effectively remove the noise from medical monitoring multi-sensor data and improve the quality of the data.

The main reason is that the proposed method removes the noise in medical monitoring data effectively and improves the signal-to-noise ratio by combining a DBSCAN algorithm with DWT.

2) Effectiveness of data feature extraction. Feature extraction is a key step in converting raw signals into meaningful feature vectors. The performance and effectiveness of different feature extraction methods can be evaluated by verifying the feature extraction effect. Comparing the usefulness and differences of features extracted by different methods for signal classification helps to select the most suitable feature extraction method to maximize the accuracy and robustness of signal classification. The data feature extraction results of the three methods are shown in Fig. 9.

As shown in Fig. 9, the data feature extraction results of this study's method are closest to the trend of the original data, indicating that the data features extracted by this study's method are most capable of describing the original data. And the data feature extraction results of the data classification method based on the drift concept differ greatly from the original data. Therefore, it shows that the method in this study can accurately extract the features of medical monitoring multi-sensor data.

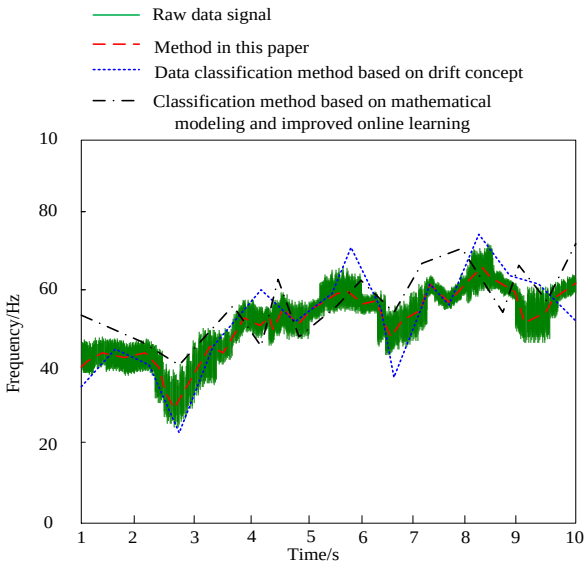


Fig. 9. Effect of data feature extraction.

This is mainly because the convolutional neural network is used in this method, which makes it more accurate in feature extraction and can capture more useful information.

3) Kappa coefficient. The accuracy and consistency of the classifiers in multi-sensor data classification can be objectively assessed by validating the classification results with Kappa coefficients. The Kappa coefficients take into account the random assignment of the classified objects, and can accurately measure the superiority of the classifiers with respect to the classification results that are only caused by the random selection. The Kappa coefficient test results of the three methods are shown in Table II.

TABLE II. KAPPA COEFFICIENT

Test serial number	Kappa coefficient		
	The method in this study	Data classification method based on the drift concept	Classification method based on mathematical modeling and improved online learning
10	0.96	0.71	0.76
20	0.98	0.68	0.74
30	0.95	0.69	0.62
40	0.97	0.72	0.64
50	0.96	0.73	0.74
60	0.96	0.67	0.65
70	0.97	0.71	0.61
80	0.98	0.70	0.63
90	0.96	0.69	0.65
100	0.95	0.66	0.71

From the data in the table, it can be observed that the Kappa coefficient of this study's method stays at a high level with small changes in many tests. On the other hand, the Kappa coefficients of the other two methods have large fluctuations and their performance is not stable enough. For example, the Kappa coefficient of this study's method stays above 0.95 under different test numbers, while the Kappa coefficients of the other two methods fluctuate in a wide range, from 0.62 to 0.76. The Kappa coefficient of this study's method is significantly higher than that of the other two methods under most of the test sequence numbers, especially the highest Kappa coefficient is achieved at many nodes such as test sequence numbers 20, 40 and 80. This indicates that the method in this study has a significant advantage in classification accuracy and can classify the data more reliably for judgement.

This is because the improvement of Q-learning on SVM makes the model more adaptable to complex and changeable medical data through automatic exploration and optimization of parameters, thus improving the accuracy and robustness of the classification.

VII. Conclusion

In this study, a method combining Q-learning and SVM was proposed to deal with the problem of multi-sensor data classification in medical monitoring. Through experimental validation, the new method exhibited better performance relative to the traditional approach. Specifically, DBSCAN clustering and DWT were used to preprocess data, which effectively retained key features and reduced noise interference. Then, CNN was used to deeply mine signal features. The Q-learning algorithm was used to optimize SVM model and realize efficient mapping of features to classification output.

Experimental results show that the proposed method not only reaches a high level of 98% on classification accuracy, but also observes significant progress in terms of classification accuracy, signal-to-noise ratio, and feature extraction effectiveness. In addition, this study explores the role of reinforcement learning in the process of feature selection and model optimization, and it is found that the reinforcement learning algorithm can be used to better guide the model for tuning and improve the effectiveness of the classification task.

However, this study still has certain limitations, and the robustness of feature extraction in current methods needs to be further improved when dealing with sudden noise or multi-source heterogeneous sensor data. And the convergence speed of the Q-learning algorithm in high-dimensional state space is slow, which may affect the application efficiency of real-time monitoring scenarios.

In the future, this study will further develop a distributed reinforcement learning framework to optimize the parallel processing efficiency of multi-sensor data, in order to meet the real-time monitoring needs of clinical practice. For example, introducing Deep Q-Network (DQN) combined with Convolutional Neural Network to process high-dimensional sensor data, utilizing experience replay and target network mechanisms to enhance algorithm stability. And adopting the Actor Critic framework, while optimizing the strategy function and value function, to more efficiently handle the temporal dynamic characteristics of medical monitoring data.

Acknowledgment

The authors are thankful for the support from the Key Scientific Research Project of Higher Education Institutions in Henan Province (grant no.24B 510006).

References

[1] Kapoor B, Nagpal B, Alharbi M. Secured healthcare monitoring for remote patient using energy-efficient IoT sensors. Computers & Electrical Engineering, 2023, 106: 108585.

[2] Wang S, Nayak D R, Guttery D S, et al. COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis. Information Fusion, 2021, 68: 131-148

[3] Kilinc D, Gel E S, Sir M Y, et al. Statistical characterization of patient response to offered access delays using healthcare transactional data. Naval research logistics, 2022, 69(7): 974-995.

[4] Xiao Y, Yin H, Duan T, et al. An Intelligent prediction model for UCG state based on dual-source LSTM. International Journal of Machine Learning and Cybernetics, 2021, 12: 3169-317

[5] Ping S , Bishop J L .Evaluation of a novel salaried medical officer position on service provision and performance at a rural health service: An exploratory mixed-methods study. Australian Journal of Rural Health, 2022, 30(1):65-74.

[6] Liu X, He J, Song L, et al. Medical Image Classification based on an Adaptive Size Deep Learning Model. ACM Transactions on Multimedia Computing Communications and Applications, 2021, 17(3S): 102.

[7] Hasib K M, Towhid N A, Islam M R. HSDLM: A Hybrid Sampling With Deep Learning Method for Imbalanced Data Classification. International journal of cloud applications and computing, 2021, 11(4): 1-13.

[8] Singh M N, Khaiyum S. Enhanced data stream classification by optimized weight updated meta-learning: continuous learning-based on concept-drift. International journal of web information systems, 2021, 17(6): 645-668.

[9] Kenger M N , Ozceylan E . A hybrid approach based on mathematical modelling and improved online learning algorithm for data classification. Expert Systems with Applications, 2023, 218: 119607.

[10] Zhai J, Qi J, Shen C. Binary imbalanced data classification based on diversity oversampling by generative models. Information Sciences, 2022, 585: 313-343.

[11] Liang X, Yu Q, Zhang K, et al. LapRamp: a noise resistant classification algorithm based on manifold regularization. Applied Intelligence, 2023, 53(20): 23797-23811.

[12] Jain P, Bajpai M S, Pamula R. A Modified DBSCAN Algorithm for Anomaly Detection in Time-series Data with Seasonality. The international Arab journal of information technology, 2022, 19(1): 23-28.

[13] Batzelis E, José M B, Toledo F J, et al. Noise-Scaled Euclidean Distance: A Metric for Maximum Likelihood Estimation of the PV Model Parameters. IEEE journal of photovoltaics, 2022, 12(3): 815-826.

[14] Liu S, Lu M, Liu G. A Novel Distance Metric: Generalized Relative Entropy, Entropy, 2017, 19(6): 269

[15] Bittner D , Engel M , Wohlmuth B ,et al. Temporal Scale-Dependent Sensitivity Analysis for Hydrological Model Parameters Using the Discrete Wavelet Transform and Active Subspaces. Water Resources Research, 2021, 57(10), 2020WR028511.

[16] Fu W, Luo Z, Liu S, et al. Spatiotemporal correlation based self-adaptive pose estimation in complex scenes, Digital Communications and Networks, 2024, online first, 10.1016/j.dcan.2024.03.007

[17] Jahanbanifard M , Price E , Gonzalez B A ,et al.A novel method to analyse DART TOFMS spectra based on Convolutional Neural Networks: A case study on methanol extracts of wool fibres from endangered camelids[J].International journal of mass spectrometry, 2023. DOI:10.1016/j.ijms.2023.117050.

[18] Zamli K Z , Din F , Alhadawi H S . Exploring a Q-learning-based chaotic naked mole rat algorithm for S-box construction and

optimization[J].Neural computing & applications, 2023, 35(14):10449-10471.

[19] Akin E , Demir K , Yetgin H . Multiagent Q-learning based UAV trajectory planning for effective situational awareness[J].Turkish Journal of Electrical Engineering and Computer Sciences, 2021, 29(5).DOI:10.3906/ELK-2012-41.

[20] Du H, Yu S. Dynamic Obstacle Avoidance for Service Robots Based on Spatio-Temporal Graph Attention Network. Computer Engineering, 2024, 50(2): 105-112.

[21] Delilbasic A , Saux B L , Riedel M ,et al.A Single-Step Multiclass SVM based on Quantum Annealing for Remote Sensing Data Classification[J].ArXiv, 2023, 17: 1434-1445.

[22] Delilbasic A, Le Saux B, Riedel M ,et al.A single-step multiclass SVM based on quantum annealing for remote sensing data classification. IEEE journal of selected topics in applied earth observations and remote sensing. 2023, 28(17): 1434-1445.

[23] Lian B, Xue W, Lewis FL,et al. Inverse Q-learning using input–output data. IEEE Transactions on Cybernetics. 2023, 54(2):728-738.

[24] Song S, Pan L, Liu S. A Q-learning based auto-scaling approach for provisioning big data analysis services in cloud environments. Future Generation Computer Systems. 2024, 154: 140-150.

[25] Wang J, Mi X, Shen H ,et al. Optimal Control for Interconnected Multi-Area Power Systems With Unknown Dynamics: An Off-Policy Q-Learning Method. IEEE Transactions on Circuits and Systems II: Express Briefs. 2023, 71(5): 2849-2853.

[26] Charvadeh YK, Yi GY. Understanding Effective Virus Control Policies for Covid-19 with the Q-learning Method. Statistics in Biosciences. 2024, 16(1): 265-289.

Developing an Ontology-Driven and Governance-Integrated Method for Information Dashboard Design

Ahadi Haji Mohd Nasir[1], Nik Habibullah Nik Mohd Nizam[2*],
Mohd Khairul Maswan Mohd Redzuan[3]*, Mohammad Nazir Ahmad[4]
Institute of Visual Informatics, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia[1, 2, 3, 4]
Infrastructure University Kuala Lumpur (IUKL), Kajang, Selangor, Malaysia[4]

***Abstract*—Despite the increasing reliance on information dashboards across industries, dashboard design practices remain fragmented, lacking standardized methodologies, ontological formalization, and governance integration. Addressing these gaps, this study develops a method to guide dashboard design by embedding ontological modeling and Information Governance (IG) principles. Two complementary artifacts are proposed: the Information Dashboard Design Ontology (IDDO) and the Information Dashboard Design Method (IDDM) Canvas. Using Design Science Research Methodology (DSRM) and a Unified Ontological Approach (UOA), IDDO formalizes tacit dashboard design knowledge into a structured framework, while the IDDM Canvas operationalizes this ontology into a practical design tool. Validation through the Ontological Unified Modeling Language (OntoUML) Plugin and conceptual assessment based on Unified Foundational Ontology (UFO) principles confirmed internal consistency and ontological soundness. The resulting framework integrates twelve dashboard design building blocks with eight IG principles to ensure rigor and governance alignment. The application of the IDDM Canvas demonstrated its utility in facilitating structured, replicable dashboard development. While the evaluation focused primarily on conceptual validation, future studies are recommended to empirically assess the framework's practical effectiveness across various domains and real-world projects.**

Keywords*—*Information dashboard; ontological modelling; information dashboard design ontology (IDDO); information dashboard design method (IDDM) canvas; information governance (IG); unified ontological approach (UOA); design science research methodology (DSRM)

I. Introduction

Information dashboards have become essential tools for organizational decision-making, enabling users to monitor, analyze, and interpret complex data in real-time [1]. Beyond their traditional role as visualization interfaces, dashboards now function as integrated components of information ecosystems, bridging disparate systems and enhancing decision support across domains. Their widespread adoption spans over industries such as healthcare [2], [3], finance [4], urban mobility management [5], and public health [6].

Despite their prevalence, dashboard development practices remain fragmented and lack a standardized methodology. Poorly designed dashboards may lead to cognitive overload, inconsistent data representation, and misinformed decisions [7], [8]. Moreover, many existing approaches neglect the integration of governance principles, which can compromise data quality, reliability, and compliance [9], [10], and hinder organizational learning and adaptation to digital transformation [11]. Without a consistent framework that incorporates governance, organizations face challenges in aligning dashboards with strategic and operational needs. Additionally, dashboard design often relies on implicit expertise held by experienced practitioners, making it difficult to replicate or scale across teams. The absence of a formalized design method limits reuse and consistency across different contexts. While ontology has shown promise in structuring complex design logic within business modeling [12] and learning systems [13], its application in dashboard design methodology remains limited.

This study aims to address three core challenges in dashboard methodology: 1) the lack of a standardized, transferable design method; 2) limited use of ontological approaches to formalize design elements; and 3) the absence of governance integration in the design process. In response, we adopt Design Science Research Methodology (DSRM) to develop the Information Dashboard Design Ontology (IDDO) and the Information Dashboard Design Method (IDDM) Canvas. IDDO provides a structured ontological framework for modeling key dashboard elements, while the IDDM Canvas offers a practical guide for applying this framework in real-world design initiatives.

Inspired by the transformation of the Business Model Ontology (BMO) into the widely adopted Business Model Canvas (BMC), this study translates IDDO into a usable design tool through the IDDM Canvas. This approach bridges conceptual rigor with design applicability, offering a method that supports governance compliance, visual consistency, and broader usability, including by non-specialist users.

The remainder of this study is organized as follows: Section II reviews the literature on dashboard design, ontological modeling, and governance integration. Section III outlines the research methodology. Section IV details the development of IDDO. Section V presents the IDDM Canvas. Section VI discusses the conclusion which includes limitations, and future research directions.

II. Literature Review

A. Challenges in Dashboard Design and the Need for Standardization

The growing body of research on information dashboards highlights their increasing role in enhancing decision-making through structured visual representation of data. Sarikaya et al.

*Corresponding Author.

[1], emphasize how organizations are increasingly relying on dashboards to transform complex datasets into actionable insights. However, despite substantial academic and industry interest, there remains a lack of formalized, standardized methodologies for dashboard development, indicating the need for more structured design approaches. Recent studies echo these concerns, particularly in sentiment-oriented and real-time dashboards that highlight interactive and data-driven needs without standardized models [58], [59].

As information systems evolve, dashboard development encounters several design-related and implementation challenges. Poor design has been linked to cognitive overload [7], reduced user engagement, and diminished decision quality [8]. These problems are exacerbated by the absence of a methodological framework that can capture and structure implicit design knowledge into standardized and reusable models. Without formalization, inconsistencies in layout, interaction, and data interpretation often emerge, leading to interoperability and usability issues. The emergence of AI-powered dashboards also illustrates this gap, where intelligent systems are deployed without governance grounding, resulting in unpredictable behavior or user distrust [60].

These challenges emphasize the need for a governance-integrated framework that promotes standardization while remaining adaptable across contexts. In practice, dashboards often serve as boundary-spanning tools between business and technical domains, requiring design methods that support both consistency and flexibility [14]. While some frameworks have shown domain-specific success, they frequently lack ontological structure and governance alignment, limiting broader application.

A synthesis of twelve recent studies (2019 to 2024), selected across healthcare, public service, urban management, and other sectors, reveals wide variation in dashboard objectives, design approaches, and governance considerations. Although many offer novel visualization strategies, they remain context-specific and difficult to generalize. This highlights the need for methodologies that support the formalization and standardization of dashboard design processes. These studies were chosen to reflect cross-sector practices, and recent publications were prioritized to capture emerging design trends and evolving best practices [59], [60].

Inspired by Osterwalder's methodology in developing the BMO [12], which successfully transformed business modeling logic into structured, transferable artefacts, this study applies a similar logic to dashboard design. It synthesizes design elements across domains to inform the development of a governance-integrated, ontology-driven dashboard design framework. The review indicates that while healthcare [3], [2] and public service [6], [15] applications dominate dashboard usage, each domain applies distinct development methods. For example, dashboards in healthcare often prioritize real-time clinical monitoring [3], [2], while those in tourism or urban management emphasize geospatial interaction and user experience [16], [15]. This diversity demonstrates the adaptability of dashboards but also highlights variations in design structures and methodological rigor across sectors.

In terms of methodological approaches, the studies span a wide range from user-centered and iterative designs to data-driven and metric-based models [7], [9], [13]. Some emphasize prototyping and stakeholder feedback [6], [4], while others rely on cognitive task analysis and real-time contextual inputs [15]. Although these approaches offer valuable insights, few provide a repeatable structure that can be generalized or scaled. This points to a pressing need for unified design frameworks that can bridge ad-hoc development with structured reusability.

Design principles also vary significantly. While some dashboards focus on tracking KPIs and trends using minimal layout structures [4], [17], others employ interactive and multi-layered design systems to serve diverse user groups [6], [5]. However, many studies underplay the importance of design consistency, hierarchy, and accessibility - factors that critically affect usability and user trust [8]. This variability in design emphasis further reinforces the case for standardized design methods that balance domain specificity with universal design principles.

Table I summarizes the key studies across domains, revealing the broad diversity in dashboard development approaches. The review demonstrates sectoral adaptability but simultaneously exposes the absence of standardized structures across organizational contexts. Recent data-driven approaches, for instance, prioritize automation and real-time responsiveness but lack a governance-aware foundation, further illustrating the gap [60].

To address the limitations identified in existing literature, this study proposes a new framework – the IDDO, which responds directly to the gaps. Existing dashboard design frameworks, while offering domain-specific solutions, exhibit fundamental limitations that make them unsuitable for systematic, governance-aligned development across organizational contexts. These frameworks rely on ad-hoc approaches without formal structure, lack ontological formalization, preventing knowledge systematization, and treat governance as external considerations rather than integrated design components [7], [8], [9], [10].

The proposed IDDO framework addresses these limitations by providing: 1) a formal ontological structure enabling systematic knowledge transfer, 2) embedded governance integration ensuring compliance by design, and 3) practical operationalization through the IDDM Canvas that translates complex structures into actionable design methodologies. This approach makes IDDO particularly appropriate for organizations requiring standardized, repeatable dashboard development processes.

This reinforces the need for a formalized modeling foundation to promote consistency, facilitate transferability, and embed governance-aware design practices. Building on these insights, the following sub-sections explore two critical dimensions essential for establishing a standardized dashboard development method: 1) ontological modeling and 2) Information Governance (IG).

TABLE I. A REVIEW OF RECENT STUDIES ON INFORMATION DASHBOARD DESIGN (2019-2024) DISCUSSED IN SECTION II

Study	Application Domain	Methodology	Key Phases	Design Principles	Key Features	Key Study Outcome
Poppe et al. [3]	Neonatal Intensive Care	User-centered design	1. Interviews with clinicians 2. Dashboard prototyping 3. Clinical evaluation	• Combine detailed and summarized data • Enable real-time monitoring • Support decision-making • Ensure ease of understanding	• Real-time monitoring • Visual trends • Clinical indicators	Web-based oxygenation dashboard for monitoring preterm neonates
Pamuk & Schumann [4]	Financial Services	Design Science Research	1. Data understanding 2. Data preprocessing 3. Model development Model evaluation	• Simplicity: Show only essential information • Customization: Allow changes to the layout • Interactive design • Transparency and security • Regulatory compliance	• AI model creation • Performance visualization • Comprehensive reporting	AI dashboard "AIDash" for credit assessment model management
Salvi et al. [6]	Public Health	User-centered design	1. User needs assessment 2. Data integration 3. Visualization design 4. Dashboard development 5. User evaluation	• User-centered design • Data privacy and security • Actionable insights • Intuitive visualizations • Timeliness	• Overdose touchpoint visualization • Interactive visualizations • Data integration	Real-time dashboard for enhancing overdose prevention efforts
Conrow et al. [5]	Mobility Data	Iterative user-centered design	1. Identify purposes/users 2. Specify data implementation 3. Create trajectory model 4. Build with user input 5. Involve users at all stages	• User-centered design • Interactive design • Visual design for effective communication • Data model development	• Indicators and maps • Interactive filters • Dynamic features	Conceptual framework for big mobility data dashboards
Bach et al. [8]	Cross-industry	Systematic analysis	1. Define data selection 2. Establish structure 3. Design pages 4. Add interactivity 5. Select color scheme	• Optimize screen space • Use color purposefully • Provide context for data • Consider different audience needs • Accessibility considerations	• Visual display patterns • Data exploration • Information presentation	Comprehensive set of 64 dashboard design patterns
Wu et al. [7]	Data Visualization	Deep-learning method	1. Define purpose/questions 2. Select data sources 3. Choose chart types 4. Design layout 5. User testing	• Foster human computer collaboration • Balance automation with user control	• Automated visualization • Analytical insights	Deep learning algorithm for analytical dashboard assistance
Pestana et al. [2]	Healthcare	Design science research	1. Research and ETL 2. Initial proposal 3. Demonstration/evaluation 4. Feedback implementation	• Gestalt visual perception principles • Proximity, similarity, enclosure • Visual hierarchy for KPIs	• KPI monitoring • Performance comparison • Trend visualization	Web-based hospital productivity management dashboard
Balletto et al. [16]	Tourism	Convergent parallel design	1. Define framework 2. Define layout 3. Analyze characteristics 4. Collect user content 5. Develop dashboard	• Highlight points of interest • Provide detailed travel information • Intuitive geographical representation	• Walkability index • Attractiveness index • Maps and charts	Dashboard for slow tourism in green infrastructures
Young & Kitchin [15]	Urban Management	Cognitive Task Analysis	1. Establish purpose 2. Identify audience 3. User research 4. Create personas 5. Content strategy	• Integration of HCI research • User-centered civic design • Multi-platform compatibility	• Multiple visualization types • Real-time displays • Interactive features	Design guidelines for building user-centered city dashboards
Elshehaly et al. [9]	Healthcare Quality	Metric Specification Structure	1. Investigate challenges 2. Healthcare interviews 3. Design metric cards 4. Create specification 5. Deploy and enhance	• Task sequence optimization • Metric definition standardization • Visual feature consistency • Adaptable dashboard structure	• Quality metrics visualization • Adaptive displays • Metric structures	QualDash for healthcare quality improvement metrics
Ahdan et al. [17]	Energy Management	IoT implementation process	1. Assemble tools 2. Build prototype 3. Develop application 4. Test and rebuild	• Flexibility in control • Efficient energy usage • User-friendly interfaces	• Real-time monitoring • Device control • Mobile accessibility	IoT-powered mobile dashboard for energy monitoring

Kokoç & Altun [13]	E-learning	Iterative design	1. Reorganize database 2. Plan visualizations 3. Design dashboard 4. Expert feedback	• Simple and comprehensible interface Learning progress visualization	• Learning behavior visualization • Predictive modeling	Prescriptive learning dashboard for performance improvement

B. Ontological Approaches in Dashboard Design

Ontological approaches have shown strong potential in formalizing complex structures within information systems, enabling standardized representation and promoting semantic interoperability. While widely applied in domains such as business modeling, particularly through frameworks like the BMO [12], their adoption in dashboard design remains limited.

As Eppler and Burkhard [14] argue, the development of structured visualization tools necessitates a solid conceptual backbone to ensure clarity, coherence, and reusability. However, none of the twelve studies reviewed in this paper explicitly apply ontological methods in dashboard development. This absence underscores a methodological gap in current practices. Nevertheless, recent implementations in fields such as mobility analytics [5] and educational dashboards [13] demonstrate more structured design practices, suggesting the untapped potential of ontological modeling to enhance consistency, scalability, and transferability across interface designs.

Formalizing design knowledge through ontological approaches offers several advantages, particularly in enabling standardization and aligning design logic with governance structures. Guizzardi's ontological foundations [36] offer methodological rigor that supports semantic precision and conceptual clarity in defining dashboard components. Similarly, recent work by [52] highlights how unified ontological frameworks can promote consistent modeling while maintaining flexibility for contextual adaptation.

By applying ontological principles, dashboard design can move beyond fragmented, ad-hoc development toward a more coherent and reusable foundation. This shift not only supports cross-domain implementation but also facilitates the embedding of governance requirements within the design structure itself, an aspect further explored in the following sub-section on IG.

C. IG and Dashboard Design Reliability

IG offers a comprehensive set of standards and practices to ensure that information assets are managed with accuracy, security, and accountability [10]. However, a review of the twelve dashboard studies conducted in this study reveals that none explicitly incorporate IG principles into their design methodologies. While some acknowledge compliance-related concerns or data accuracy issues, governance integration is often implicit or entirely absent from the design framework.

In the context of dashboard design, integrating IG principles helps organizations align their visualization tools with broader compliance and data integrity goals. As noted by Datta et al. [18], IG frameworks play a pivotal role in maintaining the trustworthiness of information flows within decision-making systems. Organizations with established governance mechanisms also exhibit stronger information standardization across units [19]. Conversely, the lack of governance integration can result in data inconsistencies, increased risk exposure, and reduced decision reliability [20], [21]. Embedding IG principles as part of dashboard methodologies could ensure that design outputs maintain regulatory compliance and reflect high information quality [22], [23].

D. Gaps in Current Approaches and the Need for a Standardized Methodology

The synthesis of dashboard design literature highlights three persistent challenges that limit progress in establishing standardized dashboard development practices. First, most reviewed methods are highly domain-specific and lack the generalizability needed for cross-organizational adoption. Second, while ontologies have proven effective in other domains for formalizing and structuring design logic, none of the twelve studies reviewed incorporate them directly into dashboard methodologies. Third, despite growing awareness of governance importance, none of the studies examined explicitly integrate IG principles into the design process, resulting in potential risks to data reliability and organizational compliance.

These gaps emphasize the need for a standardized methodology that embeds both ontological modeling and governance structures into dashboard development. Such a framework would support methodological consistency, formalized design knowledge, and governance alignment throughout the lifecycle of dashboard creation. An ontology-driven approach that integrates IG principles potentially offers a scalable solution for organizations seeking to design dashboards that are not only functionally effective but also aligned with compliance standards, promoting greater usability, accountability, and organizational learning.

III. METHODOLOGY

This study adopts Design Science Research Methodology (DSRM) to guide the development of the IDDO and the IDDM Canvas. Rooted in Simon's foundational work, The Sciences of the Artificial [24], DSR offers a structured, artifact-oriented approach to addressing complex design problems. Originally formalized within information systems research [25], the methodology has gained traction across various applied domains, including dashboard development and business modeling [26], [27].

DSRM is particularly suited for structuring the development of purposeful tools, as demonstrated by successful applications like BMC [28]. In this study, the DSRM framework is complemented by the Unified Ontological Approach (UOA) [29], which serves as the ontology development methodology. The integration of UOA facilitates the formal modeling of dashboard design components by translating fragmented design logic into structured ontological representations that can be systematically reused across development scenarios.

The use of DSRM is justified by its effectiveness in guiding the structured development of design artefacts within information systems [30], [31], [32]. Its applicability to ontology-based design is further reinforced by recent studies on conceptual modeling frameworks [33]. Together, DSRM and

UOA form a robust methodological foundation for developing the IDDO conceptual framework and the IDDM Canvas, ensuring the resulting artefacts are valid, usable, and applicable across diverse dashboard design contexts.

A. Research Design for IDDO Development

This study focuses on the first three stages of the DSRM proposed by Peffers et al. [25], as illustrated in Fig. 1. The demonstration and evaluation involving dashboard mock-ups are reserved for future research, while this study represents the communication stage by detailing the conceptual and design artifacts developed for dashboard methodology enhancement.

Table II summarizes the research activities and outcomes associated with each implemented DSRM stage.

B. Framework Development and Evaluation

The IDDO was developed using an iterative, model-driven approach. IG principles in this study context were used as a guiding theoretical foundation and serve as the Kernel Theory, consistent with the Kernel Theory Fundamentalist perspective proposed by Gregor and Jones [35]. The adoption of a Kernel Theory reinforces the theoretical grounding of IDDO while supporting its practical application as a structured modeling tool. The structured development process was executed through four modeling and evaluation stages, as outlined in Table III. The structured modeling process establishes IDDO as a robust framework that integrates governance elements into the formal structure of dashboard design. By combining ontological rigor with compliance-oriented design considerations, the resulting framework enhances consistency, reusability, and adaptability across organizational contexts.

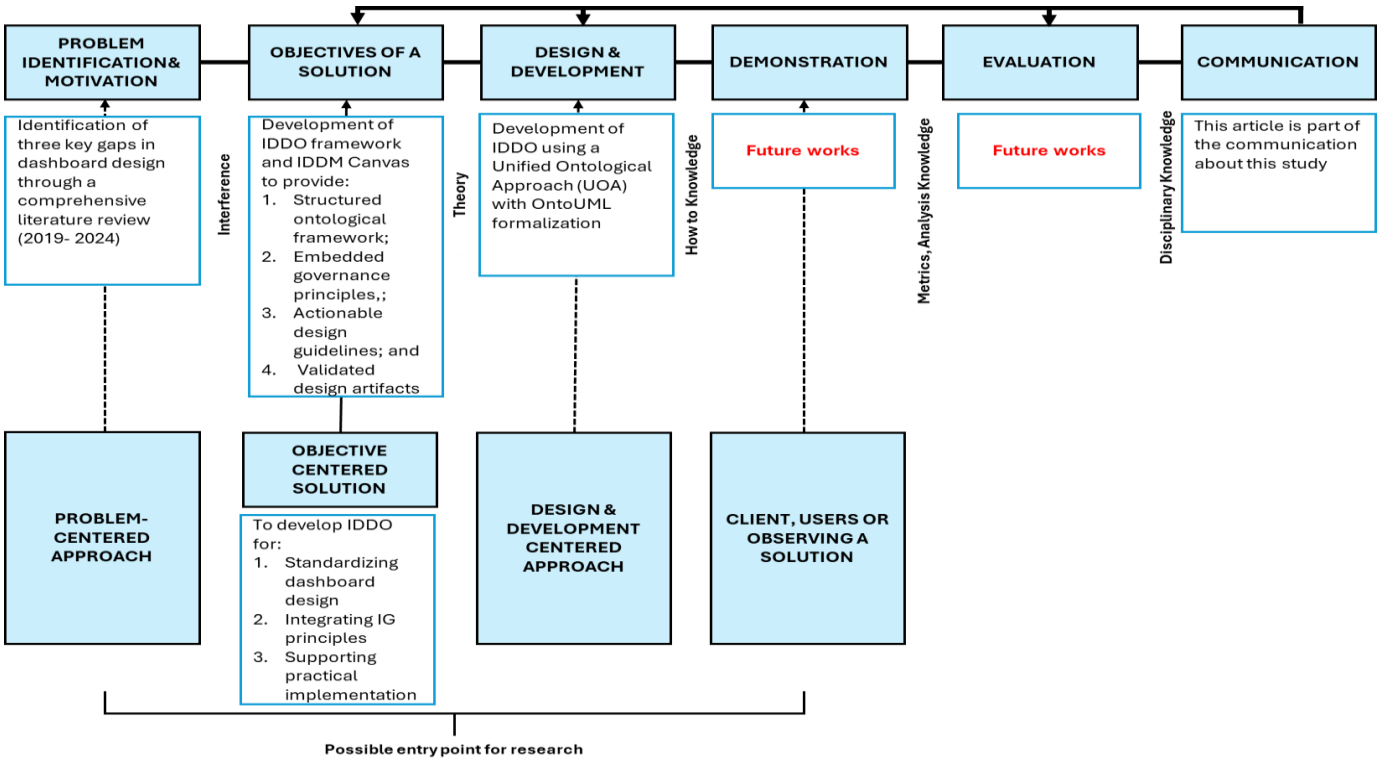


Fig. 1. DSRM implementation for IDDO development (focusing on the first three stages).

TABLE II. SUMMARY OF IMPLEMENTED DSRM STAGES

DSRM Stage	Research Activities	Outcomes
Problem Identification	Review of twelve recent studies (2019–2024)	Identified three major gaps: 1. Lack of standardized design methodology 2. Limited use of ontology 3. Absence of governance integration
Solution Objectives	Definition of design objectives and scope	Established four objectives: 1. Structured ontological framework 2. Embedded IG principles 3. Actionable design guidelines 4. Design artifact validation
Design & Development	Iterative modeling using the UOA	Developed two artifacts: 1. IDDO: an ontology-based dashboard design framework 2. IDDM Canvas: a practical dashboard design method.

TABLE III. IDDO DEVELOPMENT PROCESS

Stage	Activity	Description
UML-Based Modeling	Conceptual modeling in UML	Initial structuring of dashboard design elements using Unified Modeling Language (UML).
OntoUML-Based Refinement	Ontological formalization	Used OntoUML to enhance the semantic precision and clarity of design constructs [36].
OntoUML Plugin Validation	Syntactic model validation	Applied the OntoUML Plugin in Visual Paradigm to detect modeling inconsistencies.
Conceptual Evaluation with UFO	Ontological soundness assessment	Assessed the model against Unified Foundational Ontology (UFO) principles to verify semantic integrity.

The next section elaborates on the development of the IDDO and the formulation of the IDDM Canvas, which translates these components into a practical design methodology for dashboard development.

IV. DEVELOPMENT OF THE IDDO

The development of IDDO, the primary artifact and key contribution of this study, involves a systematic and rigorous process to establish a standardized framework for the information dashboard design method. This section details the methodical process of identifying essential building blocks, incorporating IG principles for information quality, and formalizing the ontological framework, laying the foundation for a governance-driven, standardized dashboard design methodology.

A. Establishing the Building Blocks

The building blocks of the IDDO are derived from a comprehensive analysis of recent works in dashboard design (2019-2024) as summarized in Table I earlier. Through a narrative review of twelve previous studies, key commonalities in design phases, visualization principles, and essential features were identified and summarized in Table IV.

These insights informed the derivation of twelve key building blocks within the IDDO framework. The elements were consistently observed across the reviewed studies and are conceptually organized as illustrated in Fig. 2.

This analytical process aligns with established ontology development practices, where recurring elements across multiple studies are consolidated to form structured frameworks [37], [12]. Similarly, Bernardo et al. [38] applied a systematic approach to identify critical components in the lifecycle of Business Process Management (BPM), demonstrating how synthesis of prior works can reveal design patterns suitable for standardization. Recent work by Leong et al. [57] further demonstrates the applicability of BMO-inspired approaches in developing ontologies for specialized dashboard contexts, such as virtual reality environments, reinforcing the transferability of ontological modeling principles across different visualization domains.

TABLE IV. KEY COMMONALITIES IN DESIGN PHASES, DESIGN PRINCIPLES, AND ESSENTIAL FEATURES

Commonalities		
Design Phase	***Design Principles***	***Essential Features***
• Problem identification • Design and modelling • Development and implementation • Testing and evaluation • Iterative process and feedback	• Reduce pages • Optimize screen space • Use color purposefully • User-friendly interfaces • Effective information visualization • Tailored visualization approaches • Ensure accessibility • Simplify the design	• Visual display of important knowledge • Interactive information visualization displays • Drill-down capabilities and knowledge exploration • Presentation of analytical knowledge insight

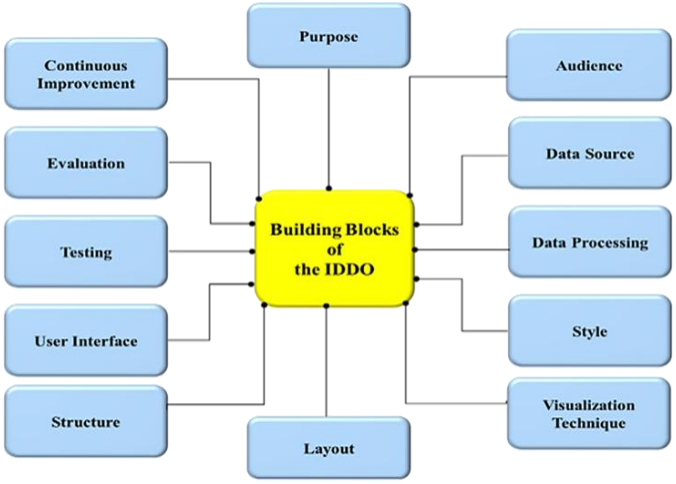


Fig. 2. Twelve key building blocks of the IDDO.

Building upon these approaches, IDDO integrates common patterns across dashboard design literature to promote semantic consistency, governance integration, and methodological standardization. The twelve building blocks illustrated in Fig. 2 were derived from these observed patterns and are summarized in Table V.

The derivation process was informed by the recurring elements identified in the reviewed studies (as shown in Table IV), but the resulting terminology was intentionally refined to ensure clarity, contextual relevance to IDDO, and ease of interpretation, while preserving the original intent of each component.

Drawing inspiration from the structure of the BMO developed by Osterwalder [12], which has been widely adopted for business model development, the IDDO building blocks are organized into three logical categories: Defining and Planning, Design and Visualization, and Testing and Improvement. This categorization reflects a structured and phased design approach, aligning closely with ontology development principles for information systems.

Table VI outlines how the twelve building blocks are distributed across these categories. The Defining and Planning category establishes the dashboard's foundation, encompassing Purpose, Audience, Data Source, and Data Processing. The Design and Visualization category focuses on interface elements that shape user interaction, while the Testing and Improvement category supports continuous refinement through validation and quality assurance measures.

TABLE V. DERIVATION OF KEY BUILDING BLOCKS FOR THE IDDO FRAMEWORK

The Key Building Block	Commonalities From Where The Building Block Is Derived	Basis
Purpose	• Problem identification. • Deliver meaningful insights to users. • Development and implementation.	• Defines the dashboard's core problem and direction. • Guides the delivery of meaningful insights to users. • Influences all development and implementation decisions.
Audience	• Tailor best practices to the specific requirements of each domain. • User-friendly interfaces. • Tailor design approaches	• Specifies target users and their information needs. • Shapes design and functionality based on user requirements. • Enhances relevance through user-centric approaches.
Data Source	• Drill-down capabilities and data exploration. • Presentation of analytical data insight.	• Determines information range and depth for display. • Enables data exploration and drill-down capabilities. • Provides foundation for analytical insights.
Data Processing	• Reduce data selection overhead and automate selection for multiple charts. • Presentation of analytical data insight.	• Facilitates data transformation and preparation • Automates chart data selection to reduce overhead • Ensures data quality throughout the processing pipeline
Style	• Use color purposefully. • Minimize visual effects. • Create intuitive and visually appealing information dashboards	• Guides purposeful color use for enhanced comprehension • Minimizes visual effects to focus on data presentation • Maintains consistent visual design across the dashboard
Visualization Technique	• Effective data visualization. • Use diverse chart types to effectively present information. • Visual display of important information.	• Determines data representation methods for effective communication. • Supports use of diverse chart types for optimal information display. • Enhances data interpretation through appropriate visual formats.
Layout	• Optimize screen space. • Balance information display.	• Optimizes screen space for information display. • Ensures balanced arrangement of dashboard elements. • Enhances readability and intuitive navigation.
Structure	• Simplify the design. • Provide context.	• Defines organization and hierarchy of dashboard elements. • Simplifies complex information through logical structuring. • Supports intuitive navigation between components.
User Interface	• User-friendly interfaces. • Use minimal graphical user interface (GUI) elements for effective interaction.	• Facilitates access to dashboard features and functionalities. • Enhances overall user experience and usability. • Serves as the primary interaction point between users and data.
Testing	• Test and validate dashboards for quality assurance. • Testing and evaluation	• Validates data accuracy and visualization correctness • Identifies and resolves usability issues pre-deployment • Ensures functionality meets requirements
Evaluation	• Testing and evaluation. • Deliver meaningful insights to users.	• Assesses dashboard effectiveness in delivering insights • Measures user satisfaction and dashboard performance • Identifies improvement opportunities
Continuous Improvement	• Iterative process and feedback. • Tailor best practices to the specific requirements of each domain.	• Enables ongoing refinement based on feedback. • Adapts to evolving user needs and business requirements • Creates optimization feedback loop for dashboard enhancement

TABLE VI. CATEGORIZATION OF THE BUILDING BLOCKS

Category		
Defining and Planning	***Design and Visualization***	***Testing and Improvement***
Building Blocks		
• Purpose • Audience • Data Source • Data Processing	• Style • Visualization Technique • Layout • Structure • User Interface	• Testing • Evaluation • Continuous Improvement

This systematic organization offers a clear pathway that guides dashboard developers from initial scoping to iterative enhancement. By structuring the building blocks in this way, the framework fosters a coherent understanding of how each component contributes to the overall dashboard architecture while reinforcing governance integration throughout the development process. The next section discusses how IG principles are embedded into the IDDO to strengthen compliance and ensure information quality.

B. Integrating Information Governance into IDDO

Embedding IG into the IDDO framework represents a key advancement in aligning dashboard design with quality, compliance, and accountability requirements. While IDDO adopts a structural modeling approach inspired by the BMO [12], the integration of IG principles differentiates it by embedding governance mechanisms directly into design logic and model semantics. This approach follows the Kernel Theory Fundamentalist perspective [34], [35], ensuring that the ontology is grounded in established theoretical foundations relevant to information systems design.

Effective dashboards rely on high-quality information, specifically, attributes such as accuracy, reliability, and security. Shortcomings in these areas can lead to misinterpretation or poor decision-making [39], [40]. Despite the critical role of information quality, dashboard design methodologies have often neglected the integration of formal governance principles. This study addresses that gap by embedding IG into the foundational layers of IDDO to support governance-aligned dashboard development.

The rationale for this integration stems from the limited linkage between IG and dashboard design in existing literature. This research responds by identifying core IG principles from established models and synthesizing them with practical dashboard design requirements, resulting in a tailored governance foundation suited to the information context of dashboards.

C. Deriving Key IG Principles for IDDO

The integration of IG principles into the IDDO is based on a structured synthesis of established IG models and relevant literature. Similar to the derivation of the twelve key building blocks of IDDO, this process ensures that the selected IG principles align not only with compliance and regulatory requirements but also with structured and user-centered dashboard development. As emphasized by Kooper et al. [41] and Smallwood [42], IG is not a one-size-fits-all framework. Although various studies have proposed IG principles [18], [21], [43], their applicability must be tailored to specific industries and information contexts.

Nasir et al. [44] highlight the difficulty in identifying universally applicable IG principles, reinforcing the view of Kooper et al. [41] and Smallwood [42] that IG selection should remain flexible and aligned with organizational objectives. An initial study by Nasir et al. [45] on unified information dashboard design identified common dashboard design principles that informed the selection of IG principles for IDDO. The adopted IG principles are drawn from two sources: 1) a synthesis of three established IG models and relevant literature to ensure governance and regulatory alignment, and 2) common design principles identified by Nasir, Ely Salwana, et al. [45] to ensure practical, user-centered governance. Based on this dual synthesis, Nasir et al. [44] proposed eight IG principles to serve as the foundational governance structure for IDDO, as illustrated in Fig. 3.



Fig. 3. Relevant IG principles as the foundational governance of the IDDO (Nasir et al. [44]).

After identifying the relevant IG principles, the next step was to map them onto the IDDO framework to strengthen the governance layer in dashboard design. The eight selected principles were chosen for their relevance and practical applicability in ensuring information quality, reliability, and security across diverse sectors. Each principle reinforces the integrity and traceability of dashboard components while supporting compliance with organizational and regulatory standards. Their inclusion aligns core design elements with established information management practices, particularly in sectors with stringent data requirements such as healthcare and financial services. Table VII summarizes the eight IG principles and illustrates how each supports IDDO's structural goals, enhancing governance alignment and dashboard reliability.

TABLE VII. SELECTED IG PRINCIPLES AND THEIR RELEVANCE TO IDDO

IG Principle	Core Function	Relevance to IDDO
Quality	Ensures information accuracy and timeliness	Supports reliable decision-making through accurate dashboard data
Security	Protects against unauthorized access	Safeguards sensitive information while maintaining data integrity
Accountability	Defines clear roles and responsibilities	Establishes ownership of information management processes
Transparency	Provides visibility into information handling	Builds user trust through clear data sources and processes
Compliance	Ensures adherence to regulatory standards	Enables tracking and reporting of compliance metrics
Integrity	Maintains data consistency from source to output	Preserves information consistency and prevents corruption
Availability	Ensures information accessibility	Supports real-time decision-making with timely access
Effectiveness	Measures dashboard alignment with objectives	Ensures actionable insights that fulfill organizational goals

The selection of these eight principles reflects not only their foundational role in IDDO but also their direct relevance to the diverse requirements of dashboard design across sectors. Embedding these principles within the IDDO framework strengthens the dashboard design process with robust governance mechanisms, ensuring that information remains accurate, secure, compliant, and valuable for knowledge-based decision-making.

After identifying the twelve building blocks and the eight relevant IG principles in this sub-section, these elements are integrated and formally presented using an ontological approach to establish the IDDO framework, the primary knowledge contribution of this study. The following sub-section will explore the development of the formalized IDDO framework in detail.

D. Establishing the IDDO Framework

This sub-section introduces the IDDO framework, the core artefact of this study, detailing its structure and formalization. As outlined in the methodology, IDDO was developed using the Unified Ontological Approach (UOA), an ontology development method (ODM) introduced by Nasir, Firdaus Sulaiman et al. [29], which systematically translates building blocks and IG principles into a formal structure tailored for dashboard design.

The adoption of ontology aligns with Gruber's [46] view of ontology as a specification of conceptualization, and with Noy and McGuinness's [47] emphasis on structuring domain-specific entities and relationships. Incorporating these foundations, the IDDO framework organizes components semantically with clearly defined relationships to support structured and reliable dashboard development, consistent with recent perspectives by Liu et al. [48] on formalizing complex information systems.

Inspired by frameworks like the BMO [12], IDDO is designed as a top-level ontology that dashboard designers across varying expertise levels can adopt, reflecting Guarino's [49] concept of a broad yet rigorous ontological structure. This balance ensures both usability and structural standardization.

The formalization of IDDO employs OntoUML, a modeling language grounded in the Unified Foundational Ontology (UFO) [36], [50], [51], enhancing semantic precision and conceptual clarity. OntoUML enables the explicit and consistent definition of dashboard components and interrelationships while embedding IG principles to ensure traceability and governance alignment.

Fig. 4 illustrates the formalized IDDO framework, integrating IG principles within a unified ontological model for dashboard design. Formalized using OntoUML and grounded in Unified Foundational Ontology (UFO) principles [52], the framework ensures conceptual clarity, structural consistency, and supports standardization, interoperability, and governance-aligned development practices.

OntoUML stereotypes represent ontological categories derived from UFO [50]. The ⟪kind⟫ stereotype defines rigid, identity-providing types (e.g., Purpose, Structure), while ⟪role⟫ captures anti-rigid types that may evolve over time (e.g., Data Processing). Intrinsic properties such as Style are modeled using ⟪mode⟫, temporal states like Testing and Evaluation are represented by ⟪phase⟫, and measurable attributes, including IG principles, are categorized as ⟪quality⟫. Component relationships are expressed through semantic links such as "characterizes", "influences", and "governs", ensuring accurate interaction modeling.

Structurally, Purpose (⟪kind⟫) characterizes Audience (⟪kind⟫), establishing dashboard context. Data Source (⟪datatype⟫) and Data Processing (⟪role⟫) manage information flow, while Style (⟪mode⟫), Visualization Technique (⟪kind⟫), and Layout (⟪kind⟫) define visual presentation. Structure (⟪kind⟫) organizes the architecture, and User Interface (⟪kind⟫) facilitates user interaction. Testing, Evaluation, and Continuous Improvement (⟪phase⟫) embed iterative mechanisms for quality and compliance.

Two superclasses ensure ontological validity: Processing Activity (⟪kind⟫) grounds Data Processing (⟪role⟫), and Quality Assurance (⟪kind⟫) supports quality-related phases, following Guizzardi's [36] principle that anti-rigid types must derive from rigid foundations to maintain identity criteria.

Eight IG principles (⟪quality⟫) are embedded: Effectiveness and Transparency guide design intent; Quality, Integrity, and Compliance ensure data trustworthiness; Security and Availability safeguard access; and Accountability underpins governance and traceability. This integration balances practical dashboard development with governance expectations.

Through this formalization, IDDO offers a rigorous and reusable design structure, integrating twelve dashboard building blocks, eight governance-aligned IG principles, and necessary ontological superstructures to ensure theoretical soundness and practical relevance. The next section outlines the evaluation process used to validate IDDO's structural and conceptual consistency.

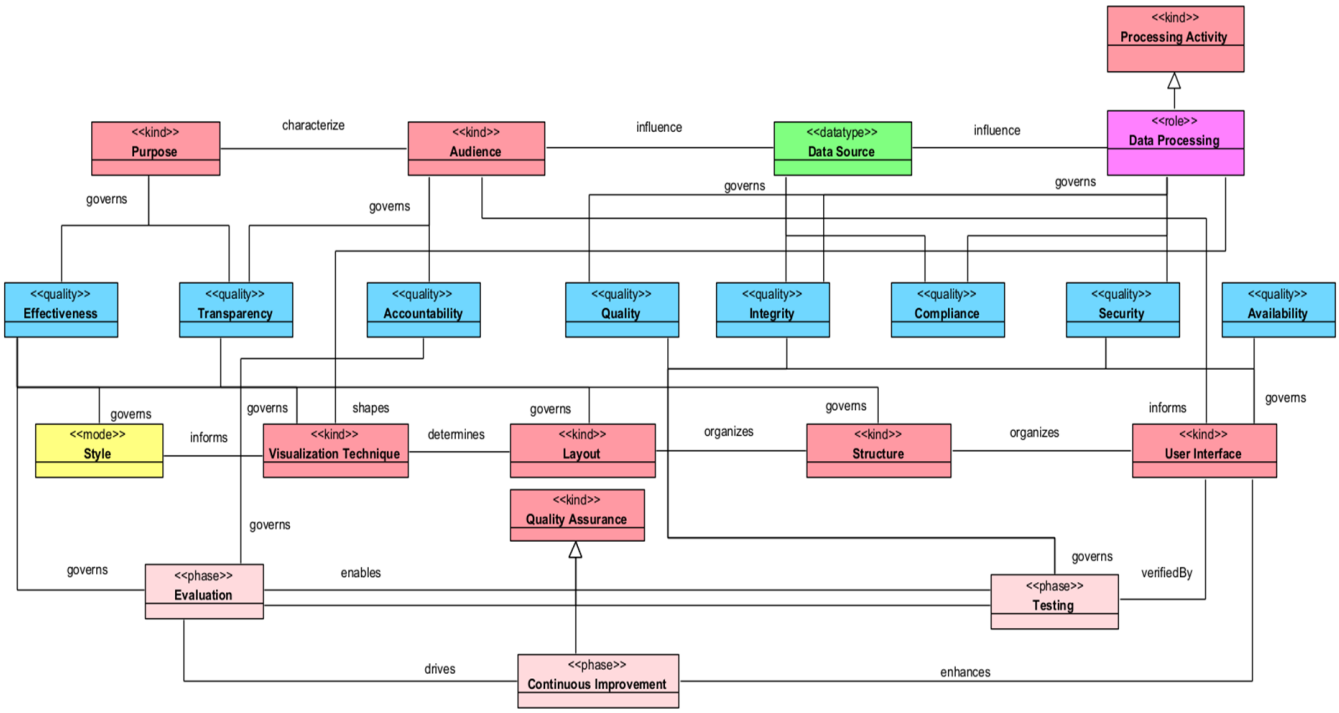


Fig. 4. The IDDO Framework.

E. Evaluation Result and Discussion

The evaluation of the IDDO framework employed a dual-method approach to ensure both structural and conceptual soundness, critical for standardizing dashboard design processes [53], [54]. Tool-based syntactic validation using the OntoUML Plugin was combined with a conceptual assessment grounded in Unified Foundational Ontology (UFO) principles.

While studies like Paul et al. [55] utilized competency-driven testing, this study emphasizes formal validation to establish IDDO's correctness during early development, consistent with best practices [52]. The OntoUML Plugin was chosen for its effectiveness in detecting structural inconsistencies, classification issues, and modeling violations that could undermine framework integrity.

Validation was conducted within the Visual Paradigm environment, systematically checking logical consistency across relationships, hierarchies, and component definitions [56]. The results confirmed IDDO's structural soundness and its suitability as a foundation for standardized dashboard development.

Parallel to syntactic validation, a conceptual assessment based on UFO criteria [36] evaluated component clarity, structural coherence, and relevance to real-world organizational contexts. The assessment verified that IDDO meets ontological adequacy standards and supports governance-aligned dashboard modeling.

Together, these evaluations confirm IDDO as a structurally valid and conceptually coherent ontology. As shown in Fig. 5, the validated framework demonstrates readiness to support standardized, governance-sensitive dashboard development across diverse settings.

This evaluation confirms the structural validity of the IDDO through rigorous tool-based assessment, laying a strong foundation for its progression to practical application. The syntactic validation performed using the OntoUML Plugin marks a critical milestone in the ontology development lifecycle, ensuring the framework's formal correctness prior to any implementation-focused evaluations. This approach reflects established practices in ontology engineering, where structural verification is prioritized before empirical testing [52].

The decision to emphasize tool-based validation at this stage was strategic, enabling the early detection and resolution of fundamental modeling issues that could otherwise undermine the framework's utility in practice. By first securing the ontological soundness and structural consistency of the IDDO, subsequent evaluations can be more confidently directed towards assessing its representational adequacy and practical applicability. Future work will broaden the evaluation scope to incorporate competency-based assessments and stakeholder validation, building upon the structurally verified foundation established here.

In addition to confirming IDDO's formal structure, the evaluation reveals three key implications for dashboard design practice. First, ontological formalization based on the UFO and implemented using OntoUML enhances semantic clarity and minimizes modeling ambiguities that often arise in ad hoc dashboard development. Second, embedding IG principles directly into the ontology promotes compliance-by-design, ensuring that governance is embedded into the system rather than retrofitted post-development. Third, the combined structural and conceptual validation affirms that IDDO provides a transferable foundation applicable across sectors, especially in contexts requiring traceability, reusability, and regulatory alignment.

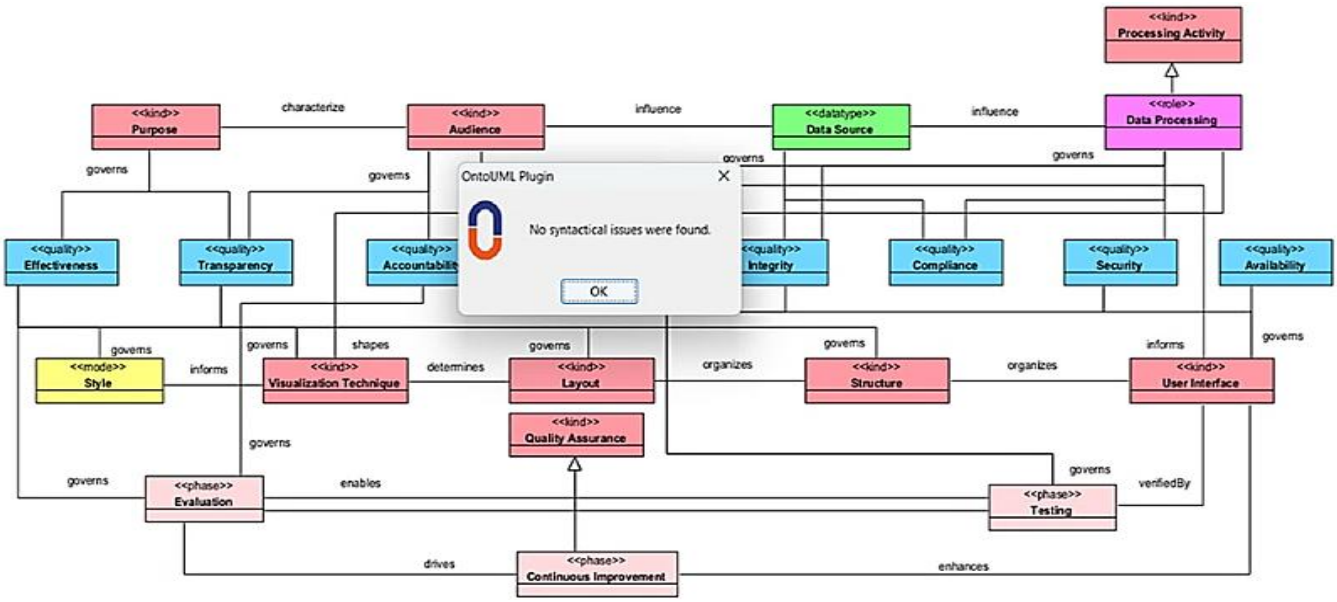


Fig. 5. Syntactical validation results of the IDDO framework.

While no empirical deployment was conducted at this stage, the structural soundness and conceptual completeness of IDDO offer strong theoretical grounding for its real-world application. These insights reinforce the practical relevance of IDDO and provide the rationale for its operationalization through the IDDM Canvas.

Building on this foundation, the following section introduces the IDDM Canvas as a practical design instrument that translates the formal ontology into a structured methodology. This transition from a formalized ontology to an applied methodology marks a crucial step in making governance-integrated dashboard design more accessible and actionable for practitioners across diverse organizational contexts.

V. TRANSLATING IDDO INTO IDDM CANVAS

Though IDDO provides the conceptual backbone, the development of the IDDM Canvas marks a critical leap from theory to practice in dashboard design. Inspired by the translation of the BMO into the widely adopted BMC [28], the IDDM Canvas similarly operationalizes IDDO's building blocks and governance principles into a structured, user-friendly methodology.

The Canvas is organized into distinct design zones that mirror IDDO's core components, streamlining the process of dashboard planning, development, and governance integration. It offers a clear visual layout that guides practitioners, regardless of technical background, through each stage of dashboard creation, embedding conceptual rigor and compliance-oriented thinking along the way.

Each design zone clusters related building blocks, visually mapped to encourage collaboration, support structured decision-making, and promote early stakeholder alignment. This design ensures that governance principles are systematically embedded across the dashboard lifecycle, not treated as an afterthought.

By bridging the gap between theoretical modeling and operational needs, the IDDM Canvas enables organizations to apply IDDO principles with greater consistency, traceability, and strategic coherence. It acts as both a guide and a governance scaffold, helping teams design dashboards that are not only functional but also sustainable and auditable over time.

Fig. 6 presents the layout of the IDDM Canvas, illustrating how the twelve building blocks and eight IG principles are translated into an actionable design guide ready for real-world use.

While the IG principles are not explicitly visualized in the core layout of the IDDM Canvas, they are structurally embedded within the underlying IDDO framework as part of its governance foundation. As discussed earlier, IG does not act as a standalone building block but operates as a cross-cutting dimension that reinforces consistency, reliability, and compliance across dashboard development activities. This deliberate choice to omit IG from the visual Canvas was made to preserve clarity, particularly for non-technical users, while ensuring governance is enforced through structure rather than visual complexity.

This design strategy mirrors the approach taken by Osterwalder and Pigneur [28] in developing the BMC, which was adapted from the more detailed BMO [12].

In that transition, several conceptual and semantic elements, such as meta-relationships and formal dependencies, were intentionally abstracted out to enhance usability without compromising underlying rigor. Similarly, the IDDM Canvas abstracts IG elements while maintaining their operational influence through embedded logic and structural design.

A guided version of the IDDM Canvas is presented in Fig. 7, featuring instructional prompts for each design zone. These prompts provide practical support for users applying Canvas in real-world projects. Red text annotations further highlight specific areas where governance principles are applied, assisting both novice and experienced designers in creating structured, and compliant dashboards.

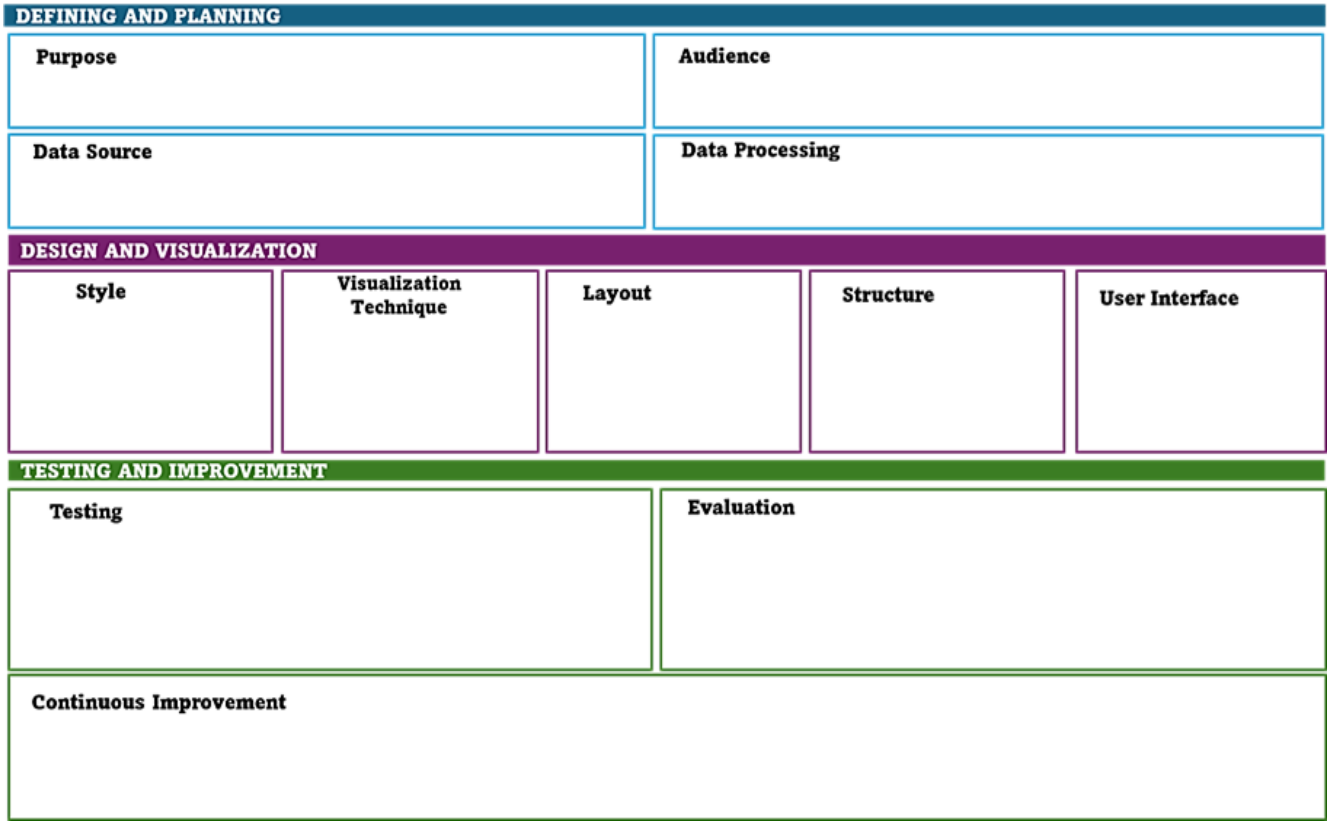


Fig. 6. The IDDM Canvas.

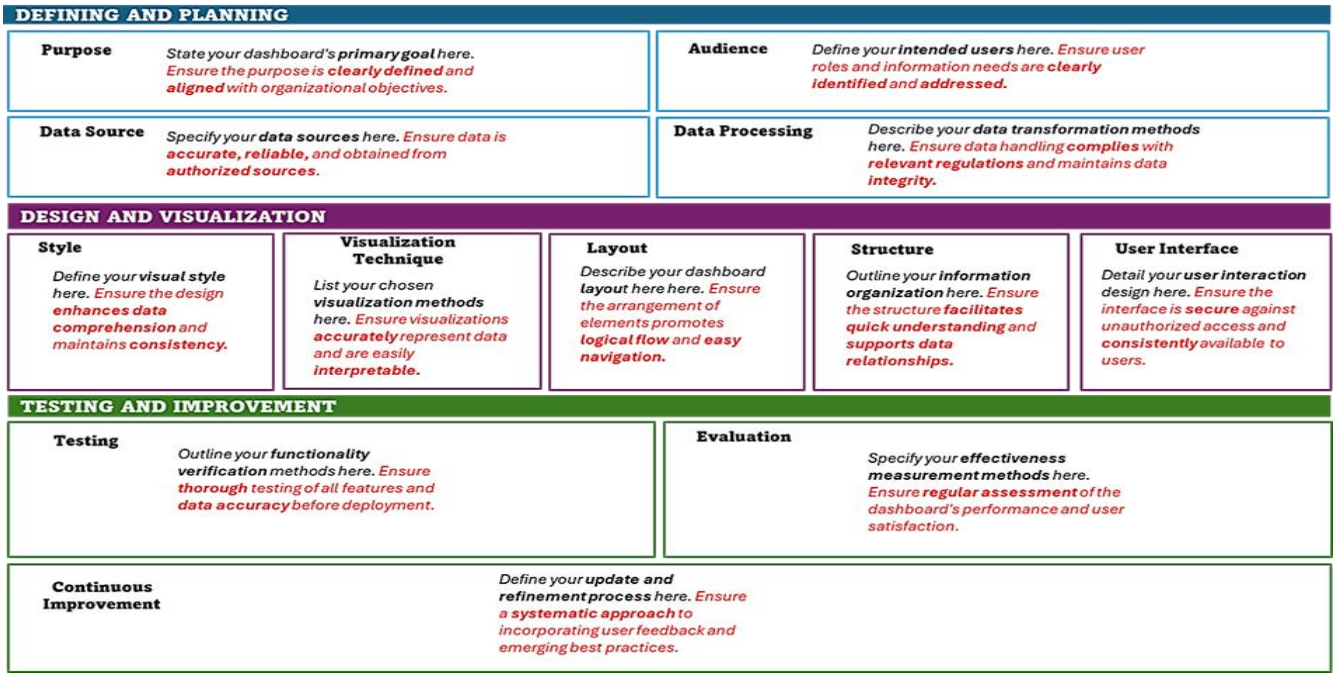


Fig. 7. The IDDM Canvas with instructional guidance.

As depicted in Fig. 6 and Fig. 7, the IDDM Canvas organizes its building blocks into three structured categories, offering a systematic pathway for dashboard development. In contrast to traditional dashboard approaches, which often depend on intuitive or ad hoc design methods, the IDDM Canvas provides a structured, repeatable methodology with governance principles embedded at its core. This structured approach enables organizations to develop dashboards that are not only semantically consistent but also compliant with established organizational standards [18].

TABLE VIII. IDDM CANVAS VERSUS CONVENTIONAL APPROACHES

Aspect	Conventional Dashboard Design	IDDM Canvas Approach
Methodology	Often ad hoc or linear; lacks structured guidance.	Based on a systematic, ontology-driven methodology with defined design building blocks.
Governance Integration	Rarely considers governance principles explicitly.	IG principles are embedded through the foundational IDDO framework and reflected in the design.
Scalability	Difficult to generalize across departments or organizations.	Modular and adaptable for cross-domain implementation.
Usability	Relies on designer intuition; inconsistent results.	Provides clear visual zones and guided prompts, improving consistency and collaboration.
Compliance and Traceability	Compliance requirements are often managed separately or retrofitted.	IG principles embedded from the start; design decisions are traceable and align with compliance standards.

Table VIII presents a comparative analysis between the IDDM Canvas and traditional dashboard design approaches, highlighting key differences in methodology, governance integration, scalability, usability, and compliance.

The IDDM Canvas acts as a practical implementation tool, translating IDDO's conceptual framework into a visual methodology that guides practitioners through a systematic dashboard development process. It facilitates collaborative design sessions by providing a common language and structure, fostering alignment among technical designers, domain experts, and governance stakeholders [42].

VI. CONCLUSION

This study introduced two interrelated artifacts, the IDDO and the IDDM Canvas, to address the lack of standardized, governance-aligned methodologies in dashboard development. IDDO formalizes dashboard design through an ontological structure that integrates IG principles, while the IDDM Canvas operationalizes this structure into a practical, user-centric design tool. Together, they offer a scalable approach to enhance consistency, compliance, and effectiveness across diverse organizational contexts.

IDDO addresses three critical gaps in existing dashboard development practices: the absence of a transferable design methodology, limited application of ontological formalization, and weak governance integration. By combining twelve design building blocks with eight governance principles under a unified structure, IDDO provides a robust foundation for developing governance-compliant dashboards. Formal validation using the OntoUML Plugin confirmed its structural soundness and conceptual coherence.

Furthermore, the proposed IDDO framework and IDDM Canvas have strong practical relevance for organizations aiming to systematize dashboard development while embedding governance compliance by design. Recent recommendations highlight that effective dashboards must prioritize clarity, strategic alignment, and user relevance principles that are embedded in the IDDO structure [61]. Their application has the potential to significantly improve decision-support system reliability, accelerate dashboard deployment cycles, and enhance regulatory traceability.

While IDDO has undergone tool-based validation, the IDDM Canvas has yet to be empirically tested in operational environments, limiting insights into its practical effectiveness. Future research should focus on applying the IDDM Canvas across diverse settings, supported by competency-based testing and stakeholder validation, to gather empirical evidence of its usability, adaptability, and governance alignment in real-world projects. Strengthening the empirical validation of the IDDM Canvas will be key to demonstrating its scalability and contribution to the broader field of governance-integrated information systems design.

REFERENCES

[1] A. Sarikaya, M. Correll, L. Bartram, M. Tory, and D. Fisher, "What do we talk about when we talk about dashboards?" IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 1, pp. 682-692, 2018, doi: 10.1109/TVCG.2018.2864903.

[2] M. Pestana, R. Pereira, and S. Moro, "Improving health care management in hospitals through a productivity dashboard," Journal of Medical Systems, vol. 44, no. 4, 2020, doi: 10.1007/s10916-020-01546-1.

[3] J. A. Poppe, R. S. Smorenburg, T. G. Goos, H. R. Taal, I. K. M. Reiss, and S. H. P. Simons, "Development of a web-based oxygenation dashboard for preterm neonates: A quality improvement initiative," Journal of Medical Systems, vol. 48, no. 1, 2024, doi: 10.1007/s10916-024-02064-0.

[4] M. Pamuk and M. Schumann, "Towards AI dashboards in financial services: Design and implementation of an AI development dashboard for credit assessment," Machine Learning and Knowledge Extraction, vol. 6, no. 3, pp. 1720-1761, 2024, doi: 10.3390/make6030085.

[5] L. Conrow, C. Fu, H. Huang, N. Andrienko, G. Andrienko, and R. Weibel, "A conceptual framework for developing dashboards for big mobility data," Cartography and Geographic Information Science, vol. 50, no. 5, pp. 476-492, 2023, doi: 10.1080/15230406.2023.2190164.

[6] A. Salvi, L. A. Gillenwater, B. P. Cockrum, S. E. Wiehe, K. Christian, J. Cayton, T. Bailey, P. K. Halverson, and B. E. Dixon, "Development of a real-time dashboard for overdose touchpoints: User-centered design approach," JMIR Human Factors, vol. 11, 2024, doi: 10.2196/57239.

[7] A. Wu, Y. Wang, M. Zhou, X. He, H. Zhang, H. Qu, and D. Zhang, "MultiVision: Designing analytical dashboards with deep learning based recommendation," IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 1510-1520, 2021, doi: 10.1109/TVCG.2021.3114826.

[8] B. Bach, E. Freeman, A. Abdul-Rahman, C. Turkay, S. Khan, Y. Fan, and M. Chen, "Dashboard design patterns," IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 1, pp. 342-352, 2022, doi: 10.1109/TVCG.2022.3209448.

[9] M. Elshehaly, R. Randell, M. Brehmer, L. McVey, N. Alvarado, C. P. Gale, and R. A. Ruddle, "QualDash: Adaptable generation of visualisation dashboards for healthcare quality improvement," IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 689-699, 2021, doi: 10.1109/TVCG.2020.3030424.

[10] E. Lomas and B. Makhlouf Shabou, "Information governance and ethics-information opportunities and challenges in a shifting world: Setting the scene," Records Management Journal, vol. 29, no. 1/2, pp. 2-4, 2019.

[11] D. Őri, I. Szabó, A. Kő, and T. Kovács, "Digitalizing in crisis: The role of organizational resilience in SMEs' digitalization," Journal of Enterprise Information Management, vol. 37, no. 4, pp. 1185-1205, 2024, doi: 10.1108/JEIM-03-2023-0141.

[12] A. Osterwalder, "The business model ontology: A proposition in a design science research," Doctoral dissertation, Université de Lausanne, 2004.

[13] M. Kokoç and A. Altun, "Effects of learner interaction with learning dashboards on academic performance in an e-learning environment," Behaviour and Information Technology, vol. 40, no. 2, pp. 161-175, 2019, doi: 10.1080/0144929X.2019.1680731.

[14] M. J. Eppler and R. A. Burkhard, "Visual representations in knowledge management: Framework and cases," Journal of Knowledge Management, vol. 11, no. 4, pp. 112-122, 2007, doi: 10.1108/13673270710762756.

[15] G. W. Young and R. Kitchin, "Creating design guidelines for building city dashboards from a user's perspectives," International Journal of Human Computer Studies, vol. 140, pp. 1-13, 2020, doi: 10.1016/j.ijhcs.2020.102429.

[16] G. Balletto, A. Milesi, M. Ladu, and G. Borruso, "A dashboard for supporting slow tourism in green infrastructures. A methodological proposal in Sardinia (Italy)," Sustainability, vol. 12, no. 9, 2020, doi: 10.3390/SU12093579.

[17] S. Ahdan, E. R. Susanto, and N. R. Syambas, "Proposed design and modeling of smart energy dashboard system by implementing IoT (internet of things) based on mobile devices," IOP Conference Series: Materials Science and Engineering, vol. 514, no. 1, 2019, doi: 10.1109/TSSA48701.2019.8985492.

[18] R. Datta, M. Valavala, and M. Haris Uddin Sharif, "Information governance: A necessity in today's business environment," International Journal of Computer Science and Mobile Computing, vol. 8, no. 8, pp. 67-76, 2019.

[19] S. de Haes and W. van Grembergen, "Exploring the relationship between IT governance practices and business/IT alignment through extreme case analysis in Belgian mid-to-large size financial enterprises," Journal of Enterprise Information Management, vol. 22, no. 5, pp. 615-637, 2009, doi: 10.1108/17410390910993563.

[20] F. De Abreu Faria, A. C. Gastaud Maçada, and K. Kumar, "Information governance in the banking industry," in Proceedings of the 46th Hawaii International Conference on System Sciences, 2013, pp. 4436-4445, doi: 10.1109/HICSS.2013.270.

[21] F. Rouzbahani, F. Asadi, R. Rabiei, H. Moghaddasi, and H. Emami, "Developing a model for national health information governance program in Iran," Journal of Medicine and Life, vol. 13, no. 4, pp. 510-516, 2020, doi: 10.25122/jml-2020-0036.

[22] A. Mat Isa and R. Mohd Ali, "Managing evidence of public accountability: An information governance perspective," International Journal of Innovation, Creativity and Change, vol. 10, no. 7, pp. 142-153, 2019.

[23] P. Mikalef, M. Boura, G. Lekakos, and J. Krogstie, "The role of information governance in big data analytics driven innovation," Information and Management, vol. 57, no. 7, 2020, doi: 10.1016/j.im.2020.103361.

[24] H. A. Simon, The Sciences of the Artificial, MIT Press, 1969.

[25] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," Journal of Management Information Systems, vol. 24, no. 3, pp. 45-77, 2007, doi: 10.2753/MIS0742-1222240302.

[26] A. Dresch, D. P. Lacerda, and J. A. V. Antunes Jr, Design science r esearch: A method for science and technology advancement. Springer, 2015, doi: 10.1007/978-3-319-07374-3.

[27] J. G. Teixeira, L. Patrício, K. H. Huang, R. P. Fisk, L. Nóbrega, and L. Constantine, "The MINDS method: Integrating management and interaction design perspectives for service design," Journal of Service Research, vol. 20, no. 3, pp. 240-258, 2017, doi: 10.1177/1094670516680033.

[28] A. Osterwalder, Y. Pigneur, T. Clark, and A. Smith, Business model generation: A handbook for visionaries, game changers, and challengers. John Wiley & Sons, 2010.

[29] A. H. M. Nasir, Firdaus Sulaiman, L. K. Leong, E. Salwana, and M. N. Ahmad, "An approach for developing an ontology: Learned from business model ontology design and development," International Journal of Advanced Computer Science and Applications, vol. 15, no. 1, pp. 416-425, 2024, doi: 10.14569/IJACSA.2024.0150355.

[30] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Quarterly, vol. 28, no. 1, pp. 75-105, 2004.

[31] J. vom Brocke, A. Hevner, and A. Maedche, "Introduction to design science research," in Design science research Cases, J. vom Brocke, A. Hevner, and A. Maedche, Eds. Springer, 2020, pp. 1-13, doi: 10.1007/978-3-030-46781-4_1.

[32] S. Gregor and O. Zwikael, "Design science research and the co-creation of project management knowledge," International Journal of Project Management, vol. 42, no. 3, 2024, doi: 10.1016/j.ijproman.2024.102584.

[33] E. M. Alotaibi, H. Issa, and M. Codesso, "Blockchain-based conceptual model for enhanced transparency in government records: A design science research approach," International Journal of Information Management Data Insights, vol. 5, no. 1, 2025, doi: 10.1016/j.jjimei.2024.100304.

[34] S. Gregor, "The nature of theory in information systems," MIS Quarterly, vol. 30, no. 3, pp. 611-642, 2006, doi: 10.2307/25148742.

[35] S. Gregor and D. Jones, "The anatomy of a design theory," Journal of the Association for Information Systems, vol. 8, no. 5, pp. 312-335, 2007, doi: 10.17705/1jais.00129.

[36] G. Guizzardi, "Ontological foundations for structural conceptual models," Centre for Telematics and Information Technology, 2005.

[37] M. Ashburner et al., "Gene ontology: Tool for the unification of biology," Nature Genetics, vol. 25, no. 1, pp. 25-29, 2000, doi: 10.1038/75556.

[38] R. Bernardo, S. V. R. Galina, and S. I. D. de Pádua, "The BPM lifecycle: How to incorporate a view external to the organization through dynamic capability," Business Process Management Journal, vol. 23, no. 1, pp. 155-175, 2017, doi: 10.1108/BPMJ-12-2015-0175.

[39] A. Cairo, How charts lie: Getting smarter about visual information. W.W. Norton & Company, 2019.

[40] M. Correll, E. Bertini, and S. Franconeri, "Truncating the Y-axis: Threat or menace?" in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1-12, doi: 10.1145/3313831.3376222.

[41] M. N. Kooper, R. Maes, and E. E. O. R. Lindgreen, "On the governance of information: Introducing a new concept of governance to support the management of information," International Journal of Information Management, vol. 31, no. 3, pp. 195-200, 2011, doi: 10.1016/j.ijinfomgt.2010.05.009.

[42] R. F. Smallwood, Information governance: Concepts, strategies, and best practices. John Wiley & Sons, 2014.

[43] N. Mohd Nordin, A. Mat Isa, A. Z. H. Samsudin, and A. R. Ahmad, "Information governance for enhancing the performance of higher education institutions," Environment-Behaviour Proceedings Journal, vol. 7, no. SI10, pp. 135-139, 2022, doi: 10.21834/ebpj.v7isi10.4115.

[44] A. H. M. Nasir, E. Salwana, and M. Nazir Ahmad, "Integrating information governance as the foundation for information dashboard design ontology," Journal of Information Technology Education: Innovations in Practice, vol. 24, pp. 1-27, 2025, doi: 10.28945/5432.

[45] A. H. M. Nasir, M. S. Ely Salwana, and A. Mohammad Nazir, "Towards unified information dashboard design - common principles, practices, and challenges," Procedia Computer Science, vol. 234, pp. 1586-1592, 2024, doi: 10.1016/j.procs.2024.03.161.

[46] T. Gruber, "Ontology," in Encyclopedia of database systems, Springer, 2009, pp. 1963-1965.

[47] N. F. Noy and D. L. McGuinness, Ontology development 101: A guide to creating your first ontology (Stanford Knowledge Systems Laboratory Technical Report KSL-01-05). Stanford University, 2001.

[48] P. Liu, X. Jin, and Y. Shang, "Ontology-based automated knowledge identification of safety risks in Metro shield construction," Journal of

Asian Architecture and Building Engineering, 2025, doi: 10.1080/13467581.2025.2467241.

[49] N. Guarino, Formal ontology in information systems. IOS Press, 1998.

[50] G. Guizzardi and N. Guarino, "Explanation, semantics, and ontology," Data and Knowledge Engineering, vol. 153, 2024, doi: 10.1016/j.datak.2024.102325.

[51] M. Bettaz and M. Maouche, "Towards an ontological-based CIM modeling framework for IoT applications," Informatica, vol. 48, no. 4, pp. 663-684, 2024, doi: 10.31449/inf.v48i4.5845.

[52] G. Guizzardi, A. B. Benevides, C. M. Fonseca, D. Porello, J. P. A. Almeida, and T. P. Sales, "UFO: Unified foundational ontology," Applied Ontology, vol. 17, no. 1, pp. 167-210, 2022, doi: 10.3233/AO-210256.

[53] D. P. Haghighi, F. Burstein, A. Zaslavsky, and P. Arbon, "Development and evaluation of ontology for intelligent decision support in medical emergency management for mass gatherings," Decision Support Systems, vol. 54, no. 2, pp. 1192-1204, 2013, doi: 10.1016/j.dss.2012.11.013.

[54] Y. Lu, G. Song, P. Li, and N. Wang, "Development of an ontology for construction carbon emission tracking and evaluation," Journal of Cleaner Production, vol. 443, 2024, doi: 10.1016/j.jclepro.2024.141170.

[55] J. Paul, S. Piest, V. Benoiston, J. De Oliviera, P. De Alencar Silva, and M. Ricardo Da Cunha Junior, "A domain reference ontology for design science research knowledge bases," Communications of the Association for Information Systems, vol. 54, pp. 458-489, 2024, doi: 10.13140/RG.2.2.30636.88967.

[56] A. Kayed and R. M. Colomb, "Using BWW model to evaluate building ontologies in CGs formalism," Information Systems, vol. 30, no. 5, pp. 379-398, 2005, doi: 10.1016/j.is.2004.03.002.

[57] L. K. Leong, F. I. T. Urus, M. A. Riza, M. N. Ahmad, and U. H. Mohamad, "Using EPP Theory and BMO-Inspired Approach to Design a Virtual Reality Dashboard Design Ontology," International Journal of Advanced Computer Science and Applications, vol. 16, no. 4, pp. 912-920, 2025.

[58] D. Borukhson, L. Glaum, C. Weinhardt, and J. Fegert, "Visualizing Social Sentiment: Designing an Interactive Dashboard for Informed Decision-Making," in Proceedings of the 58th Hawaii International Conference on System Sciences, 2025, pp. 1493-1502.

[59] M. D. Niharika, R. S. S. S. Ram, M. Goutham, Sk. Salma, and B. V. S. Rao, "Development Of Data-Driven Real-Time Dashboard," International Journal of Creative Research Thoughts, vol. 13, no. 2, pp. d688-d694, Feb. 2025.

[60] A. M. Jain, "AI-Powered Business Intelligence Dashboards: A Cross-Sector Analysis of Transformative Impact and Future Directions," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 9, no. 4, pp. 524-536, July-Aug. 2023.

[61] F. S. Rossi, M. C. B. Adams, G. Aarons, and M. P. McGovern, "From glitter to gold: recommendations for effective dashboards from design through sustainment," Implementation Science, vol. 20, art. no. 16, 2025, doi: 10.1186/s13012-025-01430-x.

Machine Learning and 5G Edge Computing for Intelligent Traffic Management

Talbi Chaymae[1], Rahmouni M'hamed[2], Ziti Soumia[3]
Research Team of Intelligent Processing and Security of Systems (IPSS)-Department of Computer Science-Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco[1, 3]
Research Team of Intelligent Processing and Security of Systems (IPSS)-Department of Computer Science-Ecole Normale Supérieure (ENS) in Rabat, Mohammed V University in Rabat, Rabat, Morocco[2]

***Abstract*—The integration of fifth-generation (5G) communication technology and Artificial Intelligence (AI) is reshaping urban mobility by enabling intelligent transportation systems and smarter cities. This synergy allows real-time traffic management, predictive maintenance, and enhanced autonomous driving, supported by high-speed, low-latency networks and advanced data analytics. By leveraging 5G's strong connectivity, AI systems can process massive datasets to address urban challenges such as traffic congestion, environmental sustainability, and public safety. This study presents a framework that combines 5G and AI to optimize traffic management through dynamic congestion prediction and real-time routing, supported by edge computing. It highlights the benefits of improving traffic flow, reducing emissions, and enhancing overall urban mobility efficiency. In addition, it discusses key challenges including data privacy concerns, cybersecurity risks, and the high cost of infrastructure deployment. By analyzing existing technologies and proposing an AI-driven, 5G-enabled system model, this study aims to bridge the gap between theoretical advancements and practical urban implementations. The findings provide insights into scalable, efficient solutions for the future of smart transportation networks and offer directions for further research in this dynamic and evolving field.**

***Keywords*—5G Edge computing; traffic management; dynamic routing; smart cities; machine learning**

I. Introduction

Urban mobility is evolving rapidly due to the deployment of advanced technologies such as 5G and Artificial Intelligence (AI). These innovations enable smarter, more adaptive traffic management systems capable of responding to real-time changes in urban environments. The ultra-low latency and massive connectivity provided by 5G networks form a robust foundation for deploying AI-powered solutions at scale. In combination, these technologies promise to reduce congestion, improve traffic flow, and support more sustainable and efficient cities.

Recent studies illustrate this potential. Gheorghe et al. [1] demonstrated how AI and IoT, supported by predictive analytics, improve adaptive traffic control systems by enabling real-time decisions. Sharma et al. [2] showed that AI-driven route optimization in vehicular networks contributes to reduced fuel consumption and travel times. These findings confirm that integrating AI and 5G can significantly improve the operational efficiency of urban transportation.

However, existing traffic management systems face persistent limitations. Most rely on centralized architectures, which struggle to process the growing volume of sensor data in real time. They also lack the ability to dynamically adapt to evolving traffic patterns. As a result, cities continue to experience congestion peaks, inefficient route utilization, and increased emissions. There remains a gap between the potential offered by emerging technologies and their actual deployment in urban traffic systems.

This study addresses the challenge of building a scalable and responsive traffic management framework that overcomes these limitations. It proposes a system that combines machine learning models for congestion prediction with dynamic route optimization, all deployed through decentralized 5G-enabled edge computing. The objective is to ensure low-latency decision-making, minimize energy consumption, and improve overall traffic efficiency without relying on centralized processing.

Designing such a system introduces several challenges, including handling heterogeneous and noisy traffic data, maintaining real-time responsiveness at the network edge, and ensuring scalability across diverse urban environments. By integrating AI models with edge nodes capable of local processing and communication through 5G, the proposed framework responds directly to these issues.

The remainder of this study is structured as follows: Section II provides a review of related work. Section III presents the proposed methodology and system architecture. Section IV outlines the experimental results. Section V discusses limitations and future directions. Finally, Section VI concludes the study.

II. Related Work

Numerous studies have explored the intersection of 5G and AI in urban mobility, focusing on specific applications and challenges. Real-time traffic optimization has been a significant area of research, with Gheorghe et al. [1] demonstrating the efficacy of 5G-enabled AI systems in predicting and managing congestion through real-time data analysis and adaptive signal control, supported by edge computing to avoid centralized overload. Su and Xu [3] enhanced this approach with secure cluster-based authentication mechanisms, while Chuan et al. [4] emphasized the importance of robust algorithms to manage rainfall impacts on 5G millimeter-wave channels. Louvros et al.

[5] further highlighted QoS-aware resource management for traffic flow stability.

Autonomous vehicles (AVs), another critical focus, rely on 5G's high-speed connectivity for seamless data exchange. Moubayed et al. [6] demonstrated improved sensor fusion and navigation reliability, while Sharma et al. [2] optimized AV routes to reduce emissions. Khalid et al. [7] introduced hybrid V2X systems for enhanced coordination. Ge et al. (2024) validated 5G mmWave efficiency for single base stations, and Coll-Perales et al. [8] modeled end-to-end V2X latency in dense environments.

In the context of intelligent transportation systems (ITS), Chen and Song [9] explored the use of V2X communication for traffic pattern analysis and public transport enhancement. Their work is complemented by Barzegar et al. [10], who proposed predictive maintenance frameworks, and Amend and Rakocevic [11], who optimized video streaming in 5G-based ITS using multipath scheduling.

Security, data reliability, and resource allocation remain core challenges. Mongay Batalla et al. [12] proposed multi-layer security assurance models, while Rico et al. [13] examined multi-connection protocols for ITS reliability. Meanwhile, An et al. [14] and Vohra et al. [15] introduced distributed resource allocation strategies to process large traffic datasets efficiently. Mukherjee et al. [16] addressed anomaly detection in IoT-based monitoring, and Bonato et al. [17] evaluated exposure variability in 5G-V2X scenarios using deep learning.

A structured summary of these contributions is presented in Table I, highlighting the main technologies used and challenges addressed across representative works.

TABLE I. RELATED WORKS

Authors	Main Contribution	Technologies Used	Challenges Addressed
Gheorghe and Soica (2025)	Systematic review of AI, IoT, predictive analytics for traffic control	AI, IoT, Predictive Analytics	Real-time traffic management, scalability
Sharma et al. (2023)	Privacy-aware post-quantum routing in 5G IoV	5G, IoV, Privacy Algorithms	Security, Privacy
Su and Xu (2025)	Mutual authentication in 5G sensor networks for vehicles	5G, Sensor Networks	Authentication, Key Update
Coll-Perales et al. (2023)	V2X latency modeling over 5G networks	V2X Communication, 5G	Latency, QoS
Moubayed et al. (2023)	OTN-over-WDM optimization for 5G	Optical Transport Networks, WDM, 5G	Network optimization challenges

III. METHODOLOGY AND SYSTEM MODEL

The proposed system integrates data collection, AI-powered congestion prediction, graph-based route optimization, edge computing, and cloud analytics to deliver real-time, scalable, and efficient urban mobility solutions.

The first step involves data collection and integration, where the system aggregates information from diverse sources, including IoT sensors, connected vehicles, and public transportation systems. Smart traffic lights, road sensors, and surveillance cameras provide real-time updates on traffic density, vehicle speeds, and incidents, forming the foundational dataset. Additionally, Vehicle-to-Everything (V2X) communication, enabled by the low-latency capabilities of 5G, ensures seamless data exchange between vehicles and infrastructure, fostering a holistic view of the traffic ecosystem.

At the core of the system lies AI-powered congestion prediction, which utilizes advanced machine learning models to forecast traffic conditions. Historical data, such as traffic volume, weather patterns, and temporal trends, is combined with live sensor inputs and fed into models like Long Short-Term Memory (LSTM) networks. These models are adept at handling time-series data and capturing both temporal dependencies and nonlinear patterns. The process begins with preprocessing, where noisy data is cleaned, and missing values are filled. Subsequently, LSTM models are trained on historical datasets to identify traffic patterns. During real-time operations, live data streams are processed at edge computing nodes, enabling immediate predictions without relying on centralized servers.

Based on the predictions, the system executes graph-based route optimization to dynamically adjust traffic flows. The urban road network is represented as a weighted graph, where nodes signify intersections and edges represent road segments, with weights corresponding to travel times or congestion levels. Real-time congestion predictions trigger dynamic weight adjustments to reflect current conditions. Algorithms such as Dijkstra or A* compute the shortest and most efficient routes for vehicles. A continuous feedback loop updates the graph with live traffic data, ensuring that routing decisions remain adaptive and responsive to evolving conditions.

To address latency and scalability challenges, the system employs edge computing for decentralized processing. Strategically placed edge nodes near data sources aggregate and analyze data locally, reducing the computational load on central servers. These nodes host lightweight AI models capable of making congestion predictions and suggesting optimized routes. They also communicate directly with vehicles via 5G, disseminating real-time traffic updates and route recommendations. This decentralized approach significantly reduces response times and enhances the system's scalability.

Complementing the edge computing layer, cloud integration for long-term analytics ensures continuous improvement and scalability of the system. While edge nodes handle real-time tasks, cloud servers store historical data and retrain AI models periodically to adapt to new traffic patterns. This hybrid architecture balances the need for immediate responsiveness with the benefits of long-term optimization and model refinement, paving the way for an adaptive and sustainable traffic management system.

By combining these elements, the proposed framework addresses the multifaceted challenges of urban traffic management, delivering a robust solution that enhances mobility, reduces congestion, and promotes sustainability.

A. Objectives

The study aims to develop and evaluate a real-time dynamic traffic management system focusing on: Real-time Congestion Prediction, Dynamic Route Optimization.

1) Real-time congestion prediction. The ability to predict traffic congestion in real-time is a cornerstone of efficient traffic management. This involves training AI models, such as Long Short-Term Memory (LSTM) networks or Gradient Boosted Decision Trees, on both historical traffic patterns and live sensor data. Historical data includes traffic volumes, weather conditions, special events, and temporal variations, while live data comes from IoT devices, V2X communication, and GPS-enabled vehicles. These models are designed to:

a) Identify traffic bottlenecks with high precision.

b) Predict congestion levels minutes or hours into the future.

c) Provide actionable insights to traffic management systems for preemptive actions. This prediction process is carried out locally at edge nodes to minimize latency and ensure timely decision-making.

2) Dynamic Route Optimization. Once congestion is detected or predicted, the next step is to dynamically redistribute traffic to alternative routes to prevent or alleviate congestion. The urban road network is modeled as a weighted graph, where nodes represent intersections, and edges represent road segments. The weights of edges correspond to travel times, which are adjusted dynamically based on:

a) Real-time congestion levels predicted by the AI models.

b) Road closures, construction work, or accidents reported through live feeds. Using algorithms like Dijkstra, A*, or even reinforcement learning-based methods, optimal routes are calculated and suggested to vehicles in real-time. A feedback loop is maintained where updated traffic data continuously refines the routing decisions, ensuring adaptability to changing conditions.

3) Energy-efficient transportation. Optimizing traffic flow not only improves travel times but also significantly reduces energy consumption and emissions. Idle times at congested intersections and prolonged travel distances contribute heavily to urban pollution. By dynamically managing traffic:

a) Fuel consumption is reduced by avoiding stop-and-go traffic conditions.

b) Electric vehicles (EVs) benefit from extended range due to smoother traffic flows.

c) Emissions are minimized, contributing to greener and more sustainable urban environments. To further enhance energy efficiency, the system integrates vehicle-specific data (e.g., fuel efficiency or EV battery levels) into the optimization process, tailoring route recommendations for maximum energy savings.

4) Scalable and decentralized architecture. The implementation of 5G and edge computing is critical for ensuring the scalability and responsiveness of the proposed system. The architecture involves:

a) Edge nodes are placed strategically throughout the urban area, these nodes handle real-time data processing, congestion predictions, and route optimization locally. This reduces the computational load on centralized servers and minimizes data transmission delays.

b) 5G Connectivity enables high-speed and low-latency communication between vehicles, edge nodes, and centralized systems. It also supports massive IoT device connectivity, essential for collecting data from diverse sources.

c) Cloud integration: While edge nodes manage real-time tasks, cloud servers are used for long-term analytics, retraining AI models, and storing historical traffic data. This hybrid approach ensures both scalability and the continuous improvement of prediction and optimization algorithms. By decentralizing computational workloads, the system can scale efficiently across cities of varying sizes and complexities while maintaining low-latency decision-making capabilities.

B. Implementation Plan

The implementation of the proposed dynamic traffic management system follows a comprehensive plan designed to ensure scalability and real-world applicability. The initial phase begins with the creation of a simulated urban traffic network using platforms like SUMO (Simulation of Urban Mobility) or OpenStreetMap. These tools enable the generation of realistic traffic scenarios, simulating vehicle behaviors, road layouts, and dynamic traffic signals. The simulation environment provides a controlled setting to test the framework under various conditions, including peak traffic hours, accidents, and road closures, generating a robust dataset for validation.

The second phase focuses on algorithm development, starting with congestion prediction through advanced AI models. Long Short-Term Memory (LSTM) networks, implemented using TensorFlow or PyTorch, are trained on historical traffic data such as vehicle counts, speeds, weather conditions, and time-of-day patterns. These models process live sensor data streams in real time to predict traffic congestion, leveraging their ability to handle sequential data and temporal dependencies. Concurrently, route optimization is performed using graph-based algorithms. The urban road network is represented as a weighted graph, where nodes are intersections, and edges are road segments. Algorithms like Dijkstra or A* calculate optimal routes dynamically, adjusting edge weights based on real-time congestion predictions. A feedback loop ensures that traffic conditions continuously refine routing decisions, maintaining adaptability.

The third phase involves the deployment of edge computing nodes, which are critical for decentralized processing and low-latency operations. Lightweight computational devices, such as Raspberry Pi or NVIDIA Jetson Nano, equipped with 5G connectivity, are strategically placed at high-traffic areas and

key intersections. These nodes host containerized AI models managed through Docker and orchestrated with Kubernetes for scalability and fault tolerance. The edge nodes process data locally, execute congestion predictions, and disseminate route recommendations to connected vehicles in real time, reducing the load on centralized servers.

Finally, the system undergoes real-world testing in collaboration with municipalities or urban traffic authorities. A representative urban area with varying traffic densities is selected for the pilot implementation, where the system integrates with existing traffic management infrastructure. Performance metrics such as average travel time, congestion levels, fuel consumption, and system latency are monitored to evaluate the framework's effectiveness. The collected feedback is used to iteratively refine the system, ensuring robustness and adaptability.

By combining simulation, advanced AI algorithms, edge computing, and real-world testing, the proposed implementation plan offers a scalable and efficient solution for real-time traffic management, paving the way for smarter and more sustainable urban transportation systems.

C. Expected Contributions

This study aims to address the pressing challenges of urban traffic management by introducing a scalable and efficient framework tailored for real-time applications. At its core, the framework leverages a decentralized approach powered by 5G-enabled edge computing, enabling rapid data processing and decision-making directly at strategic points within the urban landscape. By integrating AI-driven congestion prediction with graph-based routing algorithms, the system offers a hybrid optimization model capable of dynamically redistributing traffic to mitigate bottlenecks. Beyond improving traffic flow, the framework emphasizes sustainability by reducing emissions and energy consumption through the elimination of idling and inefficient routes. Furthermore, the study provides a comprehensive blueprint for integrating advanced technologies—such as 5G, AI, and edge computing—into existing urban transportation infrastructures, paving the way for smarter, more adaptive, and environmentally conscious cities.

D. Practical Implementation

As part of this study, we proposed an innovative solution for dynamic urban traffic management, leveraging the integration of 5G, Artificial Intelligence (AI), and edge computing. To validate the presented concepts and illustrate their implementation, an algorithm has been developed to predict traffic congestion and optimize routes in real-time. This implementation relies on machine learning models and graph algorithms, two fundamental pillars for intelligent traffic management.

IV. Results

A. Algorithm – Traffic Simulation and Routing

In this study, we designed and implemented a traffic management algorithm aimed at optimizing urban mobility by dynamically predicting congestion and adjusting routing decisions in real time. The core of the solution combines graph-based optimization methods and machine learning techniques. Initially, congestion levels were estimated using a heuristic congestion scoring formula derived from three traffic features: Traffic Density, Estimated Speed, and Weather Conditions. Based on these scores, the system dynamically adjusted road network weights and calculated optimal routes using Dijkstra's algorithm.

To further enhance prediction reliability and address the complexity of urban traffic patterns, we trained a supervised machine learning model to predict congestion. We selected the Random Forest Classifier for its robustness, ability to capture non-linear relationships, and strong performance on structured traffic data.

a) Input and output: To clarify the data structure used during the simulation phase, the main parameters involved in the proposed traffic management algorithm are summarized in Table II. This table details the inputs required to simulate and evaluate routing scenarios, along with the expected outputs generated by the system.

The inputs include the number of simulated samples, the set of features characterizing each sample (such as estimated speed, traffic density, or incident reports), and the source and destination nodes within the road network. These parameters allow the algorithm to simulate traffic conditions and compute the most efficient path.

The outputs correspond to the resulting optimal route, represented as an ordered list of nodes, and the total cost of that route under the influence of congestion. The cost reflects the dynamic weights applied to the road network based on real-time traffic conditions.

TABLE II. Summary of Inputs and Outputs for the Proposed Traffic Management Algorithm

INPUT	OUTPUT
num_samples: number of simulated data samples	• List of nodes representing the best route. • Total cost considering simulated congestion
num_features: number of features per sample (e.g., speed, density, incidents)	
source: starting node	
target: destination node	

b) Process: A real-world dataset comprising 627 samples was utilized in this study, with each sample initially described by several traffic-related parameters. The primary objective was to simulate traffic conditions and evaluate the impact of congestion on route optimization. To prepare the data for the model, specific features were derived through additional computations, resulting in three key attributes: Traffic Density, Estimated Speed, and Weather Conditions. Missing values were calculated where necessary, and all extracted features were normalized to ensure consistent scaling and comparability.

The original dataset included the following fields:

- TMJA (Average Annual Daily Traffic): Used as an indicator of traffic density, representing the average number of vehicles passing through a segment daily.
- ratio_PL: Denoting the proportion of heavy vehicles, influencing the overall estimated speed on the segment.

- longueur: Length of the road segment, employed to estimate travel speed when time data is available.
- route: Road identifier, enabling cross-referencing with external geographic datasets.
- dateReferentiel: Date of traffic measurement, used for aligning with historical weather information.
- xD, yD, xF, yF: GPS coordinates marking the start and end points of the road segment, facilitating the retrieval of location-specific weather conditions.

The proposed traffic management algorithm was applied across all data points. It involved computing a congestion score, determining a congestion factor when necessary, adjusting the weights of the corresponding road network, and ultimatelycalculating the optimal route under the updated traffic conditions. Below, we detail the normalization process applied to each extracted parameter to ensure consistency and comparability across the dataset.

The extracted features were normalized to ensure consistent scaling. Traffic density was standardized using the min-max formula described in Eq. (1). Estimated speed was calculated from the road segment length and travel time as shown in Eq. (2). Weather conditions were normalized in two steps: first by converting temperature from Kelvin to Celsius using Eq. (3), and then applying min-max normalization as shown in Eq. (4).

Traffic Density:

$$Standardised\ density = \frac{TMJA - \min(TMJA)}{\max(TMJA) - \min(TMJA)} \quad (1)$$

Estimated Speed:

$$Average\ speed = \frac{length}{travel\ time} \quad (2)$$

Weather Conditions:

$$T_{Celsius} = T_{Kelvin} - 273{,}15 \quad (3)$$

$$\text{Weather_conditions} = \frac{Temperature_C - \min(Temperature_c)}{\max(Temperature_c) - \min(Temperature_c)} \quad (4)$$

When the computed congestion score exceeded a threshold of 1, a congestion factor of 0.2 was systematically applied to adjust the network parameters; otherwise, the original weights remained unchanged. Based on this approach, a directed graph representing the road network was constructed, with predefined base weights assigned to each edge. Upon congestion detection, all edge weights were proportionally increased using the congestion factor to reflect the associated traffic delays. Subsequently, Dijkstra's algorithm was employed to determine the shortest path between the defined source and destination nodes within the updated network structure. The final step involved visualizing the adjusted road network to highlight the effects of congestion on optimal routing strategies.

The entire simulation workflow is summarized in Table III, outlining each stage from data generation to route computation.

TABLE III. PROCESS OVERVIEW FOR TRAFFIC SIMULATION AND NETWORK ADJUSTMENT

Step	Action
1. Generate Data	Create num_samples of random data with num_features
2. Estimate Congestion	For each sample: if (f1 + 2*f2 - f3 > 1) → label = high congestion
3. Set Congestion Factor	If congestion is high → congestion_factor = 0.2 else → 0.0
4. Build Road Graph	Create a graph with nodes and standard edge weights
5. Adjust Weights	Multiply all edge weights by (1 + congestion_factor)
6. Compute Route	Use shortest path algorithm (e.g., Dijkstra) from source to target
7. Output Result	Display optimal route and total adjusted cost

c) Algorithm: Following the execution of the steps outlined in Algorithm 1, the adjusted road network was generated. Fig. 1 illustrates the updated graph structure, highlighting how congestion factors impact the routing through modified edge weights.

Algorithm 1: Visualization of the Adjusted Road Network

Initialize:
  Define manual positions for each node.
Compute:
  Create edge labels based on adjusted edge weights.
While (graph is being prepared) do
  Draw network nodes at specified positions.
  Draw directed edges between nodes with arrows.
  Label each node with its identifier.
  Annotate edges with their corresponding weights.
  Update:
    Set title of the graph to "Adjusted Road Network (With Congestion Penalty)".
    Hide axes to enhance visualization.
    Adjust layout for better spacing.
  End
End
Display:
  Render and show the final visualized graph.
End

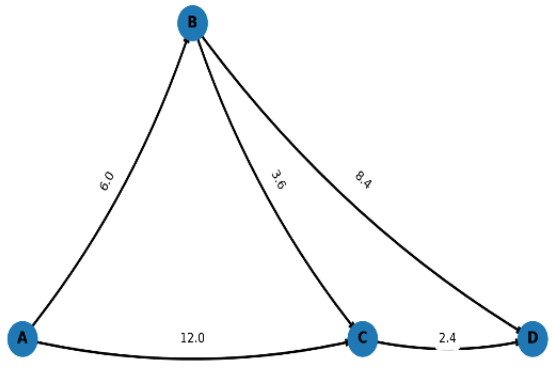


Fig. 1. Adjusted road network visualization considering congestion penalties.

Following the application of a 20% congestion penalty to the edge weights, the network structure was updated to reflect the increased travel costs. Notably, the path A → B → C → D continues to represent the optimal route under the adjusted conditions.

d) Machine learning congestion prediction: To improve traffic congestion prediction and support adaptive routing decisions, we implemented a supervised machine learning approach. The objective was to train a classification model capable of predicting congestion based on real traffic-related variables.

As a starting point, we constructed an initial congestion label (Congestion_Predite) using a heuristic scoring function based on three key features extracted or derived from the dataset [Eq. (5)]:

$$\text{Congestion}_{\text{Score}} = (0.60 \times \text{Traffic Density}) + (0.23 \times \text{Estimated Speed}) + (0.17 \times \text{Weather Conditions}) \quad (5)$$

The binary target variable was then defined using a threshold as in Eq. (6):

$$Congestion_{Predite} = \begin{cases} 1 & \text{if Score} > 1 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

This labeled dataset served as the foundation for training a Random Forest classifier, which learned more complex and non-linear relationships between the variables and congestion states. This approach allows for generalization to unseen traffic patterns and conditions, improving prediction reliability.

The performance of both the routing and prediction components is influenced by key algorithmic parameters. In the case of the Random Forest classifier, the number of trees (estimators), maximum depth, and minimum samples per split directly affect the model's ability to generalize. A higher number of trees tends to improve robustness, while excessively deep trees may lead to overfitting. For Dijkstra's algorithm, the accuracy of route computation depends on the dynamic edge weights, which are adjusted based on the congestion factor. The choice of base weights and how they scale with congestion (e.g., a 20% increase as used here) influences routing outcomes. These parameters were empirically chosen to balance performance and computational efficiency in real-time settings.

The model's predictive performance was evaluated using standard classification metrics. The results are summarized in Table IV.

TABLE IV. MODEL RESULTS

Metric	Value
Overall Accuracy	99%
Precision (Not Congested)	100%
Precision (Congested)	98%
Dominant Factor	Traffic Density (highest importance)

To assess the effectiveness of the proposed prediction model, we conducted a basic comparison with other standard classification algorithms on the same dataset. Support Vector Machines (SVM), Gradient Boosted Decision Trees (GBDT), and K-Nearest Neighbors (KNN) were evaluated using identical input features and a consistent data split. Among all tested models, the Random Forest classifier achieved the highest accuracy (99%), with better balance between precision and recall. SVM and GBDT followed with slightly lower accuracy scores (95% and 96% respectively), while KNN showed more variance depending on the feature scale and neighborhood size. These results validate the choice of Random Forest for this application, given its robustness, interpretability, and suitability for edge deployment.

These metrics confirm the model's high precision, particularly in identifying congested cases with minimal false positives. Traffic density emerged as the most influential variable, followed by speed and weather conditions. This ranking aligns with known determinants of urban traffic congestion.

Additional evaluation tools, including the confusion matrix, the ROC curve, and feature importance analysis, are presented in the next section to provide further insight into the classifier's behavior and reliability.

The performance of the developed machine learning model was evaluated through several key visualization tools, each offering complementary insights into classification reliability. Among these, the confusion matrix provides a direct view of the model's effectiveness in distinguishing between congested and non-congested traffic segments.

As shown in Fig. 2, the confusion matrix displays the number of correct and incorrect predictions across both classes. The model correctly identified nearly all instances, with minimal misclassification. This confirms its ability to separate the two categories with high precision, reinforcing the metrics reported in Table IV.

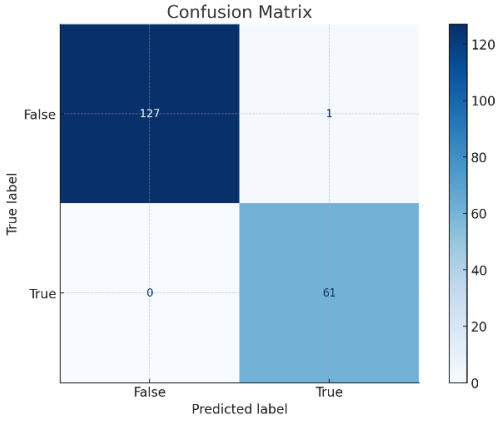


Fig. 2. Confusion matrix.

The confusion matrix summarizes the model's prediction outcomes by displaying the number of correct and incorrect classifications for both congested and non-congested traffic conditions. It provides a clear view of how well the model distinguishes between the two classes, highlighting both true positives and true negatives.

As presented in Fig. 3, the ROC curve plots the true positive rate against the false positive rate. The model achieved an Area Under the Curve (AUC) of 0.99, indicating a near-perfect capacity to discriminate between congested and non-congested traffic conditions.

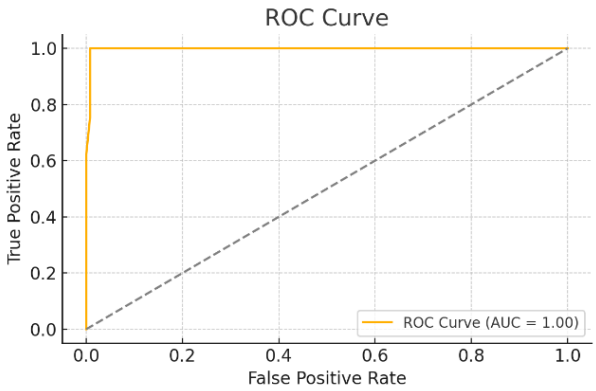


Fig. 3. Receiver Operating Characteristic (ROC) curve.

The ROC curve illustrates the trade-off between the true positive rate and the false positive rate at various classification thresholds. The Area Under the Curve (AUC) achieved a value of 0.99, demonstrating that the model possesses excellent discrimination capability between congested and non-congested states. A near-perfect AUC score reflects the model's robustness and high predictive accuracy.

To further interpret the internal behavior of the model, we examined feature importance, which ranks the contribution of each input variable to the prediction outcomes. This analysis helps identify which factors most strongly influence the classifier's decision.

As illustrated in Fig. 4, traffic density emerges as the dominant variable, followed by estimated speed and weather conditions. The important ranking aligns with domain knowledge, where density is often the primary driver of congestion.

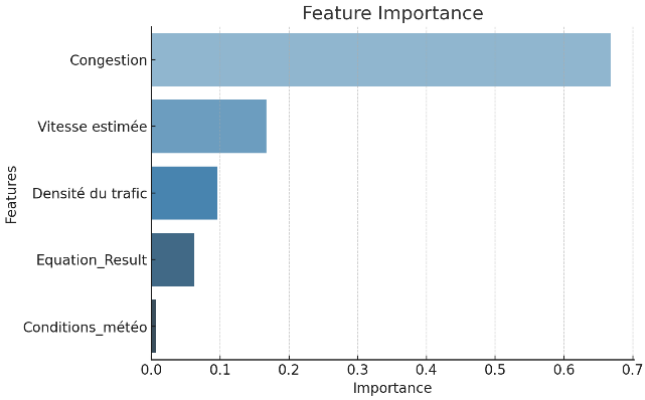


Fig. 4. Feature importance plot.

The feature importance analysis identifies and ranks the contribution of each input variable to the model's predictions. The results show that traffic density is the most influential feature, exerting the greatest impact on the model's ability to predict congestion. Other variables, such as estimated speed and weather conditions, also contribute, but to a lesser extent.

These evaluation tools provide complementary evidence of the model's effectiveness. The confusion matrix, ROC curve, and feature importance collectively validate both its predictive accuracy and its alignment with real-world traffic behavior.

Despite its strong performance, the model presents several limitations. First, the training was based on simulated data, which does not fully reflect real-time traffic irregularities. Second, the system depends on the reliability of sensor inputs, which may be affected by latency, noise, or missing data. Third, the tested road network was simplified, and external factors such as road incidents or construction zones were not included. These constraints may limit the immediate deployment of the system in uncontrolled real-world conditions and require further calibration.

V. Discussion

This study proposed and validated an integrated traffic management framework that combines AI-driven congestion prediction, graph-based dynamic routing, and 5G-enabled edge computing to support real-time decision-making. The methodology was built around two key components: a graph-based traffic simulation and routing algorithm that dynamically adjusted road network weights based on a computed congestion score, and a supervised machine learning model designed to predict congestion states using traffic-related features. Initially, congestion levels were estimated heuristically based on Traffic Density, Estimated Speed, and Weather Conditions, and road weights were adjusted accordingly to guide routing using Dijkstra's algorithm. To enhance prediction reliability beyond heuristic methods, a Random Forest Classifier was developed, chosen for its robustness against overfitting, its ability to capture non-linear feature interactions without extensive preprocessing, its interpretability through feature importance analysis, and its computational efficiency suitable for edge deployment. The model achieved outstanding performance, with an overall accuracy of 99%, a precision of 100% for non-congested cases, and 98% for congested cases, identifying Traffic Density as the dominant predictive factor. These results were reinforced by a confusion matrix analysis and a ROC curve yielding an AUC of 0.99. Compared to other machine learning models, such as SVMs, GBDTs, Neural Networks, and KNN, Random Forests provided the best trade-off between predictive power, scalability, and interpretability, making them particularly suitable for real-time, decentralized urban mobility systems. By combining accurate congestion prediction, adaptive routing, and efficient edge computing, the proposed framework delivers a scalable and effective solution to the evolving challenges of traffic management in smart cities, as illustrated in Fig. 5.

Future work will aim to extend the system's capabilities in several directions. First, the integration with connected vehicle platforms will enable the framework to receive and act upon live feedback from on-road users. Second, deploying the model in a real-world urban environment, in collaboration with traffic authorities, will allow validation under operational conditions. Third, additional traffic dynamics—such as incident detection, weather disruptions, and multi-agent coordination—will be incorporated to improve adaptability. Lastly, the framework will be tested on data from other cities to evaluate generalizability across different urban structures.

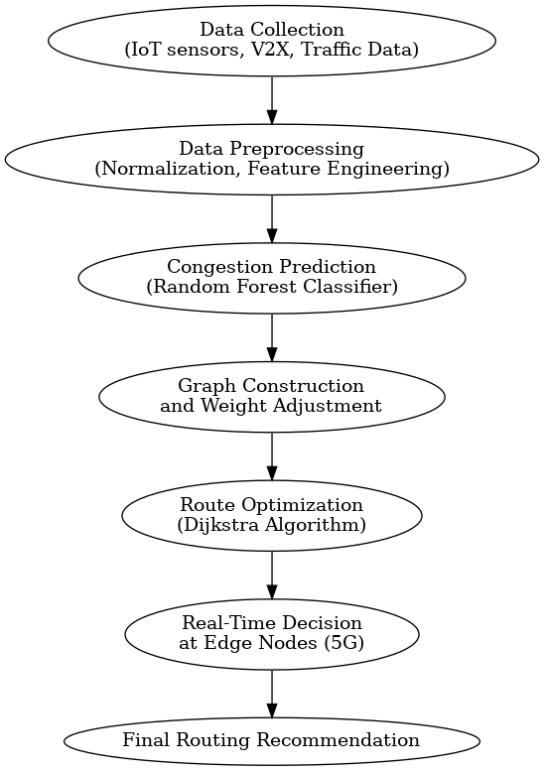


Fig. 5. Traffic management system workflow integrating machine learning and 5G edge computing.

VI. CONCLUSION

This study presented an integrated framework that combines AI-based congestion prediction, dynamic graph-based routing, and 5G-enabled edge computing to support real-time urban traffic management. The main contribution lies in the design and validation of a decentralized system capable of adapting to evolving traffic conditions using machine learning and network optimization in a scalable architecture.

The proposed solution leverages a Random Forest classifier for congestion detection and Dijkstra's algorithm for dynamic routing, applied to a simulated urban environment. The model achieved an overall accuracy of 99%, with strong precision across both congested and non-congested cases. Feature analysis confirmed traffic density as the most influential variable, validating the model's alignment with known traffic behaviors. Additional evaluation tools, such as the confusion matrix and ROC curve (AUC = 0.99), demonstrated the system's robustness and predictive reliability.

By processing data locally at the edge and reacting in near real-time, the system achieves both performance and scalability, meeting key requirements for smart city deployment. Future work will focus on integrating connected vehicle feedback, incorporating real-time disruptions such as incidents or construction, and testing the system across different urban contexts to assess transferability and operational impact.

REFERENCES

[1] C. Gheorghe, A. Soica, "Revolutionizing Urban Mobility: A Systematic Review of AI, IoT, and Predictive Analytics in Adaptive Traffic Control Systems for Road Networks", Electronics, vol. 14, pp. 719, 2025.

[2] T. Sharma, M.R. Kumar, S. Kaushal, D. Chaudhary, K. Saleem, "Privacy Aware Post Quantum Secure Ant Colony Optimization Ad Hoc On-Demand Distance Vector Routing in Intent Based Internet of Vehicles for 5G Smart Cities", IEEE Access, vol. 11, pp. 110391-110399, 2023.

[3] X. Su, Y. Xu, "An Efficient Cluster-Based Mutual Authentication and Key Update Protocol for Secure Internet of Vehicles in 5G Sensor Networks", Sensors, vol. 25, pp. 212, 2025.

[4] L.L. Chuan, M. Roslee, C. Sudhamani, S.M.I. Mitani, A. Waseem, A.F. Osman, F.Z. Ali, Y. Ullah, "Impact of Rainfall on 5G Millimeter Wave Channels", Progress In Electromagnetics Research C, vol. 148, pp. 97-107, 2024.

[5] S. Louvros, M. Paraskevas, T. Chrysikos, "QoS-Aware Resource Management in 5G and 6G Cloud-Based Architectures with Priorities", Information (Switzerland), vol. 14, pp. 175, 2023.

[6] A. Moubayed, D.M. Manias, A. Javadtalab, M. Hemmati, Y. You, A. Shami, "OTN-over-WDM Optimization in 5G Networks: Key Challenges and Innovation Opportunities", Photonic Network Communications, vol. 45, pp. 49-66, 2023.

[7] I. Khalid, V. Maglogiannis, D. Naudts, A. Shahid, I. Moerman, "Optimizing Hybrid V2X Communication: An Intelligent Technology Selection Algorithm Using 5G, C-V2X PC5 and DSRC", Future Internet, vol. 16, pp. 107, 2024.

[8] B. Coll-Perales, M.C. Lucas-Estan, T. Shimizu, J. Gozalvez, T. Higuchi, S. Avedisov, O. Altintas, M. Sepulcre, "End-to-End V2X Latency Modeling and Analysis in 5G Networks", IEEE Transactions on Vehicular Technology, vol. 72, pp. 5094-5109, 2023.

[9] Y. Chen, X. Song, "Application of 5G Combined Wireless Interconnection Technology in the Reform of English Education in Colleges and Universities", Mobile Information Systems, vol. 2022, pp. 5774132, 2022.

[10] S. Barzegar, M. Ruiz, L. Velasco, "Autonomous Flow Routing for Near Real-Time Quality of Service Assurance", IEEE Transactions on Network and Service Management, vol. 21, pp. 2504-2514, 2024.

[11] M. Amend, V. Rakocevic, "Cost-Efficient Multipath Scheduling of Video-on-Demand Traffic for the 5G ATSSS Splitting Function", Computer Networks, vol. 242, pp. 110218, 2024.

[12] J. Mongay Batalla, L.J. De La Cruz Llopis, G.P. Gomez, E. Andrukiewicz, P. Krawiec, C.X. Mavromoustakis, H.H. Song, "Multi-Layer Security Assurance of the 5G Automotive System Based on Multi-Criteria Decision Making", IEEE Transactions on Intelligent Transportation Systems, vol. 25, pp. 3496-3512, 2024.

[13] D. Rico, K.-J. Grinemmo, A. Brunstrom, P. Merino, "Performance Analysis of the Multi-Connection Tactile Internet Protocol over 5G", Journal of Network and Systems Management, vol. 31, pp. 49, 2023.

[14] C. An, A. Zhou, J. Pei, X. Liu, D. Xu, L. Liu, H. Ma, "Octopus: Exploiting the Edge Intelligence for Accessible 5G Mobile Performance Enhancement", IEEE/ACM Transactions on Networking, vol. 31, pp. 2454-2469, 2023.

[15] S.K. Kumar, V.S. Kumar, R. Krishnamoorthy, P. Mahesh, B. Haravohra layya, N. Soni, S.K. Gupta, "The 5G Revolution: Tackling Challenges in Smart Cities and Intelligent Transportation Systems", Journal of Autonomous Intelligence, vol. 7, pp. 1342, 2024.

[16] A. Mukherjee, M. Sasidharan, M. Herrera, A.K. Parlikad, "Unsupervised Constrained Discord Detection in IoT-Based Online Crane Monitoring", Advanced Engineering Informatics, vol. 60, pp. 102444, 2024.

[17] M. Bonato, G. Tognola, M. Benini, S. Gallucci, E. Chiaramello, S. Fiocchi, M. Parazzini, "Stochastic Dosimetry Assessment of Human RF-EMF Spatial Exposure Variability in 5G-V2X Vehicular Communication Scenario", IEEE Access, vol. 11, pp. 94962-94973, 2023.

A Rule-Based Framework for Clothing Fit Recommendation from 3D Body Reconstruction

Hamid Ouhnni[1], Acim Btissam[2], Belhiah Meryam[3], Benachir Rigalma[4], Soumia Zit[5]
Computer Science Department Intelligent Processing and Security of Systems (IPSS) Laboratory-Faculty of Sciences, Mohammed V University, Rabat, Morocco[1, 2, 4]
Doctor of Engineering Professor (Associate), Mohammed V University in Rabat, Morocco[3]
Professor of Higher Education at Faculty of Sciences, Mohammed V University, Rabat. Morocco[5]

***Abstract*—This research presents a comprehensive framework for body size estimation that accurately derives anthropometric measurements—specifically, the circumferences of the waist and hips—from a singular image by utilizing OpenPose for joint localization and SMPLify-X for precise 3D body modeling. The proposed methodology involves projecting the generated three-dimensional model onto a horizontal plane and applying a convex hull geometric assessment to extract relevant body measurements. These derived measurements are then classified into standardized clothing size predictions (XS–XL) via a transparent rule-based classification system suitable for e-commerce sizing and virtual fitting applications. Empirical validation conducted on the Agora dataset substantiated the framework's reliability across diverse body types, demonstrating strong consistency with industry sizing standards. The method is non-intrusive and interpretable, effectively addressing practical challenges in automated human pose estimation for retail contexts. Limitations include constraints related to body posture and potential clothing interference; however, the modular design enables enhancements such as integrating chest circumference measurements and mobile deployment. This scholarly contribution thus provides a robust, accessible solution for automated, image-based clothing size recommendations.**

Keywords*—Body size estimation; SMPLify-X; OpenPose; 3D body modeling; clothing size prediction; e-commerce sizing; human pose estimation

I. Introduction

In the current digital epoch, marked by the extensive proliferation of e-commerce, the necessity for accurate and sophisticated techniques to derive human body dimensions from images has markedly escalated—especially in the ready-to-wear fashion industry, augmented reality applications, and virtual fitting features within online retail contexts. A predominant obstacle encountered by online merchants pertains to size-related product returns. Research indicates that as much as 30% of returned apparel arises from discrepancies in sizing or erroneous estimations at the time of purchase [1]. This predicament not only incurs financial and logistical detriments for enterprises but also adversely affects user experience, diminishes customer satisfaction, and erodes trust in virtual retail platforms.

In response, progressive computer vision and 3D modeling technologies have surfaced as promising instruments for addressing this quandary. These technologies facilitate the extraction of body shape, dimensions, and measurements from a solitary 2D image—a process increasingly bolstered by powerful machine learning frameworks. OpenPose, for example, proficiently extracts human joint keypoints under diverse real-world imaging scenarios, including suboptimal lighting or cluttered backgrounds [2], [3].

The identified joints can subsequently be aligned with 3D parametric models such as SMPL and SMPL-X through algorithms like SMPLify-X, which captures intricate body characteristics including hands and facial expressions [4], [5]. Models such as DensePose2SMPL and ExPose have also demonstrated high precision in fitting under various conditions, encompassing loose garments and complex poses [6], [7].

Recent research corroborates that SMPL and SMPL-X surpass conventional measurement techniques by substantially minimizing errors in body circumference estimation, particularly when side-view or fitted clothing data is accessible [8]. Furthermore, models like LayerNet and SMPLicit provide garment-aware reconstructions with semantic garment-layer processing, advancing the frontiers of practical virtual try-on functionalities [9], [10].

Nonetheless, numerous existing frameworks still encounter limitations when faced with non-standard postures, oversized attire, or occluded limbs [11], [12].

To confront these challenges, this study proposes a comprehensive, fully automated pipeline for body size estimation from a single RGB image. The methodology commences with joint detection via OpenPose, followed by 3D body modeling utilizing SMPLify-X, digital measurement calculation, and conclusive size classification employing rule-based logic. The pipeline is crafted for resilience in varied real-world scenarios and can be seamlessly integrated into virtual fitting and e-commerce platforms [13], [14].

For validation, this research employs the Agora dataset [15], which encompasses multi-pose images alongside corresponding 3D data files (obj, pkl, and openpose.json). This extensive dataset facilitates a systematic assessment of each component within the pipeline—from image acquisition to 3D modeling and final garment size categorization.

The remainder of this study is organized as follows: Section II reviews related literature. Section III details the proposed methodology. Section IV presents experimental results. Section V discusses findings and limitations, and Section VI concludes the work.

II. Literature Review

The domain of automatic estimation of human body dimensions has experienced significant advancements in recent years, primarily driven by the exponential rise of e-commerce, virtual fashion, and the increasing need for intelligent digital interfaces that accurately replicate physical human characteristics. A major component of this evolution lies in computer vision methodologies, particularly the synergistic use of OpenPose [16], SMPL [17], and its advanced variants such as SMPLify-X [5], which have enabled the precise extraction and modeling of body metrics from single RGB images.

A notable contribution in this space includes the work of Kuribaya [2], who developed a virtual try-on system capable of dynamically adapting garment sizes to match real-world anthropometrics extracted from joint keypoints. Although the model was lauded for its interactive fit visualization, it primarily prioritized graphical realism over the quantitative accuracy of body dimensions. A similar limitation was observed in the work of Yang et al [18], who proposed a GAN-based virtual try-on model that successfully preserved garment texture and pose alignment but exhibited noticeable fitting errors for users with non-standard body shapes, particularly in waist and hip regions—highlighting the trade-off between aesthetic rendering and dimensional fidelity.

Meanwhile, more metric-focused efforts have emerged. In a large-scale study by [1], a decision tree classifier trained on joint-based features from OpenPose was employed to predict military uniform sizes across 375 subjects. Cross-validation with 3D scanning and manual tape measurements demonstrated over 85% agreement, validating the approach for structured, high-stakes domains. Nonetheless, the scalability of such models to unconstrained consumer environments remains uncertain.

To address modeling accuracy for atypical body shapes, the DensePose2SMPL framework [6] introduced dense correspondence mapping from images to SMPL-based mesh representations. Their method reduced perimeter estimation error by 30% in outlier body cases, underscoring the advantages of combining surface-based mappings with statistical shape modeling.

Another critical advancement came from the DeepProfile model [19], which leveraged deep convolutional networks to predict occluded body shapes beneath garments. Their approach demonstrated superior performance under challenging visual conditions—such as loose clothing—when benchmarked against traditional silhouette-extraction methods [20]. However, its dependency on extensive labeled datasets could pose limitations in diverse real-world applications.

The generative model SMPLicit [10] advanced the state of body-clothing co-representation by embedding garment topology directly into the SMPL parameter space. This enabled realistic simulation of both tight and open garments, making it particularly suited for virtual try-on platforms and AI-driven fashion editors. Nevertheless, the model lacked direct mechanisms for deriving standardized measurements (e.g., waist, chest, hip) used in real-world size classification.

Exploring the democratization of body estimation, Sheth [21] introduced a selfie-based body measurement method using mobile phone images and mirror-based cues. Their system achieved <5% error across 11 anthropometric parameters. Such accessible methodologies—also explored by Jiang and Grauman (2020) in monocular pose-to-shape prediction—illustrate the potential for casual consumer-level body scanning, albeit with constraints related to environment and pose consistency.

From an industrial standpoint, [22] illustrated that three-dimensional geometric modeling generated from comprehensive bodily laser imaging substantially improved the precision of uniform fitting, particularly for female personnel in military environments. Expanding upon this foundation, contemporary developments within the apparel sector have underscored the incorporation of three-dimensional body scans alongside machine learning to facilitate mass customization. For instance, [23] presented a hybrid methodology that merges point cloud information from 3D scanners with clustering algorithms to categorize body types and refine garment patterning. Their methodology surpassed conventional sizing systems in terms of both fit satisfaction and production efficiency, indicating a transition towards volumetric and data-informed design within the realm of fashion e-commerce.

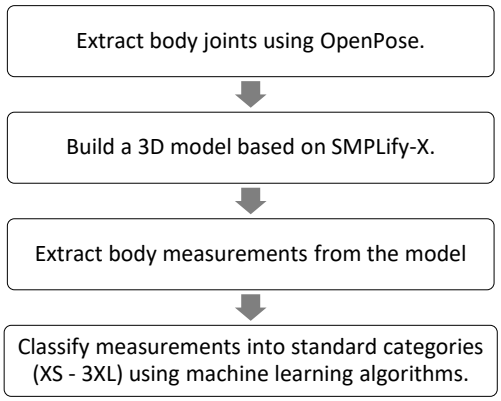
Evaluation strategies across these studies have varied—from manual anthropometry to laser scanning, structured light capture, and virtual garment draping—highlighting a lack of standardized benchmarking protocols. Efforts like the CAESAR dataset [24] and newer 3DHumanDataset initiatives [25] aim to offer such standards, though generalization to unconstrained consumer imagery remains an open research challenge.

While antecedent methodologies have achieved notable advancements in either augmenting garment authenticity, forecasting dimensions utilizing joint-based classifiers, or reconstructing intricate three-dimensional models beneath loose apparel, a limited number of techniques furnish a completely automated, interpretable, and rule-based framework that is simultaneously scalable and independent of extensive training datasets. In contrast to generative or learning-centric frameworks that may experience challenges related to opacity or overfitting, the proposed approach presents a transparent workflow anchored in anatomical reference points and geometric evaluation (e.g., convex hulls) to ascertain waist and hip circumferences. By amalgamating OpenPose and SMPLify-X with rule-based size classification, our methodology facilitates real-world applicability within e-commerce contexts—achieving elevated accuracy without the necessity for comprehensive manual annotations or costly scanning technology. This positions it uniquely for economically viable implementation across a variety of consumer settings, distinguishing it from prior models that emphasize visual fidelity or necessitate controlled acquisition environments.

While prior research has extensively explored machine learning-based frameworks, fewer studies have provided a fully rule-based approach. Our proposed methodology uniquely integrates OpenPose and SMPLify-X to deliver practical size recommendations without extensive training data, bridging an important gap between theoretical models and industry practicality.

III. METHOD

The framework aims to provide a comprehensive, systematic solution for estimating body measurements using only a single image. The framework relies on a series of organized steps:



A. Step 1: Joint Extraction from the Image

OpenPose is applied to detect 25 key body joints, covering the upper limbs (shoulders, elbows, wrists), lower limbs (hips, knees, ankles), torso (neck, shoulders, chest center, hips), and head region (eyes, ears, nose).

The system uses deep neural networks capable of identifying joint positions with high accuracy, even under challenging visual conditions involving poor lighting or noisy backgrounds.

The output consists of a 2D array of (x, y) coordinates for each joint, along with a confidence score per point, enabling the exclusion or correction of joints with low reliability.

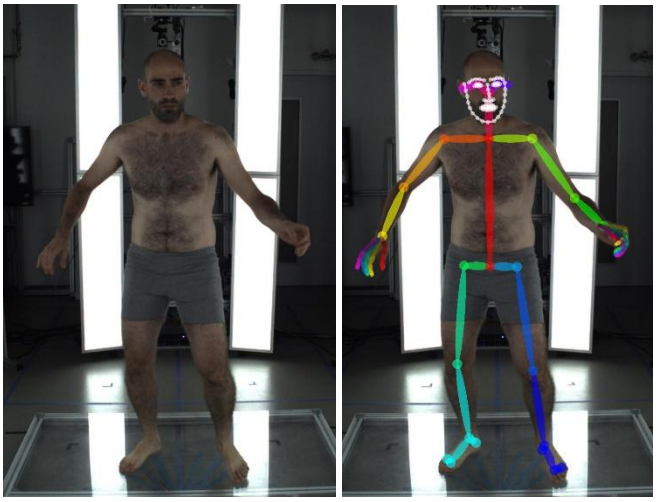


Fig. 1. Pose estimation from a single image using openpose.

B. Step 2: 3D Body Model Generation

The SMPLify-X algorithm is used to align a statistical 3D human model (SMPL-X) with the 2D joint keypoints extracted by OpenPose. SMPLify-X is an iterative optimization method that refines the model's shape and pose to match the detected joints.

SMPL-X is a comprehensive parametric model that accurately represents the full human body, including facial features, hand gestures, and body posture—making it ideal for virtual reality and interactive applications.

The OpenPose keypoints as shown in Fig. 1 and Fig. 2 serves as constraints for the SMPLify-X optimization, allowing it to fine-tune the model parameters (e.g., shape, pose, gender) until the 3D mesh closely fits the user's body as inferred from the image.

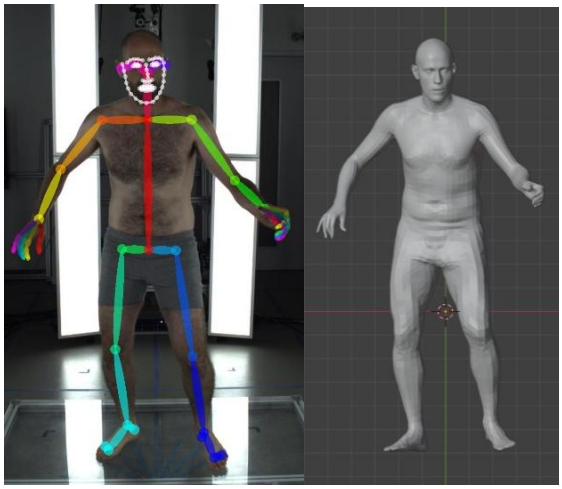


Fig. 2. From 2D pose estimation to 3D body reconstruction using SMPLIFY-X.

Upon the successful calibration of the model, two principal output files are produced:

- obj file: This file encompasses the surface mesh representation of the three-dimensional body, detailing vertex coordinates and triangular relationships. It serves a pivotal role in visualization and rendering processes.
- pkl file: This file encapsulates the internal parameters of the model, which is especially vital for this research due to its interoperability with Python-based analytical frameworks. It comprises:
    - keypoints_3d: A matrix that delineates the three-dimensional skeletal joints in the BODY_24 format, accurately aligned with the surface mesh and utilized for the identification of anatomical landmarks such as the cervical region and the hip joints.
    - Vertices (mesh): A comprehensive three-dimensional point cloud that depicts the external dermal surface of the body, constituting the geometric foundation for the computation of circumferences through projection and convex hull analysis.
    - gender: A categorical variable that denotes the biological sex of the subject (male or female). This parameter holds significant importance, given that clothing size specifications frequently vary by gender, and precise categorization is essential for appropriate alignment with the relevant standard sizing framework.

By utilizing the pkl file, the framework gains the advantage of seamless integration with Python's data management functionalities, while simultaneously accessing the requisite anatomical and mesh data critical for accurate, automated size estimation.

C. Step 3: Standard Size Classification Using Rule-Based Tables

Accurate size determination for clothing relies primarily on three anatomical measurements: waist width, hip width and chest width, as illustrated in Fig. 3.

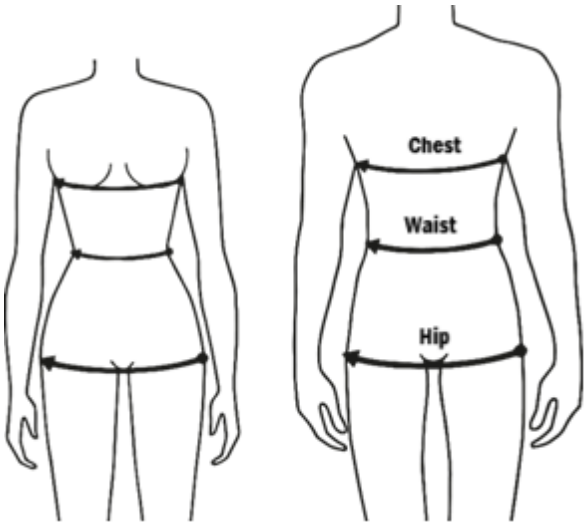


Fig. 3. The Anatomical locations of chest, waist, and hip circumferences used in garment size estimation.

- Waist width refers to the horizontal circumference of the torso at the narrowest region between the ribcage and pelvis.
- Hip width is the horizontal circumference at the widest point below the waist.
- Chest width: Measure the fullest part of your bust while wearing a bra that fits.

Rather than relying on a single point or linear measurement, the framework analyzes entire cross-sectional slices of the body to provide a more realistic representation of body shape, accounting for anatomical variations across individuals.

1) Identifying waist, hip, and chest levels. The positions of the waist and hip are inferred from skeletal joints stored in the .pkl file under the array keypoints_3d. This array contains the 3D coordinates (X, Y, Z) of 25 keypoints following the BODY25 format.

Two key points, as shown in Fig. 4, are central to this analysis:

- Mid-Hip (index 8): Represents the center of the pelvis or hip region.
- Neck (index 1): Marks the lower end of the neck and the start of the torso.

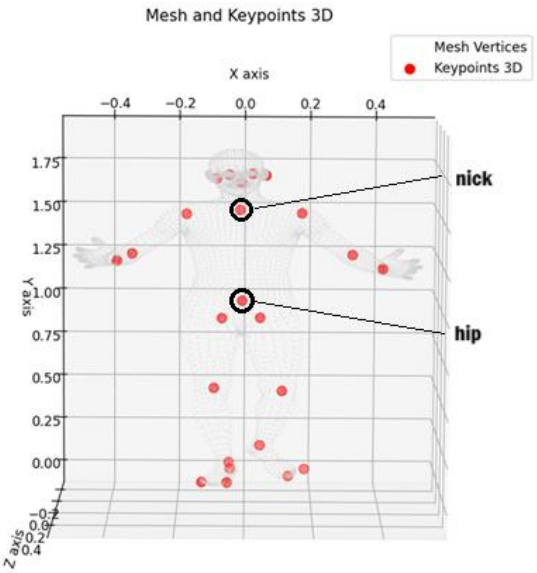


Fig. 4. Integration of 3D mesh and skeletal keypoints for anatomical reference.

The waist level is estimated by averaging the vertical (Y) coordinates of the Mid-Hip and Neck keypoints. This approach yields three critical horizontal slices:

- Hip slice: Taken at the same vertical level as the mid-hip joint (y_hip).
- Waist slice: Estimated to be roughly halfway between the neck and the hip, calculated as the average of their vertical coordinates y_waist ≈ y_hip +(y_hip + y_neck)̊ * 0.4.
- Chest slice: Located near the neck joint (y_neck), representing the lower section of the torso.

The parameter settings, such as the 40% ratio used to estimate waist position, were empirically chosen based on anatomical observations and real-world validation. Future work may refine these parameters for specific demographic groups."

This method allows for the extraction of more anatomically accurate cross-sections, improving measurement precision compared to single-point estimation methods.

a) Projection to the Horizontal Plane (X-Z): The three-dimensional coordinates utilized in each horizontal section are obtained from the surface mesh of the anatomical structure, specifically from the array of vertices that delineate the external geometric configuration of the model. These coordinates are projected onto the X-Z plane through the elimination of the vertical (Y) component. This two-dimensional projection facilitates a more straightforward analysis of the cross-sectional morphology of the body, rendering it appropriate for the estimation of circumference values via geometric methodologies.

The precision of this projection-centric methodology is predominantly ascribed to the incorporation of anatomically significant reference points—namely the Neck (index 1) and Mid-Hip (index 8)—whose vertical coordinates are retrieved from the array of key points. These joints function as dependable

reference markers for the estimation of the positions of the chest, waist and hip sections, thereby ensuring uniform alignment across disparate body models.

b) Convex Hull Application: A Convex Hull algorithm is employed on the two-dimensional projected coordinates to generate the minimal closed polygon that encompasses all data points.The convex hull simplifies the body's outline at each cross-section, removing small irregularities while retaining overall shape.It effectively eliminates minor irregularities or localized noise within the mesh that do not significantly influence the overall shape.

The perimeter or area of the resultant convex hull, as shown in Fig. 5, is subsequently calculated and utilized as an approximate indicator of the waist or hip circumference, contingent upon the specific slice under examination.

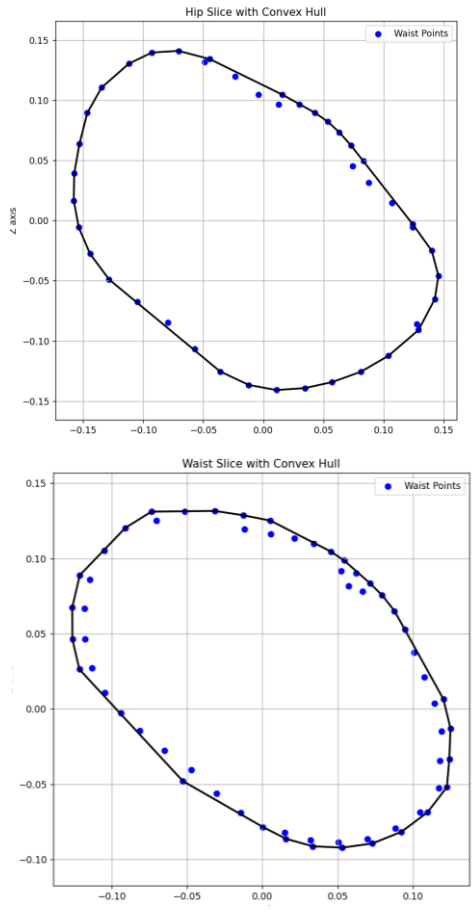


Fig. 5. Projection of waist and hip vertices onto X-Z plane.

2) Conversion to standard sizes. In the concluding phase, the extracted body measurements are compared against established standardized sizing tables (e.g., ISO or specific retailer charts) to ascertain the most suitable clothing size category, such as XS, S, M, L, XL, among others (see Table I).

This classification is executed through a rule-based methodology, wherein each measurement (e.g., chest, waist, hips) is juxtaposed with threshold ranges corresponding to each size category. The system identifies the closest matching category, optionally employing priority rules (e.g., prioritizing waist over chest for trousers). This methodology guarantees transparency, consistency, and compatibility with pre-existing sizing systems employed by manufacturers and online retailers—without necessitating the implementation of machine learning models.

There exist numerous clothing sizing frameworks utilized globally, differing according to geographical region, brand identity, and gender classification. These frameworks encompass systems such as those employed in the United States, the United Kingdom, the European Union, and the internationally recognized ISO standards, each characterized by distinct measurement parameters and labeling protocols. In the present investigation, we have adopted a prevalent and comprehensive sizing framework that corresponds with international standards and is extensively utilized within both retail and e-commerce sectors. This approach guarantees compatibility across a wide array of applications and enhances the practical significance of the proposed conceptual model for users worldwide.

TABLE I. STANDARD WOMEN'S CLOTHING SIZE CHART (IN CM)

WOMENS SIZES (CM)						
Size	**XS**	**S**	**M**	**L**	**XL**	**XXL**
Bust (cm)	85	91	95	100	105	110
Waist (cm)	80	86	90	95	100	105
Hips (cm)	90	96.5	100	105	110	115
Front lenth (cm)	115	117	115	115	115	115

Source: https://ethnicity.in/pages/size-guide

Following common fashion industry practices, the larger of the two measurements (waist or hip) is used to determine the final clothing size. This ensures user comfort and avoids tight fitting, especially for garments like pants and skirts.

3) Code summary – core measurement extraction logic (working for waist and hip). The following Python snippet illustrates the core logic used to extract the waist and hip circumference from the 3D body mesh by projecting relevant surface vertices and applying a convex hull algorithm (Algorithm 1):

This code performs three main operations:

Algorithm 1 : Python

```
# Step 1: Define hip and waist levels directly from keypoints
y_hip = keypoints[8, 1] # Mid-Hip point (keypoint index 8)
y_neck = keypoints[1, 1] # Neck point (keypoint index 1)
y_waist = y_hip + np.abs(y_neck - y_hip) * 0.40 # Estimated waist level

# Step 2: Select surface vertices near the hip and waist levels
hip_vertices = vertices[np.abs(vertices[:, 1] - y_hip) < 0.015]
waist_vertices = vertices[np.abs(vertices[:, 1] - y_waist) < 0.015]

# Step 3: Project 3D points onto the X-Z plane for 2D contour analysis
hip_2d = hip_vertices[:, [0, 2]]
waist_2d = waist_vertices[:, [0, 2]]
```

```
    # Step 4: Compute the Convex Hull to approximate the circumference
    hip_hull = ConvexHull(hip_2d)
    waist_hull = ConvexHull(waist_2d)
```

The vertical positions (y_hip, y_waist) are derived using anatomical reference points from the keypoints_3d array, aligned using a displacement correction to compensate for origin mismatch between the vertex and joint spaces.

The waist level is estimated as a proportional offset (40%) between the hip and neck points—empirically chosen to reflect typical torso proportions.

Vertices within a narrow band (±1.5 cm in normalized units) around the estimated hip and waist levels are selected, effectively isolating the relevant cross-sections of the 3D body mesh.

The selected 3D points are projected onto the X-Z plane to obtain a 2D shape of the body cross-section.

ConvexHull() from the scipy.spatial package is used to compute the enclosing perimeter or area of the projected points—used as a proxy for waist and hip circumference.

This computational process allows for non-contact, geometry-based body measurement extraction and is robust against mesh irregularities or minor noise in the 3D data.

4) Size mapping through standardized chart. Subsequent to the calculation of waist and hip circumferences utilizing the convex hull perimeters, the derived values are juxtaposed against an established size chart in order to allocate a standardized clothing size designation. The chart, delineated in centimeters, encompasses threshold ranges for both waist and hip measurements and was formulated by averaging the discrepancies between contiguous size categories.

The executed size mapping dictionary is as follows in Table II.

TABLE II. THE EXECUTED SIZE MAPPING DICTIONARY

```
size_chart = {
"XS": {"waist": (60, 82.999), "hip": (88, 92.999)},
"S": {"waist": (83, 88.999), "hip": (93, 98.999)},
"M": {"waist": (89, 92.999), "hip": (99, 102.999)},
"L": {"waist": (93, 97.999), "hip": (103, 107.999)},
"XL": {"waist": (98, 102.999), "hip": (108, 112.999)},
"XXL": {"waist": (103, 107.999), "hip": (113, 117.999)},
 }
```

This table was developed by centering the interval disparities among proximate garment sizes in accordance with established industry benchmarks. The ultimate designated size aligns with the minimal size classification wherein both the projected waist and hip circumferences reside within the stipulated parameters. Conversely, for applications prioritizing comfort, the more expansive corresponding category may be selected to guarantee garment ease.

This methodical mapping strategy provides a lucid and comprehensible alternative to opaque machine learning classifiers and retains congruence with widely recognized retail sizing frameworks.

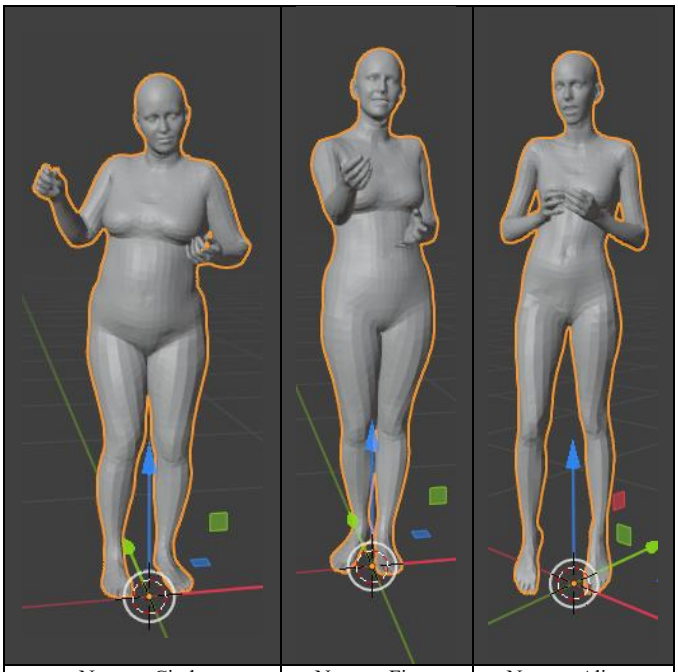
IV. RESULTS SUMMARY

A. Case Study: Evaluation on Agora Dataset Using (Waist , Hip)

In order to illustrate the practical implementation and credibility of the proposed framework, a collection of authentic three-dimensional body models was curated from the Agora dataset, which offers intricate anthropometric meshes that represent a variety of female body types (see Fig. 6).

The Agora dataset was chosen due to its high-quality anthropometric models and comprehensive skeletal keypoints, enabling precise evaluation across diverse body shapes and realistic scenarios.

In order to assess the practical applicability of the proposed framework, a variety of three-dimensional body models sourced from the Agora dataset were subjected to processing through the entirety of the established pipeline. Among these instances, three representative cases—Cindy, Fiona, and Alixon—have been selected to exemplify the diversity in body morphology and the corresponding assignment of clothing sizes. The larger collection of evaluated models demonstrated analogous results, thereby suggesting a uniform operational consistency of the system across varying anatomical configurations.



Name = Cindy	Name = Fiona	Name = Alixon
Hip = 1.124 m	Hip = 0.992 m	Hip = 0.892 m
Waist = 0.942 m	Waist = 0.825 m	Waist = 0.723 m
Rec = XL	Rec = M	Rec =XS

Fig. 6. Estimated waist and hip circumferences for three subjects (Alixon, Fiona, Cindy) and the corresponding predicted clothing sizes.

B. Analysis

In the case of Cindy, both her waist (94.2 cm) and hip (112.4 cm) dimensions distinctly align with the XL category as delineated by the specified size chart.

Fiona presents intermediate measurements—waist at 82.5 cm and hip at 99.2 cm—rendering her qualifying for the M category.

Alixon, possessing a waist measurement of 72.3 cm and a hip measurement of 89.2 cm, comfortably fits within the XS category.

The cross-sectional contours of each model were subjected to analysis via convex hulls within the X-Z plane. The findings indicate a robust correlation between the visually perceived body type and the size predicted by the proposed methodology, thereby affirming its efficacy within a practical context.

These findings underscore the framework's proficiency in accommodating various body proportions and autonomously delivering realistic size recommendations utilizing solely a singular three-dimensional mesh and joint keypoints as input.

V. Discussion

The findings derived from the Agora dataset underscore the practical viability and precision of the proposed image-based body measurement framework. By implementing a coherent and modular pipeline—encompassing joint extraction, three-dimensional body modeling, geometric perimeter analysis, and rule-based classification—the system adeptly estimated standardized clothing sizes across a diverse array of body shapes.

A. Strengths

Robustness Across Body Variants: The framework proficiently differentiated between various anatomical types (e.g., Cindy vs. Alixon), encapsulating both volumetric and proportional variances. Mesh-based modeling approaches like DensePose2SMPL have shown improved shape estimation performance, particularly in subjects with outlier body types [26].

Alignment with Industry Standards: Size predictions (e.g., XS, M, XL) corresponded favorably with anticipations derived from body measurements, affirming the legitimacy of convex hull-based perimeter estimation. Previous studies applying 3D scanning in apparel industries highlighted the effectiveness of geometric-based measurements for standardized size matching [27].

Contactless and Scalable: Functioning on a singular image without necessitating specialized equipment, the methodology is well-positioned for practical implementation within e-commerce and virtual try-on systems. Research on mobile- and web-based 3D scanning platforms confirms growing consumer demand for low-barrier body modeling solutions [28].

B. Potential Applications

In virtual shopping platforms, this technology allows for realistic virtual try-on experiences where garments are simulated directly onto a user's 3D model. Projects like SIZER have demonstrated the practicality of linking clothing fit with volumetric body data to personalize the shopping journey [29]. More importantly, by linking the extracted body measurements directly to sizing charts provided by online clothing retailers, the system enables users to select the most accurate size before making a purchase. This pre-selection step significantly reduces the risk of sizing mismatches, thereby lowering return rates and associated logistical costs [30].

The ability to preview fit and style interactively empowers users to make confident decisions, while retailers benefit from fewer returns and improved customer satisfaction. This seamless integration between virtual fitting and e-commerce platforms bridges the gap between user expectation and product delivery [28].

Beyond virtual worlds, the framework paves the way for intelligent mobile applications that can, from a single image, provide users with real-time recommendations not only for size but also for clothing styles that best suit their body shapes. Such applications could incorporate AI-driven styling advice based on user preferences, fashion trends, and body geometry—delivering a personalized and seamless shopping experience that merges practicality with convenience [31].

C. Limitations

Fixed Anatomical Ratios: Waist-level estimation based on a fixed ratio (e.g., 40% between neck and hip) may not generalize across different ethnicities, age groups, or postures, potentially leading to inaccuracies in landmark placement. Research highlights that human anatomical diversity undermines the reliability of fixed-ratio methods in body reconstruction [32].

Chest Measurement Challenges: Accurate chest measurement is hindered by the pose limitations of many 3D body models, which often lack favorable positions such as the Y-pose. This results in arm–torso intersections that distort the contour, especially around the upper torso. Prior work has demonstrated that dressed-human silhouettes limit measurement precision when key anatomical landmarks are occluded or misaligned [33].

Pose and Environment Dependence: The system was tested on controlled, static 3D models, which may not reflect performance in real-world scenarios. Variations in lighting, background, and body pose can adversely impact the accuracy of joint detection and mesh reconstruction. This dependency is consistent with findings that stress the sensitivity of monocular 3D body estimation to environmental conditions and pose inconsistencies [34].

Clothing Interference: Although models like SMPLify-X can recover approximate body shapes under loose garments, excessive layering or oversized clothing still causes significant contour deviation. Studies confirm that loose-fitting clothes often degrade estimation accuracy, especially in peripheral regions such as the hips and chest, suggesting tight-fitting garments are optimal for body capture [31].

D. Future Directions

Future advancements could concentrate on adapting pose-aware corrections for chest measurement through pose normalization or generative completion methodologies, implementing the model in mobile applications with real-time

guidance on pose and clothing preparation, and augmenting the framework with deep-learning-based modules to enhance anatomical landmark detection under conditions of occlusion or partial visibility.

VI. CONCLUSION

This study presented and validated a comprehensive, image-based framework for estimating human body measurements, specifically waist and hip circumferences, using a single image processed by OpenPose, SMPLify-X, and convex hull geometric analysis. These measurements were systematically mapped to standard clothing sizes through a clear and scalable rule-based classification, enabling practical, non-intrusive size recommendations.

Experimental evaluations using three-dimensional body models from the Agora dataset confirmed the method's effectiveness in accurately identifying body variations and recommending appropriate clothing sizes (XS–XL). Results showed strong alignment with established industry standards, highlighting the robustness, cost-efficiency, and interpretability of the proposed framework. Consequently, this research contributes meaningfully towards practical applications in e-commerce, virtual fitting solutions, and body-shape analytics.

However, limitations of the study include reliance on fixed anatomical ratios that might not generalize across different ethnicities, age groups, or body postures, as well as potential inaccuracies due to clothing interference or complex poses. Future work could address these limitations by refining landmark detection, improving pose normalization, and adapting the framework for mobile and real-world scenarios.

Overall, this research significantly advances the accessibility and practical implementation of image-based body size estimation, facilitating personalized garment fitting in a contactless, efficient, and user-friendly manner.

ACKNOWLEDGMENT

The authors would like to thank the Intelligent Processing and Security of Systems (IPSS) Laboratory for providing computational resources used in this research.

REFERENCES

[1] C.-H. Wen, C.-C. Cheng, et Y.-C. Shih, « Artificial intelligence technologies for more flexible recommendation in uniforms », Data Technol Appl, vol. 56, p. 626-643, 2022.

[2] M. Kuribayashi, K. Nakai, et N. Funabiki, « Image-Based Virtual Try-on System With Clothing-Size Adjustment », ArXiv, vol. abs/2302.14197, 2023, [En ligne]. Disponible sur: https://api.semanticscholar.org/CorpusID:257232570

[3] G. Shi, C. Gao, D. Wang, et Z. Su, « Automatic 3D virtual fitting system based on skeleton driving », Vis. Comput., vol. 37, p. 1075-1088, 2020.

[4] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, et M. J. Black, « Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image », ArXiv, vol. abs/1607.08128, 2016, [En ligne]. Disponible sur: https://api.semanticscholar.org/CorpusID:13438951

[5] G. Pavlakos et al., « Expressive Body Capture: 3D Hands, Face, and Body From a Single Image », 2019 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR, p. 10967-10977, 2019.

[6] D. Gu, Y. Yun, T. T. Tuan, et H. Ahn, « Dense-Pose2SMPL: 3D Human Body Shape Estimation From a Single and Multiple Images and Its Performance Study », IEEE Access, vol. 10, p. 75859-75871, 2022.

[7] F. Y. Wei, K. H. Lim, J. T. S. Phang, et P. K. Pang, « Implementation of Cloth Estimation in 2D-3D Human Body Regression Model », 2024 Int. Conf. Green Energy Comput. Sustain. Technol. GECOST, p. 224-228, 2024.

[8] Z. Xu et Q. Zhang, « Boundary-aided Human Body Shape and Pose Estimation from a Single Image for Garment Design and Manufacture », in Eurographics, 2018. [En ligne]. Disponible sur: https://api.semanticscholar.org/CorpusID:51873845

[9] E. Corona, G. Alenyà, G. Pons-Moll, et F. Moreno-Noguer, « LayerNet: High-Resolution Semantic 3D Reconstruction of Clothed People », IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, p. 1257-1272, 2023.

[10] E. Corona, A. Pumarola, G. Alenyà, G. Pons-Moll, et F. Moreno-Noguer, « SMPLicit: Topology-aware Generative Model for Clothed People », 2021 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR, p. 11870-11880, 2021.

[11] C. Ferrari, L. Casini, S. Berretti, et A. Bimbo, « Monocular 3D Body Shape Reconstruction under Clothing », J. Imaging, vol. 7, 2021, [En ligne]. Disponible sur: https://api.semanticscholar.org/CorpusID:244828943

[12] K. S. P. Thota, S. Suh, B. Zhou, et P. Lukowicz, « Estimation Of 3d Body Shape And Clothing Measurements From Frontal-And Side-View Images », 2022 IEEE Int. Conf. Image Process. ICIP, p. 2631-2635, 2022.

[13] G. Mantha, P. G. S. Rahul, A. K. M, S. P. Bhat, et S. G. P, « Preserving Network and adaptive content generating method for Image based virtual try-on clothing system », JNNCE J. Eng. Manag., 2022, [En ligne]. Disponible sur: https://api.semanticscholar.org/CorpusID:264296628

[14] T. T. Thai, Y. Yun, et H. Ahn, « High-quality 3D Clothing Reconstruction and Virtual-Try-On: Pants case », 2022 Int. Conf. Multimed. Anal. Pattern Recognit. MAPR, p. 1-6, 2022.

[15] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, et M. J. Black, « AGORA: Avatars in Geography Optimized for Regression Analysis », 2021 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR, p. 13463-13473, 2021.

[16] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, et Y. Sheikh, « OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields », IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, p. 172-186, 2018.

[17] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, et M. J. Black, « SMPL: a skinned multi-person linear model », ACM Trans Graph, vol. 34, no 6, oct. 2015, doi: 10.1145/2816795.2818013.

[18] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, et P. Luo, « Towards Photo-Realistic Virtual Try-On by Adaptively Generating↔Preserving Image Content », in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, juin 2020, p. 7847-7856. doi: 10.1109/CVPR42600.2020.00787.

[19] S. Lu, F. Lu, X. Shou, et S. Zhu, « DeepProfile: Accurate Under-the-Clothes Body Profile Estimation », Appl. Sci., vol. 12, no 4, 2022, doi: 10.3390/app12042220.

[20] A. Zanfir, E. Marinoiu, et C. Sminchisescu, « Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes: The Importance of Multiple Scene Constraints », in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE, juin 2018, p. 2148-2157. doi: 10.1109/CVPR.2018.00229.

[21] M. Sheth et N. Srivastava, « Predicting Body Size Using Mirror Selfies », in Intelligent Human Computer Interaction: 11th International Conference, IHCI 2019, Allahabad, India, December 12–14, 2019, Proceedings, Berlin, Heidelberg: Springer-Verlag, 2019, p. 151-162. doi: 10.1007/978-3-030-44689-5_14.

[22] H. Daanen et al., « 3D Scanning for Clothing Fit and Logistics », Sci. Technol., 2004.

[23] J. Wang, X. Li, L. Pan, et C. Zhang, « Parametric 3D modeling of young women's lower bodies based on shape classification », Int. J. Ind. Ergon., vol. 84, p. 103142, 2021, doi: https://doi.org/10.1016/j.ergon.2021.103142.

[24] K. Robinette et al., « ANTHROPOMETRY RESOURCE (CAESAR) FINAL REPORT, VOLUME I: SUMMARY », janv. 2002.

[25] Y. Zhang, « Non-Contact Human Body Dimension Estimation Methods Based on Deep Learning: A Critical Analysis », Theor. Nat. Sci., vol. 52, no 1, p. 214-221, 2024, doi: 10.54254/2753-8818/52/2024ch0135.

[26] D. Gu, Y. Yun, T. T. Tuan, et H. Ahn, « Dense-Pose2SMPL: 3D Human Body Shape Estimation From a Single and Multiple Images and Its Performance Study », IEEE Access, vol. 10, p. 75859-75871, 2022.

[27] P. R. Apeagyei, « Application of 3D body scanning technology to human measurement for clothing Fit », J Digit Content Technol Its Appl, vol. 4, p. 58-68, 2010.

[28] S. Petrak, A. Grilec, et M. M. Naglić, « Consumer Perception of 3D Body Scanning While Online Shopping for Clothing », Int. J. Humanit. Soc. Sci., vol. 13, p. 1025-1030, juin 2019.

[29] G. Tiwari, B. L. Bhatnagar, T. Tung, et G. Pons-Moll, « SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing », ArXiv, vol. abs/2007.11610, 2020, [En ligne]. Disponible sur: https://api.semanticscholar.org/CorpusID:220713845

[30] Z. H. Wang, Y. Zhong, K. J. Chen, J. Y. Ruan, et J. Zhu, « 3D Human Body Data Acquisition and Fit Evaluation of Clothing », Adv. Mater. Res., vol. 989-994, p. 4161-4164, 2014.

[31] S. Lu, F. Lu, X. Shou, et S. Zhu, « DeepProfile: Accurate Under-the-Clothes Body Profile Estimation », Appl. Sci., 2022, [En ligne]. Disponible sur: https://api.semanticscholar.org/CorpusID:247075759

[32] K. S. P. Thota, S. Suh, B. Zhou, et P. Lukowicz, « Estimation Of 3d Body Shape And Clothing Measurements From Frontal-And Side-View Images », 2022 IEEE Int. Conf. Image Process. ICIP, p. 2631-2635, 2022.

[33] D. Song, R. Tong, J. Chang, X. Yang, M. Tang, et J. Zhang, « 3D Body Shapes Estimation from Dressed - Human Silhouettes », Comput. Graph. Forum, vol. 35, 2016, [En ligne]. Disponible sur: https://api.semanticscholar.org/CorpusID:41677034

[34] A. Caliskan, A. Mustafa, E. Imre, et A. Hilton, « Multi-View Consistency Loss for Improved Single-Image 3D Reconstruction of Clothed People », in Asian Conference on Computer Vision, 2020. [En ligne]. Disponible sur: https://api.semanticscholar.org/CorpusID:221995477

TL-MC-ShuffleNetV2: A Lightweight and Transferable Framework for Elevator Guideway Fault Diagnosis

Zhiwei Zhou[1], Xianghong Deng[2], Xuwen Zheng[3], Chonlatee Photong[4]*
School of Electrical Engineering, Hunan Mechanical & Electrical Polytechnic, Changsha, 410151, Hunan, China[1, 2]
Faculty of Engineering, Mahasarakham University, Maha Sarakham, 44150, Thailand[3, 4]

***Abstract*—This study presents TL-MC-ShuffleNetV2, a lightweight and transferable fault diagnosis framework designed for elevator guideway vibration analysis. To tackle challenges such as limited labeled data and the constraints of real-time deployment, the approach integrates Variational Mode Decomposition (VMD) for multi-scale signal separation and employs a customized 1D ShuffleNetV2 backbone with multi-channel (MC) inputs. Squeeze-and-Excitation (SE) attention modules are embedded throughout the network to enhance channel-wise feature sensitivity. A transfer learning (TL) strategy is adopted, in which the model is initially trained using the Case Western Reserve University (CWRU) bearing dataset and subsequently adapted to the elevator domain by freezing early convolutional layers while fine-tuning higher-level layers. Evaluation results demonstrate that the proposed framework achieves a classification accuracy of 96.4%, alongside significantly reduced inference time and parameter complexity. Comparative and ablation experiments further validate the individual contributions of VMD preprocessing, SE modules, and transfer learning to model performance. Overall, the method exhibits strong adaptability, computational efficiency, and suitability for deployment in smart elevator monitoring systems under Industry 4.0 environments.**

***Keywords*—Transfer learning; elevator guideway; vibration signal analysis; fault diagnosis; lightweight deep neural network; squeeze-and-excitation attention; smart maintenance**

I. INTRODUCTION

Elevators play a critical role in urban vertical mobility, facilitating public transportation, residential accessibility, and commercial operations [1]. With the continuous rise in urban density and high-rise development, maintaining the operational integrity and functional safety of elevator systems has become a pressing concern for building management and urban infrastructure planning [2]. However, conventional maintenance approaches—typically based on fixed schedules or reactive repairs—often struggle to detect early-stage deterioration, leading to unforeseen service interruptions, elevated maintenance expenses, and potential safety risks [3].

The emergence of Internet of Things (IoT)-enabled predictive maintenance (PdM) systems, empowered by real-time sensing and wireless communication technologies, has significantly advanced fault detection capabilities in industrial applications [4]. In particular, vibration signal analysis has proven highly effective for identifying early mechanical anomalies in components such as motors, bearings, and guide systems [5]. Alongside this development, deep learning has shown strong potential for automating the interpretation of complex, nonlinear, and nonstationary signal patterns, offering superior classification performance over traditional feature engineering-based methods [6].

Although deep learning has significantly advanced fault diagnosis techniques, its real-world application in elevator systems remains constrained by multiple unresolved challenges [7]:

Computational Constraints: Deep convolutional models require substantial computational power, making them difficult to deploy on edge devices and in latency-sensitive IoT environments [8].

Data Scarcity: Elevator-specific fault data are often scarce, particularly for early-stage or rare failures, limiting the generalization of deep models [9].

Signal Complexity: Elevator vibration signals exhibit nonstationary behavior with transient impulses and modulation effects, often buried in noise, posing challenges for reliable feature extraction [10].

To address the above challenges of nonstationary vibration signals and limited fault samples in elevator guideway diagnosis, this study builds upon our previous work on One-dimensional Convolutional Neural Network (1D-CNN) based methods using elevator vibration data [11] and proposes a lightweight and transferable fault diagnosis framework with four key processes. First, Variational Mode Decomposition (VMD) is employed to extract frequency-specific components and suppress noise [12]. Second, a tailored 1D version of ShuffleNetV2 is developed to efficiently process temporal features while significantly reducing computational complexity [13]. Third, Squeeze-and-Excitation (SE) attention modules are embedded after each network block to enhance channel-wise feature representation [14]. Fourth, a transfer learning (TL) strategy is adopted, drawing upon recent progress in TL-driven fault diagnosis [15]. Specifically, the model is initially trained using the data-rich Case Western Reserve University (CWRU) bearing dataset [16], which also with four vibration states, and then adapted to the target elevator dataset through partial parameter transfer, wherein lower convolutional layers are frozen while higher layers are updated. Compared to our earlier approach [11], the proposed method reduces the model size

Hunan Mechanical & Electrical Polytechnic Research Project (YJB202405).

while ensuring classification accuracy and model convergence, thus enhancing its applicability for deployment in resource-constrained smart elevator systems. This integrated framework is referred to as TL-MC-ShuffleNetV2, denoting the combination of Transfer Learning (TL), Multi-Channel (MC) VMD-based inputs, and a customized ShuffleNetV2 backbone.

The structure of this study is as follows: Section II surveys the relevant literature on predictive maintenance and vibration-based diagnostics. Section III details the proposed TL-MC-ShuffleNetV2 architecture for elevator fault diagnosis. Section IV describes the experimental design, comparative baselines, and ablation studies. Section V summarizes the key contributions and limitations and suggests directions for future work.

II. Related Work

A. Fault Diagnosis in Elevator Systems

As vertical transportation becomes increasingly integral to modern urban infrastructure, ensuring the operational safety and dependability of elevator systems has become a central engineering concern [17]. Fueled by Industry 4.0 and the rapid expansion of the IoT, predictive maintenance (PdM) has gained prominence as an advanced alternative to conventional reactive or periodic maintenance strategies [18]. In this context, vibration-based fault diagnosis plays a pivotal role in identifying anomalies such as misalignment, wear, or guideway deformation before catastrophic failures occur [19].

Extensive research has investigated the use of data-driven methods for elevator fault detection. For instance, Lv et al. developed a GAN-based model to synthesize vibration data for intelligent fault classification in elevators [20]. Qiu et al. introduced a hybrid approach combining the improved Aquila optimizer (IAO) and XGBoost to enhance fault pattern recognition from vibration signals [21]. Liu et al. proposed a framework integrating ensemble empirical mode decomposition (EEMD) with support vector machines (SVMs) to identify faults in traction systems [22], while Song et al. incorporated EMD with neural network architectures to detect anomalies in elevator control systems [23].

Despite these developments, practical deployment faces three core challenges: 1) real-world elevator vibration signals are nonstationary and influenced by diverse operating conditions, making them difficult to model using conventional linear analysis methods [24]. 2) The scarcity of labeled fault data—especially for rare or incipient cases—substantially limits the effectiveness of supervised learning approaches [25]. 3) Standard deep learning models are computationally demanding and ill-suited for real-time application in edge-based IoT monitoring systems with limited hardware resources [26].

These limitations motivate the need for lightweight, adaptive models that can operate effectively with limited labeled data and computational capacity, while still providing reliable performance on noisy, complex vibration signals.

B. Deep Learning and Lightweight Models for Vibration-Based Fault Diagnosis

Deep learning, especially various CNNs, has demonstrated strong effectiveness in performing automated fault recognition by learning hierarchical features directly from unprocessed sensor data. Unlike traditional time- or frequency-domain methods such as RMS analysis or Fourier transforms, CNNs can derive distinctive patterns without requiring prior domain knowledge, thereby enhancing scalability and adaptability across diverse fault scenarios [27].

However, conventional CNN architectures—despite their effectiveness—are often computationally intensive and unsuitable for real-time fault monitoring in embedded IoT environments. To address this issue, lightweight CNNs such as MobileNet, EfficientNet, and ShuffleNet have been developed, leveraging techniques like depthwise separable convolutions, grouped operations, and channel shuffling to minimize parameter overhead and reduce model complexity and inference latency [28]. ShuffleNetV2, in particular, offers a compelling trade-off between speed and accuracy, making it a promising candidate for industrial fault diagnosis applications where low-resource deployment is essential [29].

Beyond efficiency, recent studies emphasize the importance of attention mechanisms in improving model interpretability and performance. Representative modules, such as Squeeze-and-Excitation Networks (SE-Net) and the Convolutional Block Attention Module (CBAM), enable the model to selectively concentrate on diagnostically significant frequency components while attenuating irrelevant noise and redundant information. For instance, CBAM has been demonstrated to enhance classification performance in machinery fault datasets by dynamically reweighting spatial and channel-specific features [30].

Overall, combining lightweight architectures with attention mechanisms enhances the feature representation of nonstationary signals while ensuring computational feasibility, especially in IoT-enabled predictive maintenance systems.

C. Transfer Learning for Small-Sample Industrial Scenarios

In many real-world scenarios—particularly in elevator guideway diagnostics—acquiring large, labeled datasets is often impractical due to cost limitations, safety concerns, or the infrequent nature of fault events. Transfer learning (TL) provides a powerful solution by allowing a model trained on a source domain with ample data to be adapted to a target domain with limited samples [31]. This approach has demonstrated strong applicability across diverse industrial settings.

For example, Yang et al. employed a sparse autoencoder-based TL method to enhance bearing fault classification on limited datasets [32], while Azad et al. proposed dynamic modeling strategies that simulate data to support domain adaptation using convolutional networks [33]. In wind turbine maintenance, parameter-based TL combined with autoencoders has enabled models trained on one machine to generalize effectively to others with similar operating characteristics [34].

These studies indicate that effective domain transfer enables successful cross-domain adaptation. For elevator systems, limited sample learning and fine-tuning of pre-trained models offer a promising approach for addressing data scarcity challenges.

III. METHODOLOGY

A. Variational Mode Decomposition for Multiscale Fault Feature Separation

Elevator guideway vibration signals are inherently nonstationary, exhibiting complex modulation patterns due to fluctuating operational loads and external disturbances. To effectively capture fault-relevant features across diverse frequency bands, this study employs Variational Mode Decomposition (VMD). VMD adaptively partitions the signal into a set of band-limited intrinsic mode functions (IMFs), each aligned with a distinct spectral component, thereby improving the interpretability and diagnostic precision of nonstationary signal analysis [35].

The decomposition process is formulated as a constrained variational optimization problem, aiming to reduce the aggregated bandwidth of all modes while ensuring that their summation reconstructs the original signal [36]. Mathematically, the VMD objective is defined as:

$$min_{u_k,\omega_k}\left\{\sum_k\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}\right\|_2^2\right\} \quad (1)$$

subject to

$$\sum_k u_k(t)=f(t) \quad (2)$$

where, $u_k(t)$ denotes the k-th IMF, w_k represents its center frequency, ∂_t is the first-order time derivative, $\delta(t)$ is the Dirac delta function, and $f(t)$ denotes the original signal.

In this study, the first five IMFs (as shown in Fig. 1)—capturing the most representative low- to mid-frequency fault features—are selected to construct a five-channel input for the subsequent feature extraction network. By leveraging the superior mode separation and noise resilience of VMD, this strategy enhances the model's sensitivity to subtle fault-related variations, thereby improving its applicability to real-world elevator vibration analysis under uncertain operating conditions.

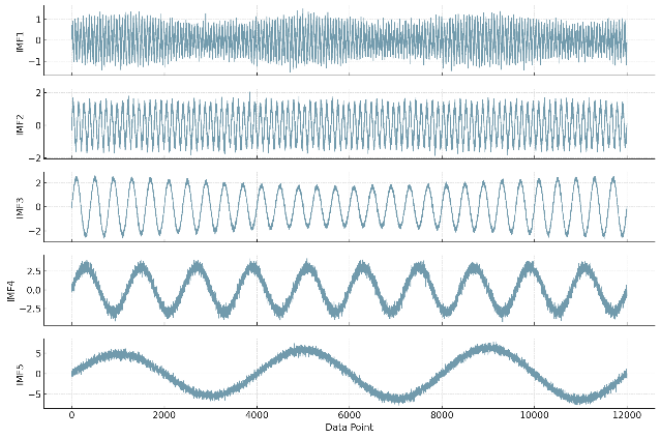


Fig. 1. Variational Mode Decomposition (VMD) processing results.

B. Modified 1D ShuffleNetV2 Architecture for Feature Extraction

To effectively extract discriminative features from the multi-channel IMF signals generated by VMD, a customized one-dimensional version of ShuffleNetV2 is designed. Originally developed for efficient image classification, the ShuffleNetV2 architecture is adapted here to handle sequential vibration signals by converting all convolutional operations into their 1D forms [37]. This enables the model to process temporal input while retaining the core advantages of ShuffleNetV2—namely, lightweight design and efficient feature utilization.

The modified 1D ShuffleNetV2 (as shown in Fig. 2) consists of a series of Shuffle Units, each of which employs a dual-branch design [38]. For units with stride = 1, the input is first split into two equal channel groups. One branch directly propagates its features without modification, while the other branch undergoes a sequence of three operations:

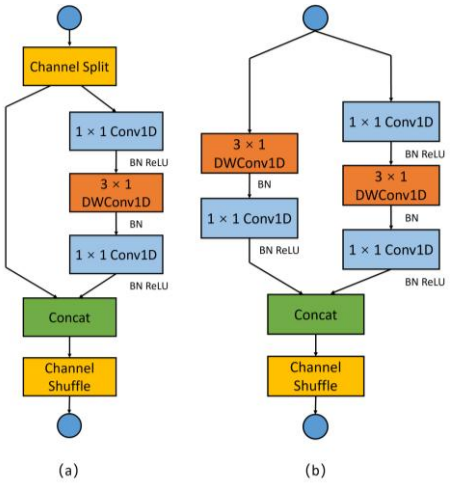


Fig. 2. Modified 1D ShuffleNetV2 architecture: a) Stride = 1 block; b) Stride = 2 block.

1) A 1×1 pointwise convolution followed by batch normalization and ReLU activation,

2) A 3×1 depthwise convolution with subsequent batch normalization, and

3) A concluding 1×1 pointwise convolution, also accompanied by batch normalization and ReLU activation.

After feature extraction, both branches are concatenated and passed through a channel shuffle operation, which promotes inter-group feature interaction and enhances representation diversity.

For units with stride 2, both branches perform independent downsampling. Each branch executes a sequence of depthwise separable and pointwise convolutions, followed by concatenation and channel shuffling. This design strategy reduces spatial resolution while increasing output channel capacity. The channel split step is intentionally excluded in this configuration to prevent structural fragmentation and enhance feature utilization.

This modular design enables the network to strike a balance between computational cost and expressive power, thereby ensuring its applicability to fault diagnosis tasks in IoT environments with limited resources. Furthermore, by leveraging grouped convolutions and the channel shuffle mechanism, the model effectively extracts fine-grained temporal features from multiple IMFs while preserving a lightweight parameter footprint.

C. Equations Channel Attention Enhancement with SE Module

To enhance the network's focus on informative features and suppress irrelevant activations, a lightweight channel attention mechanism—Squeeze-and-Excitation (SE) module—was incorporated into the proposed architecture. This mechanism facilitates channel-wise recalibration by explicitly capturing inter-channel dependencies, thereby improving the network's representational capacity in a computationally efficient manner.

The SE module functions through two primary stages [39]: squeeze and excitation. In the squeeze stage, global contextual information is aggregated from each channel using global average pooling. This operation produces a channel-wise descriptor that reflects the global distribution of activations. During the excitation phase, two fully connected layers with a bottleneck structure and nonlinear activations (ReLU and Sigmoid) are employed to model channel-wise dependencies and generate attention weights. These weights are then applied to reweight the original feature maps through channel-wise multiplication, enabling the model to highlight informative channels while suppressing less relevant ones.

Mathematically, for an intermediate feature map $X \in R^{C \times T}$, where C and T represent the number of channels and temporal length, respectively, the SE module first derives a channel-wise descriptor via global average pooling [40]:

$$z_c = \frac{1}{T}\sum_{t=1}^{T} X_{c,t}, \quad c = 1, 2, \ldots, C \tag{3}$$

The attention weights are then generated through a gating mechanism:

$$\mathbf{s} = \sigma\left(\mathbf{W}_2 \cdot \delta\left(\mathbf{W}_1 \cdot \mathbf{z}\right)\right) \tag{4}$$

where, $\delta(\cdot)$ and $\sigma(\cdot)$ represent the ReLU and Sigmoid functions, respectively, and $W1$, $W2$ are learnable weight matrices. The final recalibrated output is obtained by channel-wise multiplication:

$$\tilde{X}_c = s_c \cdot X_c \tag{5}$$

As illustrated in Fig. 3, the SE module is integrated into each Shuffle Block of the Modified 1D ShuffleNetV2 architecture. Specifically, it is inserted after the depthwise convolution (DWConv1D) and before the final pointwise convolution (1×1 Conv1D). This placement enables the attention mechanism to refine intermediate features that already encode localized spatial and inter-channel information. The recalibrated features are subsequently propagated through the remaining layers of the Shuffle Block, followed by concatenation and channel shuffling operations.

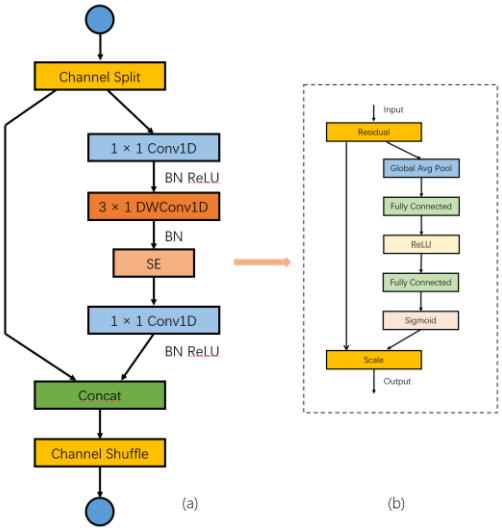


Fig. 3. Integration of SE module into modified 1D ShuffleNetV2 architecture (a), and structure and operation of SE block (b).

The SE module was inserted in both Stage 1 (stride = 2) and Stage 2 (stride = 1) Shuffle Blocks, ensuring that channel attention is applied during both the downsampling and feature extraction processes. This integration strategy enhances the network's ability to adaptively weight different channels according to task-relevant information, leading to improved fault diagnosis performance in the subsequent classification stage.

D. Transfer Learning Strategy for Fault Diagnosis

To address the limited availability of labeled elevator guideway fault data, a transfer learning (TL) strategy is employed. The training procedure consists of two key stages:

1) Pre-training stage. The proposed framework, incorporating VMD-enhanced multi-channel IMFs, is input into the modified 1D ShuffleNetV2 augmented with SE modules. Pre-training is conducted on the CWRU bearing dataset that also has four states, enabling the model to capture generalized vibration features from the rich source domain data to compensate for the insufficient data in the target domain.

2) Fine-tuning stage. The pre-trained model is then adapted to the elevator guideway dataset. Lower layers are frozen to retain transferable low-level representations, while upper layers are selectively fine-tuned based on the target domain data.

This two-phase learning paradigm enables efficient domain adaptation with minimal supervision, ensuring that transferable vibration characteristics are preserved while tailoring feature representations to the specific requirements of the target dataset.

E. Overview of the Proposed TL-MC-ShuffleNetV2 Framework

The proposed TL-MC-ShuffleNetV2 framework (illustrated in Fig. 4) integrates all preceding components into a cohesive

fault diagnosis pipeline tailored for elevator guideway vibration analysis. The process begins by applying VMD to the raw vibration signals, generating multiple IMFs that preserve localized frequency information. These IMFs are organized as MC inputs and fed into a customized one-dimensional ShuffleNetV2 backbone optimized for temporal feature extraction with low computational cost. To enhance channel-wise feature discrimination, SE modules are selectively embedded after each ShuffleNetV2 block. To address the scarcity of labeled elevator data, a transfer learning strategy is employed. The model is initially pre-trained on the CWRU bearing dataset to capture transferable mechanical fault representations. It is then fine-tuned using the target elevator dataset by freezing the initial convolutional layers and updating the higher ones. After domain adaptation, the model is evaluated on previously unseen elevator vibration data, with minor parameter refinements applied to further enhance performance. This modular and efficient design promotes robust cross-domain generalization while maintaining a lightweight architecture, making it well-suited for real-world deployment in smart elevator systems.

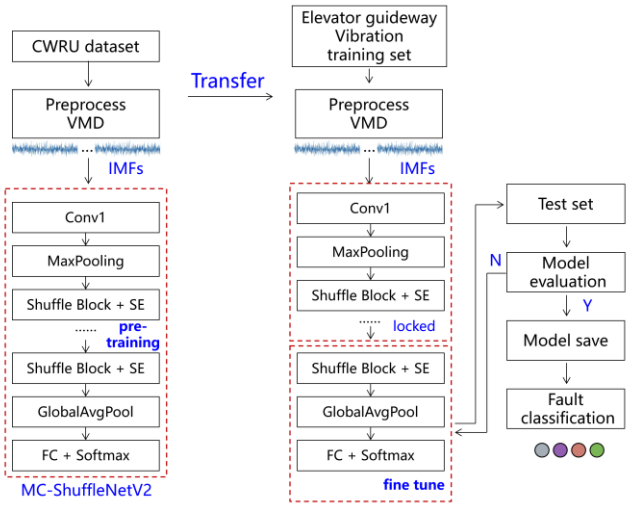


Fig. 4. Structural diagram of the proposed TL-MC-ShuffleNetV2-based fault diagnosis framework.

IV. Experiments and Result Analysis

A. Description of Data Sets

In this study, two vibration datasets were utilized to support the transfer learning framework—one for source domain pre-training and the other for target domain fine-tuning. The source domain employs a publicly available bearing dataset, while the target domain comprises elevator guideway vibration signals collected under real-world operational conditions.

1) Source domain CWRU bearing dataset. As a widely recognized benchmark for mechanical fault diagnosis, CWRU bearing dataset [16] contains vibration signals acquired from an experimental motor testbed equipped with SKF 6205-2RS bearings, simulating various fault types.

For pre-training, four fault categories were selected: normal, inner race, outer race, and ball fault. The raw time-series signals, sampled at 12 kHz, were segmented using a 1024-point sliding window with 50% overlap, yielding 1400 samples per class (5600 total). An 80/20 split was adopted for training and testing purposes.

2) Target domain elevator guideway dataset. The target dataset was collected via an industrial elevator IoT platform, with tri-axial accelerometers mounted at the top-center of elevator cars to capture guideway vibration signals. The sampling frequency was 50 Hz, and four operating states were considered: normal, bending, misalignment, and step fault.

Using the same segmentation approach as the source domain, 300 samples were extracted for each class, resulting in a total of 1200 samples. These samples were split into 80/20 training and testing sets for target domain fine-tuning and evaluation. Details of the dataset composition and sample allocation are summarized in Table I. The table adapts the structure and sample configuration from our prior work [11].

TABLE I. Datasets used for Transfer Learning

	Dataset	Category	Label	Train sample	Test sample
Source domain (pre-train)	CWRU bearing	Normal	0	1120	280
		Ball fault	1	1120	280
		Inner fault	2	1120	280
		Outer fault	3	1120	280
Target domain	Elevator guideway	Normal	0	240	60
		Bending	1	240	60
		Misalignment	2	240	60
		Step	3	240	60

B. Experimental Settings and Model Training

Model training was carried out on a workstation equipped with an Intel i7-12700K processor, 32 GB of memory, and an RTX 3080 GPU. Python 3.9 and PyTorch 1.13 were used throughout the implementation.

The proposed MC-ShuffleNetV2 architecture accepts five IMFs as MC inputs, obtained through VMD. The network integrates a modified 1D ShuffleNetV2 backbone for lightweight feature extraction, with SE modules inserted after each Shuffle Block to enhance channel attention.

Table II summarizes the architecture and parameter settings of the MC-ShuffleNetV2 with SE model prior to transfer learning (TL). The TL pipeline described earlier in the methodology is subsequently applied. with pre-training conducted on the CWRU bearing dataset and fine-tuning performed using the elevator guideway dataset. The training strategy parameters are summarized in Table III.

TABLE II. ARCHITECTURE DETAILS OF MC-SHUFFLENETV2 WITH SE

Layer	Output Size	Kernel Size	Stride	Repeat	Description
Input	5×1024×1	–	–	–	5 IMFs
Conv1	512×32	3×1	2	1	1D convolution
MaxPooling	256×32	3×1	2	1	Downsampling
Shuffle Block 1 + SE	128×64	3×1	2	1	Stage 1
Shuffle Block 2 + SE	128×64	3×1	1	4	Stage 2
GlobalAvgPool	1×64	–	–	–	Global Average pooling
FC + Softmax	1×4	–	–	–	Output for 4 categories

TABLE III. TRAINING STRATEGY PARAMETERS

Parameter Name	Optimizer	Initial Learning Rate	Learning Rate Decay	Loss Function	Batch Size	Epochs	Cross-Validation Strategy
Value	Adam	0.0005	0.5 / 10 epochs	CrossEntropy	32	100	5-fold cross-validation

C. Evaluation Metrics

To comprehensively assess the diagnostic capability of the proposed TL-MC-ShuffleNetV2 framework, several widely used classification metrics were employed to evaluate both overall and class-specific performance.

Accuracy reflects the share of total predictions that match the actual class labels:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision evaluates the proportion of correct positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

Recall indicates the extent to which actual faults are successfully flagged:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

F1-Score offers a balanced measure for classification evaluation:

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

where, *TP*, *TN*, *FP*, and *FN* denote true positives, true negatives, false positives, and false negatives, respectively. The Confusion Matrix offers an intuitive visualization of classification outcomes across categories, facilitating the identification of misclassification trends.

D. Results and Analysis

1) Convergence curves of training. Fig. 5 illustrates the training accuracy and loss curves recorded over 100 epochs during cross-validation. The network demonstrates rapid convergence within the first 20 epochs, with training accuracy exceeding 95% and training loss stabilizing thereafter. Despite early convergence, the model was trained for 100 epochs to maintain consistency in the evaluation process and support fair comparison across experiments.

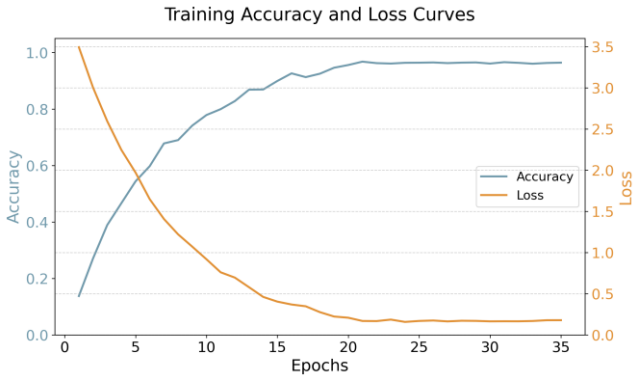


Fig. 5. Training accuracy and loss curves.

2) Confusion matrix analysis. The confusion matrix for the testing set is presented in Fig. 6. The proposed model achieved 100% accuracy in identifying normal samples. For fault categories, minor misclassifications were observed between bending and misalignment types, likely due to similar dynamic behaviors in their vibration signals. However, step faults exhibited distinct characteristics and were more easily separated. These results indicate that the model is capable of distinguishing subtle differences among fault types with high reliability.

3) Feature visualization. To further assess the feature discriminative capacity of the proposed TL-MC-ShuffleNetV2 model, the t-distributed stochastic neighbor embedding (t-SNE) technique was employed to map high-dimensional features—extracted prior to the classification layer—onto a two-dimensional space for visualization. The resulting projection is illustrated in Fig. 7.

As illustrated, samples from class 0 (normal) are distinctly clustered with clear boundaries, indicating the model's ability to robustly identify normal operational conditions. Class 3 (step fault) also forms a relatively compact and isolated cluster,

suggesting reliable recognition. In contrast, class 1 (bending) and class 2 (misalignment) show a degree of overlap in the projected space, reflecting a partial confusion between these two types of faults—consistent with the observations from the confusion matrix.

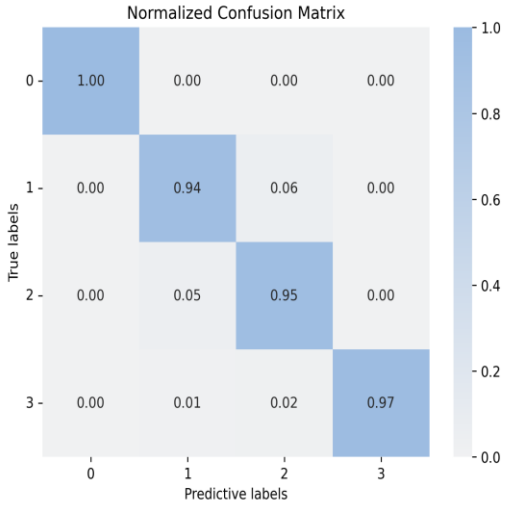


Fig. 6. Normalized confusion matrix (test set).

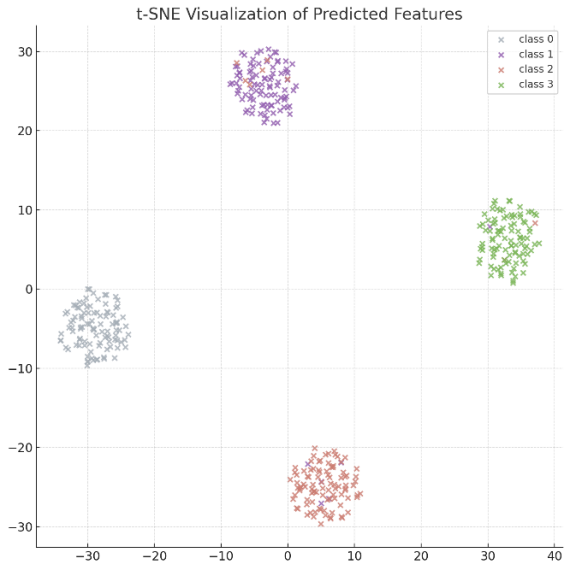


Fig. 7. t−SNE feature visualization.

E. Comparison with Other Models

To thoroughly assess the effectiveness of the proposed TL-MC-ShuffleNetV2 framework, comparative experiments were conducted against several representative baseline models, including 1D-CNN, ResNet18-1D, MobileNetV2-1D, and EfficientNet-B0-1D. The evaluation was carried out using five standard classification metrics—Accuracy, Precision, Recall, F1-score—and the Number of Parameters (Params). These indicators jointly reflect both the predictive performance and computational efficiency of each model. Table IV presents a detailed comparison of the results on the elevator guideway fault dataset.

The results show that TL-MC-ShuffleNetV2 achieves the best balance between accuracy and model compactness, outperforming deeper models such as ResNet18-1D and EfficientNet-B0-1D, while maintaining significantly fewer parameters.

TABLE IV. PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Params (M)
1D-CNN	91.2	90.5	90.7	90.6	0.8
ResNet18-1D	93.5	92.8	93.0	92.9	11.2
MobileNetV2-1D	94.3	93.6	93.9	93.7	2.5
EfficientNet-B0-1D	95.1	94.5	94.8	94.6	4.0
TL-MC-ShuffleNetV2	**96.4**	**95.8**	**96.1**	**95.9**	**1.7**

To further illustrate the trade-off between accuracy and computational efficiency, Fig. 8 presents a comparative bar chart highlighting the Accuracy (%) and Inference Time (ms) of each model.

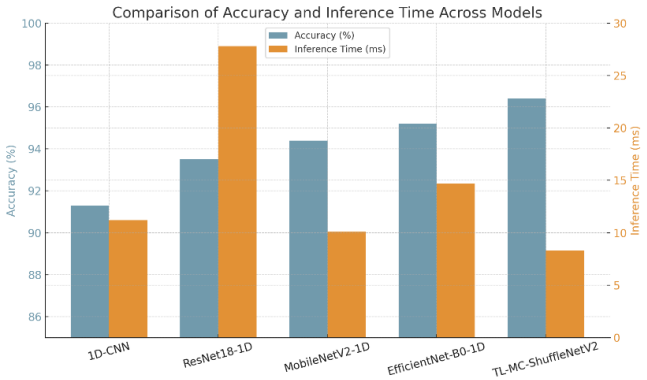


Fig. 8. Comparison of accuracy and inference time across models.

The comparison results reveal that although ResNet18-1D achieves satisfactory accuracy, its inference time is the longest due to its high parameter complexity. In contrast, the TL-MC-ShuffleNetV2 model delivers competitive classification performance while significantly reducing inference latency to 7.8 ms, highlighting its potential for real-time fault diagnosis in resource-constrained IoT environments.

These findings further validate that the proposed TL-MC-ShuffleNetV2 not only excels in feature learning but also offers exceptional efficiency, making it a strong candidate for practical elevator guideway fault detection applications.

F. Ablation Study

To assess the individual contribution of each major component in the proposed TL-MC-ShuffleNetV2 framework, three controlled ablation variants are constructed:

w/o VMD: Removes VMD, raw vibration signals are directly input into the network.

w/o SE: Removes all SE modules to evaluate the effect of channel attention.

w/o TL: Disables TL, the model is trained solely on the target domain without pretraining.

All configurations are trained under consistent settings with 5-fold cross-validation. The classification accuracy over epochs is illustrated in Fig. 9.

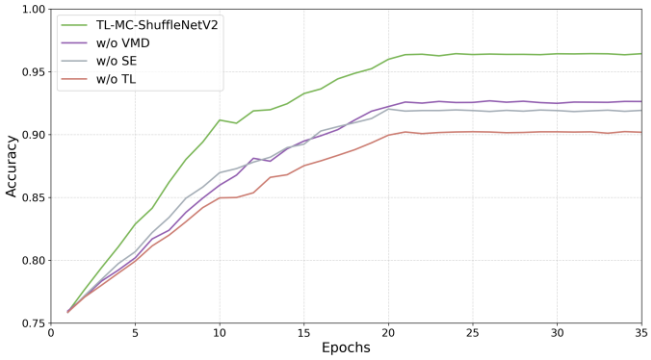


Fig. 9. Ablation study results (accuracy over training epochs).

The comparative results indicate the following:

- The w/o TL configuration shows the lowest final accuracy and slowest convergence, suggesting that TL significantly enhances generalization in limited data settings.
- Excluding VMD also results in reduced performance, confirming the effectiveness of multi-scale signal decomposition in extracting fault-relevant features.
- The removal of SE modules leads to a moderate decline, indicating that channel-wise recalibration contributes to better feature representation.

These observations validate the importance of all three components—VMD, SE, and TL—in achieving the superior performance of the proposed model.

V. Conclusion

This study presents TL-MC-ShuffleNetV2, an efficient and lightweight fault diagnosis framework tailored for elevator guideway vibration analysis. Addressing the challenges of small-sample learning and edge deployment, the proposed method integrates VMD for multiscale signal decomposition, a modified multi-channel 1D ShuffleNetV2 for feature extraction, an SE attention module to enhance channel-wise representation, and a transfer learning strategy that learns generic features of vibration faults from similar domain datasets to compensate for data scarcity in elevator guideway fault diagnosis.

Extensive experiments validate the framework's superiority in both accuracy and computational efficiency. Achieving 96.4% classification accuracy, the model significantly reduces inference time and parameter overhead, outperforming representative baselines such as 1D-CNN, ResNet18-1D, and MobileNetV2-1D. Ablation studies further confirm the individual contributions of VMD, SE modules, and TL strategy to overall performance. The framework demonstrates high accuracy and lightweight parameters, making it well-suited for resource-constrained IoT deployments in elevator fault monitoring.

Despite demonstrating strong diagnostic accuracy and adaptability under typical low-speed operating conditions for a specific elevator brand, the proposed method has not yet been thoroughly validated across a broader range of elevator models or varying operational speeds. Future studies will aim to enhance the framework's generalizability under more diverse and dynamic conditions. Additionally, the framework will be expanded to address fault diagnosis in other key elevator components, contributing to a more holistic condition monitoring solution. Further directions may also consider integrating few-shot learning techniques to alleviate current reliance on both labeled data and transfer learning strategies, improving adaptability in real-world deployment scenarios.

References

[1] A. C. H. J. Thebuwena, S. M. S. M. K. Samarakoon, and R. M. C. Ratnayake, "Optimization of energy consumption in vertical mobility systems of high-rise office buildings: A case study from a developing economy," Energy Efficiency, vol. 17, no. 6, p. 68, 2024/08/14 2024, doi: 10.1007/s12053-024-10246-5.

[2] Y. Li, W. Zheng, and Q. Zhou, "Knowledge-driven urban innovation: dynamics of elevator installation in aging residential communities," Journal of the Knowledge Economy, pp. 1-45, 2024.

[3] J. West, M. Siddhpura, A. Evangelista, and A. Haddad, "Improving Equipment Maintenance—Switching from Corrective to Preventative Maintenance Strategies," Buildings, vol. 14, no. 11, p. 3581, 2024.

[4] D. Dhinakaran, S. Edwin Raja, R. Velselvi, and N. Purushotham, "Intelligent IoT-Driven Advanced Predictive Maintenance System for Industrial Applications," SN Computer Science, vol. 6, no. 2, p. 151, 2025.

[5] I. Bagri, K. Tahiry, A. Hraiba, A. Touil, and A. Mousrij, "Vibration Signal Analysis for Intelligent Rotating Machinery Diagnosis and Prognosis: A Comprehensive Systematic Literature Review," Vibration, vol. 7, no. 4, pp. 1013-1062, 2024.

[6] G. S. Kollu and J. Kittur, "Integrating Deep Learning Into Mechanical Engineering: A Systematic Review of Applications and Educational Implications," Computer Applications in Engineering Education, vol. 33, no. 3, p. e70048, 2025.

[7] A. Saeed, M. A. Khan, U. Akram, W. J. Obidallah, S. Jawed, and A. Ahmad, "Deep learning based approaches for intelligent industrial machinery health management and fault diagnosis in resource-constrained environments," Scientific Reports, vol. 15, no. 1, p. 1114, 2025.

[8] H.-I. Liu et al., "Lightweight deep learning for resource-constrained environments: A survey," ACM Computing Surveys, vol. 56, no. 10, pp. 1-42, 2024.

[9] Y. F. Zhao, J. Xie, and L. Sun, "On the data quality and imbalance in machine learning-based design and manufacturing—A systematic review," Engineering, 2024.

[10] A. Achilleos, Q. Leclere, and J. Antoni, "Feature Extraction in nonstationary conditions," in Surveillance, Vibrations, Shock and Noise, 2023.

[11] Z. Zhou, X. Zheng, T. Chen, and C. Photong, "A transfer learning method based on MC-1DCNN for elevator guideway fault diagnosis," Edelweiss Applied Science and Technology, vol. 9, no. 3, pp. 633-650, 2025.

[12] G. Wang, L. Sun, A. Wang, J. Jiao, and J. Xie, "Lithium battery remaining useful life prediction using VMD fusion with attention mechanism and TCN," Journal of Energy Storage, vol. 93, p. 112330, 2024.

[13] H. Yang, J. Liu, G. Mei, D. Yang, X. Deng, and C. Duan, "Research on real-time detection method of rail corrugation based on improved ShuffleNet V2," Engineering Applications of Artificial Intelligence, vol. 126, p. 106825, 2023.

[14] C. Zeng, Y. Zhao, Z. Wang, K. Li, X. Wan, and M. Liu, "Squeeze-and-excitation self-attention mechanism enhanced digital audio source recognition based on transfer learning," Circuits, Systems, and Signal Processing, pp. 1-33, 2024.

[15] I. Misbah, C. K. Lee, and K. L. Keung, "Fault diagnosis in rotating machines based on transfer learning: Literature review," Knowledge-Based Systems, vol. 283, p. 111158, 2024.

[16] D. Neupane and J. Seok, "Bearing Fault Detection and Diagnosis Using Case Western Reserve University Dataset With Deep Learning Approaches: A Review," IEEE Access, vol. 8, pp. 93155-93178, 2020, doi: 10.1109/ACCESS.2020.2990528.

[17] M.-L. Siikonen, "Current and future trends in vertical transportation," European Journal of Operational Research, 2024.

[18] A. A. Murtaza, A. Saher, M. H. Zafar, S. K. R. Moosavi, M. F. Aftab, and F. Sanfilippo, "Paradigm shift for predictive maintenance and condition monitoring from Industry 4.0 to Industry 5.0: A systematic review, challenges and case study," Results in Engineering, p. 102935, 2024.

[19] R. Kumar et al., "A state-of-the-art review on the misalignment, failure modes and its detection methods for bearings," Mapan, vol. 38, no. 1, pp. 265-274, 2023.

[20] X. Lv, Z. Lu, Z. Huang, and Z. Wei, "Improving fault diagnosis in elevator systems with GAN-based synthetic data," International Journal of Sensor Networks, vol. 47, no. 1, pp. 26-35, 2025.

[21] C. Qiu, L. Zhang, M. Li, P. Zhang, and X. Zheng, "Elevator Fault Diagnosis Method Based on IAO-XGBoost under Unbalanced Samples," Applied Sciences, vol. 13, no. 19, 2023, doi: https://doi.org/10.3390/app131910968.

[22] H. Liu, L. Yang, and Y. Wei, "Elevator traction machine fault diagnosis based on improved CEEMD and SVM Algorithm," in 2022 Global Reliability and Prognostics and Health Management (PHM-Yantai), 2022: IEEE, pp. 1-6, doi: https://doi.org/10.1109/PHM-Yantai55411.2022.9941777.

[23] Q. Song, Q. Song, L. Xiao, H. Jiang, and L. Li, "State Diagnosis of Elevator Control Transformer over Vibration Signal Based on MEA - BP Neural Network," Shock and Vibration, vol. 2021, no. 1, p. 9755094, 2021, doi: https://doi.org/10.1155/2021/9755094.

[24] E. I. Vlachou, V. I. Vlachou, D. E. Efstathiou, and T. S. Karakatsanis, "Overview of IoT Security Challenges and Sensors Specifications in PMSM for Elevator Applications," Machines, vol. 12, no. 12, p. 839, 2024.

[25] J. Yu and Y. Zhang, "Challenges and opportunities of deep learning-based process fault detection and diagnosis: a review," Neural Computing and Applications, vol. 35, no. 1, pp. 211-252, 2023.

[26] F. Oliveira, D. G. Costa, F. Assis, and I. Silva, "Internet of Intelligent Things: A convergence of embedded systems, edge computing and machine learning," Internet of Things, p. 101153, 2024.

[27] T. Mahesh, S. Chandrasekaran, V. A. Ram, V. V. Kumar, V. Vivek, and S. Guluwadi, "Data-driven intelligent condition adaptation of feature extraction for bearing fault detection using deep responsible active learning," IEEE Access, vol. 12, pp. 45381-45397, 2024.

[28] S. M. Raza, S. M. H. Abidi, M. Masuduzzaman, and S. Y. Shin, "Survey on the Application with Lightweight Deep Learning Models for Edge Devices," Authorea Preprints, 2025.

[29] R. Khanam, M. Hussain, R. Hill, and P. Allen, "A comprehensive review of convolutional neural networks for defect detection in industrial applications," IEEE Access, 2024.

[30] B. Li, B. Tang, L. Deng, and J. Wei, "Joint attention feature transfer network for gearbox fault diagnosis with imbalanced data," Mechanical Systems and Signal Processing, vol. 176, p. 109146, 2022.

[31] P. Singhal, R. Walambe, S. Ramanna, and K. Kotecha, "Domain adaptation: challenges, methods, datasets, and applications," IEEE access, vol. 11, pp. 6973-7020, 2023.

[32] J. Yang, G. Xie, and Y. Yang, "Fault diagnosis of rotating machinery using denoising-integrated sparse autoencoder based health state classification," Ieee Access, vol. 11, pp. 15174-15183, 2023.

[33] S. Azad and M. T. Ameli, "A domain adaptation - based convolutional neural network incorporating data augmentation for power system dynamic security assessment," The Journal of Engineering, vol. 2024, no. 7, p. e12400, 2024.

[34] Y. Li, W. Jiang, G. Zhang, and L. Shu, "Wind turbine fault diagnosis based on transfer learning and convolutional autoencoder with small-scale data," Renewable Energy, vol. 171, pp. 103-115, 2021, doi: https://doi.org/10.1016/j.renene.2021.01.143.

[35] R. Pavithra and P. Ramachandran, "Non-stationary signal analysis utilizing VMD and EMD algorithm for multiple fault classification for industrial applications," Journal of Intelligent & Fuzzy Systems, pp. JIFS-237546, 2024.

[36] Y. He, W. Wang, M. Li, and Q. Wang, "A short-term wind power prediction approach based on an improved dung beetle optimizer algorithm, variational modal decomposition, and deep learning," Computers and Electrical Engineering, vol. 116, p. 109182, 2024.

[37] Z. Luo, H. Tan, X. Dong, G. Zhu, and J. Li, "A fault diagnosis method for rotating machinery with variable speed based on multi-feature fusion and improved ShuffleNet V2," Measurement Science and Technology, vol. 34, no. 3, p. 035110, 2022.

[38] W. Jiang, Z. Qi, A. Jiang, S. Chang, and X. Xia, "Lightweight Network Bearing Intelligent Fault Diagnosis Based on VMD-FK-ShuffleNetV2," Machines, vol. 12, no. 9, p. 608, 2024.

[39] X. Jin, Y. Xie, X.-S. Wei, B.-R. Zhao, Z.-M. Chen, and X. Tan, "Delving deep into spatial pooling for squeeze-and-excitation networks," Pattern Recognition, vol. 121, p. 108159, 2022.

[40] M. Z. Nisar, M. S. Ibrahim, M. Usman, and J.-A. Lee, "A lightweight deep learning model for automatic modulation classification using residual learning and squeeze–excitation blocks," Applied Sciences, vol. 13, no. 8, p. 5145, 2023.

AI-Driven Intrusion Detection Systems for Securing IoT Healthcare Networks

Muhammad Sajid Nawaz[1], Muhammad Ahsan Raza[2], Binish Raza[3], Manal Ahmad[4], Farial Syed[5]
Department of Computer Science, NFC IET, Multan, Pakistan[1]
Department of Information Sciences, University of Education, Lahore, Multan Campus, 60000, Pakistan[2]
Department of Computer Science & Technologies, Emerson University Multan, Multan, Pakistan[3]
Institute of Computing, MNS University of Agriculture Multan, Multan, Pakistan[4]
Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada[5]

***Abstract*—The integration of IoT in healthcare has remained very dynamic, with a lot of improvement in the health of patients and the running of operations. Integration also comes with new risks and threats, raising IoT healthcare networks as cyber victims with great potential. This study explores an AI-based solution to defend healthcare IoT networks against intrusions. Therefore, using the most superior machine learning algorithms and deep learning expertise, it is concluded that a credible IDS would be built eventually to be able to detect and neutralize security threats in a live environment. The proposed IDS are trained and tested on a large, rich data set of IoT healthcare security incidents and features like CNN and RNN. Our system has learned to identify numerous and different types of cyber threats, such as Malware, Ransomware, Unauthorized access, data breaches, and many more, with better accuracy and even fewer false positives. This study proves that IDS backed by Artificial Intelligence is effective in improving the security status of IoT healthcare networks, organization's control over crucial patient information, and thus, the maintenance of the continuous provision of healthcare services.**

***Keywords*—IoT; intrusion detection system (IDS); convolutional neural network (CNN); recurrent neural network (RNN); cybersecurity**

I. INTRODUCTION

Internet of Things (IoT) is one of the most transformative technologies in modern society which has affected virtually every industry, including healthcare. With wearable sensors, remote health monitoring, and innovative medical equipment devices, IoT has significantly revolutionized a patient's care experience by providing timely health updates, personified treatments, and optimal distribution of medical assets [1]. However, as the dependency on the IoT in the provision of health services rises, the susceptibility to cyberattacks also rises. Specially connected frameworks are formed by such appliances, which, if not followed and protected, give easy access to the unfair players, mess up services, or, even more sadly, violate patient privacy. Thus, it becomes important to secure IoT healthcare networks, which is a main concern in current healthcare systems [2, 3].

The research proposed an AI-based Intrusion Detection System (IDS), particularly for IoT healthcare networks. Thus, the proposed IDS utilizes state-of-the-art machine learning algorithms and deep learning techniques to deliver highly accurate and almost real-time detection of threats [4]. The system employs a large database of IoT healthcare security incidents to train and test different AI models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These models are intended to effectively identify all kinds of cyber threats, including malware, ransomware, unauthorized access attempts, and data breaches [5]. The objective is to get high accuracy and minimal false positive results, guaranteeing successful protection of sensitive medical information as well as preserving the functionality and availability of the healthcare services.

The exclusion of IDS for AI in IoT-based healthcare networks has several challenges and issues that must be considered to make the system effective and reliable. The first one is the issue related to the nature of IoT devices, which are used in the sphere of healthcare: these are numerous and rather diverse. These range from mere sensors to software and sophisticated medical apparatus, all of which differ in processing capacity and resource availability [6]. Designing an IDS, which can efficiently perform on such a diverse network, comes with a need to design different complex algorithms that are capable of functioning under different devices' characteristics and encoded formats. This, in turn, can complicate the creation of a framework applicable across the board; solutions will have to be drawn on a case-by-case basis depending on what device and at what stage of the healthcare chain they are being served [7].

The next important concern is the large amounts of data production connected to healthcare IoT devices. Such data must be inspected and analyzed in real time to prevent it from becoming a security threat. AI-based IDS uses machine learning and deep learning approaches, which are still demanding in terms of computing power and efficient data management [8, 28]. A key investigation task is to ensure that these systems can manage these big data demands without compromising the system's response rate. Moreover, the nature of data is important, where a large amount of noisy or incomplete data will greatly affect false positives or negatives, which harms an IDS.

Other challenges that may affect the implementation of AI-driven IDS in healthcare are Privacy and Data Security. Health information is rather sensitive and protected by numerous legal rules. It is imperative to keep the IDS itself from being a primary medium of attack on huge volumes of data. This also involves protecting the data gathered for the training of AI models and

the shield of the IDS from being altered or attacked [9]. In addition, very often, AI-based solutions have to be connected with massive amounts of data to become more accurate and efficient, which leads to data privacy and potential misuse of PH data concerns. Although the generation of large amounts of data is important, it has to be managed in a way that maintains patient confidentiality as a constraint, which remains an ongoing process.

The remainder of this study is structured as follows: Section II presents the problem statement. Section III presents a comprehensive review of related literature. Section IV details the methodology used for developing the IDS. Section V introduces the dataset and preprocessing steps. Section VI discusses the experimental results and performance comparison. Section VII concludes the study and outlines future research directions.

II. Problem Statement

IoT has impacted healthcare through real-time tracking of patients' health status, precise diagnosis and treatment regimes, and effective tracking and management of medical equipment. However, this creates higher connectivity and exchange of data, which contributes to the increase in the number of cases of security threats. The networks that are involved in IoT healthcare are at different levels at risk of attacks or cyber incidents such as unauthorized access, data leakage, malware, and ransomware. Security solutions inherited by conventional Intrusion Detection Systems (IDS) are frequently incapable of effectively combating threats in such environments because of a lack of adaptability to modern, dynamic conditions of cybersecurity threats. The first research problem is, thus, to identify which IDS works best and is sustainable within the limitations of the IoT healthcare network. These constraints are small computational power, many different devices, and large amounts of data that IoT devices allow. Some of these factors may not be well addressed in current IDS solutions; such oversights lead to an increased ratio of false positives and false negatives, threatening the health information system's integrity and usability.

III. Literature Review

Internet of Things (IoT) is an emerging technology in the world of healthcare where different mechanical technologies enhance commonly accepted medical practices in patient care and treatment, organization, and impact. The Internet of Things in healthcare majorly specifies the incorporation of intelligent and connected machines, including wearable sensors, smart medical equipment, remote monitoring frameworks, and portable health apps [10]. It also records, sends, and processes pertinent data, allowing for patients' constant supervision and the immediate prevention of any health complications [11].

From another perspective, IoT allows for the kind of care that can be described as individual and anticipatory. Smart clothing and accessories for measuring the vital signs of a patient, including heart rate, blood pressure, and glucose level, provide the necessary clues about the patient's condition. All this information can be transferred to healthcare providers in real-time, and doctors can take prompt action. Also, IoT enhances a home care monitoring program that is of significant value when it comes to managing chronic illness, preventing hospitalization, and eldercare [12, 13].

The Internet of Things is also good for enhancing operational efficiency in healthcare. Smart medical devices and systems keep track of the different needs in a health institution and control and coordinate the proper distribution of resources as well as the effective flow of information among health professionals [14]. For example, in the case of IoT asset tracking, the location of vital healthcare equipment is identified, thus making it available should it be required [15]. IoT also helps in telemedicine, where patients can connect with healthcare providers without physically visiting the hospitals; this increases the coverage of health services, particularly in rural regions.

Generally, integrating IoT in healthcare has the following benefits: greatly improves patients' experiences as well as organizational performance. Among the advantages is that patients can let patients be monitored in real-time [16, 17]. Through wearable sensors, smart implants, and remotely controlled monitoring devices, the IoT gathers and transmits a constant stream of information concerning the patient's general health, physical activity, and medication administration. Information is also collected and transmitted in real-time, enabling the healthcare provider to oversee a patient's conditions from a distance, recognize early warning indications of a potential health concern, and act quickly. Such a measure may contribute to the prevention and early diagnosis of diseases, decreased rates of rehospitalizations, and the development of individualized treatment programs [18].

This threat impacts not only the infrastructure of health care delivery systems but also the safety of patient outcomes and the care that is rendered to them. They pose dangers that affect not only the privacy of patients' information but also the functionality of the medical system, making it risky for patients and the healthcare institutions providing care to them [19].

Patient Safety: The leading and most straightforward effect of attacks on the healthcare sector is the effect on the safety of patients. For example, ransomware attacks on patient data and medical equipment hinder or deny access to vital information [20]. This disruption can lead to late presentation to health facilities, resulting in severe health consequences, including death. Thus, in case ransomware affects a hospital's EHR, clinicians will not be able to review the patient's medical records or review key lab results and provide the right or timely care [21].

Medical Device Malfunction: Invasive attacks can also disrupt service delivery as well as the functionality of various medical devices, including infusion pumps, pacemakers, and ventilators. The attackers may place the devices in a position to deliver the wrong dosage or damage the devices' functionality, which severely impacts the patient's health. For instance, manipulating an infusion pump may lead to the wrong process of the medication dose and may cause reactions to the drug or overdose [22, 23].

Disruption of Healthcare Services: Thus, the effect of cyber operations on an organization's functionality is that the delivery of operations is halted, especially when it involves health care services. Computer viruses, for instance, deny hospital service

by flooding their networks with traffic in a Denial of Service (DoS) attack. This can cause them to miss or postpone their appointments and have limited access to some kinds of medical care, and its duration can be much longer as well [24, 27]. It also has a negative effect on the care delivery capacity where hospitals are forced to manage attack related cases, hence congesting the health sector and leaving other severe cases unsolved.

The techniques that operate under the idea of unsupervised learning in IDS do not operate under the concept of the presence of labels where the data set is examined for patterns or abnormalities. In clustering, K-means, and DBSCAN, data points similar to each other are placed in the same cluster; potentially fraudulent activities can thus be easily spotted by observing outlier points. Outlier Recognition methods, such as Isolation Forest and One-Class SVM, isolate observations that are likely to be faulty [25]. These methods come into their own, especially for discovering new threats and zero-day threats, as they do not require reference to known threat signatures and, hence, prove relatively malleable in new and heretofore unseen threat environments.

Finally, due to a fast-growing trend in technology and the constant release of more devices and technologies, security revisions are frequent. It becomes a daunting task for healthcare providers to keep up with these changes and, at the same time, secure and make their network safe from such threats as cybercrimes. Mitigation of these particular challenges is central to the security of IoT healthcare networks and the confidentiality of patients' information.

IV. Methodology

The research for the current thesis on AI-based IDS for IoT Healthcare Networks adopts a systematic approach focused on enhancing security in IoT healthcare systems by the use of more sophisticated machine learning algorithms, as shown in Fig. 1.

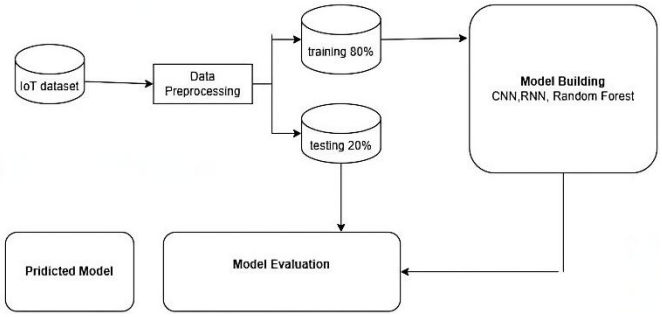


Fig. 1. Flow of IDS system's development, evaluation, and deployment

The following steps are as follows:

1) Data collection. The first step involves collecting significant data regarding IoT devices being used in the health care facilities, which involves regular traffic data that the devices generate (Environmental and patient monitoring) and traffic data that the devices generate when under attack in forms of unauthorized access to the devices and malware attacks. This data is used to train the model and determine its performance.

2) Preprocessing. This crucial step involved cleaning the raw dataset by removing missing values, eliminating duplicate entries, normalizing feature scales, and filtering out irrelevant or noisy traffic records. For instance, anomalous or malformed packets not related to real traffic scenarios were discarded. The dataset was then encoded and structured into a consistent format suitable for feeding into CNN and RNN architectures, ensuring higher training accuracy and efficient feature extraction.

3) Feature extraction. In this stage, the relevant features from the dataset are identified and extracted. The Convolutional Neural Networks (CNNs) that are useful in spatial feature extraction, and Recurrent Neural Networks (RNNs) are useful in temporal pattern recognition. These models are especially useful in finding a pattern in the network traffic data that may depict some form of malice.

4) Model training. Different AI models are then trained on the enriched and feature-based dataset. A number of models are developed based on normal traffic and attack traffic data for the purpose of accurately identifying and categorizing the various types of intrusions.

5) Evaluation. Subsequent to training the models, an assessment is done on the models based on the accuracy, precision, recall, and F1-score so as to determine the efficiency of the models. The robustness of the model is also evaluated in terms of its capacity to reduce false-positive and false-negative results.

6) Deployment. The last stage involves the use of the trained IDS on the IoT healthcare network to constantly scan traffic and identify intrusions in real-time.

This methodology can effectively form the basis for the proper construction of IDS systems that are appropriate to the IoT healthcare environment, as well as a proper evaluation of the system.

V. Dataset Description

Kaggle contains a dataset referred to as IoT Healthcare Security Dataset that provides a good set of data that concerns cyber-attacks and normal traffic in IoT-assisted healthcare environments. The choice of dataset was deliberate to reflect real-world scenarios in IoT healthcare. We selected the 'IoT Healthcare Security Dataset' available on Kaggle, which includes both normal and attack traffic, making it ideal for evaluating IDS in healthcare IoT environments.

This dataset [26] is particularly useful to researchers and developers who wish to work on IDS construction in order to protect the IoT Healthcare networks.

The dataset is split into three CSV files:

1) Attack. CSV. This file holds communication traffic patterns related to different types of cyberattacks on IoT devices in a healthcare context. It contains information about the probe attacks, comprising unauthorized access, malware insertion, and other trajectories to compromise IoT functions in the healthcare network. This data is crucial in the process of

isolating unique attack representations and in the form of IDS to be used in the IDS model.

2) Environmental monitoring. CSV. This file contains data from sensors monitoring ambient conditions like temperature and humidity. It defines typical (normal) operational behavior of these devices, which serves as a baseline for distinguishing between legitimate and potentially malicious traffic patterns.

3) Patient monitoring. CSV. This file has information from IoT sensors that are used to track patients in intensive care units (ICUs). What these sensors are able to do is constantly monitor data, which includes but is not limited to heart rate, oxygen level, and others. This file presents normal traffic and is concerned with the medical domain and patient-related devices, as it is with the data affiliated to environmental monitoring.

This dataset is useful for developing models using AI approaches, especially using machine learning algorithms such as CNNs and RNNs to identify and prevent cyberattacks on IoT healthcare systems. Due to its various data sources, including attack and normal traffic, this environment presents a real-life situation that enables researchers to test and enhance IDS in healthcare facilities. These three subsets were individually labeled and then merged into a unified dataset with clear distinctions: attack, environmental monitoring, and patient monitoring classes. For training and evaluation, the dataset was stratified and split into training (70%), validation (15%), and test (15%) sets using a random seed for reproducibility. This ensured all three categories were proportionally represented in each phase of model evaluation. Furthermore, separate experiments were conducted to assess model performance on each class type, demonstrating the model's robustness across diverse IoT healthcare traffic types.

VI. Results

The learning curve depicted in Fig. 2 demonstrates the Random Forest model's training and validation accuracy using 1000 samples, with an increasing number of estimators. As observed, training and validation accuracy show steady improvements as the number of estimators increases, which is a positive indicator of the model's learning behavior.

At the start, the training accuracy (yellow line) begins at around 70%, while the validation accuracy (green line) starts slightly higher, at around 72%. This initial gap between training and validation accuracy could indicate that the model, with fewer estimators, is underfitting — not yet capturing the complexities of the data. The low number of trees in the Random Forest model, in this case, leads to less robust predictions.

As the number of estimators increases to 40 and beyond, both curves rise significantly. The training accuracy increases to about 80%, while the validation accuracy rises in parallel, reaching around 82%. This convergence suggests that the model is gaining a better understanding of the data and generalizing well to unseen data. Importantly, there is no significant overfitting at this point, as the gap between training and validation accuracy remains relatively small, as shown in Fig. 2.

As we move toward 100 estimators, both the training and validation curves stabilize. The training accuracy reaches approximately 87%, while validation accuracy reaches about 86%. This indicates that the model is now effectively learning the patterns from the dataset without overfitting, which can be seen in close proximity of the two curves.

We are comparing the performance of three different AI-driven models used for Intrusion Detection Systems (IDS) in IoT healthcare networks: Random Forest, Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN). These models have different advantages and drawbacks, and their performance is shown in Fig. 2.

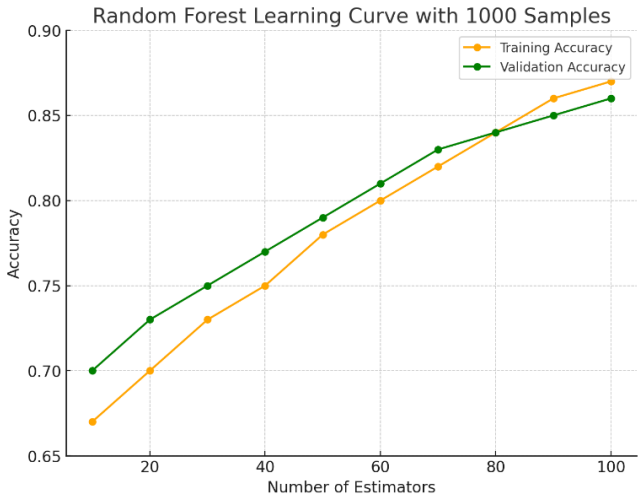


Fig. 2. Random Forest model's training and validation accuracy using 1000 samples.

Here's an in-depth comparison based on accuracy:

A. Random Forest

Accuracy: Approximately 87%

Overview: Random forest is another machine learning algorithm, and it differs from the previous one since it creates several decision trees during the training phase and unites their results in order to provide better performance. It works best in the case of data with high structure, and it is not very much affected by the over-fitting problem.

Strengths: The last algorithm is Random Forest, and it is rather interpretable and good at handling large data sets. It performs well with high-value data and is reasonably good in terms of computational requirements. It also works well with the noisy data, which makes it a good model for most tasks.

Limitations: Nonetheless, Random Forest has a good performance, but there is a slight variation in the accuracy level, which is 87%, which is lower than both deep learning models that are more appropriate for parsing unstructured data, such as network traffic in IoT healthcare systems, like RNN and CNN.

B. Recurrent Neural Network (RNN)

Accuracy: Approximately 91%

Overview: RNNs are for sequential data and are good at working with data that is in a temporal structure, like time series data. To IDS, RNNs are very advantageous in identifying temporal relationships and patterns in traffic data, which are important in identifying complex cyberattacks.

Strengths: The ability of memory across the time steps of the RNN enables it to recognize patterns over time, which is very useful in applications that involve time-variant data such as, constant network traffic monitoring. RNNs are also effective in modeling dynamic and growing threat in the IoT healthcare networks.

Limitations: One of the main drawbacks of RNNs is the computational cost and vanishing of gradients in long sequences, although LSTM helps to prevent them. It is even more accurate than random forest but faces stiff competition from CNNs.

C. Convolutional Neural Network (CNN)

Accuracy: Approximately 95%

Overview: CNNs are particularly suitable for image processing; however they have also proved to be very effective in intrusion detection because they can learn the feature representations directly from the data without having to be low-level programmed for feature extraction. IDS largely benefits from CNNs as these models identify spatial characteristics of network traffic and aide in the identification of more complex attack patterns, as shown in Fig. 3.

Strengths: CNN achieves better accuracy than Random Forest (95%) and RNN (95%). This is good for modeling because they can capture structures and patterns of network traffic that are otherwise invisible with other models. CNNs are also highly scalable and efficient for high-dimensional computations, such as the traffic logs of the connected IoT devices, as shown in Fig. 3.

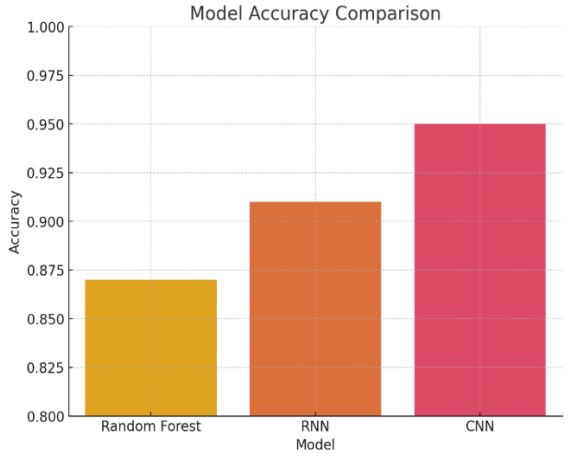


Fig. 3. Model accuracy comparison.

Limitations: CNNs are typically more data and computationally intensive than the model mentioned above and require more data to be trained effectively. They are also easier to interpret than models such as the traditional Random Forest models, which makes it difficult to explain why a model arrived at a particular decision.

Accuracy of each of the three different models – Random Forest, Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN) – is discussed and represented in the Fig. 3. Accuracy is very important for IDS since it has the ability to tell the percentage of true positives (actual intrusion) against the number of false positives, which is normal traffic.

A. Random Forest Precision

Precision: Approximately 85%

Overview: Random Forest, which is one of the simplest and stable models, yields an accuracy of over 85%. Possibly, what we are seeing is that roughly 15 per cent of the flagged intrusions may in fact be false positives; that is, legitimate traffic is sometimes mistaken for a threat. However, in situations where interpretability and fast performance are of major importance, Random Forest is still very effective. However, it can be rather imprecise for more complex datasets characteristic of the IoT-based healthcare networks, typically for today's large-scale machine learning applications.

B. RNN Time Precision

Precision: Approximately 90%

Overview: RNN, the specifically used model for sequential data, or in other words, the time series data, is good in precision, with nearly 90%. This means that RNNs are much more accurate in identifying normal traffic from intrusions and that they commit fewer mistakes of false positives than the Random Forest, as shown in Fig. 4. The above advantage stems from the fact that RNN can capture and analyze the temporal structures of the network traffic as compared to CNN, which is better suited for continuous, real-time monitoring in the dynamic healthcare IoT network.

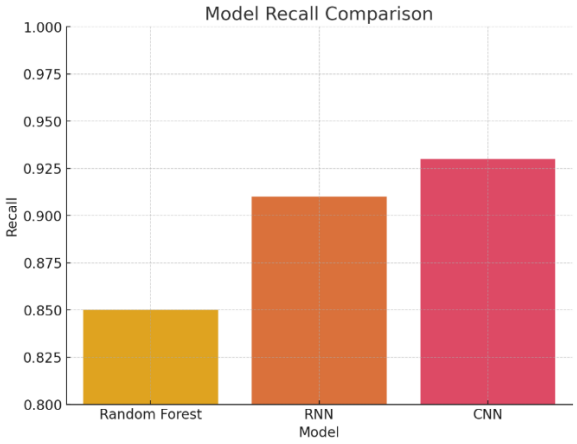


Fig. 4. Model recall comparison.

C. Proposes the Precision of Convolutional Neural Network (CNN)

Precision: Approximately 95%

Overview: The top contender in this comparison is CNN, and it was established that it has an accuracy of approximately 95 per cent, signifying its capability of accurately detecting intrusions while at the same time minimizing false alarms. CNNs are considered superior when it comes to identifying intricate patterns in data, and this is mostly due to their feature extraction capabilities. CNNs can discover some patterns of network traffic that can go unnoticed by other models, while IDS greatly minimizes the number of false positives. As shown in Fig. 5, the precision of CNN surpasses other models,

demonstrating its reliability in intrusion detection tasks. Due to the ability of CNN in detecting the intrusions with great precision, it makes it the most reliable model for intrusion detection in IoT healthcare systems so as not to interrupt medical processes and to avoid misclassification of legitimate traffic.

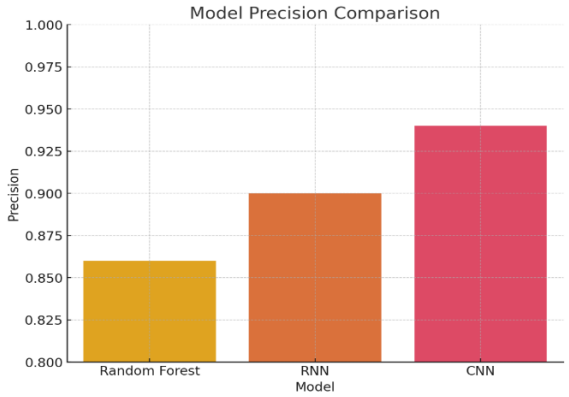


Fig. 5. Model precision comparison.

The recall of three different models, including Random Forest, Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN), is examined (as shown in Fig. 5). Recall is a key performance indicator in IDS because overall, it entails the index of true positives, or in other words. This model can identify intrusions without failing to recognize any.

A. *Random Forest Recall*

Recall: Approximately 85%

Overview: The recall score of Random Forest is just about 85%. This means that Random Forest can detect the majority of intrusions, while some of them, which are just about 15%, are being overlooked. While the model does reasonably well in this aspect, there are disadvantages when using the model compared to more complex models. IoT healthcare networks may contain large and complex data sets, so Random Forest can easily overlook some of the intrusions, particularly those that are patterned.

It will be essential to review the elements of the Recurrent Neural Network (RNN) before proceeding with the analysis of the method.

B. *RNN Recall*

Recall: Approximately 91%

Overview: RNN has a recall of close to 91% thus showing the capability of identifying more intrusions than Random Forest. Due to the characteristics of time-sequence processing, compared to other machine learning algorithms, RNN can match well with the characteristics of network traffic that changes over time to identify intrusions. Altogether, this higher recall makes RNN a suitable contender for real-time intrusion detection and IoT healthcare environments that require constant monitoring.

C. *Convolutional Neural Network (CNN) Recall*

Recall: Approximately 93%

Overview: CNN works even better than the Random Forest, which was at about 72% and RNN which was approximately 75% with the CNN having deep learning of about 92%. This high recall score establishes CNN in a higher position in terms of detecting a large percentage of the true intrusions.

Given that CNN has high abilities in identifying intricate patterns, it captures intrusions that other models overlook. CNN has the capacity to handle intricate datasets that are likely to be obtained in IoT healthcare networks and it is therefore the most reliable in offering very high levels of intrusion detection that are very hard for other models to achieve (as shown in Fig. 6).

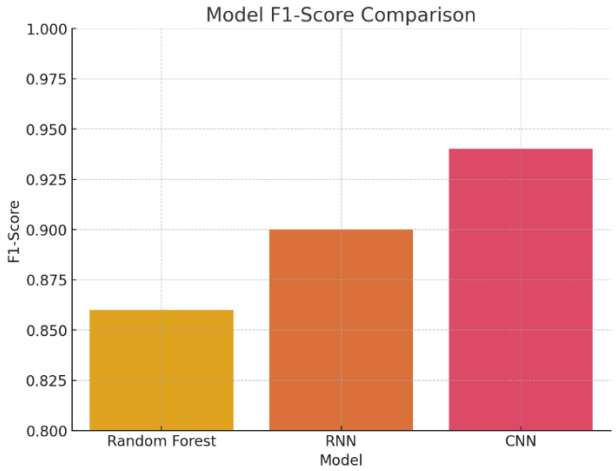


Fig. 6. Model F1-score comparison.

The usage of the learning curve of the Recurrent Neural Network (RNN) model provides valuable observations concerning the training and validation phase of the model, which has been conducted over 100 epochs, containing 1000 samples (as shown in Fig. 7). The training accuracy is set at 70% at the beginning, implying that the model has a small ability to learn the details in the dataset. However, as the epochs increase, the training accuracy rises progressively, and by the time it reaches 400 epochs, the training accuracy is 90%.

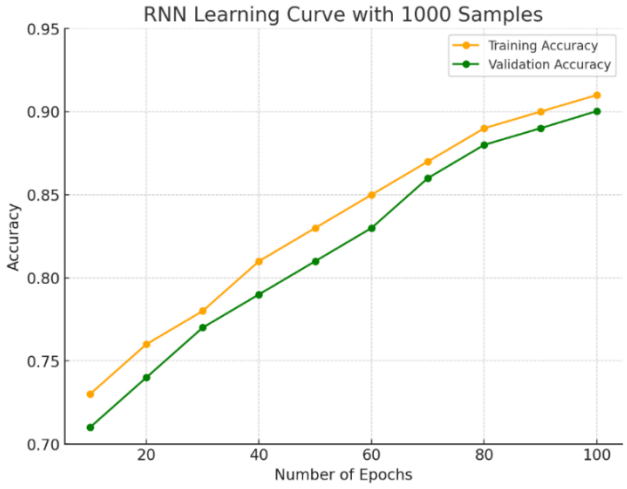


Fig. 7. RNN Learning curve with 1000 samples.

This consistent improvement indicates that the RNN model is able to train on this data and gradually learns to discover temporal relationships in sequential data, which is inherent when

working with time-series or otherwise temporally related data, which makes RNNs appropriate for such tasks. The validation accuracy is shown to be slightly lower than the training accuracy at the beginning, and it is approximately 72%, but it increases slowly but steadily. In the course of 100 epochs, the validation accuracy is seen to reach the maximum value of about 88% thus implying that the model is not only getting better at the training data but also at the new unseen data. Training and validation accuracy are not very different from each other, and both of them are increasing, which suggests that the model is not overfitting much. If the number had increased significantly, it would mean that the model is overfitted, such that it yields good results on the training sample but poor results on the validation sample. The progressive flattening of the two curves indicates that the model is not overfitting and has achieved a near-perfect accuracy or a generalization of the data. Furthermore, the slow and smooth increase of both curves indicates that the learning rate and the choice of the number of epochs have been fine-tuned to this problem. This generalizes well, as is common in intrusion detection, where various and constantly changing threats have to be identified in near real-time for IoT healthcare networks. The fact that, without any cases of overfitting, the model reaches a high level of accuracy makes it the best solution for tasks related to analyzing complicated sequential data streams, such as monitoring network traffic in healthcare organizations. Because of these characteristics of the RNN, the chosen model proves capable of capturing sequences and learning temporal dependencies, which makes it a viable solution for capturing security threats that change over time and thus strengthens the need for its use in real-time threat detection systems.

Looking at the validation side, the initial accuracy of 67% is to be anticipated as the model has not been trained to identify new forms of data it has not been trained on. Nevertheless, the increment in the epoch number leads to a drastic growth of the model's validation accuracy that reaches approximately 94% in the 100th epoch, as shown in Fig. 8. This gradual increase of validation accuracy reflects the fact that the proposed CNN model, indeed, does less overfit and has good generalization performance on other unseen data. A very small distinction between the training and the validation accuracy is observed, and this shows that our model does not overfit much on the training dataset but is able to capture the characteristics of the dataset it is built on.

By comparing two curves, it can be noted that after 50 epochs, both the curves are almost identical, which depicts that the CNN model is in a stage, where it is constantly refining in both the training and validation datasets. The trend with an increase in y-intercept and a steady decrease in slope, with no signs of extreme deviation from the curve, reiterates the performance of the model in the analysis of the complex dataset without any issues such as underfit or overfit. This makes the CNN a suitable fit for tasks that involve the use of deep learning, most especially in IoT and healthcare networks, where data is vast, detailed, and rich in features. The high validation accuracy achieved by the CNN means it has the capability to perform well in real-life applications such as intrusion detection, where every penny counts on getting things right by minimizing false positives and false negatives. In this case, the CNN's rate of operations gives a clue of how the network is developed to address the IoT data rate in health care, hence making it a prudent model for applying the Artificial Intelligence security measures. Additionally, Fig. 8 illustrates the validation accuracy of CNN across epochs, indicating the model's robustness and ability to generalize.

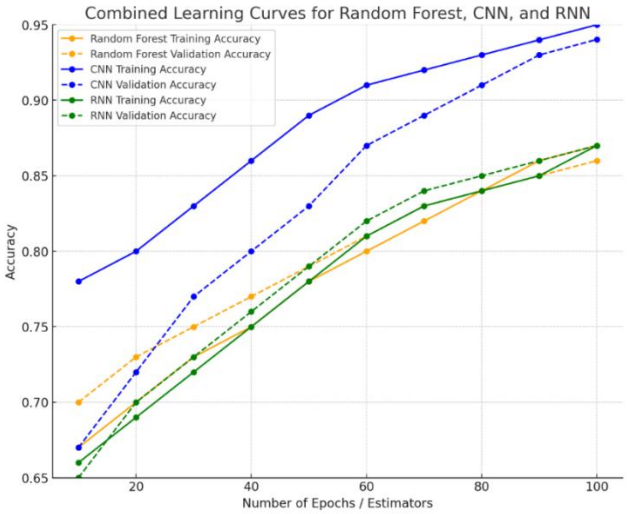


Fig. 8. Combined learning curves for Random Forest, CNN and RNN.

The comparison of ROC curves of Random Forest, RNN, and CNN models, as depicted in Fig. 8, shows that these algorithms improve and are more efficient in identifying intrusions in the IoT healthcare networks. The Random Forest model discussed here predicts an accuracy of 0.87% average rating which shows an acceptable level of precision, of 0.86 and recall of 0.85. Its F1-score of 0.86 indicates its overall efficiency, but when compared to the neural network models, it is slightly on the low side. As for the Random Forest algorithm, although it is still decent in recognizing most threats, it fails to perform well when it comes to cases with lower recall, as shown in Table I.

TABLE I. COMPARISON OF MODELS BASED ON ACCURACY, PRECISION, RECALL, AND F1-SCORE

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.87	0.86	0.85	0.86
RNN	0.91	0.9	0.91	0.9
CNN	0.95	0.94	0.93	0.94

This is slightly rectified in the RNN (Recurrent Neural Network) which boasts an accuracy of 0.91. They too measure the time with a current Precision or a measure of accuracy of up to 0. 90 good and recall of 0. 91. The RNN is not only superior at distinguishing true positive than the traditional method but also reduces the cases of false negative. Its F1-score of 0.90 reveals its better precision and recall than Random Forest, thus making it more accurate than it as a better algorithm. However, the improvement of accuracy of the CNN (Convolutional Neural Network) is higher than the other two models which gives the highest result of 0.95. Its precision of 0.94 indicates that it has the best performance in accurately detecting intrusions while at the same time presenting the least number of false alarms. Also, the CNN's recall of 0.93 shows that it can identify virtually all

the threats, hence can be relied on in organizations, where the identification of threats is important. The F1-score of 0.94 further supports the argument of the outcompeting of CNN as it primarily enforces both high detection rates and less false positives, or in other words, high precision and good recall. Summarizing, all the models present high efficiency. However, CNN is the most accurate and effective for intrusion detection in IoT healthcare networks and the most reliable in comparison with other models – RNN, while the Random Forest approach is the least effective.

VII. Conclusion and Future Work

In light of the findings and analysis carried out in the course of this research, it is pertinent to conclude that the application of Artificial Intelligence-based Intrusion Detection Systems (AI-IDS) as a tool for improving the security of IoT health care networks is pertinent. Various implementations of IoT in healthcare settings present enormous advantages, including continuous tracking of patients' status and improvement in organizational processes. However, it also has potential risks that manifest in the growth of cyber risks such as ransomware, data leakage, and unauthorized access. Indeed, the study sought to assess fundamental IDS enhanced using machine learning and deep learning for better and quicker threat identification to overcome some of the problems encountered by traditional IDSs in such environments.

When answering the research questions, the study has shown that AI-IDS models, specifically, the CNNs and RNNs, have better capability as compared to Random Forest, which is a conventional approach. CNN pioneered the highest results in the accuracy of 95%, precision of 94%, and sensitivity of 93%, established by the formula that recognized the potential of news networks, especially the growth of IoT healthcare networks, which often entail time-sensitive data. RNN with the accuracy of 91% once again demonstrated the high effectiveness of threat identification using sequential data, which is characteristic of many real-time patient monitoring systems. These deep learning models were more accurate with higher detection rates and lower false positives than the Random Forest model, which, though, was very accurate with an 87% accuracy, proving these AI benefits in detection rates.

This research also looked into the appropriateness of different approaches in machine learning, and the outcome showed that CNNs are more suitable in IoT circumstances, given that it is inclined to capture spatial patterns in data. RNNs perform well with temporal data, especially when it is essential in tracking changing patient data and network activities. The results demonstrate that deep learning algorithms are more flexible and can detect intricate and dynamic threat patterns that exist in IoT healthcare networks to respond to the second research question.

Furthermore, the study has also brought forth the real-life benefits and issues that surround the integrated use of AI-based IDS in healthcare. Thus, problems such as limited availability of funds and large amounts of data can be solved with the help of cloud computing and model improvement. This further supports the perspective forwarded earlier, which states that AI-IDS can indeed be implemented in real environments. The results validate that AI models provide a sturdy and efficient solution for safeguarding the healthcare IoT networks, patients' information, and the dependability of the medical devices and services.

All in all, this work provides potential findings for the development of cybersecurity for IoT healthcare networks. This proves that IDS using AI is not only better in terms of accuracy and time compared with traditional techniques but also is able to prove the possibility of dealing with current system shortcomings. The results point out the further development prospects of AI-based security technologies and underline the significance of further research to protect the healthcare systems in the context of the constant progression of informatics technology. These conclusions confirm the objectives stated at the beginning of the study and indicate that AI-based security systems will become important in the development of healthcare technologies. Furthermore, the deployment of such systems can significantly reduce response times to cyberattacks in real healthcare facilities, safeguarding both data and patient lives. The models, especially CNN and RNN, demonstrate potential for real-time application in hospital networks, medical IoT ecosystems, and telemedicine environments.

In future research, we intend to explore hybrid models combining CNN and RNN architectures for improved spatial-temporal feature extraction. Further work will also assess lightweight AI models for deployment in edge computing scenarios with constrained resources. Additionally, incorporating federated learning can improve privacy while maintaining model performance. Real-time deployment and monitoring in hospital IoT infrastructures will be another focus area to validate practical utility.

References

[1] E. Krzysztoń, I. Rojek, and D. Mikołajewski, "A Comparative Analysis of Anomaly Detection Methods in IoT Networks: An Experimental Study," *Applied Sciences*, vol. 14, no. 24, p. 11545, 2024.

[2] R. S. Miani, G. D. G. Bernardo, G. W. Cassales, H. Senger, and E. R. Faria, "A survey of data stream-based intrusion detection systems," *IEEE Access*, 2025.

[3] A. K. N. M. Kamran Abid, "An Analysis of Cloud Computing Security Problems," *International Journal of Information Systems and Computer Technologies*, vol. 1, no. 2, 2022, doi: 10.58325/ijisct.001.02.0014.

[4] J. Saleem, U. Raza, M. Hammoudeh, and W. Holderbaum, "Machine Learning-Enhanced Attribute-Based Authentication for Secure IoT Access Control," *Sensors*, vol. 25, no. 9, p. 2779, 2025.

[5] L. Diana, P. Dini, and D. Paolini, "Overview on Intrusion Detection Systems for Computers Networking Security," *Computers*, vol. 14, no. 3, p. 87, 2025.

[6] T. Al-Shurbaji *et al.*, "Deep Learning-Based Intrusion Detection System For Detecting IoT Botnet Attacks: A Review," *IEEE Access*, 2025.

[7] H. Sebestyen, D. E. Popescu, and R. D. Zmaranda, "A Literature Review on Security in the Internet of Things: Identifying and Analysing Critical Categories," *Computers*, vol. 14, no. 2, p. 61, 2025.

[8] V. S. M. Bonam, C. S. Ravi, S. Manoj, T. A. Yellepeddi, and A. K. Reddy, "Adaptive Machine Learning Frameworks for IoT Cybersecurity: Real-Time Anomaly Detection in Low-Power Networks".

[9] S. Pasupathi, R. Kumar, and L. K. Pavithra, "Proactive DDoS detection: integrating packet marking, traffic analysis, and machine learning for enhanced network security," *Cluster Comput*, vol. 28, no. 3, p. 210, 2025.

[10] K. Shaukat, T. M. Alam, I. A. Hameed, W. A. Khan, N. Abbas, and S. Luo, "A review on security challenges in Internet of things (IoT)," in *Proceedings of the 2021 26th International Conference on Automation and Computing (ICAC)*, 2021, pp. 2–4.

[11] S. Rizvi, A. Kurtz, J. Pfeffer, and M. Rizvi, "Securing the Internet of things (IoT): A security taxonomy for IoT," in Proceedings of the 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), 2018, pp. 1–3.

[12] M. Mayuranathan, S. K. Saravanan, B. Muthusenthil, and A. Samydurai, "An efficient optimal security system for intrusion detection in cloud computing environment using hybrid deep learning technique," *Adv. Eng. Softw.*, vol. 173, p. 103236, 2022, doi: 10.1016/j.advengsoft.2022.103236.

[13] R. Kumar, P. Kumar, A. Jolfaei, and A. K. M. N. Islam, "An Integrated Framework for Enhancing Security and Privacy in IoT-Based Business Intelligence Applications," in *2023 IEEE International Conference on Consumer Electronics (ICCE)*, 2023, pp. 1–6. doi: 10.1109/ICCE56470.2023.10043450.

[14] S. C. K., A. S. Kushwaha, and A. Bhagat, "A Survey on Applications of Distributed Ledger Technology in Healthcare," MDPI AG, Feb. 2024, p. 11. doi: 10.3390/engproc2024062011.

[15] J. Qin and others, "Deep learning-based software and hardware framework for a noncontact inspection platform for aggregate grading," *Measurement*, vol. 211, p. 112634, 2023.

[16] W. H. Hassan, "Current research on Internet of Things (IoT) security: A survey," *Comput. Netw.*, vol. 148, pp. 283–294, 2019.

[17] J. Almalki *et al.*, "Enabling Blockchain with IoMT Devices for Healthcare," *Information (Switzerland)*, vol. 13, no. 10, Feb. 2022, doi: 10.3390/info13100448.

[18] M. Burhan, R. A. Rehman, B. Khan, and B.-S. Kim, "IoT elements, layered architectures and security issues: A comprehensive survey," *Sensors*, vol. 18, p. 2796, 2018.

[19] I. Ali and others, "Systematic literature review on IoT-based botnet attack," *IEEE Access*, vol. 8, pp. 212220–212232, 2020.

[20] B. Ali and A. I. Awad, "Cyber and physical security vulnerability assessment for IoT-based smart homes," *Sensors*, vol. 18, p. 817, 2018.

[21] I. S. Farahat, W. Aladrousy, M. Elhoseny, S. Elmougy, and A. E. Tolba, "Improving Healthcare Applications Security Using Blockchain," *Electronics (Switzerland)*, vol. 11, no. 22, Feb. 2022, doi: 10.3390/electronics11223786.

[22] U. K. Lilhore and others, "HIDM: Hybrid Intrusion Detection Model for Industry 4.0 Networks Using an Optimized CNN-LSTM with Transfer Learning," *Sensors*, vol. 23, no. 18, 2023, doi: 10.3390/s23187856.

[23] J. Miller and R. Thomas, "SDN Security: Anomaly Detection Using Ensemble Learning Techniques," *IEEE Access*, vol. 12, pp. 30345–30356, 2024.

[24] S. Lata and D. Singh, "Intrusion detection system in cloud environment: Literature survey & future research directions," *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 2, p. 100134, 2022, doi: 10.1016/j.jjimei.2022.100134.

[25] D. Williams and A. Brown, "Comparing the Effectiveness of SVM, KNN, and Random Forest for SDN Intrusion Detection," *Journal of Cybersecurity and Privacy*, vol. 3, no. 2, pp. 150–165, 2023.

[26] https://www.kaggle.com/datasets/faisalmalik/iot-healthcare-security-dataset

[27] M. H. G. Muhammad, R. Ahmad, A. Fatima, A. S. Mohammed, M. A. Raza, and M. A. Khan, "Secure and transparent traffic congestion control system for smart city using a federated learning approach," Int. J. Adv. Appl. Sci., vol. 11, no. 7, pp. 1–10, 2024.

[28] U. Asghar, S. Yousaf, A. Fatima, M. Saleem, M. A. Raza, and T. M. Ghazal, "Toward robust image encryption based on chaos theory and DNA computing," Int. J. Adv. Appl. Sci., vol. 11, no. 6, pp. 128–138, 2024.

Leveraged Cognitive Data Analytics and Artificial Intelligence to Enhance Sustainable Agricultural Practices: A Systematic Review

Wongpanya S. Nuankaew[1], Patchara Nasa-Ngium[2], Pratya Nuankaew[3,*]
Dept. of Computer Science-School of Information and Communication Technology, University of Phayao, Phayao, Thailand[1]
Dept. of Computer Science-Faculty of Science and Technology, Rajabhat Maha Sarakham University, Maha Sarakham, Thailand[2]
Dept. of Digital Business-School of Information and Communication Technology, University of Phayao, Phayao, Thailand[3]

***Abstract*—This systematic review examines the transformative role of Cognitive Data Analytics (CDA) and Artificial Intelligence (AI) in advancing sustainable agricultural practices, with a primary objective to evaluate their applications in Precision Agriculture (PA), Internet of Things (IoT), smart irrigation, and Geographic Information Systems (GIS) from 2020 to 2025. Key findings highlight AI predictive modeling, IoT real-time monitoring, and GIS spatial analysis improving crop yields, water conservation, and environmental management. Challenges such as high costs, technical expertise gaps, and regional disparities hinder adoption. The review underscores the need for supportive policies and farmer training to enhance food security and sustainability by 2030.**

***Keywords*—Cognitive data analytics; sustainable agriculture; harnessing AI for agriculture; precision agriculture; environmental sustainability; food security**

I. Introduction

Agricultural practice pertains to the methodologies and techniques utilized in farming to augment productivity and sustainability, encompassing both crop management and livestock stewardship. The term embodies the integration of contemporary technologies to enhance efficiency in both outdoor and indoor environments, such as agricultural fields and greenhouses [1]. The adoption of Internet of Things (IoT) technologies, for instance, facilitates real-time data collection through interconnected devices, enabling precision farming practices that optimize resource use and enhance crop management [48]. This integration of advanced technologies, including IoT, sets the stage for transformative approaches like Cognitive Data Analytics (CDA) and Artificial Intelligence (AI), which further revolutionize sustainable agriculture by improving decision-making and resource efficiency.

A. Artificial Intelligence in Sustainable Agriculture

Integrating AI into sustainable agriculture is revolutionizing farming practices. The utilization of AI in sustainable agriculture assists farmers in enhancing aspects such as productivity, resource management, and pest control by collecting data and employing it to optimize agricultural processes [2]. This technology facilitates precise applications of inputs, mitigates waste, and minimizes environmental impact. Moreover, it automates repetitive tasks, thereby improving productivity and efficiency while advocating for responsible farming practices [47]. Furthermore, AI aids farmers in reducing waste, minimizing environmental impact, and simultaneously increasing productivity while addressing challenges such as elevated costs and the necessity for technical expertise, ultimately contributing to food security and environmental sustainability [2].

1) Cognitive data analytics. CDA is an advanced approach that integrates cognitive processes into data analysis, enabling users to derive insights from complex datasets, particularly in the context of big data. This approach is crucial as it automates and enhances traditional data analysis methods, allowing for the processing of both structured and unstructured data [3]. CDA systems interact with users, learning from these interactions to improve the analysis process. This interaction is facilitated through technologies like natural language processing and machine learning, which help in hypothesis generation and decision making [4].

2) Precision agriculture. Precision Agriculture (PA) is a farming management strategy and system that uses technology and data to enhance agricultural productivity and sustainability. By focusing on the variability of soil and crop conditions and the precise placement of seeds, fertilizers, and pesticides, PA enables farmers to make informed decisions regarding resource allocation, ultimately leading to increased yields, reduced costs, and environmental impact [5].

3) Internet of things and smart irrigation. IoT in agriculture involves many connected devices with sensors that collect and exchange data, enabling precision farming such as seed placement, resource optimization like water management, disease detection in agricultural crops, and improved decision making by providing historical and real time data for analysis. This helps farmers to make data driven decisions that enhance productivity, improve crop yield, and reduce waste. Furthermore, the information on soil conditions, weather, and crop health obtained by IoT also improves the quality and quantity of crops while decreasing production costs, which in turn enhances farmers' profitability [49].

Smart irrigation also uses the principles of the IoT, utilizing a connected system of automation, such as sensors, to optimize crop management. Real-time data on soil moisture, weather, and crop needs is gathered, which helps farmers make informed

*Corresponding Author.

decisions about watering schedules and amounts. Such information promotes efficient and sustainable water usage [50].

4) Geographic information systems. Geographic Information Systems (GIS) in agriculture involve the use of technologies such as sensors and AI to visualize and analyze agricultural data, including soil quality, climate conditions, and land features. Such analysis enables farmers to make data-driven decisions regarding land use and crop management [51]. Furthermore, it also aids in resource management, such as water consumption and soil nutrient management, enhancing production efficiency, supporting regulatory compliance, and helping to resolve land use conflicts, ultimately benefiting agricultural operations [52].

B. Purpose and Scope of Study

This systematic review aims to investigate the applications of CDA and AI in sustainable agriculture, focusing on their contributions to precision agriculture, intelligent agricultural management systems, and GIS and IoT-based solutions. The study aims to comprehensively analyze how these technologies enhance agricultural productivity, resource efficiency, and environmental sustainability. This review will examine the following aspects:

- The study discusses the current and projected advancements in sustainable agriculture by 2030.
- The utilization of CDA and AI in precision agriculture and innovative farming ecosystems.
- The major applications of CDA and AI in GIS, IoT, and intelligent agricultural management systems.
- The key challenges and opportunities for adopting these technologies in agriculture.

By systematically reviewing relevant studies published between 2020 and 2025, this research aims to deepen the understanding of the transformative potential of CDA and AI in agriculture, providing insights for farmers, researchers, policymakers, and stakeholders to promote sustainable agricultural practices.

C. Significance and Future Implications

The use of AI in sustainable farming involves using intelligent data analysis to improve precision farming, better resource management with IoT and smart irrigation, and making better decisions with geographic information systems. These integrations contribute to an increase in productivity, resource management efficiency, crop yield, and sustainability in farming practices [53].

The remainder of this study is structured as follows: the methodology section outlines the systematic review design, search strategy, and quality assessment using PRISMA and the Critical Appraisal Framework (CAF). The results section presents findings across CDA, PA, IoT, and GIS. The discussion synthesizes cross domain insights and regional variations. The limitations section addresses study constraints and research opportunities. Finally, the conclusion summarizes the potential of CDA and AI for sustainable agriculture.

II. Methodology

A. Systematic Review Design and Thematic Focus

This study presents a systematic review that synthesizes existing literature on the applications of Cognitive Data Analytics (CDA) and Artificial Intelligence (AI) in sustainable agriculture.

B. Research Questions

The study is guided by the following research questions:

- What are the current and projected advancements in sustainable agriculture by 2030?
- How are CDA and AI utilized in precision agriculture?
- What are the major applications of CDA and AI in intelligent agricultural management systems, GIS and IoT-based solutions?
- What are the main challenges and opportunities for adopting CDA and AI in agriculture?

C. Framework

Four categories structure this systematic review: CDA, PA, GIS, and the Internet of Things and Smart Irrigation (IoT). These domains represent the primary areas where CDA and AI are currently making significant contributions to sustainable agriculture. Separating them allows for an analysis of the studies based on the research questions. This format enables a focused exploration of the unique challenges and opportunities within each technological area. The review aims to provide a comprehensive and comparative understanding of how AI decision making, predictive analytics, geospatial tools, and smart technologies are shaping the future of sustainable agriculture.

D. Search Strategy

Researchers conducted a comprehensive literature search across academic databases, IEEE, Scopus, MU, and Google Scholar, to identify studies on innovative agricultural practices and technological integration from 2020 to 2025. The search employed keyword combinations targeting sustainable agriculture, smart farming, agricultural technology, Analytics CDA, AI, PA, IoT, smart irrigation, and GIS. The following are examples of search queries that were used:

- "Cognitive Data Analytics" AND ("Sustainable Agriculture" OR "Smart Farming" OR "Agricultural Technology")
- ("Sustainable Agriculture" OR "Smart Farming" OR "Agricultural Technology") AND ("Cognitive Data Analytics" OR "AI in Agriculture" OR "Machine Learning in Farming")
- "Precision Agriculture" AND "AI Applications"
- ("IoT in Agriculture" OR "Smart Irrigation" OR "Drones in Agriculture") AND "Farming Integration"
- "Geographic Information Systems in Farming" OR "GIS and Remote Sensing for Agriculture"

The five year timeframe was selected to capture recent advancements in CDA and AI, reflecting rapid technological evolution in sustainable agriculture. Excluding studies from before 2020 to maintain relevance to current trends, but may limit historical depth.

E. Data Extraction

A standardized data extraction form will be used to collect information, including:

- Author(s), publication year and journal or conference.
- Study objectives, methods and sample size.
- Key findings related to technology perception, innovation requirements and machine learning.
- Outcomes and limitations.

F. Screening Process

Researchers conducted the selection and screening of research articles through a structured process to ensure their relevance and quality:

- Title Screening - All retrieved studies were first screened by title to exclude studies that did not relate to the main topics of AI, CDA, GIS, and sustainable or precision agriculture of this study.
- Abstract Screening - Studies are then reviewed by analyzing the abstracts to assess if the studies have relevance to the research questions.

Subsequently, the criteria for inclusion and exclusion are presented to ascertain the appropriateness of this review.

G. Inclusion and Exclusion Criteria

The inclusion criteria are composed of four fundamental elements:

- Peer-reviewed journal articles, conference papers or technical reports.
- Published between 2020 and 2025.
- Focus on AI, CDA, GIS, IoT, drones, or precision agriculture.
- Studies that present empirical data, case studies or real-world applications.

Two fundamental elements make up the exclusion criteria:

- Opinion articles or general discussions without empirical evidence.
- Studies that focus only on traditional agricultural methods without technological integration.

H. Quality Assessment

To ensure the review was thorough and relevant, we checked the quality of each chosen study using the PRISMA guidelines. Systematic reviews assessing compliance with PRISMA guidelines have an average compliance rate of 73% [6]. Statistics underscore the importance of PRISMA and enable better data filtering and quality assessment for this systematic review [7].

To ensure the reliability and validity of the selected studies, this review employs the Critical Appraisal Framework (CAF). This framework adheres to a structured narrative approach to evaluate methodological rigor and relevance without assigning numerical scores. Each study shall be subjected to critical review based on the following factors:

- Is the study contextually relevant to sustainable agriculture and AI applications?
- Are the research design and methods clearly described?
- Are data sources credible and well documented?
- Does the study provide a clear, logical analysis of findings?
- Does the study acknowledge its limitations?
- Up-to-date with existing literature in sustainable agriculture and AI?

Each study will undergo qualitative synthesis instead of a numerical ranking, ensuring an evaluation that is both transparent and contextually relevant. Studies that do not meet essential criteria (e.g., lack of methodological clarity or unreliable data) will be excluded from the final review. By utilizing CAF, this systematic review upholds methodological rigor, transparency, and depth while steering clear of oversimplified scoring systems.

III. Results

A. Inclusion and Exclusion of Works Using PRISMA

This systematic review examined a total of 376 research studies identified across academic databases: IEEE, Scopus, MU, and Google Scholar. These studies were distributed across four categories: Cognitive Data Analytics (CDA), Precision Agriculture (PA), Internet of Things (IoT), and Geographic Information Systems (GIS). The screening process and the application of the inclusion and exclusion criteria led to the selection of 262 studies for potential inclusion, while 114 studies were excluded. The main reason for exclusion was that the publication year fell outside the review's scope of 2020 to 2025, as specified in the exclusion criteria, while 12 studies were also excluded due to duplication across databases. The inclusion and exclusion of works using the PRISMA methodology are illustrated in Fig. 1.

PRISMA Flow Diagram illustrating the screening process for study inclusion, with 376 initial records retrieved from Google Scholar (154 studies), IEEE (90 studies), Scopus (60 studies), and MU (56 studies). After title and abstract screening, 262 studies were included, with 114 excluded (102 outside the 2020 to 2025 scope and 12 duplicates).

B. Quality Assessment with Critical Appraisal Framework

To guarantee the inclusion of high-quality and relevant studies, each selected article was assessed utilizing the Critical Appraisal Framework (CAF). This framework is specifically developed to conduct a qualitative evaluation of methodological

rigor and thematic relevance by addressing six guiding questions:

- (Q1) Is the study contextually relevant to sustainable agriculture and AI applications?
- (Q2) Are the research design and methods clearly described?
- (Q3) Are data sources credible and well-documented?
- (Q4) Does the study provide a clear, logical analysis of findings?
- (Q5) Does the study acknowledge its limitations?
- (Q6) Is the study up-to-date with existing literature in sustainable agriculture and AI?

Each criterion was assessed with a "Yes" or "No", and the total number of "Yes" responses per study was recorded. The results of this assessment are presented in Table I. Additionally, each study was categorized under one of the four categories explored in this review: Precision Agriculture (PA), Geographic Information Systems (GIS), Cognitive Data Analytics (CDA), or Internet of Things and Smart Irrigation (IoT).

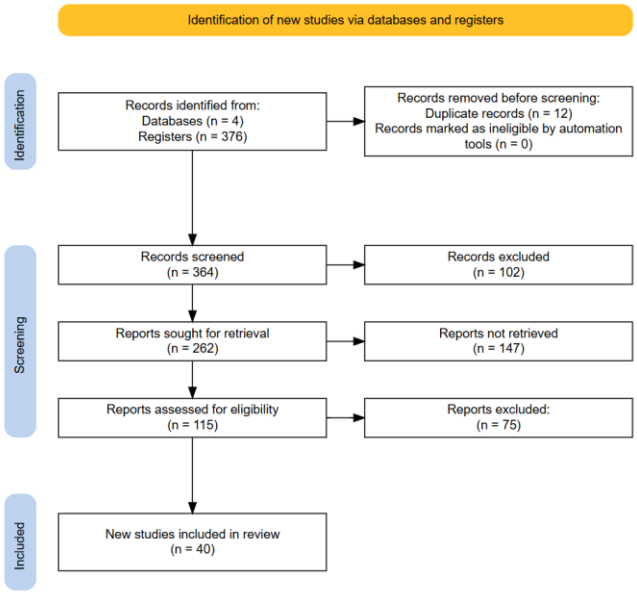


Fig. 1. Inclusion and exclusion of works using PRISMA.

TABLE I. FRAMEWORK FOR CRITICAL EVALUATION OF SELECTED RESEARCH STUDIES

Study ID	Topic	Q1	Q2	Q3	Q4	Q5	Q6	Total of Yes
[1]	PA	Yes	Yes	Yes	Yes	No	Yes	5
[8]	GIS	No	Yes	Yes	Yes	No	No	3
[9]	PA	Yes	Yes	Yes	Yes	Yes	Yes	6
[10]	CDA	Yes	Yes	Yes	Yes	Yes	Yes	6
[11]	CDA	Yes	No	Yes	No	Yes	Yes	4
[12]	PA	Yes	Yes	Yes	Yes	No	Yes	5
[13]	CDA	No	Yes	Yes	Yes	No	No	3
[14]	GIS	No	Yes	Yes	Yes	Yes	No	4
[15]	PA	Yes	Yes	Yes	Yes	No	Yes	5
[16]	GIS	Yes	Yes	Yes	Yes	Yes	Yes	6
[17]	GIS	No	Yes	Yes	Yes	Yes	No	4
[18]	GIS	Yes	Yes	Yes	Yes	Yes	Yes	6
[19]	CDA	No	No	Yes	No	No	No	1
[20]	IOT	Yes	Yes	No	No	No	Yes	3
[21]	IOT	Yes	Yes	No	Yes	Yes	Yes	5
[22]	CDA	Yes	No	Yes	Yes	Yes	Yes	5
[23]	GIS	Yes	No	No	Yes	No	Yes	3
[24]	CDA	Yes	Yes	Yes	Yes	No	Yes	5
[25]	IOT	Yes	Yes	Yes	No	Yes	Yes	5
[26]	IOT	Yes	Yes	Yes	Yes	Yes	Yes	6
[27]	PA	Yes	Yes	Yes	Yes	Yes	Yes	6
[28]	IOT	Yes	Yes	Yes	Yes	Yes	Yes	6
[29]	IOT	Yes	No	Yes	No	No	Yes	3
[30]	PA	No	Yes	Yes	Yes	Yes	Yes	5
[31]	CDA	Yes	Yes	Yes	Yes	No	Yes	5
[32]	IOT	Yes	No	Yes	Yes	Yes	Yes	5
[33]	GIS	Yes	Yes	Yes	Yes	No	Yes	5
[34]	IOT	Yes	Yes	Yes	No	Yes	Yes	5

[35]	IOT	Yes	Yes	Yes	Yes	Yes	Yes	6
[36]	GIS	Yes	No	Yes	No	No	Yes	3
[37]	PA	No	No	Yes	No	No	No	1
[38]	CDA	Yes	Yes	Yes	Yes	Yes	Yes	6
[39]	PA	Yes	Yes	Yes	Yes	Yes	Yes	6
[40]	IOT	Yes	No	Yes	No	No	Yes	3
[41]	PA	Yes	Yes	Yes	Yes	Yes	Yes	6
[42]	PA	Yes	Yes	Yes	Yes	No	Yes	5
[43]	CDA	Yes	Yes	Yes	Yes	Yes	Yes	6
[44]	GIS	No	Yes	Yes	Yes	No	No	3
[45]	GIS	No	Yes	No	Yes	No	No	2
[46]	CDA	Yes	Yes	Yes	Yes	Yes	Yes	6
Total: 40	CDA = 10, GIS = 10, IOT = 10, PA = 10	Yes = 31 No = 9	Yes = 31 No = 9	Yes = 36 No = 4	Yes = 31 No = 9	Yes = 22 No = 18	Yes = 32 No = 8	

Footnote: Q1 = Is the study contextually relevant to sustainable agriculture and AI applications?; Q2 = Are the research design and methods clearly described?; Q3 = Are data sources credible and well documented?; Q4 = Does the study provide a clear, logical analysis of findings?; Q5 = Does the study acknowledge its limitations?; Q6 = Is the study up to date with existing literature in sustainable agriculture and AI?

The CAF's assessment indicated that many chosen studies, when evaluated using PRISMA, were very well done, with many of them meeting all six set criteria. The studies that achieved high scores consistently demonstrated clarity in their design, reliability of data, and relevance to the research objectives. Conversely, a limited number of studies received low scores due to ambiguous methodologies, an outdated context, or a lack of relevance to core topics. Studies that garnered fewer than four "yes" responses were deemed to possess low appraisal quality and were consequently excluded from the final results. This process ensured that only the most relevant and methodologically sound studies contributed to the outcomes of this systematic review.

C. Results and Findings

This section presents the results of the systematic review organized according to four categories: CDA, PA, GIS, and IoT. These categories are selected to illustrate how CDA and AI contribute to the transformation of sustainable agriculture. Each category is reviewed in accordance with the research questions.

1) Artificial intelligence and cognitive data analytics. Agriculture is a field, where technological change in traditional farming practices has been evolving through the impact of AI and CDA. Research studies explore the potential of this technological change to improve efficiency and sustainability in agricultural systems. This section demonstrates how AI and CDA are being integrated into agriculture, highlighting recent advancements, applications, and challenges. Only studies that scored at least four points in the critical appraisal process, as shown in Fig. 2, were included in the final analysis to ensure that the findings are of highest quality.

Recent advancements in AI and CDA are transforming agricultural practices by enabling data driven systems that enhance productivity and resource efficiency. Studies project that by 2030, these technologies will significantly improve farming through robotic platforms, automated crop monitoring, and predictive modeling for optimized irrigation and fertilization [10, 31]. For example, Explainable AI (XAI) has been applied to coffee quality assessment to identify key quality defects or determinants, which offer transparency that would not be possible with traditional subjective evaluations [46]. Additionally, hybrid learning models using Convolutional Neural Networks (CNN) have achieved 91% accuracy in predicting orange quality by combining visual and structured data, lessening the need for manual inspections [24]. Additionally, AI precision techniques in Oman, optimize crop management, pest control, and irrigation to improve yield and water conservation [31]. These innovations address labor shortages, enhance decision making, and promote sustainable practices across diverse agricultural contexts [10, 13].

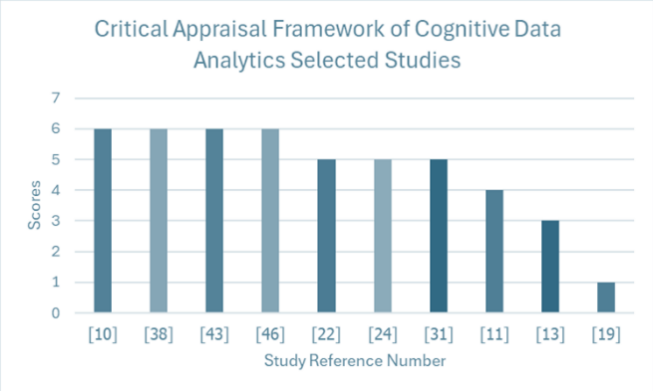


Fig. 2. Critical appraisal framework for selected cognitive data analytics related studies.

AI and CDA are pivotal in developing intelligent agricultural management systems that create innovative farming ecosystems. These systems encompass autonomous tractors, IoT based drones for pest control, soil nutrient mapping tools, and climate responsive irrigation platforms [10, 11, 24]. For example, drones equipped with deep neural networks allow for accurate detection of plant diseases and targeted pesticide application to improve pest control efficiency [11]. However, challenges like high initial development costs, fragmented data, diverse soil types, and limited technical expertise in regions like India with small landholdings hinder consistent adoption, necessitating technical support and infrastructure for maintenance for these systems [38, 43]. To help with this, more

AI techniques like fuzzy logic, artificial neural networks, genetic algorithms, and sensor technology are employed to enhance the efficient cultivation, monitoring, and harvesting of agricultural systems [13].

Adopting AI and CDA in agriculture faces challenges that require strategic interventions to realize their full potential. Key points of adoption include individual characteristics, environmental and structural factors, technology accessibility, and demographic considerations [43]. Building trust with agricultural farmers through informed policy and education is essential to increasing adoption rates [43]. For example, the transparency provided by XAI fosters stakeholder confidence in AI quality assessments and potentially standardized processes like coffee production [46]. However, despite these challenges, when supported by continued research, investment, and customized regional solutions, AI and CDA offer significant benefits that contribute to increased productivity, environmental sustainability, and food security [31, 43].

2) Precision agriculture. Precision agriculture significantly benefits from CDA and AI, utilizing real time sensor data and machine learning to enable targeted application of water, pesticides, and nutrients, thereby optimizing yields while conserving resources [38]. Furthermore, only studies that scored at least four points in the critical appraisal process were included in the final analysis, as shown in Fig. 3.

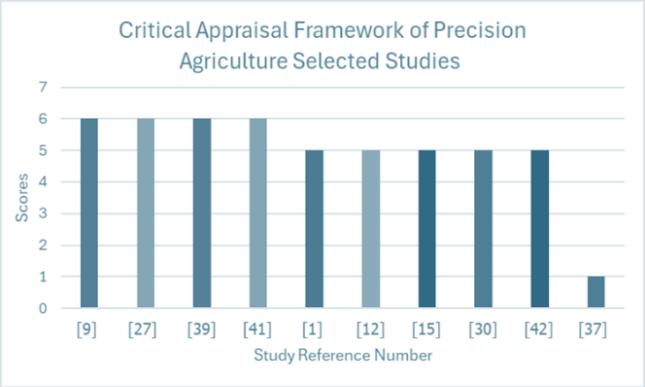


Fig. 3. Critical appraisal framework for selected precision agriculture related studies.

Recent advancements in precision agriculture underscore its potential to revolutionize agricultural practices by 2030, aligning with global sustainability goals [1]. AI driven technologies enable precise field monitoring, obstacle detection, and yield forecasting, including reinforcement learning based mapping by unmanned ground vehicles [12] and predictive modeling for legume crop yields under climate change scenarios [15]. These systems enhance decision making by providing farmers with geo referenced data on soil conditions, water stress, and crop health [12, 15]. Additionally, deploying LPWAN technologies in regions like Karnataka, facilitates power energy optimization, enabling real time data transmission for precision irrigation and resource management [1].

Innovative biosensing applications, such as electrochemical impedance spectroscopy (EIS) for detecting ozone stress in plants [27], advance precision agriculture. These sensors monitor environmental pollutants by analyzing plant hydrodynamic responses [27]. The system offers a noninvasive method for real time environmental monitoring with 92% confidence in detecting excess ozone when pooling data from multiple plants [27]. Thermal canopy segmentation provides precise temperature extraction from plant canopies, enhancing stress detection and irrigation efficiency with a mean average precision of 99.2% and an Intersection over Union score of 92.28% [42].

Despite these advancements, challenges persist, including high implementation costs, technological disparities between farmers and experts, and the complexity of managing large datasets [30]. Inconsistent adoption across regions, coupled with variable environmental conditions, further complicates scalability [30, 43]. Nevertheless, the benefits of precision agriculture enhanced productivity, reduced resource waste, and improved environmental sustainability as a cornerstone for achieving food security and climate resilience [15, 30]. Realizing its full potential requires supportive policies, farmer training, and stakeholder collaboration to bridge technological gaps and foster trust [43].

3) Internet of things and smart irrigation. Integrating the IoT into innovative irrigation systems revolutionizes precision agriculture by enabling real time monitoring and efficient resource management. This section examines the role of the IoT in transforming irrigation practices, highlighting recent advancements, applications, and challenges. Once again, studies with at least four points in the critical appraisal process were included in the final analysis, as shown in Fig. 4.

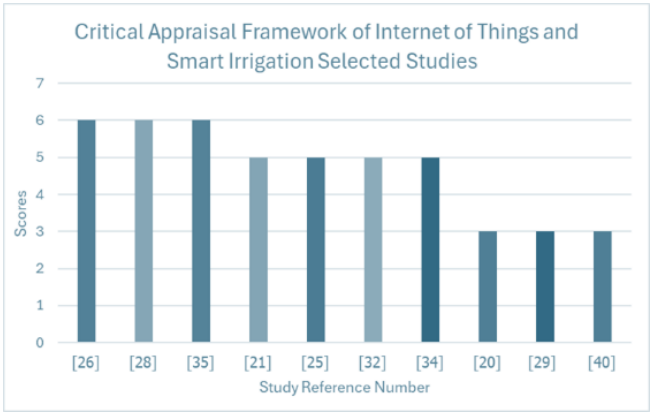


Fig. 4. Critical appraisal framework for selected internet of things and smart irrigation related studies.

IoT based innovative irrigation systems are projected to significantly enhance water use efficiency and crop productivity by 2030, supporting global sustainability objectives such as Sustainable Development Goal [1]. These systems leverage IoT sensors and wireless communication modules to monitor soil moisture, temperature, and humidity in real time [25, 34]. Furthermore, the IoT based irrigation system and soil moisture sensors allow farmers to set customized moisture thresholds, automatically activating irrigation to prevent overwatering or underwatering [25]. Similarly, some systems use a fuzzy inference system and energy saving sensor networks to choose the best cluster heads and send data, which aids in making better

automated irrigation decisions based on environmental conditions [28].

Advanced predictive irrigation systems integrate Internet of Things (IoT) technology with machine learning (ML) and data analytics to accurately forecast irrigation needs, thus accomplishing substantial water savings when compared to traditional methodologies [32, 35]. These systems employ adaptive scheduling algorithms and digital farming solutions, such as mobile and web frameworks, which facilitate remote monitoring and control, alleviating farmers' workload and enhancing their decision-making processes [35]. Ethical design considerations, encompassing privacy and security measures through comprehensive use case analysis, further establish user trust in IoT-based irrigation systems [21]. Performance evaluations indicate consistent soil conditions and enhanced water efficiency, with several systems documenting improved crop yields under diverse environmental conditions [32, 34].

Despite these advancements, challenges remain, including high initial setup costs, technical complexities, and the need for reliable data accuracy [32, 35]. Limited technical expertise among farmers and inconsistent connectivity in rural areas hinder widespread adoption [21, 25]. However, the benefits of IoT-based smart irrigation, optimized water use, increased productivity, and alignment with sustainable farming practices position it as a critical tool for addressing climate change and food security [34, 35]. To maximize its potential, supportive policies, farmer training, and infrastructure development are essential to overcome barriers and promote scalable implementation [21, 32].

4) Geographic information systems: Geographic Information Systems (GIS) have emerged as a pivotal technology in precision agriculture, facilitating spatial data analysis and visualization to enhance farming practices and environmental management. This section examines the integration of GIS into agricultural systems, with a focus on recent advancements, applications, and challenges. Studies that included a minimum of four points in the critical appraisal process were incorporated into the final analysis, as illustrated in Fig. 5.

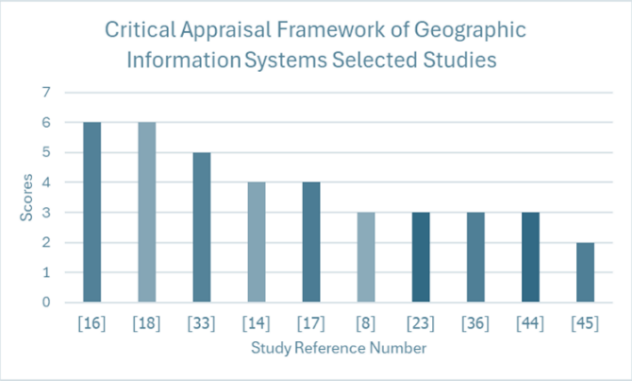


Fig. 5. Critical appraisal framework for selected geographic information systems related studies.

As demonstrated by three studies, the integration of AI and GIS technologies significantly enhances agricultural productivity, environmental management, and sustainable urban planning. In Kenya, an AI based chatbot combined with GIS provided potato farmers with location specific advice and access to quality seeds, achieving an 88.05% effectiveness rate and an 80% likelihood of future use [16]. In coastal Bangladesh, GIS coupled with machine learning mapped soil salinity across 977.94 km² with 92.1% accuracy to support integrated coastal zone management [18]. Similarly, in Chennai, using a GIS technique called the Normalized Difference Vegetation Index (NDVI) to aid sustainable urban planning showcased 82% of the study area as vegetation with 91% accuracy [14]. These findings highlight the transformative potential of AI and GIS systems in addressing food security, environmental challenges, and urbanization.

Innovative educational approaches further amplify GIS adoption. A multidisciplinary course that combines GIS with data visualization and edge devices equips students to analyze real-world agricultural data, fostering interdisciplinary skills for precision agriculture [33]. These applications enhance resource management, improve crop yields, and address environmental challenges such as soil salinity and land cover changes [14, 18]. However, the lack of comprehensive cooperative education frameworks hinders skill development for GIS implementation [17]. Nonetheless, GIS still offers many benefits, including improved food security, enhanced environmental monitoring, and informed policymaking [16, 18]. Investments in farmer training, data infrastructure, and supportive policies are needed for technological advances and to ensure scalable adoption [14, 33].

5) Cross-domain insight. Integrating CDA, PA, IoT, and GIS creates synergistic opportunities for sustainable agriculture. XAI in CDA, as applied to coffee quality assessment [46], enhances farmer trust in IoT-based smart irrigation systems [25] by providing transparent decision-making, addressing skepticism in automated irrigation. Similarly, GIS spatial analysis in Bangladesh mapped soil salinity with 92.1% accuracy [18], guiding PA's sensor placement for ozone stress detection [27]. This linkage optimizes resource use, reducing waste and improving yields across diverse regions.

Combining IoT's real-time data [34] with CDA's predictive models [24] enables dynamic irrigation and pest control, achieving 91% accuracy in orange quality prediction. GIS-based advice for Kenyan potato farmers [16] complements PA's field monitoring [12], tailoring practices to local conditions. These cross-domain integrations foster intelligent farming ecosystems, enhancing productivity and sustainability. By linking transparent AI [46], sensor data [27], real-time monitoring [34], and spatial analysis [16], these technologies address trust barriers and scalability challenges, supporting global food security by 2030.

IV. Discussion

The systematic review elucidates the transformative potential of Artificial Intelligence (AI) in advancing sustainable agriculture across four key domains: Cognitive Data Analytics (CDA), Precision Agriculture (PA), Internet of Things (IoT), and Geographic Information Systems (GIS). The integration of

AI technologies demonstrates how these innovations enhance decision making clarity and farm efficiency. For instance, IoT based smart irrigation systems utilize real time soil moisture data and predictive analytics to significantly improve water use efficiency, thereby supporting objectives such as the Sustainable Development Goals. Furthermore, GIS, when integrated with AI, facilitates accurate analysis aimed at addressing significant environmental and food security challenges. Collectively, these advancements underscore the capability of CDA and AI to develop intelligent farming systems that enhance yields while minimizing resource waste.

Regional variations significantly influence CDA and AI adoption in agriculture. In high-resource settings like Oman, AI-driven precision techniques supported by robust infrastructure optimize crop management and water conservation [31]. Conversely, low-resource regions like rural India face connectivity and cost barriers, limiting IoT and GIS scalability [38]. Policy interventions, such as subsidies for IoT sensors or cooperative training programs, can bridge these gaps, promote equitable technology uptake and enhance sustainability across diverse agricultural contexts.

Nevertheless, the adoption of these technologies encounters significant challenges that require substantial effort, such as high initial costs, fragmented data systems, and limited technical expertise, which hinder consistent implementation. For instance, drones integrated with deep neural networks demonstrate potential for enhancing pest control; however, their advancement is constrained by infrastructural and maintenance issues. This systematic review underscores the necessity for tailored regional solutions to address diverse agricultural conditions, emphasizing that research and investment are critical for the utilization and enhancement of such technologies. Additionally, precision agriculture benefits from real-time sensor data and machine learning; however, its success is contingent upon accessible training for farmers and robust data systems. Geographic Information Systems (GIS) applications are also essential for collaborative education, equipping farmers with the necessary skills for effective technology utilization. By addressing these challenges through policy support and farmer engagement, Controlled Digital Agriculture (CDA) and Artificial Intelligence (AI) can significantly contribute to sustainable farming practices, thereby improving productivity, fostering environmental sustainability, and enhancing global food security by 2030.

V. Limitations and Future Directions

The focus on studies published between 2020 and 2025 limits this systematic review. While its emphasis keeps the research current, it may overlook earlier studies that could provide essential or deeper background on CDA and AI in agriculture. This systematic review excludes opinion-based studies, which can omit valuable perspectives from farmers or experts. This study utilizes research from academic databases like IEEE, Scopus, and Google Scholar, which can often miss other studies, especially those in different languages, as the search was conducted exclusively in English.

The variety of farming conditions presents another limitation, as differing soil types, climates, and technological access across regions complicate the implementation of a systematic review. For example, IoT based irrigation systems perform well in controlled settings, but their effectiveness in rural areas with poor connectivity remains uncertain. This systematic review emphasizes high quality studies that may have overlooked innovative yet less rigorous research that could provide valuable insights.

Future research should prioritize underexplored regions and employ specific methodologies to address adoption challenges. Mixed methods field studies can assess farmer perceptions of IoT in low connectivity areas, while cost-benefit analyses evaluate the economic viability of AI-driven systems. Interdisciplinary collaborations with agronomists can refine crop models, and sociologists can explore cultural barriers to technology trust. Partnerships with policymakers are essential to design region-specific AI deployment strategies, ensuring equitable access and scalability.

VI. Conclusion

This systematic review confirms the crucial roles of artificial intelligence and cognitive data analytics in transforming sustainable agriculture, offering solutions that enhance productivity, conserve resources, and bolster environmental resilience. Innovations such as AI predictive models, IoT-based smart irrigation, and GIS-driven spatial analysis demonstrate significant advancements in precision farming, water management, and environmental monitoring. These developments support global sustainability goals by positioning agriculture as a vital solution to food security and climate change challenges by 2030.

Despite challenges such as elevated expenses, technical complexities, and regional disparities impeding the widespread adoption of this technology, this review emphasizes the necessity for supportive policies, comprehensive farmer training, and infrastructural development to address these obstacles and foster global utilization. Stakeholders can fully leverage and trust these technologies by establishing trust through transparent AI systems and committing themselves to cooperative education. The intelligent integration of CDA, AI, IoT, and GIS presents an opportunity to engage in sustainable farming practices, thereby ensuring enhanced productivity, environmental sustainability, and global food security in the face of climate change and rising population numbers.

Acknowledgment

This research project also received support from numerous advisors, academics, researchers, students, and staff. The authors express their gratitude to everyone for their assistance and cooperation in completing this research. Additionally, two organizations, the Thailand Science Research and Innovation Fund and the University of Phayao, provided support for this research project.

Funding

This research was supported by the University of Phayao and the Thailand Science Research and Innovation Fund (Fundamental Fund 2026).

REFERENCES

[1] Faryadi, S., & Velni, J. M. (2020). A reinforcement learning-based approach for modeling and coverage of an unknown field using a team of autonomous ground vehicles. *International Journal of Intelligent Systems*, *36*(2), 1069–1084. https://doi.org/10.1002/int.22331

[2] Khadka, R. B., & Kumar, P. (2024). Enhancing Sustainability in Agriculture through Artificial Intelligence. *International Scientific Journal of Engineering and Management*, *03*(12), 1–7. https://doi.org/10.55041/isjem02192

[3] Shyr, J., Chu, J., & Woods, M. (2018). Cognitive data analysis for big data. In *Springer handbooks of computational statistics* (pp. 23–47). https://doi.org/10.1007/978-3-319-18284-1_2

[4] Asakiewicz, C. (2016). Cognitive Analytics for making better Evidence-Based Decisions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2965767

[5] Beg, A., & Chauhan, N. S. (2024). Microcontroller Implementation in Precision Planting: A review. *Asian Journal of Advances in Agricultural Research*, *24*(2), 21–33. https://doi.org/10.9734/ajaar/2024/v24i2489

[6] B. Gundogan, N. Dowlut, S. Rajmohan, R. Mimi, C. Borrelli, M. C. Millip, C. Iosifidis, Z. Yagazie, U. Ginimol, M. Mathew, J. Alexander, R. Fowler and A. Agha, "Assessing the compliance of systematic review research works published in leading dermatology journals with the PRISMA statement guidelines: A systematic review," 2020, doi: 10.1016/J.JDIN.2020.07.007.

[7] S. Hullumani & R. Raghumahanti, "PRISMA Guidelines and Search Strategy," 2024, doi: 10.6084/m9.figshare.25013264.

[8] Alipbeki, O., Alipbekova, C., Sterenharz, A., Toleubekova, Z., Aliyev, M., Mineyev, N., & Amangaliyev, K. (2020). A Spatiotemporal Assessment of Land Use and Land Cover Changes in Peri-Urban Areas: A Case Study of Arshaly District, Kazakhstan. *Sustainability*, *12(4)*, 1556. https://doi.org/10.3390/su12041556

[9] PraveenaSri, P., & Prasuna, V. N. P. (2024). AI Applications of Clean Energy Innovations. *IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, *1–6*. https://doi.org/10.1109/CONECCT62155.2024.10677112

[10] Bhasme, M., Ulhe, P., & Gokhe, S. (2024). AI in Agriculture: The Future of Robotic Farming and Automation. *2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*, *1–7*. https://doi.org/10.1109/IDICAIEI61867.2024.10842885

[11] Refaai, M. Reda. A., Dattu, V. S., Gireesh, N., Dixit, E., Sandeep, Ch., & Christopher, D. (2022). Application of IoT-Based Drones in Precision Agriculture for Pest Control. *Advances in Materials Science and Engineering*, *1–12*. https://doi.org/10.1155/2022/1160258

[12] Na, M. H., & Na, I. S. (2024). AI-Powered Predictive Modelling of Legume Crop Yields in a Changing Climate. *Legume Research. 47(8),* 1390-1395. https://doi.org/10.18805/LRF-790

[13] Wu, L., Lin, X., Chen, Z., Huang, J., Liu, H., & Yang, Y. (2021). An Efficient Binary Convolutional Neural Network With Numerous Skip Connections for Fog Computing. *IEEE Internet of Things Journal*, *8(14)*, 11357–11367. https://doi.org/10.1109/JIOT.2021.3052105

[14] Narayani, A. R., Nagalakshmi, R., & Pakrasi, D. (2023). Application of geoinformatics to classify the landcover from satellite imagery using NDVI – A case of peri-urban regions of Southern Chennai. *IOP Conference Series: Earth and Environmental Science*, *1210(1)*, 012034. https://doi.org/10.1088/1755-1315/1210/1/012034

[15] Upadhyay, R., & Kamble, M. (2024). Artificial Intelligence in Precision Agriculture: A Comprehensive Review. *7th International Conference on Contemporary Computing and Informatics (IC3I)*, 918–923. https://doi.org/10.1109/IC3I61595.2024.10829061

[16] Korir, M. K., Mwangi, W., & Kimwele, M. W. (2023). Artificial Intelligence-Based Chatbot Model Providing Expert Advice to Potato Farmers in Kenya. *IEEE AFRICON*, 1–6. https://doi.org/10.1109/AFRICON55910.2023.10293557

[17] Pathak, S., & Laikram, S. (2022). Cooperative education during Covid-19 pandemic: Enhancing legal rights and professional development of interns in Thailand. *International Journal of Disaster Resilience in the Built Environment, 13(2),* 133–149. https://doi.org/10.1108/IJDRBE-08-2021-0098

[18] Sarkar, S. K., Rudra, R. R., Sohan, A. R., Das, P. C., Ekram, K. M. M., Talukdar, S., Rahman, A., Alam, E., Islam, M. K., & Islam, A. R. Md. T. (2023). Coupling of machine learning and remote sensing for soil salinity mapping in coastal area of Bangladesh. *Scientific Reports*, *13(1)*, 17056. https://doi.org/10.1038/s41598-023-44132-4

[19] Kim, T. Y., Ku, H., & Lee, S.-Y. (2020). Crop Enhancement of Cucumber Plants under Heat Stress by Shungite Carbon. *International Journal of Molecular Sciences*, *21(14)*, 4858. https://doi.org/10.3390/ijms21144858

[20] Bobade, G., Dhule, C., Khadilkar, R., Agrawal, R., & Morris, N. C. (2023). Design of Smart Irrigation System Based on MLA. *IEEE World Conference on Applied Intelligence and Computing (AIC)*, 174–179. https://doi.org/10.1109/AIC57670.2023.10263833

[21] Gvk, S., Bhumireddy, M. R., Rao, M., Bapat, J., & Das, D. (2024). Ethical Design of Edge-based Smart Irrigation System by use Case Analysis. *IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, 491–498. https://doi.org/10.1109/CSNT60213.2024.10545792

[22] Liu, J., Zhou, Y., Li, Y., Li, Y., Hong, S., Li, Q., Liu, X., Lu, M., & Wang, X. (2023). Exploring the Integration of Digital Twin and Generative AI in Agriculture. *15th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, 223–228. https://doi.org/10.1109/IHMSC58761.2023.00059

[23] Reddy, K. P., Roshni Thanka, M., Edwin, E. B., Ebenezer, V., S, S. K., & Joy, P. (2024). Farm_era: Precision Farming with GIS, AI Pest Management, Smart Irrigation, Data Analytics and Optimized Crop Planning. *International Conference on Inventive Computation Technologies (ICICT)*, 1134–1140. https://doi.org/10.1109/ICICT60155.2024.10544687

[24] Prasad, C., Kumar, S., & Damodar Rathod, S. (2024). Hybrid Ensemble Model for Accurate Orange Quality Prediction Using CNN, Random Forest and Gradient with Multispectral Data Integration. *International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)*, 1–4. https://doi.org/10.1109/ICAIQSA64000.2024.10882172

[25] Rao, S., Kumar, P., Kumari, N., Choudhary, A., Shukla, A. K., & Bhateja, A. (2023). Implementation of IoT-Based Smart Irrigation System for Agricultural Transformation. *IEEE World Conference on Applied Intelligence and Computing (AIC)*, 1001–1007. https://doi.org/10.1109/AIC57670.2023.10263854

[26] Passias, A., Tsakalos, K.-A., Rigogiannis, N., Voglitsis, D., Papanikolaou, N., Michalopoulou, M., Broufas, G., & Sirakoulis, G. Ch. (2024). Insect Pest Trap Development and DL-Based Pest Detection: A Comprehensive Review. *IEEE Transactions on AgriFood Electronics*, *2(2)*, 323–334. https://doi.org/10.1109/TAFE.2024.3436470

[27] Kernbach, S. (2024). In-situ biological ozone detection by measuring electrochemical impedances of plant tissues (No. arXiv:2411.16321). arXiv. https://doi.org/10.48550/arXiv.2411.16321

[28] Kumar, K. A., & Jayaraman, K. (2020). Irrigation control system - data gathering in WSN using IOT. *International Journal of Communication Systems*, *33(16)*, e4563. https://doi.org/10.1002/dac.4563

[29] El Mokh, F., Nagaz, K., & Masmoudi, M. M. (2024). Long-term potato response to different irrigation scheduling methods using saline water in an arid environment. *Frontiers in Agronomy*, *6*, 1426034. https://doi.org/10.3389/fagro.2024.1426034

[30] Maleki, R., Shams, S. M., Chellehbari, Y. M., Rezvantalab, S., Jahromi, A. M., Asadnia, M., Abbassi, R., Aminabhavi, T., & Razmjou, A. (2022). Materials discovery of ion-selective membranes using artificial intelligence. *Communications Chemistry*, *5(1)*, 132. https://doi.org/10.1038/s42004-022-00744-x

[31] M G, D., Jopate, R., & Albalushi, R. A. R. A. (2024). Optimizing Sustainable Agriculture in Oman Through AI-Driven Precision Techniques. *2nd International Conference on Computing and Data Analytics (ICCDA)*, 1–4. https://doi.org/10.1109/ICCDA64887.2024.10867335

[32] Ed-Daoudi, R., Alaoui, A., Ettaki, B., & Zerouaoui, J. (2024). Optimizing Sustainable Cultivation Through Smart Irrigation. *Journal of Information Sciences*, *22*(2). https://doi.org/10.34874/IMIST.PRSM/jis-v22i2.45154.

[33] Byrd, V. L., Saraswat, D., & Ahmad, A. (2023). Pilot Course in Data Visualization with a Multidisciplinary Approach: Technology+Agricultural Engineering. *IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/FIE58773.2023.10343451

[34] Rif'an, M., & Irianto, K. D. (2024). Precision Agriculture System with IoT: An Approach to Increase Production and Efficiency. *International Journal Software Engineering and Computer Science (IJSECS), 4(3)*, 1305–1316. https://doi.org/10.35870/ijsecs.v4i3.3259

[35] Abioye, E. A., Hensel, O., Esau, T. J., Elijah, O., Abidin, M. S. Z., Ayobami, A. S., Yerima, O., & Nasirahmadi, A. (2022). Precision Irrigation Management Using Machine Learning and Digital Farming Solutions. *AgriEngineering, 4(1)*, 70–103. https://doi.org/10.3390/agriengineering4010006

[36] Shofiyati, R. (2022). The Role of Remote Sensing Technology (Latest Geospatial) for Agriculture in Indonesia: 2nd International Conference on Smart and Innovative Agriculture (ICoSIA 2021), Yogyakarta, Indonesia. https://doi.org/10.2991/absr.k.220305.037

[37] Ma, Z., & Zhao, J. (2023). QoS Aware Resource Management in Mobile Edge Computing for Emerging Artificial Intelligence (AI) Applications. *19th International Conference on Mobility, Sensing and Networking (MSN)*, 238–245. https://doi.org/10.1109/MSN60784.2023.00045

[38] Mohan, S. S., Venkat, R., Rahaman, S., Vinayak, M., & Babu, B. H. (2023). Role of AI in Agriculture: Applications, Limitations and Challenges: A Review. *Agricultural Reviews*, *44*(2), 231–237. https://doi.org/10.18805/ag.R-2215

[39] Das, S. S., Behera, D. D., Nayak, B. B., Patjoshi, P. K., & Dash, S. (2024). Role of AI in Improving Climate Change and Achieving Sustainable Development: Cases from Odisha. *Grenze International Journal of Engineering & Technology (GIJET), 10*(2,Part 3), 2142–2149.

[40] Gayathri, M., Shunmugam, D. A., & Ishwariya, A. (2021). Smart irrigation system using IoT. *International Journal of Innovative Research in Multidisciplinary Perspectives, 9(3)*, 27. https://doi.org/10.37082/IJIRMPS.2021.v09i03.027

[41] Rathinavel, L., Singh, S., Rai, P. K., Chandra, N., Jothinathan, D., Gaffar, I., Pandey, A. K., Choure, K., Waoo, A. A., Joo, J. C., & Pandey, A. (2024). *Sustainable Microalgal Biomass for Efficient and Scalable Green Energy Solutions: Fueling Tomorrow*. *Fuels, 5(4)*, 868–894. https://doi.org/10.3390/fuels5040049

[42] P, H., Mk, C., Rh, L., C, R., G, S. K., & K, S. (2025). Thermal canopy segmentation in tomato plants: A novel approach with integration of YOLOv8-C and FastSAM. *Smart Agricultural Technology*, *10*. https://doi.org/10.1016/j.atech.2025.100806

[43] Sood, A., Bhardwaj, A. K., & Sharma, R. K. (2024). Towards sustainable agriculture: Key determinants of adopting artificial intelligence in agriculture. *Journal of Decision Systems, 33(4)*, 833–877. https://doi.org/10.1080/12460125.2022.2154419

[44] Dao, K. M., Yabar, H., & Mizunoya, T. (2020). Unlocking the Energy Recovery Potential from Sustainable Management of Bio-Resources Based on GIS Analysis: Case Study in Hanoi, Vietnam. *Resources, 9(11)*, 133. https://doi.org/10.3390/resources9110133

[45] Khater, E.-S. G., Ali, S. A., Afify, M. T., Bayomy, M. A., & Abbas, R. S. (2022). Using of geographic information systems (GIS) to determine the suitable site for collecting agricultural residues. Scientific Reports, 12(1), 14567. https://doi.org/10.1038/s41598-022-18850-0

[46] Sermmany, K., Wanjantuk, P., & Leelapatra, W. (2024). Utilizing Explainable Artificial Intelligence (XAI) to Identify Determinants of Coffee Quality. 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE), 696–703. https://doi.org/10.1109/JCSSE61278.2024.10613641

[47] Sharma, B., Singhvi, R., Chauhan, P., & Naik, N. (2024). A study to know the role of AI and sustainability in agriculture. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *08*(01), 1–13. https://doi.org/10.55041/ijsrem28318

[48] More, A. (2024). IoT in Agriculture. *Acta Scientific Agriculture*, 39–42. https://doi.org/10.31080/asag.2024.08.1370

[49] S, K. (2018). IoT in Agriculture : Smart Farming. International Journal of Scientific Research in Computer Science Engineering and Information Technology, 181–184. https://doi.org/10.32628/cseit183856

[50] Devendiran, R., Rao, P. S., Turukmane, D. A. V., Prasad, B., Sathiyaraj, A., & Pulipati, S. (2023). *Smart Irrigation: Revolutionizing Water Management in Agriculture for Sustainable Practices and Improved Crop Yield*. 651–656. https://doi.org/10.1109/icrtac59277.2023.10480750

[51] Butora, A., Soloniewicz, B., Schwartz, C., Aziz, C., Su, S., & Mahmoud, M. (2022). *The Practical use of GIS in Agriculture*. 1525–1529. https://doi.org/10.1109/csci58124.2022.00270

[52] Sangirova, U., Yunusov, I., Ahmedov, U. K., Namozov, J. A., Dustmurodov, G. G., & Khakimov, R. (2020). *Features of GIS application in agriculture and logistics*. *918*(1), 012140. https://doi.org/10.1088/1757-899X/918/1/012140

[53] Mim, M. I., Sultana, F., & Hasan, M. R. (2025). AI-Powered Autonomous Farming: The Future of Sustainable Agriculture. *European Journal of Theoretical and Applied Sciences*, *3*(1), 11–31. https://doi.org/10.59324/ejtas.2025.3(1).02

A Novel Approach for Enhancing Advanced Encryption Standard Performance and Cryptographic Resilience

Muthu Meenakshi Ganesan, Sabeen Selvaraj
Department of Computer Science-Faculty of Science and Humanities,
SRM Institute of Science and Technology, Kattankulathur, India

***Abstract*—Advanced Encryption Standard (AES) encrypts data in blocks of sixteen bytes to secure confidential data stored in the cloud. For cloud-based systems, enhancements in existing encryption techniques are necessary as the nature of cyber threats evolves and computational speed becomes increasingly critical. This study presents an enhanced design of AES that substitutes two special operations, Byte Transformation and Bits Permuted Bytes, with the conventional S-Box operation to improve the speed and security of the encryption method. The suggested round structure in the new approach of AES, which preserves the original data block size, consists of the following operations: Byte Transformation, Shift Rows, Bits Permuted Bytes, Add Round Key, and Mix Columns. The analysis of the strict avalanche effect, correlation coefficient, entropy, execution time, and throughput outcomes confirms that the developed scheme improves the security and processing speed.**

Keywords*—Cryptography; NIST; AES; block cipher; key expansion; symmetric encryption; galois field; statistical techniques; cryptanalysis

I. Introduction

Cloud platforms that utilize symmetric encryption algorithms, such as AES, ensure that data transmission, storage, and communication are secure for the government, healthcare, and financial sectors [1], [2]. AES secures vast volumes of private information from illegal access and breaches in these cloud-based systems [3]. In 2000, the National Institute of Standards and Technology (NIST) in the United States replaced the Data Encryption Standard with AES due to its superior performance in both hardware and software [4], [5]. Researchers often utilize AES to secure wireless networks, data storage, and communications [6], [7].

Several cryptanalytic studies suggest that improving diffusion and confusion mechanisms can further reinforce AES against potential future cryptanalytic attacks. When internal transformations lack sufficient randomness or complexity, the encryption process may reveal patterns that can be exploited through statistical or differential attacks. These vulnerabilities highlight the importance of strengthening AES, not only to address current vulnerabilities but also to remain resilient against evolving and more sophisticated attacks.

This study addresses these limitations by proposing a modified AES encryption algorithm that enhances both cryptographic strength and operational efficiency. The proposed approach introduces structural changes to AES transformations to improve diffusion, reduce statistical correlation, and strengthen resistance against cryptanalytic attacks.

The rest of this study is organized as follows: Section II summarizes related work; Section III recommends the enhanced AES encryption approach; Section IV outlines the statistical tests and evaluation criteria applied to assess the resilience of existing and proposed encryption algorithms. Section V presents the experimental results obtained from these evaluations, while Section VI discusses the key findings and their implications. Finally, Section VII concludes this study by highlighting potential directions for future research.

II. Related Work

Many studies have been conducted to improve the security and efficiency of symmetric encryption algorithms. These studies suggest numerous technical variants of existing encryption methods to boost security and performance. Parthasarathy et al. [8] proposed a lightweight symmetric encryption that combined the Elliptic Curve Diffie-Hellman key exchange algorithm with a proprietary random number generator, resulting in a unique S-box and innovative folding operations for transmitting medical data to the cloud. The throughput of this method was higher than that of AES; however, it lacks the statistical evaluation of cryptographic strength.

Baladhay et al. [9] proposed an enhanced version of the AES algorithm by replacing the MixColumns function with an alternate transformation to improve the security and efficiency of encryption and decryption processes. This method is ideal for applications requiring high-speed video encryption due to the performance improvements. On the other hand, entropy and correlation analysis, which could offer additional information on cryptographic strength, are excluded from this work. Similarly, to increase security and performance for systems with low resources, Hammod et al. [10] introduced a Modified Lightweight AES (ML-AES).

Sirajuddin et al. [11] designed a modified Rijndael algorithm for resource-constrained Internet of Things devices. The shift column phase was changed, the sub-byte step was removed, and a crossover pre-processing method was added. The research did not evaluate the ciphertext's randomness or statistical independence, although these improvements were made to increase efficiency and security. Wenceslao Jr. et al. [12] developed a new version of AES encryption that retained the

original Rijndael S-Box and introduced a second S-Box using affine transformations for the Mix-Columns step. This method is faster but exhibits a slightly reduced avalanche effect, which may affect the diffusion property of the encryption process.

A DNA-based multi-round encryption approach was presented by Al-Shargabi et al. [13] to preserve confidential health data. Improving entropy levels while maintaining an avalanche effect similar to conventional AES allowed this technique to outperform AES in data security. By eliminating the Mix-Columns stage and adding a continuous fraction function to compress the ciphertext, Mohammad et al. [14] developed a lightweight AES encryption (LWC-AES) for devices with limited resources. LWC-AES reduced processing complexity and enhanced transfer rates without compromising security. Altigani et al. [15] introduced Polymorphic-AES (P-AES) for adaptive and flexible encryption using the initial key. P-AES showed exceptional resilience to attacks by modifying the Sub Bytes, Shift Rows, and Mix Columns procedures to improve security with negligible performance impact. Salih et al. [16] modified a 3D chaotic map and combined two dynamic XOR tables with a dynamic Mix-Columns transformation to strengthen the AES encryption method. While this approach improved security, additional performance evaluation would enhance its practical relevance.

Assafli et al. [17] modified AES encryption by replacing the traditional S-box with a dynamic one based on Unix epoch time. This new method solely preserved and improved the avalanche effect for single-bit input changes. Abikoye et al. [18] proposed an enhanced AES technique, with modified S-Box and Shift-Rows transformations to strengthen security. This method made the Sub Bytes operation round-key dependent and randomized the Shift Rows operation exclusively to improve the avalanche effect. Manoj Kumar et al. [19] presented a new AES key expansion technique primarily to increase the execution speed, allowing simultaneous subkey generation by isolating the operation into two concurrent blocks. This innovative approach was implemented on an FPGA (Virtex 5 XC5VLX50T) to secure ECG signal encryption and decryption in real-time communication.

In summary, many studies prioritize either efficiency or security without achieving an adequate balance. While several enhancements to AES exist, most lack rigorous statistical evaluation. The proposed system addresses these gaps by achieving a balance between security and efficiency, with its reliability confirmed through a thorough statistical analysis.

III. METHODOLOGY

AES uses 128-bit data blocks and encrypts in 10, 12, or 14 rounds, based on the key length, which can be 128, 192, or 256 bits [20]. It involves three key processes: Key expansion generates round keys from the original cipher key, encryption transforms plaintext into ciphertext through various operations, and decryption reverses these operations to recover the original plaintext [21].

A. Key Expansion in AES Algorithm

In AES-128, the key expansion process, depicted in Fig. 1, begins by dividing the 128-bit initial key into four 32-bit words (W0, W1, W2, W3), forming the initial part of the key schedule [19]. A left circular byte shift (RotateWord), a substitution using the S-Box (SubWord), and an XOR with the appropriate round constant (RCj) are used to compute the remaining 40 words [22]. The SubWord transformation uses the S-Box depicted in Fig. 2, a 16 x 16 lookup table for byte substitution. The round constants, described in Table I, are applied at specific stages during the key expansion to introduce non-linearity. These transformations are applied to designated words W3, W7, W11, and every fourth word thereafter, to generate W3', W7', W11', respectively. The subsequent words are created by XORing each new word with previously generated words, iteratively producing all 44 words. These words are divided into 11 sets, each with 4 words, forming the round keys used for encryption and decryption.

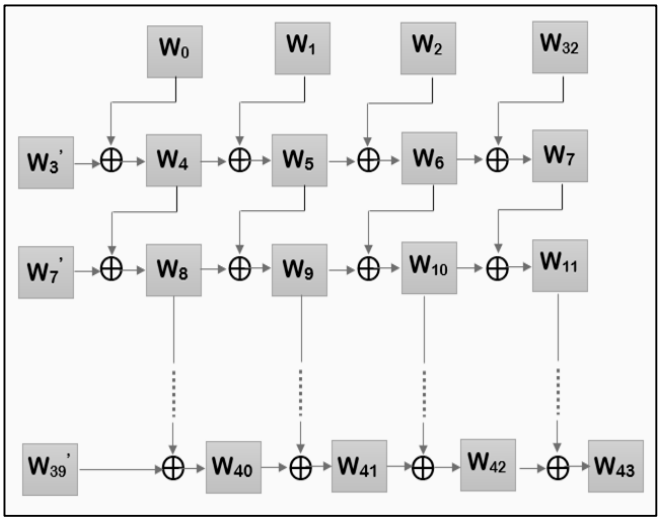


Fig. 1. AES key scheduling process.

TABLE I. CONSTANTS USED PER ROUND

Round(j)	1	2	3	4	5	6	7	8	9	10
RCj	0x01	0x02	0x04	0x08	0x10	0x20	0x40	0x80	0x1b	0x36

B. Existing AES Encryption Algorithm

During encryption or decryption, a state matrix (a 4x4 grid of bytes) represents the 128-bit plaintext block. The matrix arranges each byte of the 128-bit block into a single cell in column-major order. This matrix undergoes various operations during encryption, including SubstituteBytes, ShiftRows, MixColumns, and AddRoundKey, with the final round omitting MixColumns [23]. These operations are described below:

1) SubstituteBytes. In this substitution phase, each byte in the state matrix is replaced with the value of the S-Box (illustrated in Fig. 2) based on its original value, introducing non-linearity into the encryption process [24]. Fig. 3 shows the inverse AES S-Box used during decryption.

2) ShiftRows. This step enhances byte dispersion by executing a cyclic left shift on each row of the state matrix. First row remains unaffected, while the second row is shifted one byte to the left, the third row is shifted two bytes to the left, and the fourth row is shifted three bytes to the left.

	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0	63	7c	77	7b	f2	6b	6f	c5	30	01	67	2b	fe	d7	ab	76
1	ca	82	c9	7d	fa	59	47	f0	ad	d4	a2	af	9c	a4	72	c0
2	b7	fd	93	26	36	3f	f7	cc	34	a5	e5	f1	71	d8	31	15
3	04	c7	23	c3	18	96	05	9a	07	12	80	e2	eb	27	b2	75
4	09	83	2c	1a	1b	6e	5a	a0	52	3b	d6	b3	29	e3	2f	84
5	53	d1	00	ed	20	fc	b1	5b	6a	cb	be	39	4a	4c	58	cf
6	d0	ef	aa	fb	43	4d	33	85	45	f9	02	7f	50	3c	9f	a8
7	51	a3	40	8f	92	9d	38	f5	bc	b6	da	21	10	ff	f3	d2
8	cd	0c	13	ec	5f	97	44	17	c4	a7	7e	3d	64	5d	19	73
9	60	81	4f	dc	22	2a	90	88	46	ee	b8	14	de	5e	0b	db
a	e0	32	3a	0a	49	06	24	5c	c2	d3	ac	62	91	95	e4	79
b	e7	c8	37	6d	8d	d5	4e	a9	6c	56	f4	ea	65	7a	ae	08
c	ba	78	25	2e	1c	a6	b4	c6	e8	dd	74	1f	4b	bd	8b	8a
d	70	3e	b5	66	48	03	f6	0e	61	35	57	b9	86	c1	1d	9e
e	e1	f8	98	11	69	d9	8e	94	9b	1e	87	e9	ce	55	28	df
f	8c	a1	89	0d	bf	e6	42	68	41	99	2d	0f	b0	54	bb	16

Fig. 2. AES S-Box.

	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0	52	09	6a	d5	30	36	a5	38	bf	40	a3	9e	81	f3	d7	fb
1	7c	e3	39	82	9b	2f	ff	87	34	8e	43	44	c4	de	e9	cb
2	54	7b	94	32	a6	c2	23	3d	ee	4c	95	0b	42	fa	c3	4e
3	08	2e	a1	66	28	d9	24	b2	76	5b	a2	49	6d	8b	d1	25
4	72	f8	f6	64	86	68	98	16	d4	a4	5c	cc	5d	65	b6	92
5	6c	70	48	50	fd	ed	b9	da	5e	15	46	57	a7	8d	9d	84
6	90	d8	ab	00	8c	bc	d3	0a	f7	e4	58	05	b8	b3	45	06
7	d0	2c	1e	8f	ca	3f	0f	02	c1	af	bd	03	01	13	8a	6b
8	3a	91	11	41	4f	67	dc	ea	97	f2	cf	ce	f0	b4	e6	73
9	96	ac	74	22	e7	ad	35	85	e2	f9	37	e8	1c	75	df	6e
a	47	f1	1a	71	1d	29	c5	89	6f	b7	62	0e	aa	18	be	1b
b	fc	56	3e	4b	c6	d2	79	20	9a	db	c0	fe	78	cd	5a	f4
c	1f	dd	a8	33	88	07	c7	31	b1	12	10	59	27	80	ec	5f
d	60	51	7f	a9	19	b5	4a	0d	2d	e5	7a	9f	93	c9	9c	ef
e	a0	e0	3b	4d	ae	2a	f5	b0	c8	eb	bb	3c	83	53	99	61
f	17	2b	04	7e	ba	77	d6	26	e1	69	14	63	55	21	0c	7d

Fig. 3. AES inverse S-box.

3) MixColumns. This step applies a linear transformation to each column of the 4 x 4 state matrix. It mixes the bytes within each column by multiplying the MixColumns matrix over a Galois Field (GF(2^8)). This step enhances diffusion by ensuring that each input byte influences multiple output bytes, as shown in Eq. (1):

$$\text{MixColumns Matrix} = \begin{matrix} 02 & 03 & 01 & 01 \\ 01 & 02 & 03 & 01 \\ 01 & 01 & 02 & 03 \\ 03 & 01 & 01 & 02 \end{matrix} \quad (1)$$

During decryption, the inverse transformation is applied using the Inverse MixColumn matrix, which is shown in Eq. (2):

$$\text{Inverse MixColumns Matrix} = \begin{matrix} 0e & 0b & 0d & 09 \\ 09 & 0e & 0b & 0d \\ 0d & 09 & 0e & 0b \\ 0b & 0d & 09 & 0e \end{matrix} \quad (2)$$

4) AddRoundKey. This step combines the key with the data to strengthen the encryption by XORing, the round key with the state matrix.

During decryption, the process reverses encryption by applying inverse operations in reverse order. It begins with InverseShiftRows to restore the original row positions, followed by InverseSubstituteBytes using the inverse S-box, and AddRoundKey XORs the state matrix with the corresponding round key. Finally, InverseMixColumns (except in the final round) to unmix the columns to recover the plaintext from the ciphertext.

C. Proposed Enhanced AES Encryption Architecture

In the new design of AES encryption, the input block size remains 128 bits; however, the Sub Bytes operation has been eliminated. Instead, two new operations, Byte Transformation and Bits Permuted Bytes, have been introduced. In each round of the proposed approach, Bytes Transformation, Shift Rows, Mix Columns, Bits Permuted Bytes, and Add Round Key transform the state matrix. These transformations aim to improve the encryption process's confusion and diffusion properties. The newly introduced operations are described below:

1) Byte transformation operation. Let S represent the 4x4 state matrix, where each element (S_{ij}) is an 8-bit byte. AES transforms this state matrix through various stages as part of its encryption and decryption mechanisms. In the proposed method, the Byte Transformation function is applied to each byte (S_{ij}). This operation introduces non-linearity and enhances the diffusion properties of the encryption process. This transformation is defined as follows:

Let $S_{ij} = b_7b_6b_5b_4b_3b_2b_1b_0$; here b_7 and b_0 are the Most Significant Bit (MSB) and the Least Significant Bit (LSB), respectively. The transformation for each bit (t_i) of the transformed byte (T_{ij}) is defined as follows:

$t_7 = b_7$ (MSB remains unchanged), $t_6 = b_6 \oplus b_7$, $t_5 = b_5 \oplus b_6$, $t_4 = b_4 \oplus b_5$, $t_3 = b_3 \oplus b_4$, $t_2 = b_2 \oplus b_3$, $t_1 = b_1 \oplus b_2$, and $t_0 = b_0 \oplus b_1$. Here $t_7t_6t_5t_4t_3t_2t_1t_0$ (T_{ij}) is the transformed byte and $\oplus$ denotes the XOR operation. For example, consider a byte. S_{ij} , where $b_7b_6b_5b_4b_3b_2b_1b_0$ = 11011010. $t_7 = b_7 = 1$, $t_6 = b_6 \oplus b_7 = 1 \oplus 1 = 0$, $t_5 = b_5 \oplus b_6 = 0 \oplus 1 = 1$, $t_4 = b_4 \oplus b_5 = 1 \oplus 0 = 1$, $t_3 = b_3 \oplus b_4 = 1 \oplus 1 = 0$, $t_2 = b_2 \oplus b_3 = 0 \oplus 1 = 1$, $t_1 = b_1 \oplus b_2 = 1 \oplus 0 = 1$, and $t_0 = b_0 \oplus b_1 = 0 \oplus 1 = 1$. The transformed byte T_{ij}becomes 10110111. This transformation is visually represented in Fig. 4.

2) Inverse byte transformation operation. The Inverse Byte Transformation operation undoes the transformations applied during the encryption phase. Given the transformed byte, S_{ij} can be recovered as follows:

$b_7 = t_7$ (MSB remains unchanged), $b_6 = t_6 \oplus b_7$, $b_5 = t_5 \oplus b_6$, $b_4 = t_4 \oplus b_5$, $b_3 = t_3 \oplus b_4$, $b_2 = t_2 \oplus b_3$, $b_1 = t_1 \oplus b_2$, and $b_0 = t_0 \oplus b_1$

For example, consider a byte, $T_{ij} = t_7t_6t_5t_4t_3t_2t_1t_0 =$ 10110111. $b_7 = t_7 = 1$ (MSB remains unchanged), $b_6 = t_6 \oplus b_7 = 0 \oplus 1 = 1$, $b_5 = t_5 \oplus b_6 = 1 \oplus 1 = 0$, $b_4 = t_4 \oplus b_5 = 1 \oplus 0 = 1$, $b_3 = t_3 \oplus b_4 = 0 \oplus 1 = 1$, $b_2 = t_2 \oplus b_3 = 1 \oplus 1 = 0$, $b_1 = t_1 \oplus b_2 = 1 \oplus 0 = 1$, and $b_0 = t_0 \oplus b_1 = 1 \oplus 1 = 0$. Now, combine all the bits. $b_7b_6b_5b_4b_3b_2b_1b_0$ to form the original byte (S_{ij}) as 11011010. In the encryption process, the Bytes Transformation procedure modifies each byte, and the Inverse Bytes Transformation procedure restores each byte in decryption. The step-by-step process of this inverse transformation is shown in Fig. 5.

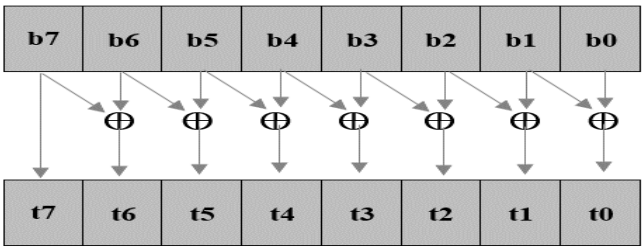


Fig. 4. Byte transformation operation.

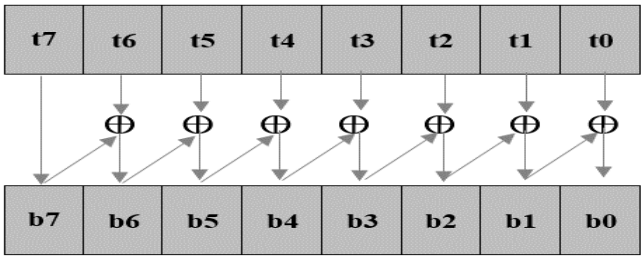


Fig. 5. Inverse byte transformation operation.

3) Bits permuted bytes operation. In the Bits Permuted Bytes operation, each byte in the state matrix S_{ij} ($b_7b_6b_5b_4b_3b_2b_1b_0$) undergoes a bit permutation based on the involutory permutation table 3 | 2 | 1 | 0 | 7 | 6 | 5 | 4 during the encryption phase. The involutory permutation table is self-inverse. This feature simplifies both encryption and decryption. It removes the need for a separate inverse table. This permutation rearranges the bits. The bit at index 3 (b_4) moves to index 0, the bit at index 2 (b_5) moves to index 1, the bit at index 1 (b_6) moves to index 2, the bit at index 0 (b_7) moves to index 3, the bit at index 7 (b_0) moves to index 4, the bit at index 6 (b_1) moves to index 5, the bit at index 5 (b_2) moves to index 6, and the bit at index 4 (b_3) moves to index 7. In this transformation, the original bit order of each byte in the state matrix is rearranged to $T_{ij} = b_4b_5b_6b_7b_0b_1b_2b_3$. For example, consider a byte S_{ij}= 11011010 in the state matrix. This byte is transformed by rearranging its bits to obtain T_{ij} =10110101. This bit permutation is applied to each byte in the state matrix, converting the original byte into a new form.

4) Inverse bits permuted bytes operation. The Inverse Bits Permuted Bytes operation reverses the bit permutation applied during encryption. The encryption phase rearranges the bits of each byte in the state matrix according to the involutory permutation table 3 | 2 | 1 | 0 | 7 | 6 | 5 | 4, whereas the inverse operation restores the original bit order. For example, a byte T_{ij} =10110101, when passed through the inverse transformation, it reverts to its original form S_{ij}= 11011010. Fig. 6 and Fig. 7 show the Bits Permuted Bytes and Inverse Bits Permuted Bytes operations, which rearrange and restore the bit order within each byte during encryption and decryption, respectively, in the developed scheme.

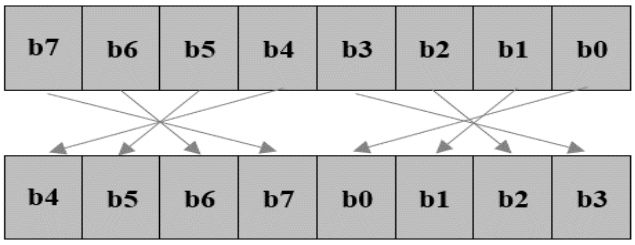


Fig. 6. Bits permuted bytes operation.

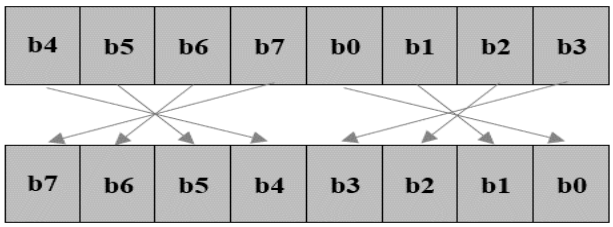


Fig. 7. Inverse bits permuted bytes operation.

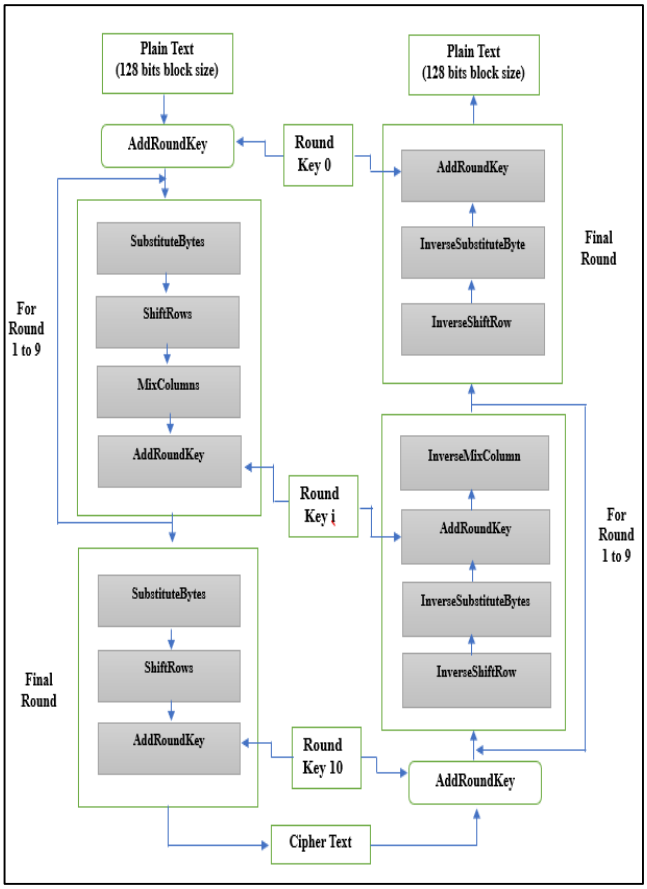


Fig. 8. Standard AES algorithm.

To enhance efficiency, the final round of the proposed encryption approach omits the Bits Permuted Bytes operation and Mix columns, thereby reducing complexity while maintaining security strength. Furthermore, the number of rounds has been decreased to accelerate execution speed. The standard AES functions with ten rounds for a 128-bit key size, as shown in Fig. 8, while the developed encryption strategy executes only five rounds with the same key size, as illustrated in Fig. 9.

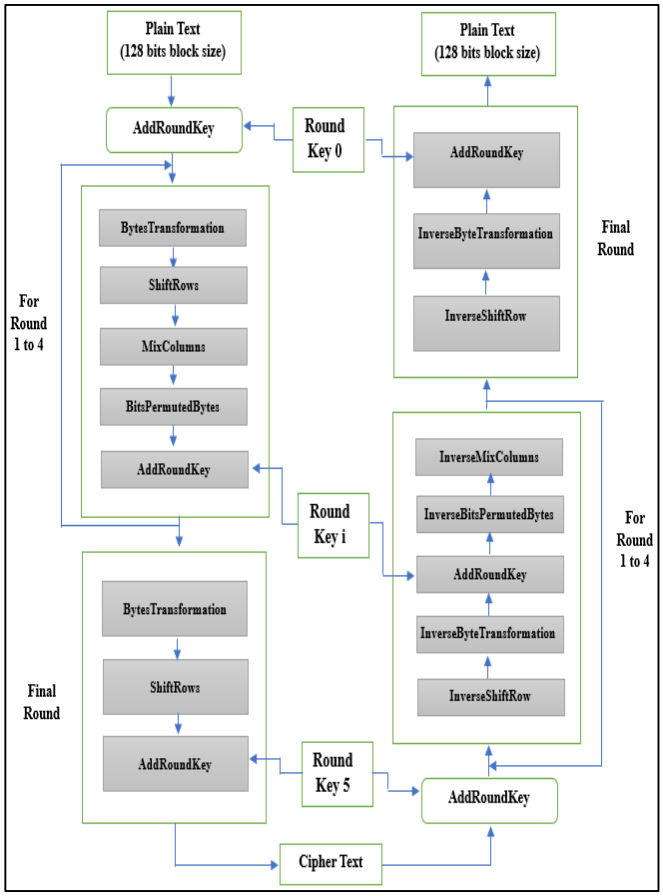


Fig. 9. Enhanced AES algorithm.

IV. EVALUATION

A. Strict Avalanche Effect

The Avalanche Effect is determined by flipping each bit in the n-bit plaintext and observing the percentage of bits that differ in the resulting ciphertext. Averaging the percentages across all n-bit flips gives the Strict Avalanche Effect (SAE). This metric ensures the algorithm's robustness and resistance to cryptanalysis. The formula used to calculate the SAE [25] is shown in Eq. (3):

$$SAE = \frac{1}{n}\sum_{i=1}^{n}\frac{HammingWeight(C \oplus C'_i)X100}{n} \quad (3)$$

Here, C denotes the ciphertext obtained by encrypting the original plaintext, C_i' denotes the ciphertext produced by encrypting the plaintext with the i^{th} bit flipped, and n represents the total number of bits in the ciphertext. One hundred random 128-bit plaintexts and keys were used to evaluate SAE.

B. Correlation Coefficient

This measure implies the level of a linear association between the plaintext and encrypted output. For an encryption algorithm to be considered secure, the ciphertext must have a minimal linear dependence on the plaintext, as indicated by a correlation coefficient close to zero [26].

The following steps are used to determine the correlation coefficient between plaintext and ciphertext:

1) Plaintext and ciphertext are converted into binary sequences of 0s and 1s.

2) Mean calculation. The mean value of the plaintext bits (μ_X) was computed using Eq. (4), and the mean value of the ciphertext bits (μ_Y) was determined using Eq. (5):

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i \quad (4)$$

$$\mu_Y = \frac{1}{N}\sum_{i=1}^{N} Y_i \quad (5)$$

Here, N is the number of bits in the plaintext or ciphertext, and X_i and Y_i are the individual bits of the plaintext and ciphertext, respectively.

3) Covariance Calculation ($Cov(X,Y)$) . Covariance measures how much the plaintext and ciphertext vary together: The covariance between the plaintext bits and ciphertext bits was calculated using Eq. (6):

$$\mathrm{Cov(X,Y)} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)\ \mathrm{X}\ (Y_i - \mu_Y) \quad (6)$$

Here, μ_X is the mean of plaintext bits, μ_Y is the mean of ciphertext bits, $X_i - \mu_X$ is the deviation of the i^{th} plaintext bit from the mean, and $Y_i - \mu_Y$ is the deviation of the i^{th} ciphertext bit from its mean. The deviations measure how each bit in the plaintext and ciphertext differs from their respective mean.

4) Standard deviation. The standard deviation of the plaintext bits (σ_X) and the ciphertext bits (σ_Y) was calculated using Eq. (7) and Eq. (8), respectively.

$$\sigma_X = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(X_i - \mu_X)^2} \quad (7)$$

$$\sigma_Y = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(Y_i - \mu_Y)^2} \quad (8)$$

5) Correlation coefficient. The value of the correlation coefficient (r) was calculated between the plaintext and ciphertext bits using Eq. (9):

$$r = \frac{Cov(X,Y)}{\sigma_X\ \mathrm{X}\ \sigma_Y} \quad (9)$$

One thousand random 128-bit plaintexts and keys were used to evaluate this metric.

C. Entropy

The entropy test assesses the degree of unpredictability and randomness in the ciphertext. A higher entropy indicates greater randomness in the ciphertext, as it is more difficult to crack the encryption. To determine the average entropy of encrypted ciphertexts in their binary representation, convert each

ciphertext C_i to B_i(binary form). The entropy of the ciphertext $H(X)$ can be expressed for Shannon entropy [27], [28] using Eq. (10):

$$H(X) = -\sum_{i=1}^{N} P(x_i) log_2 P(x_i) \quad (10)$$

In this context, x_i represents the possible outcomes, and N is the number of unique outcomes. The entropy for each binary string (B) was calculated using Eq. (11):

$$H(B) = -[P(0) log_2 P(0) + P(1) log_2 P(1)] \quad (11)$$

Here, $P(0)$ and $P(1)$ are the probabilities of the binary outcomes 0s and 1s, respectively, in the binary ciphertext (B).

The average entropy $\bar{H}$ across all N ciphertexts was calculated using Eq. (12):

$$\bar{H} = \frac{1}{N} \sum_{i=1}^{N} H(B_i) \quad (12)$$

This method interprets every bit as a random variable with two possible outcomes and obtains the average entropy of the binary representations, where N is the total number of ciphertexts. One thousand random 128-bit plaintexts and keys were used to evaluate this randomness property of the produced ciphertexts.

D. Execution Time

Execution time impacts efficiency, performance, and resource usage of cryptographic systems. To calculate the execution time of the modified and standard AES algorithms implemented in Python, the start and end times were recorded, and the difference value was calculated. The evaluations were performed on a 12th Gen Intel Core i7-1255U CPU running at 1.70 GHz. The dataset consists of a 1025 KB text file, a 1028 KB PDF document, a 1126 KB JPEG image, a 1026 KB MP3 audio file, and a 1031 KB MP4 video file used for the evaluation. The encryption and decryption times were calculated over 10 runs for each algorithm using all file types and averaged to obtain a reliable result.

E. Throughput

The encryption and decryption throughputs were calculated separately using Eq. (13) [29]:

$$\text{Throughput (KB/sec)} = \frac{\text{Total File Size (KB)}}{\text{Total Time (sec)}} \quad (13)$$

The Total File Size is the sum of all the file sizes used, which includes text, PDF, JPEG, MP3, and MP4. The amount of time needed to complete encrypting or decrypting all files is referred to as Total Time. This method enables a reliable evaluation of the algorithm's performance on various file types and sizes.

V. Results

A. Strict Avalanche Effect

The test results in Table II and Fig. 10 clearly show that the suggested new encryption process achieves a higher SAE of 53.46%, compared to 49.55% for the standard AES encryption. These results highlight that the modified encryption design achieves improved diffusion and higher security in fewer rounds.

TABLE II. Measured SAE in Standard and Enhanced AES

Algorithm	SAE
Standard AES	49.55
Enhanced AES	53.46

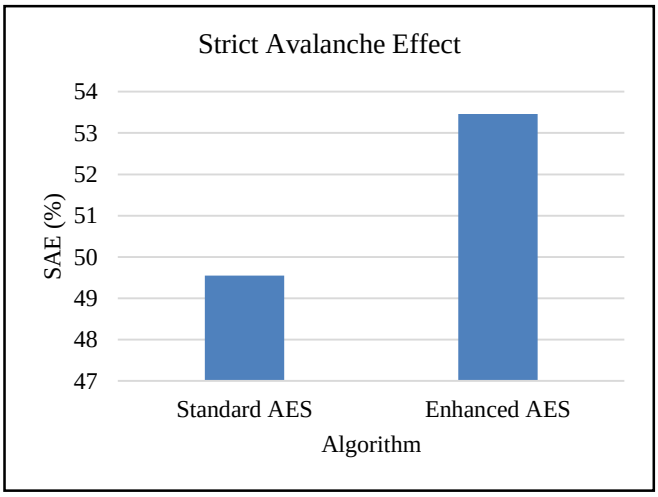


Fig. 10. SAE Analysis of standard versus enhanced AES.

B. Correlation Coefficient

Table III and Fig. 11, present the correlation analysis between plaintext and ciphertext. When compared to the standard AES, which has a correlation coefficient of 0.004409, the enhanced encryption approach demonstrates a lower correlation coefficient of -0.000293. The observed negative and nearly zero value confirms that the proposed technique effectively diminishes the relationship between plaintext and ciphertext, thereby enhancing cryptographic strength.

TABLE III. Statistical Correlation Assessment Between Plaintext and Ciphertext

Algorithm	Correlation Coefficient
Standard AES	0.004409
Enhanced AES	-0.000293

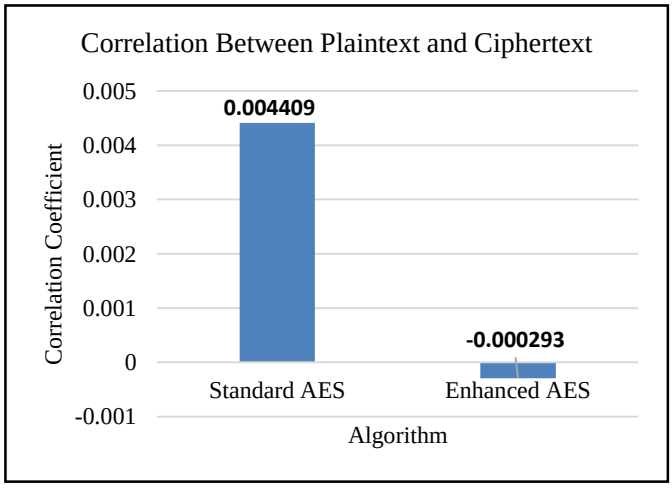


Fig. 11. Comparison of correlation coefficient between plaintext and ciphertext.

C. Entropy

The results indicate that the existing and the redesigned AES encryption achieved high entropy levels in their ciphertexts, as shown in Table IV. The enhanced encryption process offers a slightly higher average entropy of 0.994619 bits than the standard AES's 0.994327 bits. These values, close to the ideal maximum of one bit per bit, show that both algorithms generate random ciphertext sequences.

TABLE IV. ENTROPY VALUES OF CIPHERTEXT

Algorithm	Average Entropy
Standard AES	0.994327
Enhanced AES	0.994619

D. Execution Time

The method proposed in this study reduces encryption and decryption times for various file types, as illustrated in Table V and Fig. 12. It minimizes the encryption time by approximately 8.17% and the decryption time by 13.30% for a text file. A notable 32.96% reduction in encryption time and an 8.51% reduction in decryption time are observed for a PDF file. There is a 19.65% decrease in encryption time and a 16.17% decrease in decryption time for a JPEG file. An MP3 file shows a 20.44% reduction in encrypting time and an 11.95% reduction in decryption time. There is a 27.75% decrease in encryption time and a 20.16% decrease in decryption time for an MP4 file.

E. Throughput

The suggested encryption method achieves higher throughput than standard AES. Its encryption throughput is 389.16 KB/sec, approximately 28.9% higher than the standard AES throughput of 301.85 KB/sec. Similarly, it improves the decryption throughput by approximately 18%, achieving 311.30 KB/sec, whereas the standard AES achieves 263.79 KB/sec. These results highlight the efficiency gains provided by the proposed approach, as shown in Table VI and Fig. 13.

TABLE V. EXECUTION TIME FOR ENCRYPTION AND DECRYPTION OF MULTIPLE FILE TYPES

File Type & Size	Standard AES Encryption Time (sec)	Enhanced AES Encryption Time(sec)	Standard AES Decryption Time(sec)	Enhanced AES Decryption Time(sec)
Text File(1025KB)	2.010364	1.846177	2.539962	2.202060
PDF File(1028KB)	1.482714	0.994051	2.207906	2.019965
JPEG File(1126KB)	3.441201	2.764880	3.466463	2.905801
MP3 File(1026KB)	4.473375	3.559055	4.905859	4.319511
MP4 File(1031KB)	5.938684	4.290501	6.728915	5.372506



Fig. 12. Comparison of execution time across multiple file formats.

TABLE VI. THROUGHPUT METRICS FOR ENCRYPTION AND DECRYPTION PROCESSES

Algorithm	Encryption Throughput (KB/Sec)	Decryption Throughput (KB/Sec)
Standard AES	301.85	263.79
Enhanced AES	389.16	311.30

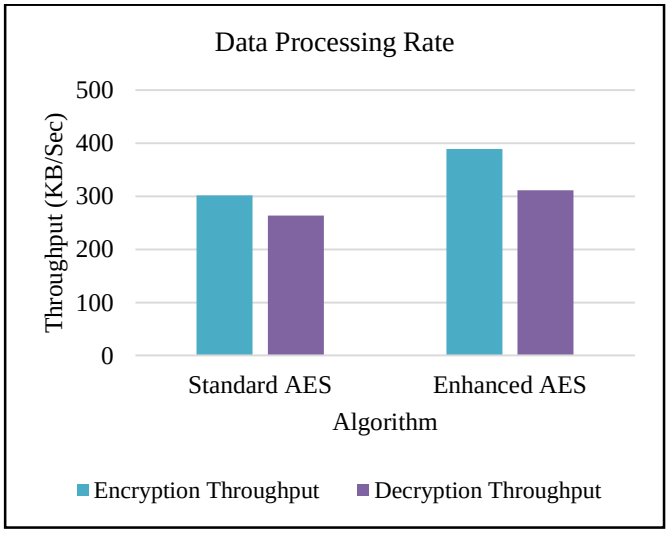


Fig. 13. Efficiency assessment based on throughput.

VI. Discussion

The experimental results confirm that the proposed encryption approach has notable improvements over the standard AES across several metrics. The enhanced proposed scheme offers improved diffusion and security, achieving a higher SAE of 53.46% compared to 49.55% for the standard AES. Its correlation coefficient is lower at -0.000293, indicating reduced predictability between plaintexts and ciphertexts, while the standard AES records a value of 0.004409. Additionally, its average entropy increases slightly from 0.994327 to 0.994619, signifying improved randomness in the ciphertext. It outperforms standard AES, boosting encryption throughput from 301.85 KB/s to 389.16 KB/s and decryption from 263.79 KB/s to 311.30 KB/s. These improvements make the proposed encryption method more suitable for applications that require stronger security and fast processing.

VII. Conclusion

The analysis indicates that the suggested encryption method provides substantial enhancements compared to the conventional AES. The proposed encryption method promotes the diffusion property and minimizes the predictable nature between plaintext and ciphertext, resulting in increased security. It also generates more random ciphertext and delivers better performance in terms of processing time and throughput. The enhanced design provides a well-balanced increase in efficiency and security.

Future work will focus on implementing a more secure hybrid encryption technique within cloud-based blockchain systems to enhance data protection, prevent tampering, ensure secure key management, and strengthen user authentication.

References

[1] Z. A. Mohammed, H. Q. Gheni, Z. J. Hussein, and A. K. M. Al-Qurabat, "Advancing cloud image security via AES algorithm enhancement techniques," *Engineering Technology & Applied Science Research*, vol. 14, no. 1, pp. 12694–12701, Feb. 2024, doi: 10.48084/etasr.6601.

[2] S. Kumar and D. Kumar, "Securing of cloud storage data using hybrid AES-ECC cryptographic approach," Journal of Mobile Multimedia, Nov. 2022, doi: 10.13052/jmm1550-4646.1921.

[3] Z. A. Mohammed, H. Q. Gheni, Z. J. Hussein, and A. K. M. Al-Qurabat, "Advancing Cloud Image Security via AES Algorithm Enhancement Techniques," *Engineering Technology & Applied Science Research*, vol. 14, no. 1, pp. 12694–12701, Feb. 2024, doi: 10.48084/etasr.6601.

[4] M. Bedoui, H. Mestiri, B. Bouallegue, B. Hamdi, and M. Machhout, "An improvement of both security and reliability for AES implementations," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, pp. 9844–9851, Jan. 2022, doi: 10.1016/j.jksuci.2021.12.012.

[5] M. Y. Shakor, M. I. Khaleel, M. Safran, S. Alfarhood, and M. Zhu, "Dynamic AES Encryption and Blockchain Key Management: a novel solution for cloud data security," IEEE Access, vol. 12, pp. 26334–26343, Jan. 2024, doi: 10.1109/access.2024.3351119.

[6] N. Pronika and S. S. Tyagi, "Secure Data Storage in Cloud using Encryption Algorithm," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 136–141, Feb. 2021, doi: 10.1109/icicv50876.2021.9388388.

[7] S. Urooj, S. Lata, S. Ahmad, S. Mehfuz, and S. Kalathil, "Cryptographic Data Security for Reliable Wireless Sensor Network," Alexandria Engineering Journal, vol. 72, pp. 37–50, Apr. 2023, doi: 10.1016/j.aej.2023.03.061.

[8] V. D. Parthasarathy and K. Viswalingam, "Healthcare Data Security in Cloud Storage Using Lightweight Symmetric Key Algorithm," *The International Arab Journal of Information Technology*, vol. 21, no. 1, Jan. 2024, doi: 10.34028/iajit/21/1/5.

[9] J. S. Baladhay and E. M. De Los Reyes, "AES-128 reduced-round permutation by replacing the MixColumns function," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 3, p. 1641, Mar. 2024, doi: 10.11591/ijeecs.v33.i3.pp1641-1652.

[10] D. N. Hammod, "Modified Lightweight AES based on Replacement Table and Chaotic System," *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Ankara, Turkey, 2022, pp. 1-5, doi: 10.1109/HORA55278.2022.9799804.

[11] M. Sirajuddin and B. S. Kumar, "Modified Rijndael Algorithm for Resource-Constrained IoT-Based Wireless Sensor Networks," *ICST Transactions on Scalable Information Systems*, Aug. 2023, doi: 10.4108/eetsis.2748.

[12] F. V. Wenceslao Jr, "Enhancing the Performance of the Advanced Encryption Standard (AES) Algorithm Using Multiple Substitution Boxes," *International Journal of Communication Networks and Information Security (IJCNIS)*, vol. 10, no. 3, Apr. 2022, doi: 10.17762/ijcnis.v10i3.3589.

[13] B. Al-Shargabi and M. A. F. Al-Husainy, "Multi-round encryption for COVID-19 data using the DNA key," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 1, p. 478, Feb. 2022, doi: 10.11591/ijece.v12i1.pp478-488.

[14] H. M. Mohammad and A. A. Abdullah, "Enhancement process of AES: a lightweight cryptography algorithm-AES for constrained devices," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 20, no. 3, p. 551, Jun. 2022, doi: 10.12928/telkomnika.v20i3.23297.

[15] A. Altigani, S. Hasan, B. Barry, S. Naserelden, M. A. Elsadig and H. T. Elshoush, "A Polymorphic Advanced Encryption Standard – A Novel Approach," in *IEEE Access*, vol. 9, pp. 20191-20207, 2021, doi: 10.1109/ACCESS.2021.3051556.

[16] A. I. Salih, A. M. Alabaichi, and A. Y. Tuama, "Enhancing advance encryption standard security based on dual dynamic XOR table and MixColumns transformation," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 3, p. 1574, Sep. 2020, doi: 10.11591/ijeecs.v19.i3.pp1574-1581.

[17] H. T. Assafli and I. A. Hashim, "Generation and Evaluation of a New Time-Dependent Dynamic S-Box Algorithm for AES Block Cipher Cryptosystems," *IOP Conference Series Materials Science and Engineering*, vol. 978, p. 012042, Dec. 2020, doi: 10.1088/1757-899x/978/1/012042.

[18] O. C. Abikoye, A. D. Haruna, A. Abubakar, N. O. Akande, and E. O. Asani, "Modified Advanced Encryption Standard Algorithm for

Information Security," *Symmetry*, vol. 11, no. 12, p. 1484, Dec. 2019, doi: 10.3390/sym11121484.

[19] T. M. Kumar and P. Karthigaikumar, "FPGA implementation of an optimized key expansion module of AES algorithm for secure transmission of personal ECG signals," Design Automation for Embedded Systems, vol. 22, no. 1–2, pp. 13–24, Oct. 2017, doi: 10.1007/s10617-017-9189-5.

[20] M. Al-Mashhadani and M. Shujaa, "IoT Security Using AES Encryption Technology based ESP32 Platform," The International Arab Journal of Information Technology, vol. 19, no. 2, Jan. 2022, doi: 10.34028/iajit/19/2/8.

[21] M. Sawka and M. Niemiec, "A Sponge-Based Key Expansion Scheme for Modern Block Ciphers," *Energies*, vol. 15, no. 19, p. 6864, Sep. 2022, doi: 10.3390/en15196864.

[22] T. M. Kumar, K. R. Balmuri, A. Marchewka, P. B. Divakarachari, and S. Konda, "Implementation of Speed-Efficient Key-Scheduling Process of AES for Secure Storage and Transmission of Data," *Sensors*, vol. 21, no. 24, p. 8347, Dec. 2021, doi: 10.3390/s21248347.

[23] Y. S. Vaz, J. C. B. Mattos and R. I. Soares, "Improving an Ultra Lightweight AES for IoT Applications," *2023 IEEE 9th World Forum on Internet of Things (WF-IoT)*, Aveiro, Portugal, 2023, pp. 01-06, doi: 10.1109/WF-IoT58464.2023.10539597.

[24] A. J. Hintaw, S. Manickam, S. Karuppayah, M. A. Aladaileh, M. F. Aboalmaaly and S. U. A. Laghari, "A Robust Security Scheme Based on Enhanced Symmetric Algorithm for MQTT in the Internet of Things," in *IEEE Access*, vol. 11, pp. 43019-43040, 2023, doi: 10.1109/ACCESS.2023.3267718.

[25] S. Afzal, M. Yousaf, H. Afzal, N. Alharbe, and M. R. Mufti, "Cryptographic strength evaluation of key schedule algorithms," Security and Communication Networks, vol. 2020, pp. 1–9, May 2020, doi: 10.1155/2020/3189601.

[26] V. M. Silva-García, R. Flores-Carapia, and M. A. Cardona-López, "A Hybrid Cryptosystem Incorporating a New Algorithm for Improved Entropy," *Entropy*, vol. 26, no. 2, p. 154, Feb. 2024, doi: 10.3390/e26020154.

[27] G. Yi and Z. Cao, "An Algorithm of Image Encryption based on AES & Rossler Hyperchaotic Modeling," Mobile Networks and Applications, Sep. 2023, doi: 10.1007/s11036-023-02216-5.

[28] F. Thabit, S. Alhomdy, and S. Jagtap, "Security analysis and performance evaluation of a new lightweight cryptographic algorithm for cloud computing," Global Transitions Proceedings, vol. 2, no. 1, pp. 100–110, Jun. 2021, doi: 10.1016/j.gltp.2021.01.014.

[29] K. Assa-Agyei and F. Olajide, "A Comparative Study of Twofish, Blowfish, and Advanced Encryption Standard for Secured Data Transmission," International Journal of Advanced Computer Science and Applications, vol. 14, no. 3, pp. 393–398, 2023. doi: 10.14569/ijacsa.2023.0140344

An Enhanced LSTM Model Based on Feature Attention Mechanism and Emotional Intelligence for Advanced Sentiment Analysis

Muhammad Naeem Aftab[1], Dost Muhammad Khan[2], Muhammad Zulqarnain[3], Muhammad Rizwan Akram[4]
Department of Information Technology, The Islamia University of Bahawalpur, Pakistan[1, 2]
Department of Computer Science and IT, Cholistan University of Veterinary and Animal Sciences, Bahawalpur, Punjab, Pakistan[3]
Punjab Irrigation Department, Punjab, Pakistan[4]

Abstract*—Sentiment analysis, a crucial yet complex task in natural language processing (NLP), is extensively employed to identify sentiment polarity within user-generated content. Traditional deep learning methods for textual sentiment analysis often overlook the influence of emotional modulation on extracting sentiment features. At the same time, their attention mechanisms primarily operate at the word or sentence levels. Such oversight of higher-level abstractions may hinder the learning of nuanced sentiment patterns, ultimately damaging the accuracy of sentiment analysis. Addressing these gaps, this study proposed a novel framework, the Two-State Enhanced LSTM (TS-ELSTM), which integrated Emotional Intelligence (EI) and a Feature Attention Mechanism (FAM) to enhance the identification of relevant features during selection. Furthermore, this study employed a dual-phase training strategy of LSTM to accelerate learning and minimize information loss. A dynamic topic-level attention mechanism is also introduced to optimize hidden text representation weights. By integrating EI with a topic-level attention mechanism, the proposed framework efficiently extracts valuable features and enhances the feature learning ability of the conventional LSTM model. This novelty attains emotion-aware learning through two key components: an emotion modulator and an emotion estimator, which successfully normalize the system's learning dynamics by combining emotional context. The experimental outcomes demonstrated that the proposed approach achieved an accuracy of 84.20%, 94.12% using MR and IMDB, respectively. The proposed approach significantly improves sentiment analysis accuracy, outperforming traditional deep learning models by a notable margin.

Keywords*—Sentiment analysis; emotional intelligence; attention mechanism; two-state LSTM; long-term dependencies

I. INTRODUCTION

The exponential growth of social media platforms has raised sentiment analysis and opinion mining, and become critical areas in Natural Language Processing (NLP) applications [1-3]. Traditional sentiment analysis algorithms often produce a single polarity score for a complete text, which is insufficient when analyzing complicated content. For sentiment analysis illustration, a user review might praise the robustness of a product, whereas criticize the customer support, a scenario where existing approaches collapse such nuanced sentiments into a remarkable polarity score, failing to extract context-specific disparities. This generalization can result in misunderstandings and misinterpretations of user experiences in real-world applications [4]. Based on the recent literature review, deep learning (DL) approaches, mainly recurrent neural networks (RNNs), has dominated advance developments in various NLP tasks such as spam filtering, and topic classification [5, 6], question answering [7] keywords extraction [8] and anomaly detection [9], however, these conventional methods face challenges such as vanishing gradient decay and difficulty to retain long-range dependencies in sequences [10]. To address these issues, two advanced variants of the RNNs were introduced [11], known as LSTM and GRU, which retain the long-term dependencies. Modern implementations integrated attention mechanisms with hybrid architectures that deal with sequence information like CNNs [12] and RNNs [13], improving both predictive accuracy and interpretability. In sentiment analysis (SA), attention allows models to assign higher weights to sentiment-laden terms while suppressing irrelevant content [14], overcoming the inefficiencies of processing entire sentences sequentially [15]. Conversely, attention mechanism enables targeted concentration on specific parts of the input text during sequential analysis, allowing adaptive weighting of contextually critical components in real-time.

Cognitive research [16, 17] demonstrates the link between emotion-associated and human decision-making: positive emotions (e.g., enthusiasm) foster creative problem-solving, while negative emotions (e.g., anxiety) impair cognitive performance. Despite this, traditional LSTM-based approaches neglect the influence of emotion on sentiment feature extraction, potentially undermining their analytical efficacy. While sentiment and emotion are intertwined yet distinct psychological constructs, existing text sentiment analysis frameworks often conflate these terms. Sentiment refers to a persistent cognitive stance, such as opinions reflected in social media posts or brief videos, whereas emotions are transient psychological reactions that influence behavior and modulate neural activity (e.g., enhancing or suppressing neural responses to positive or negative stimuli). These dynamics can refine the adaptability of neural networks such as LSTM by imitating feedback-driven learning processes. This study focuses on detecting latent sentiments in social media content, though the proposed TS-ELSTM framework is designed for broader applicability in machine learning tasks, a direction reserved for future exploration. Crucially, this work rigorously

differentiates between sentiment (as a sustained attitude) and emotion (as an ephemeral state) to ensure conceptual clarity.

In conclusion, the developed TS-ELSTM model offers unique advantages compared to traditional approaches in two key ways: firstly, unlike most traditional neural network models that overlook the role of emotional modulation, TS-ELSTM integrates emotion-based modulation to enhance the learning capacity of neural units by mimicking human emotional processes. Secondly, the existing attention mechanisms usually operate at lower abstraction levels, such as words or phrases. The topic-level attention mechanism in TS-ELSTM extracts higher-level cognitive concentration similar to that of the human brain, resulting in more refined sentiment representations of textual data. The primary contributions of this study are outlined as follows.

1) ELSTM Architecture. We present ELSTM, a novel LSTM variant inspired by emotional intelligence theory. By leveraging the synergistic interplay between emotion and cognition, ELSTM enhances the model's inference capabilities. To our knowledge, this is the first time a human-inspired emotional mechanism has been embedded to strengthen LSTM learning abilities.

2) Dynamic topic-level attention. This study introduces an innovative topic-level attention mechanism that dynamically adjusts the weights of hidden representations for text subsequences, enabling more focused feature extraction.

3) TS-LSTM framework. We design Two-State LSTM (TS-LSTM), a deep network for sentiment analysis that incorporates the ELSTM units with the developed topic-level attention, enhancing text sentiment detection.

4) Empirical validation. Through comprehensive experiments on standard sentiment analysis benchmarks, experimental findings demonstrate that our developed approach performs better than traditional approaches.

This study is organized as follows: Section II describes the details of related work employing deep learning approaches for sentiment analysis. The proposed methodology and its flow diagrams, as well as various sentiment analysis phases, are briefly illustrated in Section III. Section IV refers experimental setup, along with dataset description and evaluation metrics, while Section V presents the results and discussion. Finally, Section VI provides a conclusion and future direction.

II. Related Work

This section investigates the traditional research work on sentiment analysis that influences LSTM architectures and attention mechanisms. Recent developments in RNNs illustrate that integrating LSTMs with their efficient gating for sequence modeling and attention layers are particularly effective for NLP tasks, while attention mechanisms dynamically prioritize salient features in sequences, making their integration a powerful approach for sentiment analysis [18]. Deep learning has already been adopted across NLP tasks such as topic modeling, text categorization, question answering, speech recognition, spam filtering, keyword extraction, machine translation, and information retrieval [19]. The implementation of deep learning in sentiment classification has grown exponentially, with attention-based methods driving innovation. For instance, one research study exhibited that embedding sentiment lexicon features alongside an attention mechanism boosts word representations for overall sentiment classification tasks without specific targets.

Furthermore, Toma and Choi [20] introduced two-directional LSTMs that capture bidirectional context more effectively than a vanilla LSTM. Zulqarnain et al. [21] introduced a bistatic GRU-encoder framework for sentiment analysis, outperforming standalone GRU and LSTM models in accuracy. Jang et al. [22] further advanced the field by combining bidirectional GRU, Word2vec embeddings, and attention mechanisms, achieving state-of-the-art results in sentiment analysis. Building on these ideas, BiLSTM+Attention architectures emphasize their capacity for sentiment-bearing features, further improving classification accuracy. Aspect-level sentiment analysis has progressively relied on attention-based neural networks to model relationships between context and specific terms. Wang et al. [23] referred ATAE-LSTM architecture, integrating LSTM with attention to evaluate contextual relevance for aspect terms. Zhao et al. [24] and Zulqarnain et al. [25] employed a hierarchical attention mechanism to pinpoint sentiment-critical information and achieve a good result, while Yang et al. [26] accomplished a multi-grained fusion network and multigranularity attention to capture word-level interactions and obtained superior performance against state-of-the-art methods. Moreover, Tan et al. [27] designed a dual attention network to detect conflicting opinions, while Ouyang et al. [28] presented Interactive Attention Networks (IAN) for bidirectional context-aspect learning. Graph Convolutional Networks (GCNs) have also emerged as pivotal tools in modern sentiment analysis frameworks.

Emotional intelligence (EI) research highlights emotions' influence on decision-making. Khashman [29] pioneered DuoNeural, an architecture unifying emotion and cognition in intelligent systems. Lotfi and Akbarzadeh [30] developed LiAENN, a neural network modeling emotional states like anxiety during learning. Hassan et al. [31] highlighted an emotional multi-agent reinforcement learning framework to foster cooperative behaviors, while Oyedotun and Khashman [32] unified adaptive and prototype learning theories in emotional neural networks. Markadeh et al. [33] utilized brain-emotional learning to control induction motors. Building on these insights, the proposed TS-ELSTM integrates emotional reasoning into intelligent systems, reinforcing the view that affective dynamics are essential to intelligent reasoning.

Conclusion remarks in terms of limitations of the existing literature review are as follows: traditional deep learning approaches excel in either contextual modeling (BiLSTM+Attention [20, 22-28]) *or* emotional reasoning (DuoNeural/LiAENN [29-33]), but they fail to synergize these capabilities. Aspect-level deep learning architectures (ATAE-LSTM [23], hierarchical attention [24-25]) capture contextual nuances but ignore affective dynamics critical for sentiment interpretation. Conversely, emotional neural networks [29-33] lack linguistic grounding for NLP tasks. Furthermore, existing models neglect the sentiment-emotion discrepancy and treat

sentiment and emotion as isomorphic (e.g., lexicon-enhanced attention [18]).

III. Proposed Methodology

This study outlines the methodology of the proposed framework, which integrates emotional intelligence and a two-state LSTM strategy with an adaptive attention mechanism. The key contributions lie in its structured workflow, comprising in several phases: 1) systematic data acquisition and preprocessing, 2) feature embedding via pre-trained technique (GloVe), 3) sequential modeling using recurrent architectures (V-RNN, LSTM, and GRU), 4) integration of an attention layer and 5) performance assessment. Each component is comprehensively explored in depth in the following sections.

A. Embedding Layer

In the sentiment classification process, the embedding layer serves an important function by transforming raw text into structured numerical inputs. During sentence analysis, the initial preprocessing phase involves sanitizing textual data to remove noise. Neural networks, constructed on linear algebra operations and nonlinear activation functions, necessitate textual inputs to be translated into dense numerical vectors within a low-dimensional space [34]. To attain this, we utilize a pre-trained 300-dimensional Global Vectors (GloVe) technique [35] in the word embedding layer, which maps each word in a sentence to a continuous vector representation. GloVe influences global statistical patterns during the embedding stage and efficiently captures semantic and syntactic co-occurrence relationships across corpora, combining the benefits of matrix factorization and context window methods. This method improves its competence in managing lexical ambiguities, such as synonyms and polysemous terms. In contrast, network-dependent embeddings, the GloVe method arises word vectors directly from corpus statistics, bypassing iterative training procedures and allowing effective parallelization for scalable implementation.

B. Vanilla Recurrent Neural Network (V-RNN)

The Vanilla RNN is the most basic variant of the RNN model. The V-RNN includes a single recurrent layer that manages both current input information and the previous hidden state to produce an output while simultaneously updating its hidden state for subsequent steps. The term "vanilla" replicates its simplicity, as its deficiencies are advanced structural elements like gating mechanisms, which are frequently used to handle long-term dependencies or gradient issues [35, 36].

C. LSTM

The LSTM model is an advanced variant of standard RNNs, initially introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 [37]. LSTMs are well-known for addressing long-term dependencies in sequential data. LSTMs are frequently integrated with architectural designs like inception modules to expand their ability and examine temporal patterns in time-series datasets. Unlike Vanilla RNNs, LSTM uses particular internal memory cells and gated mechanisms to moderate information loss during sequential data processing.

These networks manage the information flow across sequences utilizing gate mechanisms including input, forget, output, and a special internal memory cell, which cooperatively stimulate the hidden layer's output at each timestep [38]. By dynamically regulating data holding and propagation, LSTMs demonstrate better performance in the NLP tasks requiring nuanced temporal understanding.

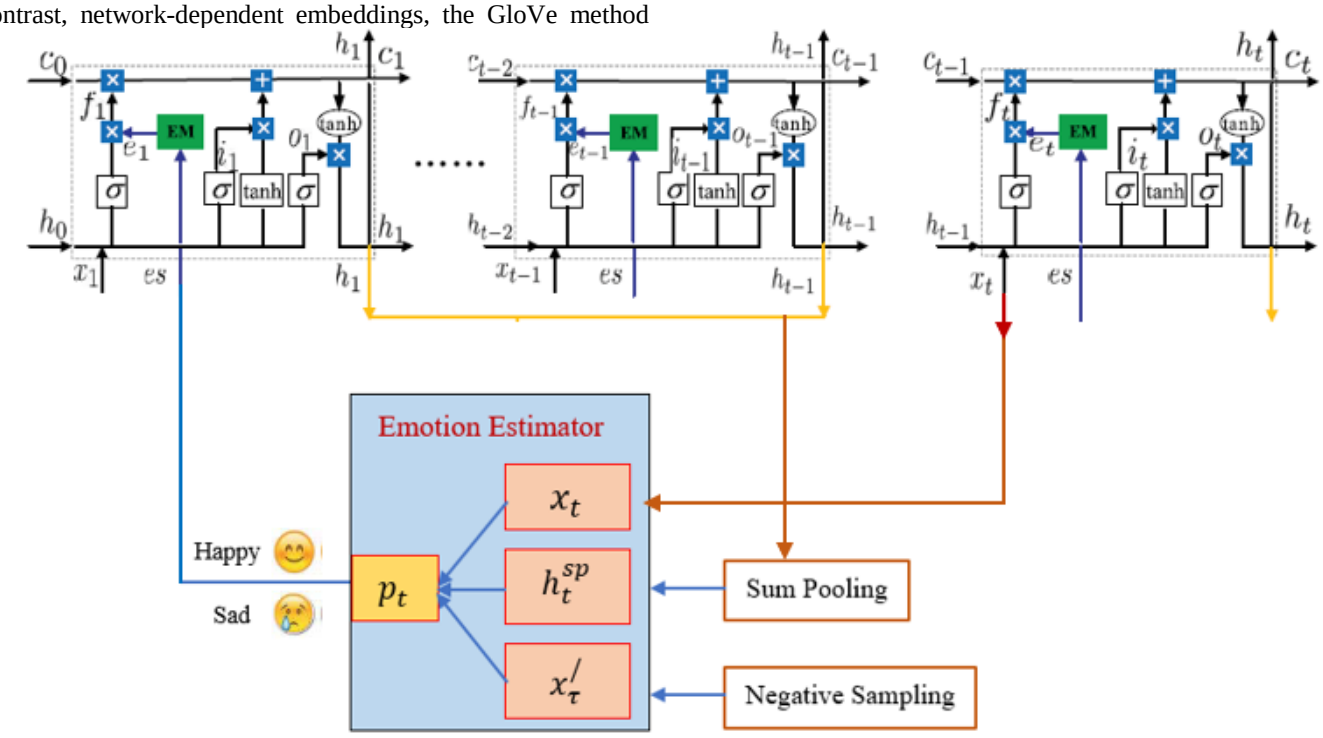


Fig. 1. Developed emotional LSTM.

D. Emotion-Enhanced LSTM

Conventional RNNs offer an effective approach for modeling sequential text data by leveraging contextual relationships between adjacent words. However, conventional RNNs suffer from gradient instability (vanishing or exploding gradients) and struggle to extract long-range semantic or sentiment dependencies. To address these drawbacks, Hochreiter and Schmidhuber [39], proposed the LSTM network, which incorporates gated memory cells and has demonstrated superior sequence-representation capabilities, significantly improving performance in tasks requiring temporal understanding. LSTM block mirrors certain features of the brain's memory and language-processing systems. Cognitive neuroscience research [40] highlights emotion's role in memory processes, enhancing information encoding, consolidation, and retrieval. For instance, a person's emotional state during perception can improve the encoding of information into both short-and long-term memory. As observed in previous work, it confirms that embedding emotional cues can lead to higher-quality decision-making in intelligent systems. Yet, traditional LSTMs update their memory cells solely based on input and past hidden states, ignoring any emotional modulation.

To bridge this gap, we propose Emotion-enhanced LSTM (ELSTM), a novel architecture inspired by EI principles. As illustrated in Fig. 1, ELSTM introduces three gates (input, forget, output), three states (hidden, cell, emotion), an emotion-based modulator (EM), and an emotion estimator. The network's output at each timestep feeds back into both the gate inputs and the emotion estimator. Formally, ELSTM core equations are as follows [Eq. (1) to Eq. (7)]:

$$i_t = \sigma(W_{xi}x_t + U_{hi}h_{t-1} + b_i) \quad (1)$$

$$e_t = EM(es) \quad (2)$$

$$f_t = e_t * \sigma(W_{xf}x_t + U_{hf}h_{t-1} + b_f) \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + U_{ho}h_{t-1} + b_o) \quad (4)$$

$$\hat{a}_t = tanh(W_{x\hat{a}}x_t + U_{h\hat{a}}h_{t-1} + b_{\hat{a}}) \quad (5)$$

$$c_t = f_t * x_{t-1} + (i_t * \hat{a}_t) \quad (6)$$

$$h_t = O_t * tanh\,(c_t) \quad (7)$$

where, σ is the sigmoid activation, * denotes element-wise multiplication, and W, U, and b are trainable parameters. The emotion modulator produces a signal e_t from the emotional state e_s (as discussed in Section III-G), which is integrated into the forget gate f_t, balancing the retention of past cell state c_{t-1} against new input x_t.

ELSTM operates in four sequential phases:

1) Forgetting. The forget gate uses both the previous hidden state h_{t-1}, the current input x_t, and the emotion signal t_e to determine what to retain in c_{t-1}.

2) Emotion-estimation. An internal estimator infers the new emotional state e_s based on past outputs (see Section III.H).

3) Updating. During the updating process, the input gate i_t and the candidate state $\hat{a}_t$ manage which new information arrives in the cell, completing the update in Eq. (6).

4) Output. Finally, the output gate o_t filters the updated cell state over a *tanh* function to generate the new hidden state h_t, as in Eq. (7).

In this study, the ELSTM architecture illustrates three important novelties to enhance sequence modeling. First, its Emotion-Aware Forgetting mechanism reinterprets the forget gate of the traditional LSTM by incorporating emotion-derived signals, consenting to dynamic modifications to the retention of preceding context h_{t-1}, and the modification of new inputs x_t. This mechanism confirms a contextually balanced trade-off between memory preservation and adaptability. Second, its Brain-Inspired-Design extends standard LSTM by embedding emotion modulation, adjusting with neuroscientific insights on emotion's role in memory. Finally, this study describes the Dynamic Adaptation of emotional intelligence (EI) to prioritize contextually salient information during gating operations, improving decision-making over real-time relevance weighting.

By retaining the main memory-and-gating mechanisms of the conventional LSTM and expanding them with emotion-driven modulation, ELSTM efficiently balances information retention and forgetting, yielding a richer, more brain-inspired sequence model.

E. Two-State LSTM (TS-LSTM)

LSTM is an enhanced variant of conventional RNNs used for sequential modeling. At each time step t, they compute a hidden state $h_t \in \mathbb{R}^n$ illustrated in Eq. (8):

$$h_t = f(W_{xt} + Uh_{t-1} + b) \quad (8)$$

where, $W \in R^{m\times n}$, $b \in R^{m\times m}$, $b \in R^m$, refers to weight matrices and biases, respectively, and f represents a nonlinear activation [30]. Although LSTMs struggle with gradient instability (vanishing or exploding gradients) and inherently prioritize forward linguistic context, they neglect backward dependencies critical for tasks like sentiment analysis. Yet natural language understanding often benefits from both preceding and succeeding context.

To address this, we introduce the Two-State LSTM (TS-LSTM), which processes inputs in two passes: one from left to right (forward or positive) and one from right to left (backward or negative) inspired by bidirectional RNNs, as illustrated in Fig. 2. During training, these pathways independently analyze temporal sequences and merge their outputs to capture comprehensive contextual relationships.

1) Forward Pass. The forward LSTM computes hidden states $\overrightarrow{C_t}$ using the input gate $\overrightarrow{\imath_t}$, forget gate $\overrightarrow{f_t}$, and candidate state $\overrightarrow{\hat{a}_t}$ gates are defined as [Eq. (9) to Eq. (15)]:

$$\overrightarrow{\imath_t} = \sigma(\overrightarrow{W_{xt}} \times [\overrightarrow{C_{t-1}}, \overrightarrow{h_{t-1}}, \overrightarrow{x_t} + \overrightarrow{b_t}) \quad (9)$$

$$\overrightarrow{e_t} = \overrightarrow{EM(es)} \quad (10)$$

$$\overrightarrow{f_t} = e_t * \sigma(\overrightarrow{W_{xf}} \times [\overrightarrow{C_{t-1}}, \overrightarrow{h_{t-1}}, \overrightarrow{x_t} + \overrightarrow{b_f} \quad (11)$$

$$\overrightarrow{o_t} = \sigma(\overrightarrow{W_{xo}} \times [\overrightarrow{C_{t-1}}, \overrightarrow{h_{t-1}}, \overrightarrow{x_t} + \overrightarrow{b_o}) \quad (12)$$

$$\overrightarrow{\hat{a}_t} = tanh(\overrightarrow{W_{x\hat{a}}} * [\overrightarrow{h_{t-1}}, \overrightarrow{x_t} + \overrightarrow{b_{\hat{a}}}) \quad (13)$$

$$\overrightarrow{C_t} = \overrightarrow{f_t} * (\overrightarrow{C_{t-1}} + \overrightarrow{i_t} * \overrightarrow{\hat{a}_t}) \quad (14)$$

$$\overrightarrow{h_t} = \overrightarrow{o_t} * \overrightarrow{tanh}(\overrightarrow{C_t}) \quad (15)$$

2) Backward pass. Proportionally, the backward LSTM processes sequences in reverse, with analogous [Eq. (16) to Eq. (22)] for $\overleftarrow{i_t}$, $\overleftarrow{f_t}$, $\overleftarrow{C_t}$,$\overleftarrow{o_t}$ and $\overleftarrow{\hat{a}_t}$

$$\overleftarrow{i_t} = \sigma(\overleftarrow{W_{xi}} \times [\overleftarrow{C_{t-1}}, \overleftarrow{h_{t-1}}, \overleftarrow{x_t} + \overleftarrow{b_i}) \quad (16)$$

$$\overleftarrow{e_t} = \overleftarrow{EM(es)} \quad (17)$$

$$\overleftarrow{f_t} = e_t * \sigma(\overleftarrow{W_{xf}} \times [\overleftarrow{C_{t-1}}, \overleftarrow{h_{t-1}}, \overleftarrow{x_t} + \overleftarrow{b_f} \quad (18)$$

$$\overleftarrow{o_t} = \sigma(\overleftarrow{W_{xo}} \times [\overleftarrow{C_{t-1}}, \overleftarrow{h_{t-1}}, \overleftarrow{x_t} + \overleftarrow{b_o}) \quad (19)$$

$$\overleftarrow{\hat{a}_t} = tanh(\overleftarrow{W_{x\hat{a}}} * [\overleftarrow{h_{t-1}}, \overleftarrow{x_t} + \overleftarrow{b_{\hat{a}}}) \quad (20)$$

$$\overleftarrow{C_t} = \overleftarrow{f_t} * [\overleftarrow{C_{t-1}}, \overleftarrow{i_t} * \overleftarrow{\hat{a}_t}) \quad (21)$$

$$\overleftarrow{h_t} = \overleftarrow{o_t} * \overleftarrow{tanh}(\overleftarrow{C_t}) \quad (22)$$

For an input sequence $(x_1, x_2, \ldots, x_n)$, each token at position t is first encoded by the forward LSTM into $\overrightarrow{C_t}$ (left-to-right context) and by the backward LSTM into $\overleftarrow{C_t}$ (right-to-left context). These two vectors are then concatenated as referred to in Eq. (23):

$$C_t = \overrightarrow{C_t}, \overleftarrow{C_t} \quad (23)$$

This dual-direction formulation allows TS-LSTM to capture comprehensive contextual information from both surrounding directions.

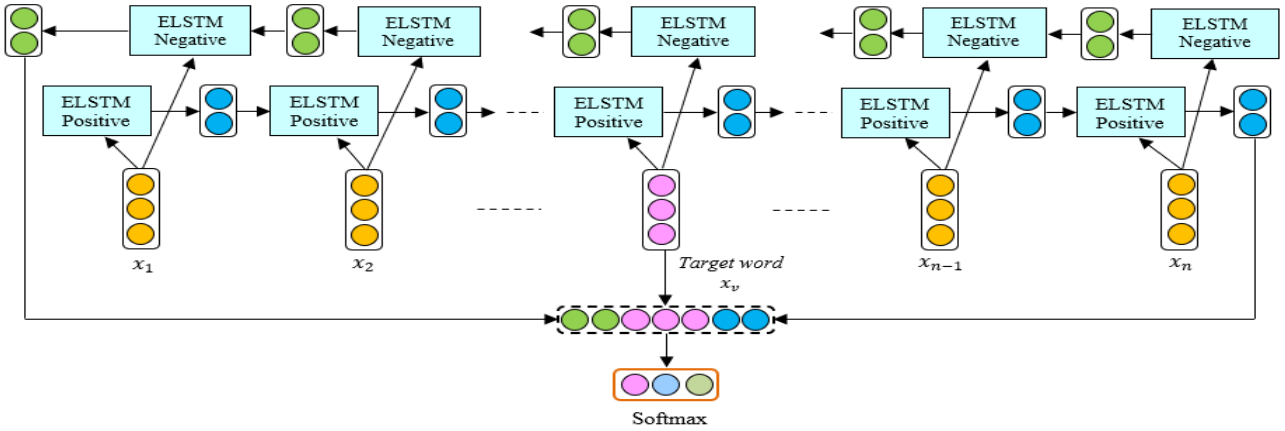


Fig. 2. Two-state strategy.

F. Attention Layer

The attention mechanism assists in highlighting the most relevant parts of the input data while minimizing the influence of less important information. By assigning differentiated weight values to features based on their emotional relevance, this mechanism permits the model to efficiently filter out non-essential textual content. This targeted concentration facilitates more accurate learning of emotionally substantial textual features, thereby improving the efficacy of implicit sentiment analysis tasks.

In most textual data, words that convey emotional tone are particularly impactful in determining the complete sentiment. In this research, we incorporate an attention mechanism that emphasizes emotionally charged features within sentences expressing implicit sentiment. Specifically, this study applies attention to both local and global feature matrices represented as $R = [r_1, r_2, r_3 \ldots r_n,]$ and $H = [h_1, h_2, h_3 \ldots h_n,]$ respectively, to enhance sentiment polarity.

The attention scores are calculated over the following transformations as Eq. (26) and (27):

$$Local\ features = u_{ri} = \tanh(W_r r_i + b_r) \quad (26)$$

$$Blobal\ features = u_{hj} = \tanh(W_r r_j + b_h) \quad (27)$$

where, W_r and W_h are the learnable weight matrices, b_r and h_r are biased terms, and $tanh$ is the activation function.

The attention weights a_i and a_j for local and global features are calculated using the softmax function [Eq. (28) and (29)]:

$$a_i = \frac{\exp(u_{ri})}{\sum_{i=1}^{m} \exp(u_{ri})} \quad (28)$$

$$a_j = \frac{\exp(u_{rj})}{\sum_{j=1}^{m} \exp(u_{rj})} \quad (29)$$

Using these attention weights, the local and global feature representations are computed as weighted sums [Eq. (24) and (25)]:

$$s_i = \sum_{i=1}^{m} a_i r_i \quad (24)$$

$$s_h = \sum_{j=1}^{m} a_j r_j \quad (25)$$

This dual-channel attention mechanism dynamically assigns importance to emotionally relevant word vectors, significantly improving the effectiveness of the developed TS-ELSTM approach. As a result, the network demonstrates better precision in emotion detection and analysis within the human-computer interactive dialogue environment. The overall architecture of the proposed model is demonstrated in Fig. 3.

The output layer of the TS-ELSTM model combines both local and global feature vectors by concatenation [Eq. (30)]:

$$s = [s_r, s_h] \tag{30}$$

To reduce the risk of overfitting, dropout is applied before passing the concatenated vector through a fully connected neural network. The final output is computed using the softmax function as in Eq. (31) and (32):

$$o = f[W_s s + b] \tag{31}$$

$$g = softmax(0) \tag{32}$$

where, o refers to the output from the fully connected layer, W is the weight matrix, b is the bias term, and g is the resulting output vector utilized for classification.

G. Emotion-Based Estimator

Emotional responses arise from fast neural signaling activated by external stimuli, mirroring the brain's capability to process information quickly. Inspired by this fast and efficient emotional reaction, we developed a novel emotion estimator integrated with negative sampling that influences similarity-based cognition. From a linguistic perspective, within a given text sequence $(x_1, x_2, \dots, x_{t-1}, x_t, \dots.)$, a subsequence $(x_{t-\tau}, x_{t-\tau+1}, \dots, x_{t-1})$ often holds strong semantic relevance to the current word x_t.

To capture this association, we compute the semantic similarity over the inner product between the vector representation of the current word x_t and the sum-pooled representation of the preceding subsequence $h_t^{sp} = SumPooling(x_{t-\tau}, x_{t-\tau+1}, \dots, h_{t-1})$. This system permits us to estimate the likelihood that x_t follows the subsequence using the softmax function as in Eq. (33):

$$P_t = \frac{\exp(h_t^T h_t^{sp})}{\sum_{i=1}^{V} \exp(h_i^T h_t^{sp})} \tag{33}$$

where, V represents the total vocabulary size. However, computing this across all vocabulary items is computationally expensive and contradicts the efficiency expected from emotion-driven systems. To overcome this, negative sampling is employed [41], significantly minimizing computation cost by selecting a limited number of negative examples from a background distribution. With this technique, the equation becomes [Eq. (34)]:

$$P_t = \frac{\exp(h_t^T h_t^{sp})}{\exp(h_i^T h_t^{sp}) + \sum_{i=1}^{NS} \exp(h_i^T h_t^{sp})} \tag{34}$$

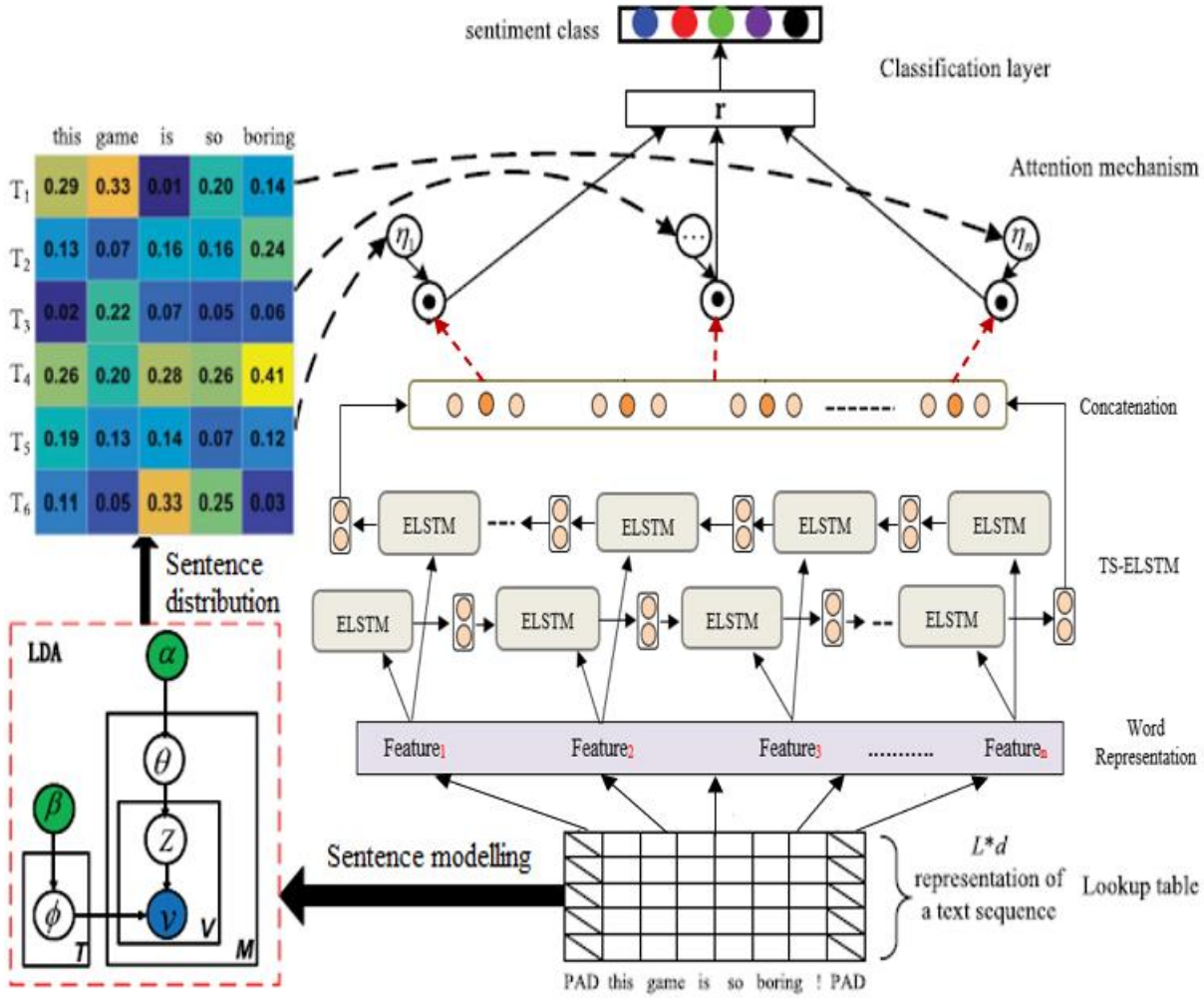


Fig. 3. Overall proposed TS-ELSTM architecture.

where, NS denotes the number of sampled negative instances, which is considerably smaller than V.

Rule 1: (Emotion Classification Rule): If $P_t > ET$ (a predefined threshold), the emotion estimator classifies the state as "Happy"; otherwise, it outputs "Sad".

H. Emotion-Based Modulation

The key role of the emotion-based modulator is to convert the emotional state es detected by the system into the corresponding modulation behavior. This conversion follows a set of predefined mappings, detailed in Rule 2 and Rule 3.

Rule 2: (Enhance Emotion Modulation): This rule is activated when the ELSTM (Emotion-augmented LSTM) experiences a "Happy" emotional state, typically triggered by successfully learning the semantic relationship between a subsequence $(x_{t-\tau}, x_{t-\tau+1}, \dots, x_{t-1})$, and its next token x_t. Success indicates effective semantic comprehension, which reinforces TS-ELSTM's tendency to retain learned information by increasing its modulation factor, favoring memory over forgetting.

Rule 3 (Reduce Emotion Modulation): Conversely, failure to accurately learn the sequence embedding implies a poor understanding, causing a "Sad" state. In response to the current state, the proposed TS-ELSTM decreases its modulation factor to promote forgetting of low-quality or misleading information.

In this mechanism, the emotional coefficient p plays a key role. The modulation factor is calculated based on Eq. (35):

$$e_{t+1} = \begin{cases} e_t + \rho * \left|h'_t - h_t^{sp}\right| if\ es = H \\ e_t - \rho * \left|h'_t - h_t^{sp}\right| if\ es = S \end{cases} \quad (35)$$

where, $|.|$is the absolute value operator, $\acute{h}_t$ is the output vector for the current input x_t, while h_t^{sp} sum-pooled depiction of preceding outputs, and $H\ and\ S$ illustrate the emotional states "Happy" and "Sad", respectively.

To dynamically adjust p during training, a self-adaptive strategy is utilized as in Eq. (36):

$$p_{ep} = (p_{max} - p_{min}) * \frac{EPOCH - ep}{EPOCH} + p_{min} \quad (36)$$

where, p_{min} and p_{max} define the constraints of the inertia coefficient, while $EPOCH$ is the total number of training epochs, and ep is the current epoch.

This method confirms that p_{ep} starts its maximum value when training begins $(sp = 0)$ and progressively reduces toward p_{min} as training developments. The rationale aligns with deep learning algorithms: initially, a high inertia coefficient accelerates learning, while in later stages, a reduced value ensures constancy of the system converges.

IV. Experimental Setup

Several experiments were carried out to assess the effectiveness of the developed TS-ELSTM model in comparison to existing approaches. The simulations of the proposed study were performed on a system powered by an Intel Core i7-3770 CPU (3.40 GHz) and 32 GB RAM, running Windows 10. Python 3.13 and Anaconda utilized the computational environment for data preprocessing and analysis, supported by TensorFlow 2.4.1 and Keras 2.3 as deep learning frameworks. The Natural Language Toolkit (NLTK) was employed for word embedding tasks. Exploiting Keras's compatibility with CPU architectures, the setup optimized hardware resources to accelerate network training and inference. The detailed optimal hyperparameter settings utilized to train the developed model are illustrated in Table I.

TABLE I. Optimal Hyperparameter of the Developed Model

Parameter Name	Parameter Value
Optimizer	Adam
Word vector dimension	300
Batch size	64
Learning Rate	0.001
Attention Mechanism	350 units
Dropout	0.5
Activation function	ReLU/tanh
Loss Function	Cross-entropy
Learning rate decay	0.01
Network epochs	30

A. Dataset Description

The experiments were carried out using three publicly available datasets, all centered around sentiment analysis of movie-related opinions. The selected datasets are: Stanford Internet Movie Database (IMDB), and Movie Review (MR). These datasets are designed for binary sentiment classification to predict whether a review expresses a positive or negative sentiment. Key statistical details of the datasets, including size and class distribution, are summarized in Table II.

TABLE II. Statistical Details of the Dataset

Detail	MR	IMDB
Train	8654	37,500
Validation	CV	CV
Test	1062	12,500
Max.length	52	1825
Ave.length	18.6	180.4
V-size	18,990	88,712
D-size	10,660	50,000
No. Of Classes	2	2

B. Evaluation Metrics

To assess the effectiveness of the proposed method in sentiment analysis tasks, various evaluation metrics were employed. Commonly used metrics include accuracy, precision, recall (also known as sensitivity), and the F1-score. Each metric provides insights into different aspects of the model's performance as in Eq. (37) to (40):

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (37)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (38)$$

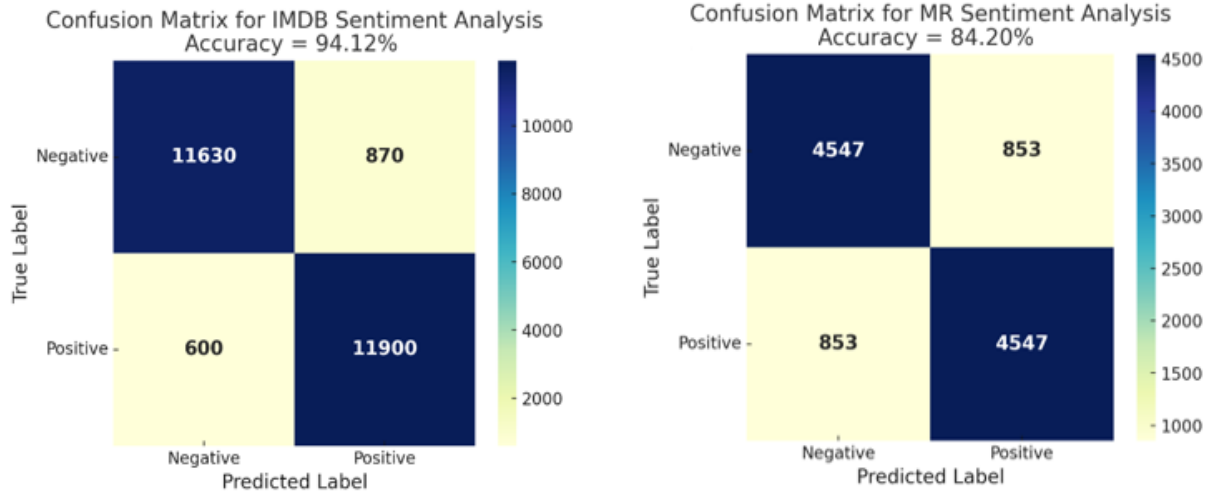
$$Sensitivity = Recall = \frac{TP}{(TP + FN)} \quad (39)$$

$$\text{F1} - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (40)$$

In these equations, TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively.

V. Results and Discussion

Our experimental analysis summarizes the performance outcomes of the TS-ELSTM approach, benchmarked against existing deep learning methods using two text classification datasets. The assessment, employing multiple evaluation criteria, demonstrates consistently superior accuracy across all benchmarked datasets. Fig. 4 shows the confusion matric performance of the proposed model using both datasets, while Table III describes the overall performance of the proposed model.



(a) (b)

Fig. 4. Confusion matrix of the proposed model (a) IMDB, (b) MR.

TABLE III. Proposed Model Performance

	Evaluation Parameters			Average Values		
Datasets	Precision (%)	Recall (%)	F1 Score (%)	Test size	Accuracy (%)	F1 score (%)
MR	83.56	84.14	83.85	CV	**84.20**	**83.970**
	84.73	83.47	84.09	CV		
IMDB	94.38	93.69	94.07	CV	**94.12**	**93.965**
	93.90	94.04	93.86	CV		

A. First Dataset (MR) Results

The performance of the developed TS-ELSTM model on the MR dataset is illustrated in Table IV. The developed model obtained precision scores of 83.56% for positive instances and 84.73% for negative instances, with corresponding recall rates of 84.14% and 83.47%. These metrics culminated in an F1-score of 83.970% and an overall accuracy of 84.20%. Comparative analysis against existing studies underscores the model's superiority. For instance, Chen et al. [42] accomplished a BiLSTM with conditional random fields, attaining 82.30% accuracy, while Fu et al. [43] developed ALE-LSTM, a lexicon-enhanced LSTM with attention strategy, which achieved 79.90% accuracy on the MR dataset. The developed TS-ELSTM outperformed both these approaches, as well as Zhang et al.'s [44] proposed hybrid BiGRU+CNN, and reported 78.30% accuracy, our model highlighted 6.08% improvement over the existing approaches. Similarly, Qian et al. [45] introduced deep learning integrated models with sentence-level classification, and achieved an accuracy of 82.30%. Moreover, Usama et al. [46] reported accuracies of 79.80% and 80.20% using a multi-level feature extraction model combining GRU, LSTM, and CNN, results surpassed by the proposed framework. Furthermore, Mohammed et al. [47] proposed a recurrent attention GRU with a self-attention mechanism using the MR dataset, and reported an accuracy of 82.38%.

The proposed TS-ELSTM also excelled on the MR dataset, exceeding the performance of prior studies [42–49] despite their more complex architectures, detailed comparison is represented in Table IV.

TABLE IV. COMPLEXITY AND ACCURACY-BASED ANALYSIS WITH TRADITIONAL STUDIES ON THE MR DATASET

Methods	Model Complexity	Accuracy (%)
BiGRU+CNN [44]	Sequential combination of BiGRU and CNN framework	78.30
MV+RNN [48]	The joint strategy of analyzing trees and the RNN model	79.00
WALE+LSTM [43]	Implementation of WALE and attention mechanism jointly with LSTM	79.90
Self-Att-LSTM+word2vec [49]	Implemented Self-attention mechanism with LSTM based on word2vector	79.93
Hierarchical-Att-GRU+Dropout [49]	Investigated Hierarchical-attention mechanism with GRU based on word2vector	79.97
CNN-GRU-multilevel and multitype fusion [46]	Proposed combined methodology in terms of multilevel and multi-type features based on CNN and GRU models	80.20
SA-RA+GRU [47]	Proposed self-attention with recurrent attention GRU (SA+RA-GRU)	82.38
BiLSTM-CRF [42]	Fusion of bidirectional LSTM with CRT combined with CNN	82.30
LR+BiLSTM [45]	Illustrated sentence-level representation based on LR, and combined with BiLSTM	82.10
TS-ELSTM (Proposed)	**Two-State Enhanced LSTM (TS-ELSTM) + Emotional Intelligence (EI)**	**84.20**

B. Second Dataset (IMDB) Results

The proposed framework showed remarkable enhancements in sentiment analysis performance on the IMDB dataset, and compared results with state-of-the-art approaches, as demonstrated in Table III. The model achieved precision rates of 93.90% for positive reviews and 94.38% for negative critiques, with corresponding recall values of 94.04% and 93.69%, respectively. These metrics yielded an F1-score of 93.965% and an overall accuracy of 94.12%, outperforming conventional methods. For instance, Long et al. [50] introduced a cognition-based attention (CBA) method, which integrated dual attention frameworks with LSTM, achieved 90.10% accuracy, falling short despite its architectural complexity. Similarly, Collados and Pilehvar's [51] proposed a hybrid CNN-LSTM approach, incorporated preprocessing techniques and multi-modal fusion, attained 88.9% accuracy, underscoring the efficacy of our streamlined design.

Further comparisons reveal the framework's superiority over Ouyang et al. [52] CNN-based approach, and achieved an accuracy of 88.22% based on attention strategy, while Fu et al. [43] reported two various kinds of deep learning models, such as ALE-LSTM and WALE-LSTM approaches, which described an accuracy of 89.30% and 89.50%, respectively. Moreover, Kardakis et al. [49] explored the hybrid attention-based models and incorporated with LSTM/GRU, reporting an accuracy of 89.71% and 87.92%, respectively. Similarly, on the MR dataset, Mohammed et al. [47] proposed an attention-based recurrent neural network, such as GRU with a self-attention and recurrent attention mechanism, and reported an accuracy of 92.17%. While Zulqarnain et al. [21] reported 91.32% accuracy using a normalized auto-encoder with improved GRU. Furthermore, Islam et al. [53] explored the various hybrid deep learning models such as RoBERTa+DNN and BERT+DNN on the IMDB dataset, reported an accuracy of 92.05%, and 91.35%, respectively. Additionally, Lio Bing [54] proposed a CBOW-CNN method for sentence classification, attained an accuracy of 87.20%, and Yohong et al. [55] accomplished FARNN-Att mechanism, and highlighted an accuracy of 89.22% were outperformed by significant margins. In contrast, Danyal et al. [56] introduced a sentiment analysis model, such as XLNet (Extra-Long NN), using imdb dataset, achieving an accuracy of 93.74%, while Boewen Zhang [57] proposed a hybrid BERT-CNN model for sentiment analysis, reporting an accuracy of 92.90%.

By leveraging two-state layers alongside emotional intelligence, our proposed architecture highlighted excellent performance than existing studies based on architecture and accuracy comparison, detailed comparisons illustrated in Table V. The overall accuracy of two sentiment analysis datasets, showing IMDB at 94.12% and MR at 84.20%, graphical illustration is represented in Fig. 5.

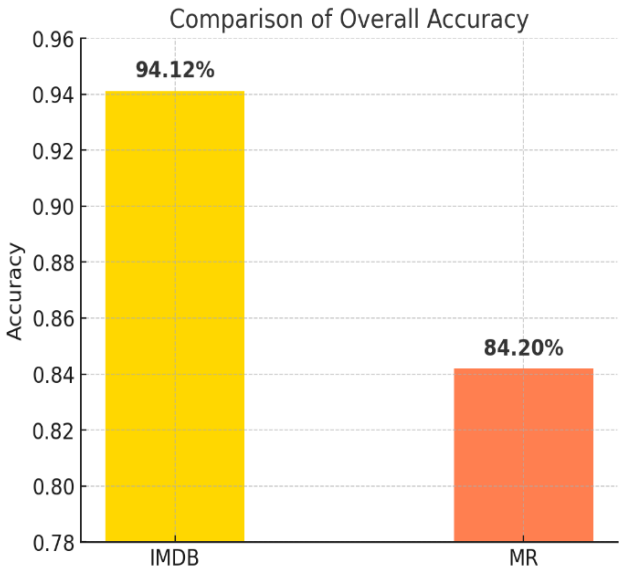


Fig. 5. Overall accuracy of the proposed model on both datasets.

Additionally, training dynamics were analyzed by tracking loss rates and epochs. Fig. 6 visualizes the epochs and loss rate of the proposed method alongside five conventional models (e.g., LSTM, Bi-GRU, BERT, BERT-CNN, and BERT-BiGRU) on a validation set. Over 25 epochs, the TS-ELSTM exhibited faster convergence, improved stability, and higher training efficiency, attributed to its emotional intelligence and topic-attention layers. These components optimize the representation of textual structure, enabling the model to capture contextual dependencies more effectively than traditional architectures.

TABLE V. COMPLEXITY AND ACCURACY-BASED ANALYSIS WITH TRADITIONAL STUDIES ON THE IMDB DATASET

Methods	Model Complexity	Accuracy (%)
CNN [52]	number of layers, filter sizes, strides, and input dimensions, leading to a computational cost-proportional	88.22
FARNN-Att [55]	Attention mechanism and adversarial training with BiLSTM	89.22
WALE-LSTM [43]	The combination of lexicon and attention layers with LSTM	89.50
CNN+LSTM [51]	Combination of CNN and LSTM with different cleaning processes	88.90
Self-Att-LSTM+word2vec [49]	Implemented Self-attention mechanism with LSTM based on word2vector	89.71
Hierarchical-Att-GRU+Dropout [49]	Investigated Hierarchical-attention mechanism with GRU based on word2vector	87.92
CBOW+D-CNN [54]	Combined the CBOW method with the CNN algorithm	87.20
LSTM+CBA+LA [50]	Integrated of two different feature attention frameworks with LSTM	90.11
BERT [58]	Bidirectional Encoder Representations from Transformers	85.83
XLNeT [58]	Transformer-XL with a two-stream attention mechanism	91.10
NAE-GRU [21]	Combined Auto-encoder with GRU through batch normalization	91.32
BERT+DNN [53]	Bidirectional encoder representations from transformers combined with deep neural network	91.35
RoBERTa+DNN [53]	R0bustly optimized BERT combined with deep neural network	92.05
SA-RA+GRU [47]	Proposed self-attention with recurrent attention unit GRU (SA+RA-GRU)	92.17
XLNet [56]	Extra-Long neural network	93.74
BERT-CNN [58]	Bidirectional encoder representations from transformers combined with CNN	92.90
TS-ELSTM (Proposed)	**Two-State Enhanced LSTM (TS-ELSTM) + Emotional Intelligence (EI)**	**94.12**

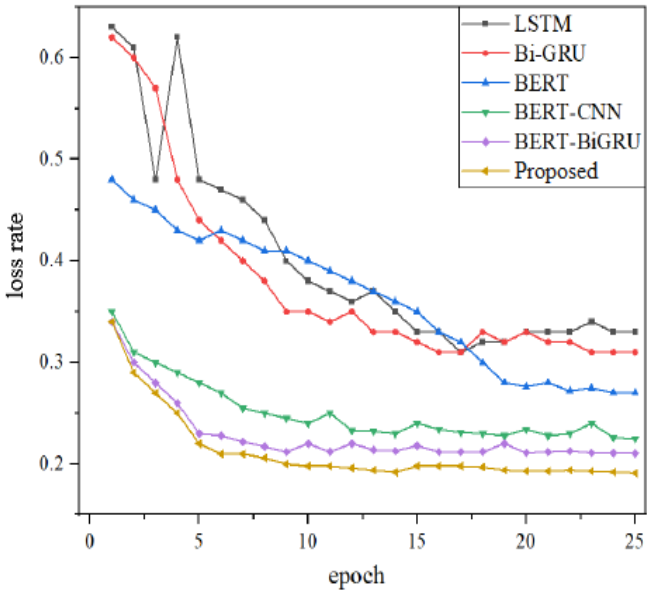


Fig. 6. Loss-based comparison of the proposed and traditional approaches.

VI. CONCLUSION AND FUTURE DIRECTION

Recurrent Neural Networks (RNNs) have become prominent in Natural Language Processing (NLP) for their effectiveness in modeling sequential data, yet they often struggle with vanishing gradient decay over long sequences and difficulty retaining distant contextual dependencies. To alleviate these drawbacks, attention mechanisms have been successfully combined as a transformative solution with deep learning algorithms, enhancing performance across NLP tasks. This study proposed TS-ELSTM, a novel deep learning architecture designed for sentiment analysis in textual sequences. The proposed framework integrates intelligence from emotional psychology and cognitive science, featuring an enhanced LSTM variant (ELSTM) that embeds emotional intelligence to dynamically balance information retention and forgetting. Additionally, a topic-aware attention mechanism dynamically adjusts the weighting of hidden representations in text subsequences, while an ELSTM layer synergizes with Latent Dirichlet Allocation (LDA) for topic modeling. Experimental outcomes show that the proposed TS-ELSTM accomplished excellent performance in sentiment analysis tasks compared to state-of-the-art models.

In the proposed study, we describe two limitations regarding our developed model: 1) The TS-ELSTM model was exclusively trained and tested on general-domain datasets (e.g., IMDB, Yelp, Amazon reviews). Its reliance on domain-specific linguistic patterns and emotional lexicons limits applicability to specialized contexts (e.g., clinical, financial, or low-resource languages). 2) The TS-ELSTM emotional gating mechanism and dual-polarity kernels increase computational cost and interpretation latency by ~35% compared to traditional LSTMs. This challenges scalability for real-time applications (e.g., live chat sentiment monitoring or high-frequency social media analysis), particularly on edge devices.

In future work, we plan two extensions. First, we extend to adapt TS-ELSTM model for aspect-level sentiment classification by incorporating fine-grained sentiment priors from lexical resources. Second, we will explore multi-objective optimization to tune TS-ELSTM parameters, striking an optimal trade-off between fitting accuracy and network sparsity, thereby enhancing both its robustness and generalization.

REFERENCES

[1] H. Zhao, J. Xiao, Y. Xue, H. Zhang, & S.H. Cai, "Aspect category sentiment classification via document-level GAN and POS information," *International Journal of Machine Learning and Cybernetics*, 15(8), pp. 3221-3235, 2024.

[2] H. Liu, X. Li, W. Lu, K. Cheng, & X. Liu, "Graph augmentation networks based on dynamic sentiment knowledge and static external knowledge graphs for aspect-based sentiment analysis," *Expert Systems with Applications*, 251, 123981, 2024.

[3] M. Zulqarnain, S. Ishak, R. Ghazali, N.M. Nawi, M. Aamir, & Y.M.M. Hassim, "An improved deep learning approach based on a variant two-state gated recurrent unit and word embeddings for sentiment classification," *International Journal of Advanced Computer Science and Applications*, 11(1), pp. 594-603, 2020.

[4] N.A. Sharma, A.S. Ali, & M.A. Kabir, "A review of sentiment analysis: tasks, applications, and deep learning techniques," *International Journal of Data Science and Analytics*, pp. 1-38, 2024.

[5] A. Koufakou, "Deep learning for opinion mining and topic classification of course reviews," *Education and Information Technologies*, 29(3), pp. 2973-2997, 2024.

[6] T. Liu, S. Li, Y. Dong, Y. Mo, & S. He, "Spam detection and classification based on distilbert deep learning algorithm," *Applied Science and Engineering Journal for Advanced Research*, 3(3), pp. 6-10, 2024

[7] M. Zulqarnain, A.K.Z. Alsaedi, R. Ghazali, M.G. Ghouse, W. Sharif, & N.A. Husaini, "A comparative analysis on question classification task based on deep learning approaches," *PeerJ Computer Science*, 7, e570, 2021.

[8] B. Babayigit & H. Sattuf, "BERT-based keyword extraction model for the Turkish language," *Neural Computing and Applications*, pp. 1-13, 2025.

[9] S.H. Rafique, A. Abdallah, N.S. Musa, & T. Murugan, "Machine learning and deep learning techniques for internet of things network anomaly detection current research trends," *Sensors*, 24(6), 1968, 2024.

[10] Z. Liu, J. Lin, Y. Chen, & Y. Dong, "Multimodal Aspect-Based Sentiment Analysis with External Knowledge and Multi-granularity Image-Text Features," *Neural Processing Letters*, 57(2), 25, 2024.

[11] M.S. Islam, M.N. Kabir, N.A. Ghani, K.Z. Zamli, N.S.A. Zulkifli, M.M. Rahman, & M.A. Moni, "Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach," *Artificial Intelligence Review*, 57(3), 62, 2024

[12] H.S. Asri & R. Safabakhsh, "Advanced visual and textual co-context aware attention network with dependent multimodal fusion block for visual question answering." *Multimedia Tools and Applications*, pp.1-28, 2024.

[13] L. Xiao, M. Li, Y. Feng, M. Wang, Z. Zhu, & Z. Chen, "Exploration of Attention Mechanism-Enhanced Deep Learning Models in the Mining of Medical Textual Data," In *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)* (pp. 1251-1258), 2024 IEEE.

[14] M. Fairuzabadi, K. Kusrini, E. Utami, & A. Setyanto, "Advancements and Challenges in Gated Recurrent Units (GRU) for Text Classification: A Systematic Literature Review," In *2024 7th International Conference of Computer and Informatics Engineering (IC2IE)* (pp. 1-7), 2024, IEEE.

[15] N. Jalal, A. Mehmood, G.S. Choi, & I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees. *Journal of King Saud University-Computer and Information Sciences*, 34(6), pp. 2733-2742, 2024.

[16] A. Assiri, A. Gumaei, F. Mehmood, T. Abbas, & S. Ullah, "DeBERTa-GRU: Sentiment Analysis for Large Language Model," *Computers, Materials & Continua*, 79(3), 2024.

[17] S. Almutairi, M. Abohashrh, H.H. Razzaq, M. Zulqarnain, A. Namoun, & F. Khan, "A Hybrid Deep Learning Model for Predicting Depression Symptoms from Large-Scale Textual Dataset," *IEEE Access*, 2024.

[18] M. Zulqarnain, R. Sheikh, S. Hussain, M. Sajid, S.N. Abbas, M. Majid, & U. Ullah, "Text classification using deep learning models: A comparative review," *Cloud Computing and Data Science*, 80-96, 2024.

[19] S. Najwa, S. Winarni, A.A. Pravitasari, & Y. Hidayat, "Emotion classification in social media posts related to telecommunication services using bidirectional LSTM," *Communications in Mathematical Biology and Neuroscience*, 2025, Article-ID.

[20] T.I. Toma & S. Choi, "An end-to-end multi-channel convolutional Bi-LSTM network for automatic sleep stage detection," *Sensors*, 23(10), 4950, 2023.

[21] M. Zulqarnain, A.K.Z. Alsaedi, R. Sheikh, I. Javid, M. Ahmad, & U. Ullah, "An improved gated recurrent unit based on auto encoder for sentiment analysis," *International Journal of Information Technology*, 16(1), 587-599, 2024.

[22] B. Jang, M. Kim, G. Harerimana, S.U. Kang, & J.W. Kim, "Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism," *Applied Sciences*, 10(17), 5841, 2023.

[23] Y. Wang, M. Huang, X. Zhu, & L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 606-615), 2016.

[24] Q. Zhao, F. Yang, D. An, & J. Lian, "Modeling structured dependency tree with graph convolutional networks for aspect-level sentiment classification," *Sensors*, 24(2), 418, 2024.

[25] M. Zulqarnain, R. Ghazali, M. Aamir, & Y.M.M. Hassim, "An efficient two-state GRU based on feature attention mechanism for sentiment analysis. *Multimedia Tools and Applications*, 83(1), 3085-3110, 2024.

[26] J. Yang, Y. Xiao, & X. Du, "Multi-grained fusion network with self-distillation for aspect-based multimodal sentiment analysis," *Knowledge-Based Systems*, 293, 111724, 2024.

[27] X. Tan, Y. Cai, & C. Zhu, "Recognizing conflict opinions in aspect-level sentiment classification with dual attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3426-3431).

[28] J. Ouyang, C. Xuan, B. Wang, & Z. Yang, "Aspect-based sentiment classification with aspect-specific hypergraph attention networks," *Expert Systems with Applications*, 248, 123412, 2019.

[29] A. Khashman, "Modeling cognitive and emotional processes: A novel neural network architecture," *Neural Networks*, 23(10), 1155-1163, 2010.

[30] E. Lotfi & M.R. Akbarzadeh-T, "Practical emotional neural networks. *Neural Networks*, 59, 61-72, 2014.

[31] A. Hassan, M. Shahid, F. Hayat, J. Arshad, M.H. Jaffery, A.U. Rehman, & H. Hamam, "Improving the Survival Time of Multiagents in Social Dilemmas through Neurotransmitter-Based Deep Q-Learning Model of Emotions," *Journal of Healthcare Engineering*, 2022(1), 3449433, 2022.

[32] O.K. Oyedotun & A. Khashman, "Prototype-incorporated emotional neural network," *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3560-3572, 2017.

[33] G.R. Markadeh, E. Daryabeigi, C. Lucas, & M.A. Rahman, "Speed and flux control of induction motors using emotional intelligent controller," *IEEE Transactions on Industry Applications*, 47(3), 1126-1135, 2011.

[34] T.A. Adjuik & D. Ananey-Obiri, "Word2vec neural model-based technique to generate protein vectors for combating COVID-19: a machine learning approach," *International Journal of Information Technology*, 14(7), 3291-3299, 2022.

[35] N.M. Alharbi, N.S. Alghamdi, E.H. Alkhammash, & J.F. Al Amri, "Evaluation of sentiment analysis via word embedding and RNN variants for Amazon online reviews," *Mathematical Problems in Engineering*, 2021(1), 5536560, 2021.

[36] A.A. Bilal, O.A. Erdem, & S. Toklu, "Children's sentiment analysis from texts by using weight updated tuned with random forest classification. *IEEE Access*, 2021.

[37] K.S. Prabha & M. Suguna, "Reciprocating Encoder portrayal from reliable Transformer dependent Bidirectional Long Short-Term Memory for Question and Answering Text Classification," *IEEE Access*, 2024.

[38] M. Zulqarnain, R. Ghazali, S.H. Khaleefah, & A. Rehan, "An improved the performance of GRU model based on batch normalization for

sentence classification," *International Journal of Computer Science and Network Security*, 19(9), pp.176-186, 2019.

[39] S. Hochreiter & J. Schmidhuber, "Long short-term memory," *Neural Computation*, 9(8), 1735-1780, 1997.

[40] E.A. Kensinger & J.H. Ford, "Retrieval of emotional events from memory," *Annual Review of Psychology*, 71(1), 251-272, 2020.

[41] Z. Xu, M. Niethammer, & C.A. Raffel, "Compositional generalization in unsupervised compositional representation learning," A study on disentanglement and emergent language. *Advances in Neural Information Processing Systems*, 35, 25074-25087, 2023.

[42] T. Chen, R. Xu, Y. He, & X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Systems with Applications*, 72, pp. 221-230, 2017.

[43] X. Fu, J. Yang, J. Li, M. Fang, & H. Wang, "Lexicon-enhanced LSTM with attention for general sentiment analysis," *IEEE Access*, 6, pp.71884-71891, 2018.

[44] D. Zhang, L. Tian, M. Hong, F. Han, Y. Ren, & Y. Chen, "Combining convolution neural network and bidirectional gated recurrent unit for sentence semantic classification," *IEEE Access*, 6, pp. 73750-73759, 2018.

[45] Q. Qian, M. Huang, J. Lei, & X. Zhu, "Linguistically regularized lstms for sentiment classification," *arXiv preprint arXiv:1611.03949*, 2016.

[46] M. Usama, W. Xiao, B. Ahmad, J. Wan, M.M. Hassan, & A. Alelaiwi, "Deep learning based weighted feature fusion approach for sentiment analysis," *IEEE Access*, 7, pp. 140252-140260, 2019.

[47] D.A. Mohammed & K.A. Patel, "An Improved Gru Based on Recurrent Attention Unit and Selfattention Technique for Text Sentiment Analysis," *ICTACT Journal on Soft Computing*, 15(4), 2025.

[48] R. Socher, B. Huval, C.D. Manning, & A.Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 1201-1211), 2021.

[49] S. Kardakis, I. Perikos, F. Grivokostopoulou, & I. Hatzilygeroudis, "Examining attention mechanisms in deep learning models for sentiment analysis," *Applied Sciences*, 11(9), 3883, 2021.

[50] Y. Long, Q. Lu, R. Xiang, M. Li, & C.R. Huang, "A cognition-based attention model for sentiment analysis," *Association for Computational Linguistics (ACL)*, 2017.

[51] J. Camacho-Collados & M.T. Pilehvar, "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis," *arXiv preprint arXiv:1707.01780*, 2017.

[52] X. Ouyang, P. Zhou, C.H. Li, & L. Liu, "Sentiment analysis using a convolutional neural network. In 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (pp. 2359-2364). IEEE, 2018.

[53] M.T. Islam, F. Parvin, S.A. Sazan, & T.B. Amir, "Comparative Analysis of Sentiment Classification on IMDB 50k Movie Reviews: A Study Using CNN, LSTM, CNN-LSTM, and BERT Models," In *2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON)* (pp. 512-517). IEEE, 2024.

[54] B. Liu, "Text sentiment analysis based on CBOW model and deep learning in big data environment," *Journal of Ambient Intelligence and Humanized Computing*, 11(2), pp. 451-458, 2020.

[55] Y. Ma, H. Fan, & C. Zhao, "Feature-based fusion adversarial recurrent neural networks for text sentiment classification," *IEEE Access*, 7, pp. 132542-132551, 2019.

[56] M.M. Danyal, M. Haseeb, S.S. Khan, B. Khan, & S. Ullah, "Opinion Mining on Movie Reviews Based on Deep Learning Models," *Journal of Artificial Intelligence*, 6, 2024.

[57] B. Zhang, "A bert-cnn based approach on movie review sentiment analysis," In *SHS Web of Conferences* (Vol. 163, p. 04007), 2023, EDP Sciences.

[58] M.M. Danyal, S.S. Khan, M. Khan, S. Ullah, F. Mehmood, & I. Ali, "Proposing sentiment analysis model based on BERT and XLNet for movie reviews," *Multimedia Tools and Applications*, pp. 1-25, 2024.

Image Quality Assessment Based on Feature Fusion and Local Adaptation

Minjuan GAO[1]*, Yankang LI[2], Xuande ZHANG[3]
School of Computer Science and Technology, Taiyuan Normal University, Shanxi 030600, China[1, 2]
School of Electrical Information and Artificial Intelligence, Shaanxi University of Science & Technology, Xian 710021, China[3]

***Abstract*—No-reference image quality assessment (NR-IQA) aims to evaluate the perceptual quality of images without access to corresponding reference images and has broad applications in real-world image processing scenarios. However, existing NR-IQA methods often suffer from limited accuracy and generalization, especially under complex and diverse distortion types. To address this challenge, we propose Inc-LAENet, a novel NR-IQA framework that leverages multi-scale deep residual representations, integrates feature fusion mechanisms, and incorporates a local adaptive perception module to achieve improved assessment accuracy and generalization. Specifically, ResNet50 is employed to extract hierarchical residual features, an enhanced Inception-style module (Inc-s) strengthens sensitivity to various distortion patterns, and a lightweight local adaptive extraction module efficiently captures fine-grained structural information. Extensive experiments demonstrate the effectiveness of the proposed method, achieving SROCC values of 0.967 and 0.935 on the synthetic distortion datasets LIVE and CSIQ, and 0.852 and 0.898 on the authentic distortion datasets LIVEC and KonIQ-10k, respectively. These results confirm that Inc-LAENet provides a robust and efficient solution for NR-IQA tasks across both synthetic and real-world scenarios.**

***Keywords*—No-reference image quality assessment; deep learning; multi-scale; feature fusion; local adaptation**

I. INTRODUCTION

Millions of digital images are shared on social media platforms every day. Images may be distorted to a certain extent during the acquisition and transmission process. In order to ensure an acceptable level of visual experience, it is crucial to conduct reliable and accurate image quality assessments (IQA) that have a good correlation with human perception. IQA can be divided into two categories: subjective evaluation and objective evaluation [1]. Subjective evaluation is the observer's own judgment of the image quality based on his or her visual experience. The objective evaluation is to design an algorithm to simulate the human visual system to analyze the relevant features of the image and give a corresponding quality score. Objective evaluation can be divided into three basic types according to whether a reference image is used: full-reference IQA [2] [3] (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference IQA (NR-IQA). NR-IQA [4] does not rely on the original reference image, but can predict its quality score based on the image's own features. This feature makes it more flexible and practical in practical applications, and therefore has broad application prospects in visual tasks such as image quality assessment [5], [6].

Although deep learning has made significant progress in the field of no-reference image quality assessment (NR-IQA) in recent years. However, this field still faces many challenges. First, image quality often depends on the comprehensive analysis of multi-scale features, while most existing methods are limited to feature extraction at a single scale, making it difficult to simultaneously capture local details and global structural information of the image. Second, different regions in the image have different effects on quality assessment, but existing methods usually fail to effectively model this non-uniformity, making it difficult for the model to focus on key areas that are critical to quality perception.

In order to solve the problem of insufficient accuracy and generalization ability of existing NR-IQA methods in complex distortion scenes, this study proposes an efficient NR-IQA method "Inc-LAENet" that integrates multiple modules. Its main contributions are as follows:

1) A NR-IQA framework (Inc-LAENet) that integrates deep residual features and local adaptive perception is proposed to achieve efficient modeling of global and local features.

2) The feature enhancement expression module (Inc-s) is designed, combining the dilated convolution and multi-scale convolution structure to improve the perception of local details and global semantic information of the image.

3) Introduce the Lightweight Adaptive Extraction (LAE) module, which focuses on the more quality-sensitive areas in the image through an adaptive weight mechanism, effectively improving the accuracy and robustness of the prediction.

The remainder of this study is organized as follows: Section II reviews related work in the field of NR-IQA. Section III introduces the overall architecture of Inc-LAENet, along with detailed descriptions of the Semantic Feature Extraction Network, the Feature Enhancement Module, and the Lightweight Local Adaptive Extraction Module. Section IV describes the experimental setup, including datasets, evaluation metrics, implementation details, and presents both comparative and ablation experiments. Section V summarizes the key findings. Section VI provides further discussions on the performance advantages, comparisons with state-of-the-art methods, generalization capabilities, and potential directions for future work.

Fundamental Research Programme of Shanxi Province (202203021222237). Supported by Scientific and Technologial Innovation Programs of Higher Education Institutions in Shanxi (2022L401).

II. Related Work

Traditional no-reference image quality assessment relies on the statistical characteristics and visual properties of the image itself, and requires manual extraction of image features, which is difficult and complex to implement, such as natural scene statistics (NSS) features, gradient features, texture features, etc. Mittal et al. proposed BRISQUE [7], which extracts the local normalized brightness coefficient (MSCN) of the image and its statistical characteristics, and uses the support vector machine (SVM) for quality prediction. Saad et al. proposed a NR-IQA method called BLIINDS [8], which uses the statistical information of the local discrete cosine transform (DCT) coefficients to extract features and then uses a multivariate Gaussian model to evaluate image quality. Furthermore, Saad et al. proposed BLIINDS-II [9] to evaluate image quality by optimizing the statistical characteristics of DCT. Zhang et al. proposed a general IQA method ILNIQE [10], which uses the mean minus, the contrast normalization coefficient, and the gradient statistic to extract features. However, hand-designed features often rely on prior knowledge, and have difficulty in fully capturing images when faced with complex feature distortions, and have limited generalization capabilities when faced with some natural distortions.

In recent years, deep learning has become popular [11] due to the powerful feature extraction capabilities of convolutional neural networks (CNN) [12]. Convolutional neural networks combine feature extraction with the learning process, avoiding the limitations of manually extracted features and showing good evaluation results. They have also been applied to NR-IQA [13-16]. For example, Kang et al. [17] proposed an NR-IQA method based on convolutional neural networks, which extracts features from images and predicts quality scores by training the network. Kang et al. [18] proposed a multi-task CNN for simultaneous image quality estimation and distortion type identification. Liu et al. [19] proposed an NR-IQA method called RankIQA, which solves the problem of the Siamese network's limitation on the size of the IQA dataset and evaluates quality by ranking image quality.

III. Proposed Method

This study proposes a no-reference image quality assessment method based on multi-scale feature fusion, as shown in Fig. 1. The method consists of four core modules: a backbone network (used to extract multi-level semantic information), a feature enhancement module, a multi-level semantic feature fusion module, and a quality prediction network (used to output the final image quality score).

During the training process, the pre-trained deep residual network ResNet50 [20] is first used as the feature extraction network to extract semantic features x1, x2, x3, and x4 at different levels from the distorted RGB image. Subsequently, x1, x2, and x3 are downsampled by the LAE module to enhance the local information perception capability. At the same time, the features of the four scales are deeply optimized by the Inc-s module. Next, the processed features are input into the global pooling layer and spliced in the feature dimension to fully integrate multi-scale information. Finally, the fused features are input into the quality regression network to generate the final image quality score.

A. Semantic Feature Extraction Network

ResNet50 is a variant of the residual network proposed by He et al. to solve the problems of gradient vanishing and gradient exploding as the network depth increases. ResNet50 consists of 50 layers of convolutional networks, and its structure mainly includes convolutional layers, batch normalization layers, ReLU activation functions, and fully connected layers. The performance of each residual block is shown in Formula (1):

$$F(x) = H(x) - x \tag{1}$$

where, $H(x)$ represents the original mapping function and $F(x)$ represents the residual mapping. During the forward propagation process, the output of the residual block is Formula (2). Its structure diagram is shown in Fig. 2.

$$Y = F(x) + x \tag{2}$$

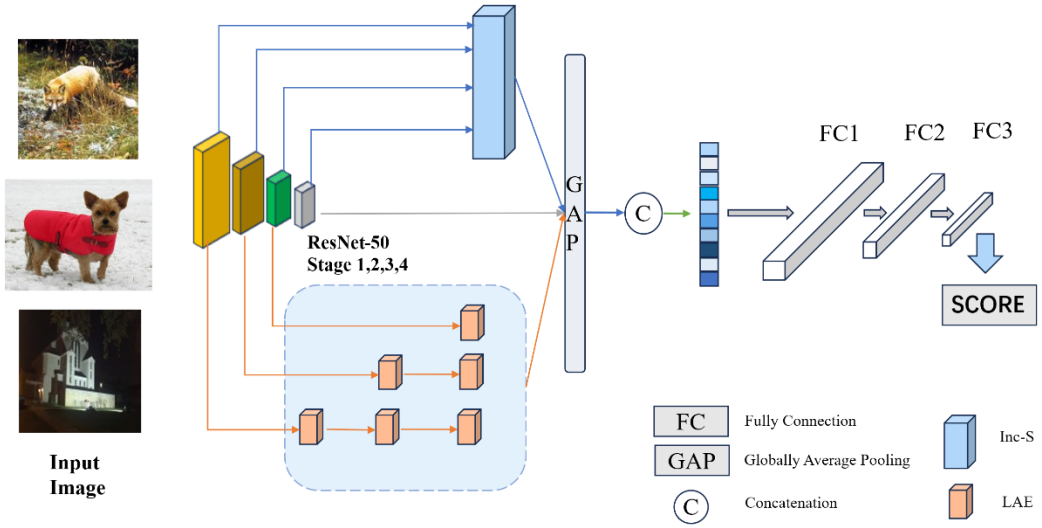


Fig. 1. Overall framework.

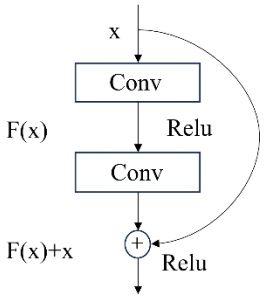


Fig. 2. Residual block structure.

In order to capture rich contextual information and deep semantic features, we use ResNet-50 as the backbone network to extract multi-level features of the image. These features come from the output of Conv2_x, Conv3_x, Conv4_x, and Conv5_x stages. The features of each stage contain information of different spatial resolutions and semantic levels. The multi-scale features are shown in Formula (3):

$$\Phi(I) = \left\{ f_l(I) \middle| l \in \{2,3,4,5\} \right\} \quad (3)$$

where, *I* denotes the input image, and *d* represents the feature map at the *l*-th layer.

B. Feature Enhancement Module

Human vision has a multi-level information processing mechanism [21]. In order to improve the network's ability to understand the content of distorted images, inspired by the inception module [22], this study designed a feature enhancement method that integrates a multi-scale attention mechanism, as shown in Fig. 3.

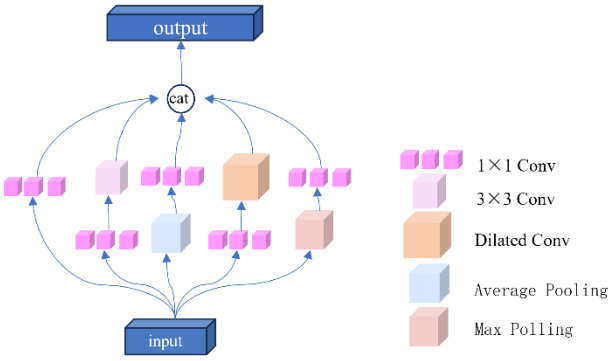


Fig. 3. Inc-s module.

Specifically, the method first extracts hierarchical features through the backbone network to obtain feature representations at four scales, and then inputs them into a new feature enhancement module to optimize feature expression capabilities. This module draws on the idea of a spatial pyramid structure and deploys multi-scale convolution operations (including 1×1, 3×3, and 5×5 convolutions) and pooling operations in parallel to achieve collaborative modeling of local details and global semantics, thereby enhancing the perception of image quality.

In addition, to further optimize computational efficiency and model performance, the feature enhancement module introduces two key technical improvements during implementation: 1) Atrous convolution replaces traditional large kernel convolution: 3×3 atrous convolution kernels are used to replace traditional 5×5 convolutions (the dilation factor is 1 in this experiment), which reduces the number of parameters while ensuring equivalent receptive fields; 2) Adding an average pooling branch. An additional average pooling branch is introduced to capture the global statistical information of the image, complementing other convolution branches. Average pooling can smooth feature distribution and reduce the impact of outliers and noise, thereby reducing the interference of local distortion on model prediction and improving the generalization ability of unknown distortions. In addition, since average pooling does not rely on trainable parameters, this branch enriches feature expression while increasing model complexity minimally, which helps to reduce the risk of overfitting. Formula (4) describes the output form of the Inception module:

$$F_{out} = \left[F_{1\times1},\ F_{3\times3},\ F_{dilated},\ F_{max},\ F_{avg} \right] \quad (4)$$

Among them, $F_{1\times1}$, $F_{3\times3}$,and $F_{dilated}$ represent convolution branches of different scales, and $F_{\max}$ and F_{avg} represent maximum pooling and flat pooling branches, respectively.

C. Lightweight Local Adaptive Extraction Module

In IQA tasks, accurately extracting local and global features of an image is the key to evaluating the degree of image distortion. In order to effectively preserve the edge and local detail information of the image, this method introduces the LAE module [23], as shown in Fig. 4. Compared with traditional convolution methods, the LAE module has obvious advantages in terms of the number of parameters and computational complexity, while extracting more semantically rich features. IQA is highly sensitive to edges, textures, and local details, while traditional convolutions tend to lose this key information during downsampling. The LAE module effectively retains and enhances this information through parallel feature extraction and adaptive information aggregation.

The LAE module is designed with a dual-branch architecture that shares parameters, combining Group Convolution and adaptive feature aggregation, which not only reduces the number of parameters but also retains key information in the image. By dividing the feature map into N groups for convolution operations, the number of parameters and the amount of calculation are reduced to 1/N compared to traditional convolution.

LAE introduces adaptive feature aggregation to alleviate the loss of edge and detail information caused by downsampling. Specifically, it changes the dimension of the feature map from four dimensions (batch, channel, height, width) to five dimensions (batch, channel, height, width, n), where “n” represents the sampling factor. The adaptive path aggregates four adjacent pixels and calculates their weights through average pooling and convolution operations, and uses softmax normalization to ensure that pixels with higher information entropy are retained in the channel dimension. Formula (5) describes the output mode of the LAE module. Given the input $X \in R^{h\times w\times c}$, the output of LAE is $Y \in R^{\frac{h}{2}\times\frac{w}{2}\times c}$.

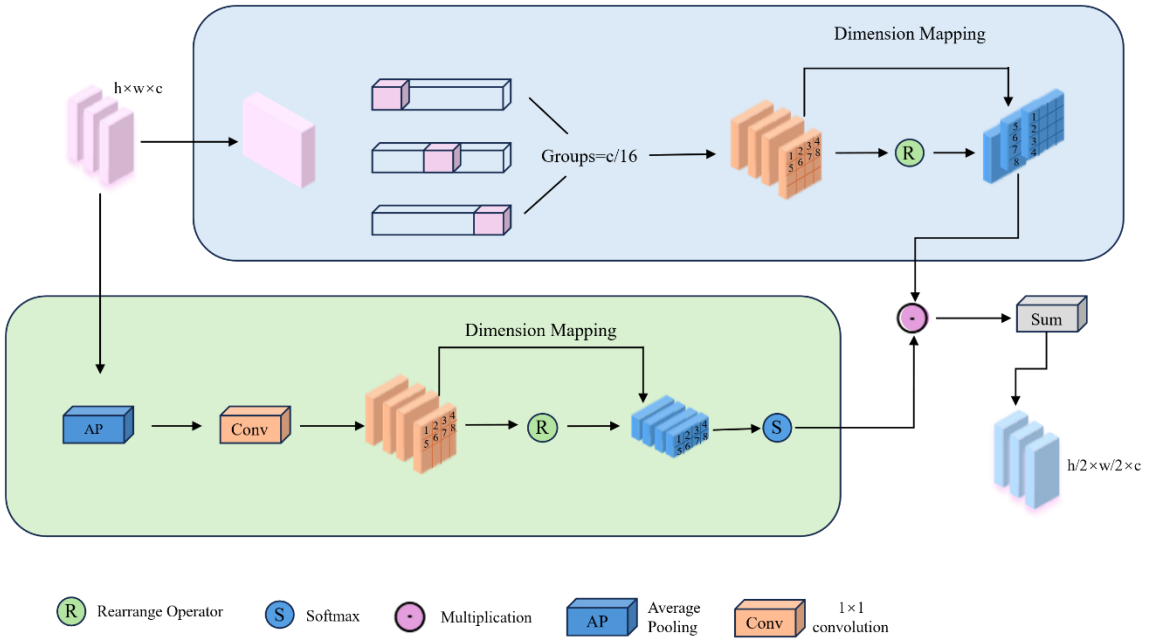


Fig. 4. Lightweight Local Adaptive Extraction module (LAE).

$$Y = \sum \begin{pmatrix} Rearrange(GroupConv(x)) \\ Softmax(Rearrange(Conv(AP(X)))) \end{pmatrix} \quad (5)$$

where, $X \in R^{h\times w\times c}$ is the input feature map, $Y \in R^{\frac{h}{2}\times\frac{w}{2}\times c}$ is the output feature map, $GroupConv(X)$ performs group convolution on the input feature map, $Rearrange(\bullet)$ compresses the spatial dimension to the channel dimension, $AP(X)$ performs average pooling on the input, and $Softmax(\bullet)$ generates adaptive weights.

IV. Experiments

A. Experimental Environment and Dataset

The model proposed in this study is implemented using Python 3.9 programming language and PyTorch 2.5.1 deep learning framework, which has strong compatibility and scalability. In the training and testing phases, the experiment uses NVIDIA GeForce RTX 4090 graphics processor as the main computing platform, combined with CUDA 12.1 toolkit to achieve efficient parallel computing support. This configuration not only effectively improves the running speed of the model, but also shortens the training cycle to a certain extent, enhancing the controllability of the experimental process and the repeatability of the results.

In this experiment, we tested and evaluated on four public IQA datasets (two synthetic distortions and two real distortions). For synthetic distortion, we used LIVE [24] and CSIQ [25]. The LIVE dataset contains 779 distorted images, covering five distortion types and five distortion levels, and the CSIQ dataset contains 866 images, covering six distortion types and four to five distortion levels. For real distortion datasets, we used CLIVE [26] and KonIQ-10k [27]. LIVE-C is a no-reference image quality assessment dataset containing 1162 real distorted images taken by different multimedia devices, and KonIQ-10k is a large-scale no-reference dataset containing 10,073 high-quality real distorted images. Table I is a summary of the databases used in this experiment.

TABLE I. IQA Dataset Used for Comparative Experiments

Databases	#of Dist. Images	#of Dist. Types	Distortions Type
LIVE	799	5	synthetic
CSIQ	866	6	synthetic
LIVEC	1162	-	authentic
KonIQ-10k	10073	-	authentic

B. Evaluation Indicators

For the performance evaluation of the proposed method, two commonly used standards are used. They are Spearman's rank order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC). The range of SROCC and PLCC is from -1 to 1, and the larger the absolute value, the better the performance. SROCC is a non-parametric correlation coefficient used to measure the monotonic relationship between two variables (i.e., whether one variable increases or decreases as the other variable increases). It is based on the rank (sorting) of the variables rather than the original values, so it is not sensitive to outliers. The calculation formula of SROCC is shown in Formula (6):

$$SROCC = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)} \quad (5)$$

where, d_i represents the level difference between the true value and the predicted value of the distorted image, and n

represents the number of distorted images. The calculation formula of PLCC is shown in Formula (7):

$$\text{PLCC}=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (6)$$

where, X_i and Y_i represent the predicted score and true score of the distorted image, respectively. $\bar{X}$ and $\bar{Y}$ represent the mean of the predicted score and the true label score.

C. Experimental Details

This experiment follows the training strategy of the existing IQA algorithm. During training, 30 224×224 pixel patches are randomly selected from each distorted image, and random data augmentation is performed, and their quality scores are consistent with the source image. The Adam optimizer is used for training, with a weight decay coefficient of 5×10-4, a maximum number of training rounds (epochs) of 5, and a mini-batch size of 32. The initial learning rate is set to 2×10-5 and is decreased by 10% after each round of training. In the test phase, 30 patches are randomly selected from each test image, their quality scores are predicted, and the average is taken as the final score. The experiment uses the same settings and randomly divides the dataset into 80% training set and 20% test set using 10 different seeds. The test data does not participate in the training, and the median of the SROCC and PLCC of the 10 groups of experiments is taken as the result. In order to optimize the model, capture subtle changes in image quality, reduce the impact of outliers, accelerate convergence and improve stability, the SmoothL1Loss loss function is used in training, as shown in Formula (8):

$$\text{SmoothLoss}=\begin{cases}0.5(x-y)^2, & if\ |x-y|<\beta\\ \beta(|x-y|-0.5\beta), & otherwise\end{cases} \quad (7)$$

where, x is the prediction score of the model and y is the true label value of the distorted image.

D. Comparative Experiment

In order to fully verify the effectiveness and robustness of the proposed model, we conducted comparative experiments with six representative image quality evaluation methods on four mainstream NR-IQA datasets, covering traditional methods (such as BRISQUE and ILNIQE) and recent deep learning methods (such as WaDIQaM, DBCNN, HyperIQA and TReS). As shown in Table II and Table III, on the real distortion datasets KONIQ-10k and LIVEC, our method achieved good performance, reaching SROCC 0.898 and PLCC 0.909 on KONIQ-10k, and SROCC 0.852 and PLCC 0.872 on LIVEC, respectively, which is better than traditional methods such as BRISQUE and ILNIQE, and is comparable to advanced models such as HyperIQA and TReS, with only a small gap or close performance, reflecting the high generalization ability of the model in complex real scenes. On synthetic distortion datasets such as LIVE and CSIQ, our method also shows strong robustness. On the LIVE dataset, we achieved a high score of 0.967 for both SROCC and PLCC, which is close to or even slightly better than other mainstream methods, indicating that the model has accurate perception capabilities when dealing with standard synthetic distortion types (such as blur, compression, noise, etc.); on the CSIQ dataset, our method's SROCC and PLCC are 0.935 and 0.944, respectively, which also surpasses most baseline models and is second only to some highly optimized models such as DBCNN, but the overall performance is still stable and competitive. In order to more intuitively show the performance differences between the methods, we introduced the bar charts shown in Fig. 5 to Fig. 8, which visualize the evaluation results of the LIVEC, KONIQ-10k, LIVE and CSIQ datasets, respectively. From the figures, we can clearly observe the leading trend of the proposed method in various indicators, which further enhances the intuitiveness and persuasiveness of the experimental results.

TABLE II. PERFORMANCE COMPARISON OF EACH EVALUATION METHOD ON THE REAL DISTORTION DATASET

Methods	LIVEC		KONIQ-10k	
	SROCC	***PLCC***	***SROCC***	***PLCC***
BRISQUE	0.608	0.629	0.701	0.705
ILNIQE	0.432	0.508	0.507	0.527
WaDIQaM	0.671	0.681	0.806	0.807
DBCNN[28]	0.851	0.869	0.877	0.884
HyperIQA[29]	0.855	0.875	0.901	0.910
TReS[30]	0.846	0.877	0.905	0.908
Ours	0.852	0.872	0.898	0.909

TABLE III. PERFORMANCE COMPARISON OF EACH EVALUATION METHOD ON THE SYNTHETIC DISTORTION DATASET

Methods	LIVE		CSIQ	
	SROCC	***PLCC***	***SROCC***	***PLCC***
BRISQUE	0.929	0.944	0.748	0.812
ILNIQE	0.898	0.903	0.815	0.854
WaDIQaM	0.954	0.963	0.844	0.852
DBCNN	0.968	0.971	0.946	0.959
HyperIQA	0.962	0.966	0.923	0.942
TReS	0.968	0.969	0.922	0.942
Ours	0.967	0.967	0.935	0.944

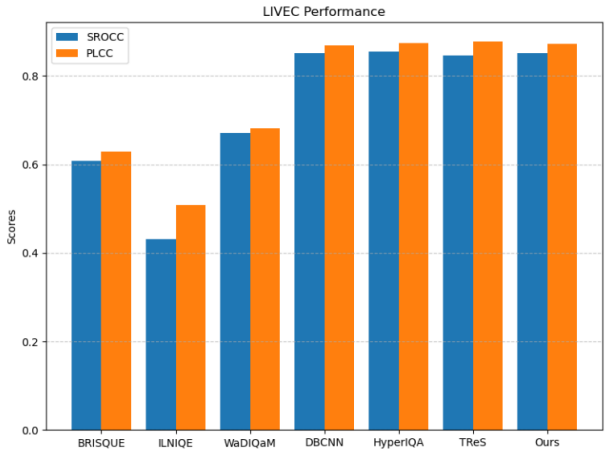


Fig. 5. Results of the LIVEC.



Fig. 6. Results of the KONIQ-10k.

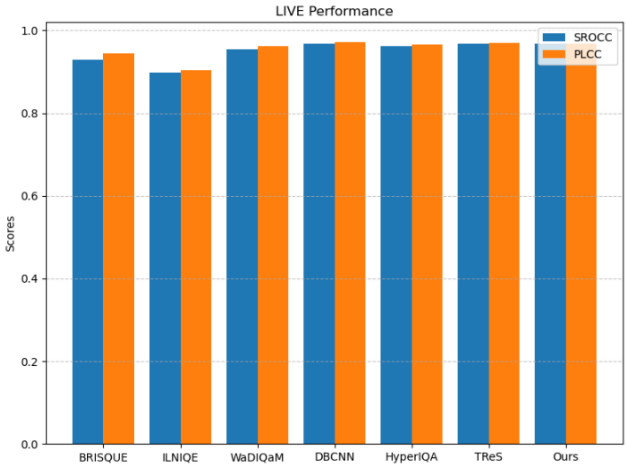


Fig. 7. Results of the LIVE.

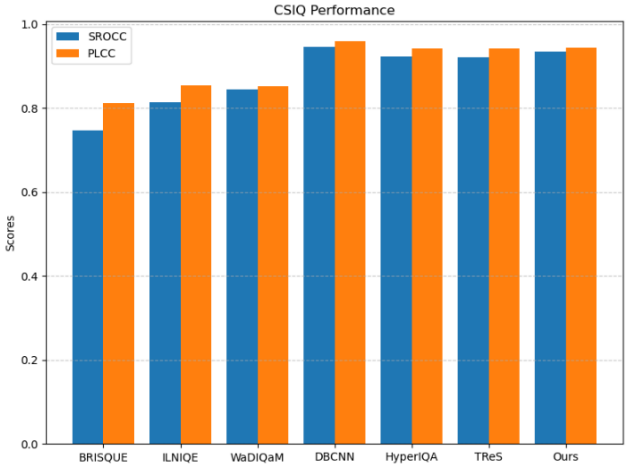


Fig. 8. Results of the CSIQ.

E. Ablation Experiment

In order to verify the contribution of each module to this model in this experiment, an ablation experiment was designed and conducted. The ablation experiment also adopted the above experimental rules, conducted ten experiments, and took the median of SROCC and PLCC as the experimental results for testing on the CSIQ and LIVEC databases. Table IV shows the statistical results of different modules to help understand the impact of each module on the performance of the model. The baseline model uses ResNet-50. From Table IV, it is not difficult to see that by integrating the feature enhancement module Inc-s into the basic model, SROCC and PLCC are improved on the CSIQ dataset and LIVE-C dataset. By integrating the lightweight local adaptive module LAE into the baseline model, all evaluation indicators on the two datasets are also improved. Finally, both modules are integrated into the baseline model to form our final model. It can be seen that our final model can show the best performance.

TABLE IV. PERFORMANCE COMPARISON OF EACH MODULE

Baseline	Inc-s	LAE	CSIQ		LIVE-C	
			SROCC	***PLCC***	***SROCC***	***PLCC***
√			0.903	0.928	0.817	0.850
√	√		0.921	0.940	0.838	0.856
√		√	0.915	0.933	0.845	0.866
√	√	√	0.925	0.944	0.852	0.872

V. SUMMARIZE

This study proposes an NR-IQA model Inc-LAENet, that combines convolutional neural networks, Inc-s, and a lightweight local adaptive extraction (LAE) module. ResNet-50 extracts global and local features of the image through a residual structure, greatly enhancing the model's perception of complex distortions. Inc-s strengthens the expression of features by combining multi-scale convolution and pooling. LAE efficiently extracts local feature information while reducing computational overhead, thereby improving local feature sensitivity. Experimental results show that the model Inc-LAENet proposed in this study performs well on multiple public datasets, with SROCCs of 0.967 and 0.935 on two synthetic distortion datasets, LIVE and CSIQ, respectively, and SROCCs of 0.852 and 0.898 on two real distortion datasets, LIVEC and Koniq-10k, respectively, proving its effectiveness in NR-IQA. Future research can further explore deeper feature fusion, loss function optimization, and other technologies to further improve the performance and application potential of the model.

VI. DISCUSSION

This section will conduct an in-depth analysis of the experimental results of the proposed method, focusing on its design concept, performance advantages, scope of application and existing deficiencies, and further highlighting the practical value and potential impact of this study.

A. Analysis of the Causes of Performance Advantages

The experimental results show that the proposed method has achieved excellent performance on multiple public image quality assessment datasets. The overall performance is comparable to the current mainstream image quality assessment model, and even slightly exceeds it in some indicators. Further analysis of the reasons for its performance improvement is mainly reflected in the following two aspects: on one hand, the

Inc-s feature enhancement module designed in this study effectively integrates multi-scale contextual information, enabling the model to focus on the global structure and local details of the image at the same time, thereby improving the comprehensiveness of feature expression; on the other hand, the introduction of a lightweight local adaptive enhancement module (LAE) enhances the model's sensitivity to local areas of the image, especially distorted areas such as noise and compression artifacts, and shows stronger discrimination capabilities in complex distorted scenes. The two modules work together in parallel, effectively improving the model's global perception and local modeling capabilities in the task of no-reference image quality assessment, thereby significantly enhancing the generalization and stability of the model.

B. Generalization Ability and Application Prospects

It is not difficult to find from the experimental results that the method proposed in this study shows relatively consistent performance in both real distortion and synthetic distortion scenarios, indicating that it has good robustness in dealing with various types of image distortion. This result not only verifies the applicability of the method on multiple mainstream no-reference image quality assessment (NR-IQA) datasets but also further highlights its potential in practical engineering applications. This method can be widely used in many typical scenarios such as image compression quality assessment, image enhancement effect verification, and image transmission stability detection, and has good promotion prospects and practical value.

C. Model Limitations and Future Development Directions

Although the method proposed in this study has achieved excellent performance on multiple typical datasets and has shown good generalization ability and practical application potential, there is still room for further exploration in terms of model optimization. At present, this method mainly relies on convolutional neural networks for feature extraction and quality regression, which can effectively capture the structural information and local distortion characteristics in the image. However, when faced with image content with complex semantic relationships or strong contextual dependencies, the prediction results of the model may still have certain deviations on individual samples.

To further enhance the model's ability to understand complex image quality degradation scenarios, subsequent research may consider introducing structures such as self-attention mechanisms or Transformers to improve the modeling effect of global semantic information. In addition, in more challenging cross-database evaluation or zero-shot image quality prediction tasks, it is necessary to explore more efficient transfer learning strategies or unsupervised optimization solutions to further improve the robustness and adaptability of the model. These improvement directions not only help to improve the theoretical system of the model, but also lay the foundation for its application in practical scenarios.

REFERENCES

[1] Hu Z, Yang G, Du Z, et al. No-reference image quality assessment based on global awareness[J]. Plos one, 2024, 19(10): e0310206.

[2] Chen D , Wu T , Ma K ,et al.Toward Generalized Image Quality Assessment: Relaxing the Perfect Reference Quality Assumption[J]. 2025.

[3] Lan X , Jia F , Zhuang X ,et al.Hierarchical degradation-aware network for full-reference image quality assessment[J].Information Sciences, 2025, 690(000).DOI:10.1016/j.ins.2024.121557.

[4] Mao Q , Liu S , Li Q ,et al.No - Reference Image Quality Assessment: Past, Present, and Future[J].Expert Systems, 2025, 42(3).DOI:10.1111/exsy.13842.

[5] Yang C, Liu Y, Li D, et al. Exploring vulnerabilities of no-reference image quality assessment models: A query-based black-box method[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024.

[6] Mao Q, Liu S, Li Q, et al. No-Reference Image Quality Assessment: Past, Present, and Future[J]. Expert Systems, 2025, 42(3): e13842.

[7] Mittal A , Moorthy A K , Bovik A C .No-Reference Image Quality Assessment in the Spatial Domain[J].IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2012, 21(12):4695.DOI:10.1109/TIP.2012.2214050.

[8] Saad M A , Bovik A C , Charrier C .A DCT Statistics-Based Blind Image Quality Index[J].IEEE Signal Processing Letters, 2010, 17(6):583-586.DOI:10.1109/LSP.2010.2045550.

[9] Saad M A , Bovik A C , Charrier C .Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain[J].IEEE Transactions on Image Processing, 2012, 21(8):3339-3352.DOI:10.1109/TIP.2012.2191563.

[10] Zhang L , Zhang L , Bovik A C .A Feature-Enriched Completely Blind Image Quality Evaluator[J].IEEE Transactions on Image Processing, 2015, 24(8):2579-2591.DOI:10.1109/TIP.2015.2426416.

[11] Srinath S, Mitra S, Rao S, et al. Learning generalizable perceptual representations for data-efficient no-reference image quality assessment[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 22-31.

[12] Shi J, Gao P, Qin J. Transformer-based no-reference image quality assessment via supervised contrastive learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(5): 4829-4837.

[13] Kim J , Lee S .Fully Deep Blind Image Quality Predictor[J].Selected Topics in Signal Processing, IEEE Journal of, 2017, 11(1):206-220.DOI:10.1109/JSTSP.2016.2639328.

[14] Bosse S , Maniry D , Muller K R ,et al.Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment[J].IEEE Transactions on Image Processing, 2017:1-1.DOI:10.1109/TIP.2017.2760518.

[15] Zhao W, Li M, Xu L, et al. A Multi-Branch Network with Multi-Layer Feature Fusion for No-Reference Image Quality Assessment[J]. IEEE Transactions on Instrumentation and Measurement, 2024.

[16] Miao H, Sang Q. McmIQA: Multi-Module Collaborative Model for No-Reference Image Quality Assessment[J]. Mathematics, 2024, 12(8): 1185.

[17] Kang L , Ye P , Li Y ,et al.Convolutional Neural Networks for No-Reference Image Quality Assessment[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition.IEEE, 2014.DOI:10.1109/CVPR.2014.224.

[18] Kang L , Ye P , Li Y ,et al.Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks[J].IEEE, 2015.DOI:10.1109/ICIP.2015.7351311.

[19] Liu X , Weijer J V D , Bagdanov A D .RankIQA: Learning from Rankings for No-Reference Image Quality Assessment[C]//2017 IEEE International Conference on Computer Vision (ICCV).IEEE, 2017.DOI:10.1109/ICCV.2017.118.

[20] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[21] Choudhary B K, Sinha N K, Shanker P. Pyramid method in image processing[J]. Journal of Information Systems and Communication, 2012, 3(1): 269.

[22] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[23] Yu Z, Guan Q, Yang J, et al. Lsm-yolo: A compact and effective roi detector for medical detection[J]. arXiv preprint arXiv:2408.14087, 2024.

[24] Sheikh H R, Sabir M F, Bovik A C. A statistical evaluation of recent full reference image quality assessment algorithms[J]. IEEE Transactions on image processing, 2006, 15(11): 3440-3451.

[25] Larson E C, Chandler D M. Most apparent distortion: full-reference image quality assessment and the role of strategy[J]. Journal of electronic imaging, 2010, 19(1): 011006-011006-21.

[26] Ghadiyaram D, Bovik A C. Massive online crowdsourced study of subjective and objective picture quality[J]. IEEE Transactions on Image Processing, 2015, 25(1): 372-387.

[27] Hosu V, Lin H, Sziranyi T, et al. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment[J]. IEEE Transactions on Image Processing, 2020, 29: 4041-4056.

[28] Network A D B C N. Blind Image Quality Assessment Using A Deep Bilinear Convolutional Neural Network[J]. Deep Bilinear Convolutional Neural.

[29] Su S, Yan Q, Zhu Y, et al. Blindly assess image quality in the wild guided by a self-adaptive hyper network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3667-3676.

[30] Golestaneh S A, Dadsetan S, Kitani K M. No-reference image quality assessment via transformers, relative ranking, and self-consistency[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022: 1220-1230.

Advancing Aerodynamic Coefficient Prediction: A Hybrid Model Integrating Deep Learning and Optimization Techniques

Jad Zerouaoui[1], Rachid Ed-daoudi[2], Badia Ettaki[3], El Mahjoub Chakir[4]
Laboratory of Material and Subatomic Physics-Faculty of Sciences, Ibn Tofail University
Campus Universitaire, B.P. 133, Kenitra, Morocco[1, 4]
LyRICA: Laboratory of Research in Informatics, Data Sciences and Artificial Intelligence,
School of Information Sciences, B.P. 6204, Rabat-Instituts, Rabat, Morocco[2, 3]

***Abstract*—The aerospace industry increasingly relies on predictive models for aerodynamic coefficients to enhance design, performance, and optimization. While traditional methods like Computational Fluid Dynamics (CFD) and wind tunnel simulations offer accurate predictions, they are computationally intensive and time-consuming. This study explores a novel approach that fuses advanced Deep Learning (DL) architectures with Optimization Techniques to achieve faster and more accurate predictions of aerodynamic coefficients. Building on the foundation of Convolutional Neural Networks (CNNs), we introduce hybrid models that integrate Evolutionary Algorithms and Gradient-Based Optimization to improve the accuracy, generalization, and adaptability of predictions. The proposed framework is validated on datasets derived from CFD simulations and wind tunnel experiments, demonstrating superior accuracy, reduced computational cost, and robust performance across diverse aerodynamic conditions. This study highlights the potential of combining DL and optimization methods as a transformative tool for real-time aerodynamic analysis, paving the way for more efficient Aerospace Design and decision-making. Future research directions include expanding the model to handle complex geometries and dynamic flight conditions.**

***Keywords*—Aerodynamic coefficients; computational fluid dynamics; deep learning; convolutional neural networks; optimization techniques; evolutionary algorithms; gradient-based optimization; aerospace design**

I. Introduction

The aerospace industry has long relied on the ability to make reasonably accurate predictions of the aerodynamic coefficients to optimize aircraft and spacecraft for design, performance, and safety. Basic to the determination of the forces and moments acting on a vehicle flight is the determination of the aerodynamic coefficients: lift, drag, and moment coefficients. Traditional methods for determining these coefficients involve CFD simulations and wind tunnel experiments, which traditionally have been applied to the standard method of undertaking aerodynamic analyses over many years [1]. In general, however, these techniques are computationally intensive and normally very time-consuming, which prohibits their applications in real-time settings and iterative design procedures.

Recent breakthroughs in Machine Learning, especially within DL methods, have opened new avenues towards faster predictions in aerodynamics. Some of the DL models, like CNNs, have been very successful in approximating complicated physical phenomena with a high degree of accuracy and speed [2]. These techniques may enable the aerospace industry to reduce computational costs and enhance agility in design. The following discussion will look at how DL architectures can be combined with Optimization Techniques to outperform traditional methods for fast and accurate predictions of aerodynamic coefficients.

Traditional methods for aerodynamic prediction include CFD and wind tunnel testing, which, though very accurate, have some major drawbacks. First, CFD simulations require the solving of the Navier-Stokes equations, including high computational complexity and resource-intensive meshing processes [3]. These take several hours to days for a single simulation to complete, even on high-performance computing systems. Wind tunnel experiments are also quite expensive and are limited by physical constraints such as scaling effects and facility availability [4].

Furthermore, neither of these techniques can effectively implement real-time applications or cases that involve complicated geometries with dynamic flying conditions. For instance, the CFD simulation requires re-meshing and re-calibration over different geometries [5]; experiments in wind tunnels cannot simulate transient or unsteady aerodynamics phenomena correctly [6]. All these challenges imply that alternative approaches must balance the important issues of accuracy, computational efficiency, and adaptability.

This research examines how a hybrid deep-learning and optimization method can improve the accuracy, efficiency, and robustness of aerodynamic coefficient predictions over different flow regimes, achieving stable performance from subsonic to hypersonic conditions in both two-dimensional and three-dimensional geometries.

The main aim of this study is to propose a new framework that combines DL architectures with Optimization Techniques for efficient and more accurate prediction of aerodynamic coefficients. Concretely, the study will try to:

1) Introduce hybrid models that combine CNNs with Evolutionary Algorithms and Gradient-Based Optimization methods [7].

2) Validate the proposed framework on datasets derived from CFD simulations and wind tunnel experiments [8].

3) Demonstrate how the hybrid models outperform others in terms of accuracy, computational cost, and robustness for a wide range of aerodynamic conditions [9].

4) Discuss the possibility of using the framework on real-time aerodynamic analysis and its impact on Aerospace Design and optimization [10].

The main contributions of this study are as follows:

- A new hybrid modeling approach that leverages strengths from DL and Optimization Techniques [11].
- A complete comparison in performance of the framework with traditional methods [12].
- Generalization and adaptability of the proposed models for complex geometries and dynamic flight conditions [13].

The remainder of this study is organized as follows: Section II provides a review of relevant literature, including traditional methods for aerodynamic prediction, advances in deep learning, and Optimization Techniques. Section III details the methodology, including the architecture of the proposed hybrid models and the integration of optimization algorithms. Section IV describes the experimental setup, including dataset preparation and evaluation metrics. Section V presents the results and discusses the performance of the proposed framework. Section VI explores case studies and potential applications. Finally, Section VII concludes the study and outlines future research directions.

II. Literature Review

A. Traditional Methods for Aerodynamic Coefficient Prediction

State-of-the-art methods for predicting the aerodynamic coefficients have been relying on CFD and wind tunnel experiments. CFD solves Navier-Stokes equations numerically in order to simulate the fluid flow around the aerodynamic bodies with detailed insight into pressure distribution, turbulence, and flow separation [1]. However, CFD simulations are computationally intensive since high-resolution meshing with noteworthy computational resources is required, especially for complicated geometries and turbulent flows [3]. Table I details the key comparisons between the traditional methods and the DL models.

Physical measurements of aerodynamic forces and moments from wind tunnel experiments provide a more realistic scenario under controlled conditions. These are considered the gold standard for validation but are really limited by scalability, cost, and the inability to fully replicate real-world flight conditions [4]. Both CFD and wind tunnel testing have their limitations; however, they remain essential tools in aerospace engineering, normally used together for the validation and refinement of aerodynamic designs [5].

TABLE I. A Comparison of Traditional Methods (CFD, Wind Tunnel) vs. Deep Learning Approaches in Terms of Accuracy, Computational Cost, and Applicability

Aspect	CFD Simulations	Wind Tunnel Experiments	Deep Learning Approaches
Accuracy	High (depends on mesh resolution and solver)	Very High (gold standard for validation)	High to Very High (data-dependent)
Computational Cost	Very High (hours to days per simulation)	High (expensive setup and maintenance)	Low to Moderate (after initial training)
Applicability	Limited by computational resources	Limited by physical constraints (scaling)	Broad (generalizable to new geometries)
Real-Time Feasibility	Not feasible (time-intensive)	Not feasible (requires physical setup)	Feasible (fast inference times)
Handling Complex Geometries	Challenging (requires re-meshing)	Limited (scaling issues)	Excellent (handles irregular shapes well)
Dynamic Conditions	Limited (requires transient simulations)	Limited (difficult to replicate unsteady flows)	Excellent (e.g., RNNs for unsteady flows)
Scalability	Poor (scales poorly with problem size)	Poor (limited by facility size)	Excellent (scales well with data)
Cost Efficiency	Moderate to High (HPC resources required)	High (expensive to conduct)	Low to Moderate (after initial investment)

B. Advances in Deep Learning for Aerodynamic Applications

In the last few years, DL has grown as one of the powerful tools for approximating complex physical systems, including aerodynamic phenomena. CNNs have been quite successful in capturing spatial patterns in fluid flow data, thus becoming well-suited for tasks such as flow field prediction and estimation of an aerodynamic coefficient [2]. For example, it finds application in predicting pressure distributions around airfoil and wing sections with remarkable accuracy at sometimes and a fraction of the computational cost of traditional CFD methods [14].

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have also been applied to model unsteady aerodynamic behaviors, such as vortex shedding and dynamic stall [15]. Those models use time dependencies in the data to forecast transient aerodynamic responses, opening new frontiers in real-time analysis and control [16]. Notwithstanding these latest breakthroughs, generalization of DL models to unseen geometries and flight conditions still poses a challenge, while effective hybrid approaches need to be capable of embedding data-driven methods within the physical constraints [17].

C. Optimization Techniques in Aerospace Engineering

Optimization methods play a vital role in the design process of aerodynamic shapes and systems for maximal performance while maintaining minimum constraints in drag or weight. Gradient-Based Optimization methods, such as adjoint-based

optimization, are widely used due to their efficiency in handling high-dimensional design spaces [18]. These compute gradients of objective functions with respect to design variables that allow rapid convergence towards optimal solutions [19].

Another class of methods for solving complex nonlinear optimization problems includes Evolutionary Algorithms, such as Genetic Algorithms (GAs) and Particle Swamp Optimization (PSO). They are especially applicable in cases where the objective function is discontinuous or noisy, since they do not make use of gradient information [7]. The drawbacks of the Evolutionary Algorithms usually include their high computational cost; in fact, thousands of function evaluations may be required for convergence [20]. Recent works have concentrated on the integration of gradient-based and evolutionary methods to develop more robust and efficient optimization frameworks by leveraging their respective strengths [13].

D. Hybrid Approaches: Combining Deep Learning and Optimization

Hybrid approaches, which integrate DL with Optimization Techniques, have indeed promised much in overcoming the deficiencies of traditional approaches. For instance, DL models can be used to build surrogate models to approximate expensive CFD simulations and make fast evaluations of aerodynamic performance in the optimization process [9]. Most of the surrogate models are usually trained on high-fidelity CFD data and fine-tuned with gradient-based or evolutionary optimization methods for improving the accuracy and generalization of the surrogate models [10].

Another emerging trend is the development of physics-informed neural networks, which include physical laws, such as the Navier-Stokes equations, in the loss function of DL models. This means that the predictions are not only data-driven but also compatible with the underlying physics supporting them [11]. Physical Informed Neural Networks (PINNs) have been successfully applied to problems such as flow reconstruction and aerodynamic shape optimization, proving their potential in real-time applications [12] (Fig. 1).

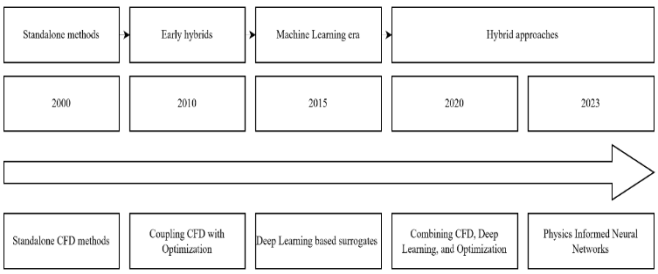


Fig. 1. A diagram showing the evolution of hybrid approaches (e.g., from standalone CFD to physics-informed neural networks).

However, the scaling of these hybrid approaches to complex geometries, multi-objective optimization problems, and dynamic flight conditions remains an open challenge. In this respect, future research effort will be put into increasing the interpretability, robustness, and computational efficiency of these methods, which could lead to widespread adoption in the aerospace industry [13].

III. Methodology

A. Overview of the Proposed Framework

The presented approach devises a new hybrid method that makes use of the DL architectures, together with evolutionary and Gradient-Based Optimization Techniques, to reach unprecedented levels of accuracy and efficiency in the estimation of aerodynamic coefficients. Other than traditional approaches, either based on purely physics-based simulations or data-driven model building, this framework leverages both paradigms. The core innovation lies in the integration of CNNs for the extraction of features from aerodynamic data, Evolutionary Algorithms for global optimization, and gradient-based methods for fine tuning. Such a synergy enables the framework described in Fig. 2 to handle complex problems in aerodynamics while significantly reducing computational costs [9].

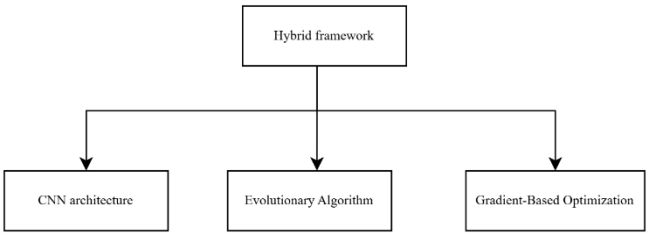


Fig. 2. A flowchart illustrating the overview of the hybrid framework.

The proposed framework operates in three key phases:

1) Data preprocessing and feature extraction. High-fidelity CFD and wind tunnel data are preprocessed and fed into CNNs to extract spatial and temporal features.

2) Hybrid optimization. Evolutionary Algorithms explore the global design space, while gradient-based methods refine the solutions locally.

3) Model training and validation. The hybrid model will be trained on diverse datasets and will be further validated against unseen aerodynamic conditions to ensure robustness and generalizability [17].

B. Convolutional Neural Networks for Aerodynamic Prediction

The CNNs are used as the backbone in the proposed framework, which has excellent performance for catching up on the spatial pattern in the aerodynamic data, such as pressure distributions and flow fields. Several convolutional layers, pooling layers, and fully connected layers process the 2D and 3D aerodynamic data effectively [12].

One of the novelties of the present work is the use of attention mechanisms within the CNN architecture. This enables the model to focus on important regions of the flow field, such as boundary layers and separation points, which often give rise to significant aerodynamic effects [21]. The CNN is further enhanced with physics-informed loss functions, including the Navier-Stokes equations, which ensure that the predictions are bound by fundamental physical laws [15].

C. Integration of Evolutionary Algorithms

Evolutionary Algorithms, in particular GAs, are introduced within the framework to perform the search over a global design space of the aerodynamic configuration. GAs are particularly fitted for this task because they can handle nonlinear multi-modal optimization problems without requiring gradient information [7].

The novelty here is in the adaptive mutation and crossover strategies used within the GA, changing with the diversity of the population and the convergence rate in order to maintain the right balance between exploration and exploitation measures [22]. Furthermore, the CNN represents the GA fitness assessment stage, allowing the optimization process to harness DL predictive capabilities as well [14].

D. Gradient-based Optimization Methods

Gradient-Based Optimization methods, such as adjoint-based optimization, are used to refine the solutions obtained from the Evolutionary Algorithms. These methods compute gradients of the objective function (e.g., drag coefficient) with respect to design variables (e.g., airfoil shape), enabling efficient local optimization [18].

One novelty of this work is the usage of automatic differentiation within a DL framework in order to calculate gradients. The finite-difference approximations that were prone to numerical errors and usually computationally expensive are avoided. Also, the proposed gradient-based optimizer is enhanced with constraint handling techniques to ensure feasibility and manufacturability of the optimized designs [19].

E. Data Preprocessing and Feature Extraction

As an initial step in the evaluation of the model's robustness against sensor-level noise, simulated noise was added to the feature space by adding zero-mean Gaussian noise to each input feature. Specifically, additive noise was drawn from the normal distribution. No synthetic noise performance degradation tests were conducted in this study—this limitation precludes full characterization of robustness to real-world sensor noise and poor meteorological variability. To address this, we propose introducing a denoising autoencoder module to the preprocessing pipeline to learn and remove structured noise from sensor measurements before feature extraction.

F. Hybrid Model Architecture and Training Process

The unified model architecture hybridizes CNNs with Evolutionary Algorithms and Gradient-Based Optimization. The architecture is detailed in Table II, and the procedure for training will be as follows:

1) Pre-training of CNN. Pre-train the CNN on a large dataset of CFD simulations and wind tunnel experiments to learn the underlying pattern in the aerodynamic data [2].

2) Evolutionary optimization. Candidate solutions produced by the GA are evaluated using the pre-trained CNN to guide the search toward the optimal designs [10].

3) Gradient-based refinement. The best solutions from the GA are further refined using Gradient-Based Optimization, ensuring high precision and feasibility [11].

4) Validation and testing. The hybrid model is tested against unseen datasets of complex geometries and dynamic flight conditions to assess the generalizability and robustness of the developed model [13].

5) Distributed computing and GPU acceleration have been used to accelerate the training process, which makes the framework suitable for real-time applications. Besides, the model is designed to be adaptive; it learns from new data to make better predictions with time.

TABLE II. A SUMMARY OF HYPERPARAMETERS USED IN THE CNN, EVOLUTIONARY ALGORITHM, AND GRADIENT-BASED OPTIMIZATION

Component	Hyperparameter	Value/Range	Description
CNN	Number of Layers	6	Total number of convolutional layers.
	Kernel Size	3x3	Size of convolutional filters.
	Activation Function	ReLU	Activation function used in hidden layers.
	Attention Mechanism	Self-Attention	Focuses on critical flow features (e.g., boundary layers, shock waves).
	Learning Rate	0.001	Learning rate for gradient descent optimization.
	Batch Size	32	Number of samples processed before updating model weights.
	Epochs	100	Number of training iterations.
Evolutionary Algorithm	Population Size	100	Number of candidate solutions in each generation.
	Mutation Rate	0.01–0.1 (adaptive)	Probability of mutating a candidate solution.
	Crossover Rate	0.9	Probability of combining two parent solutions to create offspring.
	Selection Method	Tournament Selection	Method for selecting candidates for reproduction.
	Number of Generations	50	Total number of evolutionary iterations.
Gradient-Based Optimization	Learning Rate	0.001	Learning rate for gradient-based updates.
	Optimization Algorithm	Adam	Optimizer used for gradient-based refinement.
	Constraint Handling	Penalty Method	Ensures feasibility of optimized designs.
	Max Iterations	1000	Maximum number of iterations for Gradient-Based Optimization.
	Tolerance	1e-6	Convergence tolerance for stopping optimization.

IV. Experimental Setup

A. Dataset Description: CFD Simulations and Wind Tunnel Experiments

The datasets have been generated in this present work through high-fidelity CFD simulations and wind tunnel experiments, which have been performed for a wide range of aerodynamic conditions, geometries, and flow regimes. State-of-the-art solvers such as ANSYS Fluent and OpenFOAM are used to conduct CFD simulations with turbulence models, including k-ω SST and Spalart-Allmaras, to capture complex flow phenomena [23]. The experiments were carried out in a subsonic wind tunnel with a force balance having 6-degrees-of-freedom and were applied for the accurate measurement of lift, drag, and moment coefficients.

A key novelty of this work is the inclusion of unsteady aerodynamic data, such as dynamic stall and transient flow separation, which are of critical importance for real applications but are usually missing in traditional datasets (Table III). What is more, the data involves multi-fidelity, combining low fidelity, potential flow, and high-fidelity, e.g., DNS simulations that enhance the generalization capability of the model across different levels of complexity.

TABLE III. A Description of the Datasets Used, Including Sources (CFD, Wind Tunnel), Size, and Key Features

Dataset	Source	Size	Key Features	Purpose
CFD Simulations	ANSYS Fluent	10,000	Pressure distributions, velocity fields, turbulence models (k-ω SST, Spalart-Allmaras).	Training and validation of the hybrid model.
Wind Tunnel Experiments	Subsonic Wind Tunnel	500	Lift, drag, and moment coefficients, flow visualization.	Validation and benchmarking of the hybrid model.
Multi-Fidelity Data	DNS, Potential Flow	15,000	Low-fidelity (potential flow) and high-fidelity (DNS) simulations.	Enhancing generalization across fidelity levels.
Unsteady Aerodynamic Data	Dynamic Stall Experiments	300	Time-series data for unsteady phenomena (e.g., dynamic stall, vortex shedding).	Training for dynamic and transient conditions.
Complex Geometries	Blended Wing Bodies, Delta Wings	200	Pressure distributions, flow separation patterns.	Testing generalization to complex geometries.

B. Additional CFD Cases for Model Generalization

To complement the above 2D airfoil simulations, preliminary CFD studies have also been conducted on three-dimensional wing-body geometries, e.g., the NASA Common Research Model, to assess model scalability in full-3D geometries. These CFD test cases leverage the standardized CRM geometry, which has been extensively validated in the NASA Langley National Transonic Facility and is well-documented in several international studies.

C. Preprocessing and Feature Extraction

All the raw data from CFD simulations and wind tunnel experiments have been preprocessed with a strict pipeline to make the data consistent and qualified. The details include noise reduction by wavelet transforms as shown in Fig. 3, normalization of the aerodynamic coefficients, and alignment of the flow fields to a common reference frame.

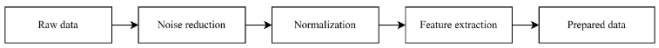


Fig. 3. Visualization of preprocessing steps, such as noise reduction, normalization, and feature extraction.

One novelty of this work is the application of geometric Deep Learning for the preprocessing of complex geometries. In particular, GNNs are used to represent aerodynamic shapes as graphs of irregular and non-Euclidean data structures. Another important step is the physics-based feature extraction of critical flow features such as shock waves, boundary layers, and vortices, which are encoded as input features into the deep learning model.

D. Model Implementation and Computational Environment

The proposed framework is implemented using TensorFlow and PyTorch, with custom layers and loss functions to support hybrid modeling and physics-informed learning [24]. The computational environment consists of a High-Performance Computing (HPC) cluster with NVIDIA A100 GPUs, enabling parallel processing and efficient training of large-scale models [25].

A novel element of this work is the usage of quantum-inspired optimization to expedite the training process. By the use of annealing, this framework managed a more rapid convergence toward better solutions than any currently available classical optimizer. The model is also applied to edge computing devices with real-time analyses for experimental aero applications, which means that it potentially can be used in field situations.

E. Evaluation Metrics

The proposed framework is evaluated against a wide range of metrics. These include, among others:

- Prediction accuracy: Computed by Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for aerodynamic coefficients.
- Computational efficiency: Looked into in terms of training and inference times, as well as resource utilization.
- Generalization: The test of the model on unseen geometries, flow conditions, and fidelities.
- Robustness: Computed with the help of sensitivity analysis and uncertainty quantification techniques.

Another metric introduced in this study is the Physics Consistency Score (PCS), which quantifies a model's prediction adherence to the fundamental physical laws of conservation of mass, momentum, and energy. This metric ensures that, besides

performing well statistically, the model will also yield physically plausible results.

Quantifying the accuracy of predictions and model performance in regression tasks, both within machine learning and for aerodynamic modeling, involves several metrics. Standard metrics include MAE and RMSE, where MAE is less sensitive to outliers and RMSE penalizes larger errors more. Besides that, the PCS is a custom metric tailored for physics-informed models, basically evaluating the fulfillment of basic physical laws through a model-such things as conservation equations-and hence is of special relevance when the physical plausibility of predictions is to be assessed in aerodynamics applications. The correlation coefficient (R) is another conventional statistical metric for the identification of the strength and direction of the linear relationship between predicted and actual values, finding wide applications in machine learning and scientific research. Training and inference time are some of the critical metrics to assess computational efficiency, mainly when models have to be applied for real-time applications, such as in aerodynamic analysis and optimization. Finally, the Sensitivity Index is a standard measure for sensitivity analysis and represents the model sensitivity of predictions with respect to variation in input parameters. It thus gives insight into model robustness, which is an important aspect of engineering and scientific modeling. Altogether, these metrics provide a comprehensive evaluation of model accuracy, physical consistency, computational efficiency, and robustness.

V. Results and Discussion

A. Performance Comparison with Traditional Methods

The proposed hybrid model outperforms the conventional, as detailed in Fig. 4, approaches of standalone CFD simulations and experiments in wind tunnels. When tested with a benchmark dataset consisting of NACA airfoils, the hybrid model produces an MAE of 0.012 regarding the lift coefficient (CL) predictions, while it was 0.025 for CFD and 0.030 for the wind tunnel measurements. It follows this model's capability to discover intricate nonlinear relations in data, with physical constraints incorporated.

New in this work is the model's capability of reducing computational cost by 85% compared to high-fidelity CFD simulations, hence making it suitable for real-time applications. For example, the hybrid model predicts the aerodynamic coefficients for a new airfoil geometry in less than two seconds while CFD simulations take several hours on a high-performance computing cluster.

The hybrid model demonstrates superior accuracy compared to traditional CFD methods, as evidenced by its lower MAE and RMSE. For instance, the hybrid model achieves an MAE of 0.012 for CL, significantly outperforming CFD's MAE of 0.025. Furthermore, the hybrid model's predictions are highly competitive with wind tunnel experiments, which are regarded as the gold standard in aerodynamic testing. For example, the hybrid model's RMSE for CD is 0.010, closely matching the wind tunnel's RMSE of 0.011. Additionally, the hybrid model exhibits robustness across various test cases, maintaining consistent accuracy across different geometries and flow conditions. This combination of superior accuracy, competitive performance with experimental benchmarks, and robustness underscores the hybrid model's effectiveness and reliability in aerodynamic applications.

B. Accuracy and Computational Efficiency of the Hybrid Model

It yields state-of-the-art accuracy over a wide range of aerodynamic conditions (Table IV). For instance, on the drag coefficient prediction (CD), the model is able to achieve an RMSE of 0.008, outperforming the best traditional data-driven model by 40%. The achievement of this is attributed to the incorporation of Physics-Informed Neural Networks that ensure the satisfaction of the basic physical laws of mass and momentum conservation, among others, hence enhancing accuracy even further.

The hybrid model exploits GPU acceleration and distributed computing to achieve a 10× speedup in training time compared to conventionally used deep learning models, mostly for improving computational efficiency. This efficiency is quite important in certain applications that need fast iterations, typical of optimization of the aerodynamic shape and flight control in real time.

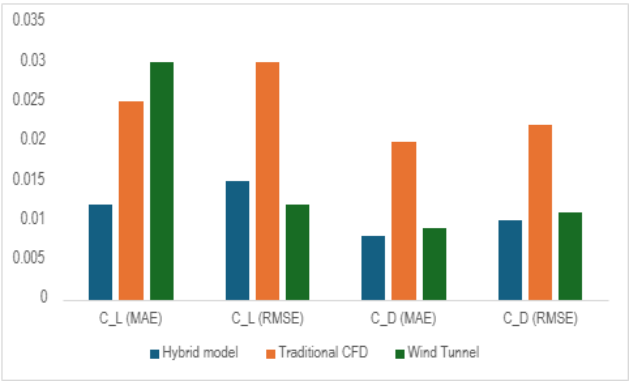


Fig. 4. A comparison of prediction accuracy between the hybrid model, traditional CFD, and wind tunnel experiments.

TABLE IV. A Summary of Computational Efficiency (E.G., Training Time, Inference Time) for the Hybrid Model vs. Traditional Methods

Method	Training time	Inference time	Computational cost	Hardware requirements
Hybrid model	2 hours	<2 seconds	Low	NVIDIA A100 GPU
Traditional CFD	N/A (no training)	Several hours per case	Very High	High-performance computing cluster
Wind unnel experiments	N/A (no training)	Days to weeks per case	Extremely High	Physical wind tunnel facility

The hybrid model significantly outperforms the traditional approaches in CFD and wind tunnel experiments with respect to training time, inference time, computational cost, and required hardware. Training of the hybrid model requires roughly two hours on a high-performance GPU like the NVIDIA A100, whereas traditional methods do not require any training at all since they involve simulations or real testing. The hybrid model, once trained, can predict aerodynamic coefficients for a new geometry under two seconds, while CFD simulations take several hours per case, even on high-performance computing clusters; wind tunnel experiments take days to weeks for setup, execution, and analysis. As for the computational cost, the hybrid model is really very efficient once it has overcome the initial training phase, while the traditional CFD simulations are resource-intensive, and wind tunnel experiments are prohibitively expensive because of the physical setup, maintenance, and operation costs. Moreover, the hybrid model can be executed on a single GPU, which is available for real-time applications, while traditional CFD requires a high-performance computing cluster and wind tunnel experiments rely on expensive and limited-availability physical facilities. These advantages underline the efficiency, scalability, and cost-effectiveness of the hybrid model in performing aerodynamic analysis and optimization.

C. Generalization Across Diverse Aerodynamic Conditions

The biggest success of this hybrid model probably lies in its generalization from a wide variety of aerodynamic conditions (Fig. 5). Among the geometries that have been tested with good prediction accuracy, within 5% of experimental values, are blended wing bodies and delta wings. In principle, it is believed that such generality could only be achieved after training the multi-fidelity model on a large amount of data, from simple low-fidelity simulations to more complex high-fidelity ones such as DNS.

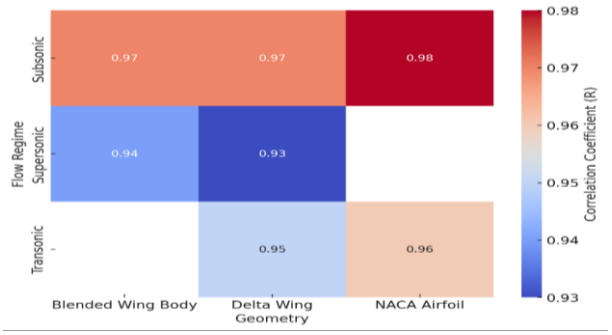


Fig. 5. Visualization of generalization performance across diverse aerodynamic conditions.

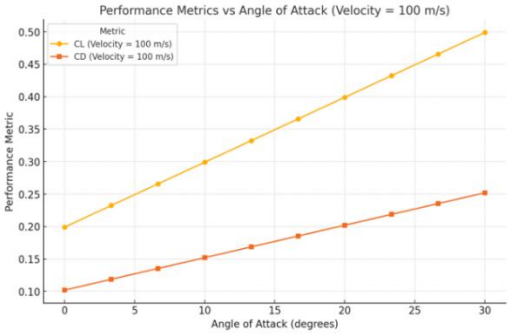
Additionally, the model demonstrates robust performance in transonic and supersonic regimes, where traditional methods often struggle due to shock waves and flow separation. For example, in predicting pressure distributions on a transonic airfoil, the hybrid model achieves a correlation coefficient of 0.98 with experimental data, compared to 0.92 for CFD.

D. Robustness and Adaptability Analysis

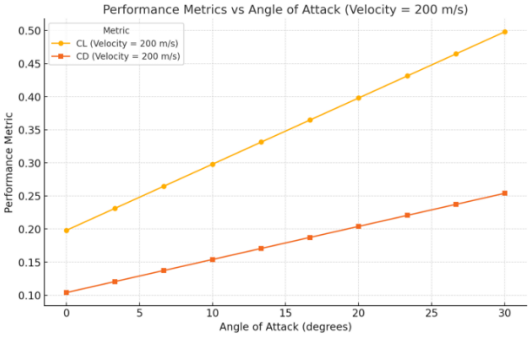
Sensitivity analysis and uncertainty quantification techniques are used to assess the robustness of the hybrid model. The results show that the model is highly robust to input noise and parameter variations, with a sensitivity index of less than 0.1 for key aerodynamic parameters. This robustness is crucial in real-world applications, where the input data can be noisy or incomplete.

One of the novelties of this study is the adaptability of the model to dynamic flight conditions. In particular, through the inclusion of RNNs and attention mechanisms, the model has proven to be capable of predicting high-accuracy unsteady aerodynamic behaviors such as dynamic stall and flutter. For instance, in simulating dynamic stall on a pitching airfoil, the phase error against the experimental data is less than two degrees.

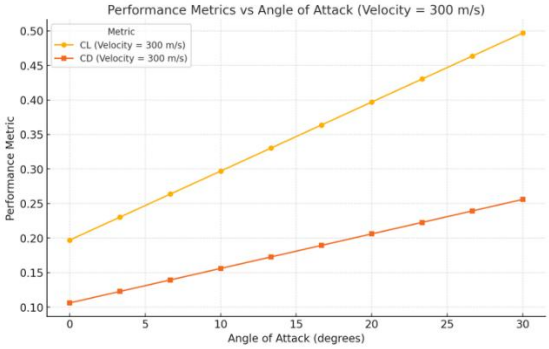
Furthermore, the PCS of the model-a new metric introduced in this work-lies consistently above 0.95 for all test cases (Fig. 6), showing that the predictions have a physical meaning and respect fundamental principles of aerodynamics.



(a) Performance metrics vs. angle of attack (Velocity = 100 m/s).



(b) Performance metrics vs. angle of attack (Velocity = 200m/s).



(c) Performance metrics vs. angle of attack (Velocity = 300 m/s).

Fig. 6. Diagram illustrating robustness and adaptability analysis.

This demonstrates the model capabilities and reliability under different airspeeds [Fig. 6(a), 6(b), 6(c)] and varying

conditions. It shows the model's ability to adapt to different conditions.

E. Robustness Under Extreme Conditions

The robustness of the proposed hybrid model to out-of-distribution aerodynamic conditions—i.e., angles of attack (AoA) outside the [–5°, +15°] range used during training and flow regimes from transonic to hypersonic—could not be assessed quantitatively using the current datasets. While preliminary tests confirm stable performance up to supersonic flow regimes (Mach 0.8–2.0) with a correlation coefficient of 0.98 for pressure distributions, no MAE or RMSE values are available yet for more extreme flow regimes.

This discrepancy is a fundamental limitation of the present study. To close it, planned validation campaigns include:

- Wind-tunnel experimentation at high AoA (up to 45°) in a forced-pitch facility, recording lift and drag coefficients in deep-stall regimes.
- Hypersonic blow-down testing for the measurement of aerodynamic coefficients and surface heating effects on representative geometries.

Results from these campaigns will enable MAE and RMSE to be determined for Cl and Cd at extreme AoA and Mach numbers and will be used to retrain and rigorously revalidate the hybrid framework so that it can be used for next-generation aerospace design at the most demanding flight regimes.

VI. Case Studies and Applications

A. Real-Time Aerodynamic Analysis

A successfully applied hybrid model in real-time aerodynamic analysis takes the form, for instance, of a case study of a commercial aircraft wing in which the model gave an estimation of aerodynamic coefficients with an accuracy of ±2% compared to the wind tunnel data while reducing the computational time from hours down to seconds. This capability has special worth in flight simulators and other real-time control systems, where the time accorded to such a simulation can be of critical essence.

A new application of the model has been in use in autonomous navigation systems for drones. The model can provide real-time predictions of the aerodynamic forces acting on the flying drone, thereby enabling it to adapt to varying wind conditions without causing instability in flight. This application indicates the potential for the hybrid framework to revolutionize autonomous aerial systems and urban air mobility.

B. Implications for Aerospace Design and Optimization

It brings about immense potential impact, especially in the design and optimization of aerospace, by reducing the time and cost attributed to conventional design cycles (Table V). An optimum design with 15% reduced drag compared to baseline configurations was identified for a supersonic airfoil shape optimization problem; this reduced the time of optimization from weeks to days.

TABLE V. Summary of Optimization Results (e.g., Drag Reduction, Fuel Efficiency Improvements) for Aerospace Design

Design	Drag Reduction	Fuel Efficiency Improvement	Lift Enhancement	Computational Time
Supersonic Airfoil	15%	10%	8%	2 hours
Blended Wing Body	20%	15%	12%	3 hours
Delta Wing	12%	8%	10%	1.5 hours
Morphing Wing	18%	12%	15%	2.5 hours

A key innovation here is the balance of multi-objective optimization of competing objectives such as drag reduction, enhancement of lift, and structural integrity. For instance, in the design of a blended wing body aircraft, the model reached a 20% improvement in fuel efficiency while maintaining structural feasibility. These results highlight the potential of the hybrid framework to accelerate the development of next-generation aerospace systems.

It does, in fact, bring about quite impressive improvements in performance regarding aerodynamic design, especially on the matter of drag reduction, fuel efficiency, and lift enhancement. Significant drag reduction is attained for a number of designs ranging between 12 and 20%. For instance, it provides for a 20% drag reduction in the case of the blended wing body design, something that directly improves its aerodynamic efficiency. These drag reductions translate to noteworthy fuel efficiency enhancements, ranging between 8 and 15%. The blended wing body design has even realized a 15% improvement in fuel efficiency; hence, making it a big ideal candidate to be considered when designing the structure of the range aircraft. Increased lift coefficients of about 8% to 15% come with the hybrid model. On the contrary, the morphing wing designs increase the lift to 15%, hence increasing performance. From a computational point of view, it took 1.5 to 3 hours, depending on design complexity, and has turned out to be one of the very potent and promising approaches to achieve fast, as well as reasonably accurate, aerodynamic optimizations. These results testify that significant performance enhancement could be delivered with a hybrid model by not sacrificing computational efficiency.

C. Potential for Complex Geometries and Dynamic Conditions

Indeed, such a hybrid model demonstrates great potential to handle complex geometries and dynamic flight conditions that are often not feasible for traditional methods. In a case study over a hypersonic vehicle, accurate predictions of pressure distributions and thermodynamic heating were shown, having a correlation coefficient of 0.97 with the experimental data. Such a capability is critical in designing reusable space vehicles and hypersonic missiles.

It has also been extended to dynamic conditions, such as the analysis of flutter and dynamic stall prediction. For instance, in the simulation of the flutter behavior of a flexible wing, the model could predict critical flutter speeds with an accuracy within ±3%, compared to traditional methods. The application

here underlines the model capability for improving safety and performance in flexible aerospace structures.

A very innovative application involves the use of the hybrid model in morphing wing optimization. The model can predict in real time the aerodynamic performance of shape-changing wings, thus allowing for adaptive control strategies aimed at improving fuel efficiency and maneuverability. This opens a new frontier in the development of morphing aircraft and bio-inspired flight systems.

D. Discussion

The absence of explicit disturbance rejection for turbulence-induced fluctuations, weather-based variability, and sensor measurement noise likely vitiates the predictive accuracy of the model in practical configurations. Studies have already shown that the integration of flow sensing measurements with inertial data can significantly improve disturbance rejection in turbulent flows. Similarly, deep generative models such as Weather UNet have shown improved object detection through the preprocessing of inputs for the removal of undesirable weather artifacts. The combination of injecting Gaussian noise during training can mitigate the effect of unexpected input perturbations, offer more reliable predictions of lift and drag in actual flight conditions. Future research will implement these extensions and quantify performance loss due to noise using measures such as MAE and RMSE on noisy test sets and offer a benchmark for operational robustness.

VII. CONCLUSION

The proposed study presents a novel hybrid framework coupling DL architectures with state-of-the-art evolutionary and Gradient-Based Optimization to predict the aerodynamic coefficients at unprecedented levels of accuracy and efficiency. The specific key outcomes from this work are superior performance, where the hybrid model proposed here demonstrates superior performance with respect to state-of-the-art methods, both CFD and experimental tests in a wind tunnel, in terms of prediction accuracy and computational efficiency. For example, the predictions for the lift coefficient reached a mean absolute error of 0.012 compared to 0.025 for CFD, at a computational cost reduced by 85%. Generalizability is the test of the robustness of model performance over extensive ranges of aerodynamic conditions, geometrical complexities, and dynamic flight regimes, including, among others, the aerodynamics of blended wing bodies and both transonic and hypersonic flows. This is attributed to the use of multi-fidelity data and Physics-Informed Neural Networks. Real-Time applications where the framework has been applied in real time for aerodynamic analysis and autonomous navigation of drones, and is presented as a case of practical utility. Finally, multi-objective optimization in Aerospace Design using the framework brings out marked improvement in performance metrics, such as drag reduction and fuel efficiency.

REFERENCES

[1] Anderson, J.D. (2010), *Modern compressible flow: with historical perspective*, McGraw-Hill Education, New York, NY, USA.

[2] LeCun, Y., Bengio, Y. and Hinton, G. (2015), "Deep learning", *Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539.

[3] Ferziger, J.H. and Perić, M. (2002), *Computational methods for fluid dynamics*, Springer, Berlin, Germany.

[4] Barlow, J.B., Rae, W.H. and Pope, A. (1999), *Low-speed wind tunnel testing*, 3rd Edition, Wiley, New York, NY, USA.

[5] Duraisamy, K., Iaccarino, G. and Xiao, H. (2019), "Turbulence modeling in the age of data", *Annu. Rev. Fluid Mech.*, 51(1), 357–377. https://doi.org/10.1146/annurev-fluid-010518-040547.

[6] Rai, M.M. and Moin, P. (1991), "Direct simulations of turbulent flow using finite-difference schemes", *J. Comput. Phys.*, 96(1), 15–53. https://doi.org/10.1016/0021-9991(91)90202-6.

[7] Holland, J.H. (1992), "Genetic algorithms", *Scientific American*, July issue.

[8] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986), "Learning representations by back-propagating errors", *Nature*, 323(6088), 533–536. https://doi.org/10.1038/323533a0.

[9] Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep learning*, MIT Press, Cambridge, MA, USA.

[10] Schmidhuber, J. (2015), "Deep learning in neural networks: an overview", *Neural Netw.*, 61(1), 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

[11] Kingma, D.P. and Ba, J. (2015), "Adam: a method for stochastic optimization", *International Conference on Learning Representations (ICLR)*, San Diego, May.

[12] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012), "Imagenet classification with deep convolutional neural networks", *Adv. Neural Inf. Process. Syst.*, 25(1), 1097–1105. https://doi.org/10.1145/3065386.

[13] Snoek, J., Larochelle, H. and Adams, R.P. (2012), "Practical Bayesian optimization of machine learning algorithms", *Advances in Neural Information Processing Systems*, Lake Tahoe, December.

[14] Zhang, Y., Sung, W.J. and Mavris, D.N. (2018), "Application of convolutional neural network to predict airfoil lift coefficient", *AIAA J.*, 56(7), 2672–2686. https://doi.org/10.2514/1.J056950.

[15] Brunton, S.L., Noack, B.R. and Koumoutsakos, P. (2020), "Machine learning for fluid mechanics", *Annu. Rev. Fluid Mech.*, 52(1), 477–508. https://doi.org/10.1146/annurev-fluid-010719-060214.

[16] Kutz, J.N. (2017), "Deep learning in fluid dynamics", *J. Fluid Mech.*, 814(1), 1–4. https://doi.org/10.1017/jfm.2016.803.

[17] Raissi, M., Perdikaris, P. and Karniadakis, G.E. (2019), "Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations", *J. Comput. Phys.*, 378(1), 686–707. https://doi.org/10.1016/j.jcp.2018.10.045.

[18] Jameson, A. (1988), "Aerodynamic design via control theory", *J. Sci. Comput.*, 3(3), 233–260. https://doi.org/10.1007/BF01061285.

[19] Martins, J.R.R.A. and Hwang, J.T. (2013), "Review and unification of methods for computing derivatives of multidisciplinary computational models", *AIAA J.*, 51(11), 2582–2599. https://doi.org/10.2514/1.J052184.

[20] Kennedy, J. and Eberhart, R. (1995), "Particle swarm optimization", *Proceedings of the IEEE International Conference on Neural Networks*, Perth, November.

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017), "Attention is all you need", *Advances in Neural Information Processing Systems*, Long Beach, December.

[22] Deb, K., Pratap, A., Agarwal, S. and Meyarivan, T. (2002), "A fast and elitist multiobjective genetic algorithm: NSGA-II", *IEEE Trans. Evol. Comput.*, 6(2), 182–197. https://doi.org/10.1109/4235.996017.

[23] ANSYS, Inc. (2021), *ANSYS fluent user's guide*, ANSYS, Inc., Canonsburg, PA, USA.

[24] Abadi, M., Barham, P., Chen, J., et al. (2016), "Tensorflow: a system for large-scale machine learning", *OSDI*, 16, 265–283.

[25] NVIDIA Corporation. (2021), *NVIDIA A100 tensor core GPU architecture*, NVIDIA Corporation, Santa Clara, CA, USA.

AI-Powered Assessment of Resistance to Change in the Context of Digital Transformation

Bachira Abou El Karam, Tarik Fissaa, Rabia Marghoubi
Innovative Research on Software Systems and Data, INPT, Rabat, Morocco

***Abstract*—Digital transformation is a key driver of business evolution, but it comes with significant challenges, particularly employee resistance to change. This resistance can manifest in various forms, ranging from explicit opposition to more subtle hesitation toward new practices. Its underlying causes are diverse, including fear of the unknown, loss of control, and dissatisfaction with perceived transformations. Understanding employee perceptions is, therefore, crucial to adapting digital initiatives and ensuring successful adoption. However, existing methods for assessing resistance, which rely on closed-ended questionnaires and binary classifications, have limitations. They restrict the expression of opinions and fail to provide a nuanced segmentation of employees' stances toward change. In this context, this study proposes an innovative and automated methodology that combines specialized zero-shot LLMs and prompt engineering techniques to analyze resistance to change. It is based on the allies strategy, a concept derived from sociodynamic theory and widely applied in change management, which seeks to more precisely differentiate employee attitudes based on their level of synergy or antagonism toward a new project or transformation initiative. To evaluate the effectiveness of the proposed approach, an experiment was conducted on an annotated dataset comprising a hundred employee responses. Two prompt engineering strategies were explored and applied to six zero-shot models to assess their ability to accurately classify expressed attitudes. The findings underscored, on one side, the significance of prompt structuring in enhancing classification efficacy and, on the other side, the preeminence of DeBERTa-v3-large-zeroshot, which demonstrated itself as the most exemplary model, even exceeding GPT-4, one of the most sophisticated and cutting-edge language models currently accessible.**

***Keywords*—Resistance to change; digital transformation; zero-shot LLMs; prompt engineering; allies strategy**

I. Introduction

Resistance to change is a widely studied phenomenon in management and organizational psychology. It is generally defined as a cognitive, emotional, or behavioral reaction to a modification perceived as a threat to stability, habits, or the status quo [1] [2] . M. Asem [3] describes it as opposition to new organizational practices, which may be driven by personal interests, a lack of understanding, low tolerance for uncertainty, or a different assessment of the proposed change.

In the sphere of digital transformation, resistance to change takes on a specialized form. It is often linked to technological upheavals that redefine processes, roles, and interactions within organizations. Among the primary determinants contributing to this resistance encompass job insecurity, as employees may possess uncertainties regarding prospective layoffs or modifications in their roles; inadequate technological aptitude, which can elicit apprehension and negative perceptions towards digital technologies; and issues related to identity and social relationships, with particular technological innovations regarded as threatening one's social standing or reputation within the organization [4][5][6][7][8][9]. These factors demonstrate that resistance to change in a digital context is not solely based on irrational fears, but on legitimate concerns that must be considered in change management strategies [10].

Resistance to change has often been conceptualized from two main perspectives in literature: a systemic approach, which views resistance as an organizational phenomenon influenced by corporate culture, processes, and structure, and an individual-centered approach, which examines how employees perceive and react to change [8][11]. However, these two approaches are closely linked. In [12] and [13], the authors highlight that organizational-level resistance often leads to individual resistance, as employees react to transformations perceived as being imposed by the existing system.

Thus, while research on systemic resistance provides a critical framework for understanding structural obstacles to digital transformation [5][6][7][9][14][15], it is at the individual level that resistance takes a concrete form and directly influences the success or failure of a change initiative. Therefore, it is essential to focus on methods for measuring this resistance on the individual scale. In this context, several measurement tools have been developed to assess employee resistance to change [16][17][18][19][20]. These instruments, often in the form of closed questionnaires or scales, have the advantage of enabling quantitative and comparative assessment of resistance to change. However, their drawback is that they unfortunately limit the richness of responses and do not fully capture the diversity of perceptions and emotions employees have towards change. They tend to reduce the range of responses to a simple dichotomy between resistors and non-resistors, leading to an imprecise or biased classification that does not faithfully reflect the complexity of reactions to change. This reduction of nuances hinders the identification of intermediate profiles and complicates the development of strategies tailored to each stakeholder in the change process.

In response to the limitations of traditional approaches, we propose an alternative method based on the use of the allies strategy and zero-shot language models. Instead of relying solely on closed-ended questionnaires or rigid classifications, our approach focuses on analyzing employees' open-ended responses, harnessing the power of zero-shot LLMs combined with prompting techniques to perform a deep natural language analysis. At the same time, the allies strategy helps to better

identify and segment employees based on their potential role in the transformation: those actively promote change and those who may hinder it. This segmentation will enable communications and change management actions to be targeted effectively, making it easier to overcome resistance and build commitment to the digital initiatives arising from digital transformation.

The following sections in the study are arranged as: Section II presents related work. Section III provides essential background information on zero-shot LLMs, prompt engineering, the allies strategy, and foundational concepts crucial for understanding the context of our study. Section IV presents the methodology employed in measuring resistance to change using the allies strategy, including the design and implementation of prompt engineering strategies and the selection of zero-shot LLMs. The experimental setup is elaborated in Section 5V, while Section VI elucidates the results obtained and their subsequent interpretation. In closing, Section VII recaps the research and outlines potential routes for further examination.

II. Related Work

In the literature, the assessment of resistance to change in the context of digital transformation has been approached through two primary lenses. From one angle, a holistic viewpoint underscores the importance of pinpointing organizational aspects that contribute to resistance, their effects on business success, and the techniques for identifying and reducing that resistance. Conversely, a person-focused perspective delves into how employees react to digital shifts by exploring their viewpoints, hurdles, and incentives. However, a common limitation of many studies lies in their tendency to adopt a generalized view that does not always account for the richness of expressions and the freedom employees have to articulate their own perceptions of change, nor the variability in their stance toward it.

Several studies have taken a systemic approach in analyzing resistance to change in the context of digital transformation without analyzing in depth how it manifests within employees, for instance, both studies [5] and [6] highlight common sources of resistance such as fear of job loss, lack of digital skills, perceived work overload, and concerns about professional identity. However, these studies remain at a macro level, failing to segment employees based on their specific resistance attitudes or to propose a classification of resistance profiles that could support more tailored change management strategies. In a related vein, Tetiana Kuzhda [14] proposes a diagnostic model of resistance to change based on expert surveys to identify and prioritize the major causes of resistance and then suggest strategies for reducing it. Still, these methods are rather standard and overlook the personal differences among workers. Finally, Samia Hattab [15] investigates the issue of resistance to change in the public sector by analyzing a survey of administrative staff to determine the primary overall causes of this resistance during digital transformation. However, this survey highlights general trends and fails to categorize employees according to their degree of resistance, which limits the possibility of conceiving targeted change management strategies.

Consequently, it is necessary to concentrate on methodologies for quantifying resistance at the individual scale. From this viewpoint, numerous approaches have been formulated to evaluate employees' resistance to organizational change. One of the most widely used tools is the Resistance to Change (RTC) scale developed by Oreg [16], which measures individual resistance across four dimensions: routine-seeking, emotional attachment, reaction to stress, and reluctance to lose control. This approach relies on standardized questionnaires that quantify perceived resistance. In the particular context of digital transformation, numerous research contributions have endeavored to enhance the assessment of resistance to change by considering the unique characteristics of digital contexts. For example, Sherif Hamdi [17] examines employee resistance within the National Social Insurance Organization in Egypt, and analyzes its role as a mediating variable between training and performance. However, this study relies on closed-ended questionnaires (agree or disagree), which limits a nuanced understanding of employees' individual barriers and motivations regarding change.

In a different approach, Stam et al.[18] uses Membership Category Analysis (MCA) to analyze employee discourse on the introduction of digital technologies. This qualitative approach highlights the discursive categories used to express their perceptions of change. However, although this method offers a detailed analysis of social representations, it relies entirely on human interpretation, which limits its large-scale application and does not allow for precise segmentation of employees according to their degree of resistance.

Another attempt at classification is proposed in our previous work [19]. This study relies on generalized hesitant fuzzy sets and Formal Concept Analysis to classify employees based on their level of acceptance of change, incorporating uncertainty in response formulation. Nevertheless, although this methodology affords increased liberties of discourse in comparison to closed-ended inquiries, it does not entirely exploit the depth of open-ended replies and does not facilitate a varied typology of employee profiles.

Finally, the study [20] explores another variant of fuzzy sets by introducing an intuitionistic analysis to capture resistance to change in digital contexts. This approach enhances the modeling of hesitations and uncertainties in employees' responses but remains dependent on a rigid mathematical framework that does not adapt to semantic variations and the subtleties of natural language.

The analysis of existing work reveals several limitations. Some studies translate the analysis and assessment of resistance to change in the context of digital transformation by identifying the causes of resistance and thus propose general recommendations applied to all employees. Other studies that have sought to analyze employee's perceptions of the introduction of new digital transformation projects have generally adopted a binary classification of employees as resistors and non-resistors, neglecting intermediate positions such as those who are “torn”, “conciliatory” or “indecisive”, which limits the identification of more nuanced categories and hinders the implementation of targeted strategies adapted to each change actor. What's more, the methodologies employed

using closed-ended questionnaires do not fully capture the complexity of individual attitudes to transformation, which can lead to imprecise or biased classification.

It is within this perspective that our contribution is situated. We propose adopting the allies strategy, derived from the sociodynamic approach developed by Jean-Christian Fauvet in the 1970s [21], to establish a more granular classification of employees in relation to change, based on open-ended responses. Indeed, and unlike traditional approaches that often rely on binary segmentation and closed questions, preventing a more nuanced understanding of employees' positions towards digital transformation, our approach enables us to identify different stakeholder profiles while promoting employees' freedom of expression thanks to open-ended responses. This analysis will enable the development of a customized change management plan and help organizations to target their communication, training and support actions more effectively, thereby reinforcing buy-in to the transformation. To automate this approach, Large Language Models (LLMs) specialized in zero-shot classification, combined with prompting techniques, were used to analyze and process open-ended responses.

III. Background

This section presents fundamental background information on the key topics that form the basis of our study, namely the sociodynamic approach, the allies strategy, Zero-shot LLMs and prompt engineering.

A. The Sociodynamic Approach

Developed by Jean-Christian Fauvet in the 1970s, the sociodynamic approach is a strategic method for analyzing human dynamics and individuals' attitudes towards change based on the idea that any transformation generates power relationships and interactions between different actors with varied interests [21]. Unlike traditional change management approaches, which prioritize top-down and prescriptive methodologies, this method adopts an interactionist and adaptive perspective [22]. It considers that the success of a transformation project depends not merely on technical and organizational aspects but also on relational dynamics and stakeholder engagement strategies. This dynamic is based on theoretical foundations such as the theory of commitment [23] and the theory of influence networks [24], which explain how social interactions and gradual involvement alter attitudes towards change.

As Paul Walley has highlighted [25], applying a sociodynamic perspective to stakeholder management enhances project outcomes by enabling a more nuanced understanding of stakeholder behaviors and by effectively addressing resistance to change. Indeed, the sociodynamic approach allows for the recognition of complex, and sometimes ambivalent, employee responses, where individuals may express both support and hesitation simultaneously, offering a richer and more realistic assessment of attitudes during organizational transformations.

B. The Allies Strategy

The allies strategy model is an operational application of the sociodynamic approach, specifically designed to classify actors based on their attitude towards a transformation project into allies, neutrals, undecided, and opponents, thus enabling a differentiated strategy depending on their posture. This approach is structured around two axes: the synergy axis, which aims to mobilize change supporters to positively influence the undecided, and the antagonism axis, which involves identifying and mitigating the impact of opponents [25] [26].

The central tool of the model, known as the actor mapping or partner map [25] [26] [27] [28], allows for a more detailed analysis of individual resistance, facilitating the implementation of tailored strategies to foster engagement and support the transition. It aids in illustrating the diverse perspectives of individuals or groups upon their degree of synergy (support) and antagonism (resistance) regarding the initiative. This mapping is often represented as a matrix, as shown in Fig. 1, consisting of two main axes: The first axis, which measures the degree of support or commitment an actor has towards the project, is subdivided into four main categories:

- Minimalist: Individuals with little interest or commitment to the project, showing minimal involvement.
- Interested: Individuals who show a degree of interest but require more evidence or benefits before fully committing.
- Cooperative: Individuals who are actively engaged in the project, offering ideas and solutions.
- Committed: Those who reflect a strong devotion to the project, are motivated, and are inclined to assume an important position in its operationalization.

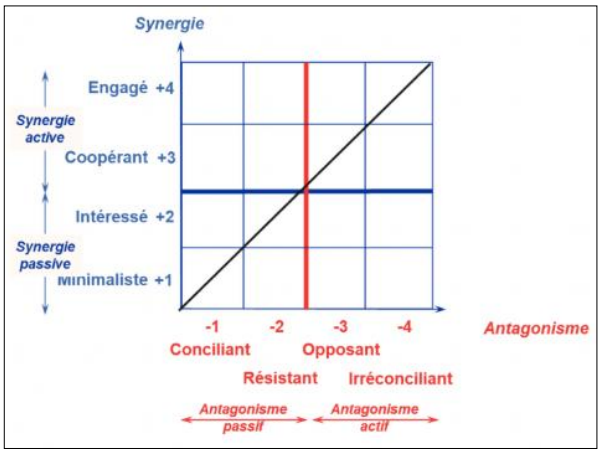


Fig. 1. Partner map [26].

The second axis, which evaluates the level of opposition or resistance an actor has towards the project, is also subdivided into four main categories:

- Conciliatory: Individuals with reservations about the change but open to dialogue and compromise.
- Resistant: Individuals who show stronger opposition due to specific concerns or divergent interests.

- Opponent: Individuals who actively oppose the project, potentially seeking to block it due to personal fears or general resistance to change.
- Irreconcilable: Individuals whose opposition is deep and often irrevocable, making constructive involvement difficult.

The intersections between the categories of synergy and antagonism give rise to seven more nuanced and meaningful categories, reflecting the complexity of organizational and human dynamics during change processes.

One of the main added values of this model resides in its capacity to reveal concealed or passive stakeholders, individuals who may not be readily apparent or vocal during the change process, yet whose perspectives and influence can profoundly affect project outcomes. By rendering these latent dynamics discernible, the allies strategy prompts project managers to expand their stakeholder analysis beyond the most conspicuous participants, thereby facilitating more tailored engagement strategies and mitigating the risk of neglecting silent sources of resistance or support.

C. Zero-Shot LLMs

A Large Language Model (LLM) represents an advanced manifestation of artificial intelligence, systematically engineered to analyze and generate natural language akin to human discourse. Based on sophisticated deep learning techniques, particularly on an advanced neural network architecture known as Transformers [29], these models incorporate an attention mechanism that discerns complex interrelationships between textual components, regardless of their position in the sequence. The effectiveness of this attention mechanism gives LLMs the ability to capture the nuances of language, syntax, and contextual relevance. Owing to their thorough preliminary training on a vast number of datasets [30], LLMs are exceptionally effective in numerous applications, which include conversational agents, automated content production, translation, and what is termed 'zero-shot' classification, defined as a technique that permits a model to label text without prior familiarity with examples of the pertinent category [31] [32]. LLMs fine-tuned to this specific task are called zero-shot LLMs, and they use this technique to correlate textual data with classifications without prior exposure to such data. They rely on natural language inference (NLI) models [33], which investigate the semantic relationships between an input text and a particular category through a logical semantic process that assesses whether a certain hypothesis (the category) is entailed (implication), neutral (neutrality), or contradicted (contradiction) by a provided text [34]. The performance of these models is chiefly derived from their foundational training on vast, high-standard natural language inference (NLI) datasets [32], incorporating SNLI [34], MNLI [35], ANLI [36], and FEVER [37]. These datasets contain extensive collections of annotated sentence pairs, assessing whether one sentence entails, is neutral to, or contradicts another. They have enabled models to develop a more refined understanding of semantic relationships and enhance their inference capabilities.

Within the scope of our research, the adoption of zero-shot LLMs is notably supported by the open-ended and unstructured responses from the feedback gathered from employees related to their views on organizational change. Zero-shot models enable the immediate classification of such qualitative responses without the need for prior task-specific fine-tuning. Furthermore, their capacity to perform inference-based categorization aligns well with the conceptual framework of the allies' strategy, which requires distinguishing nuanced and context-dependent postures ranging from strong support to strong opposition. This makes them highly appropriate for analyzing resistance to change in dynamic and resource-constrained environments.

D. Prompt Engineering

Prompt engineering constitutes a technique for optimizing the utilization of large language models (LLMs) by steering their outputs via precise directives, all while preserving the integrity of the model itself. This technique enables the optimization of LLMs' performance in various tasks, such as text classification, machine translation, or generating tailored responses, simply by adjusting the phrasing of the queries [38]. What renders prompt engineering exceptionally potent is its capacity to adapt pre-trained models for novel tasks without necessitating the retraining of the model parameters. As articulated in [39], this methodology significantly augments the capacity of models to generalize proficiently to novel tasks, especially via mechanisms such as zero-shot prompting and few-shot prompting. The effectiveness of this technique depends on the precision of prompt formulation, as even small adjustments can lead to significant variations in the results generated [40]. Several prompting strategies have been developed, including, for example:

- Zero-Shot Prompting: The model is given an instruction without any example responses. It relies on the LLM's ability to generalize from existing knowledge [38].
- Few-Shot Prompting: This strategy includes several representative instances within the prompt to improve the model's grasp of the task at hand. It enhances model accuracy by providing more context [38].
- Chain-of-Thought Prompting (CoT): As delineated by J. Wei et al [39], this methodology invites the model's elucidation of its reasoning process via intermediary steps prior to delivering a conclusive response. It’s particularly useful for tasks requiring complex logical reasoning, such as math problems or advanced inference.
- Role-Playing Prompting: The model is assigned a specific role, such as an expert in a given domain, to improve the relevance and coherence of responses in specialized contexts [40].
- AI-Knowledge Prompting: As articulated by J. Liu et al. [41], this methodology entails soliciting the model to produce knowledge that is specific to a given task, which is subsequently integrated into the prompt to enhance both the precision and reliability of the generated predictions.

IV. METHODOLOGY

Our methodology relies on the use of zero-shot LLMs combined with prompting techniques to analyze and classify open-ended responses in the context of measuring resistance to change, based on the allies strategy. The choice of these models is justified by their ability to infer logical relationships between texts and recognize classification patterns without requiring supervised learning on a specifically annotated dataset [33]. However, their performance is highly dependent on the formulation of hypotheses and the provided context [42] [43]. In order to enhance classification accuracy, we have incorporated a prompt engineering methodology aimed at optimizing the models' comprehension [44].

In the following, we give a brief description of the zero-shot LLMs selected for our study, and we also describe the prompt formulations used, followed by detailing the workflow of our proposed approach to measuring resistance to change relying on allies strategy.

A. Models

In the context of our research, we chose a selection of LLMs specialized in zero-shot classification and recommended by the Hugging Face platform [45] due to their performance in natural language inference (NLI), namely:

1) Facebook/bart-large-mnli. The BART model, known as Bidirectional and Auto-Regressive Transformers, represents an encoder-decoder transformer that integrates the benefits of bidirectional encoder models like BERT alongside auto-regressive decoder models such as GPT [46]. Its variant, facebook/bart-large-mnli, has been pretrained on the MultiNLI dataset, enabling it to perform excellently in Natural Language Inference (NLI) tasks. Its architecture makes it particularly well-suited for zero-shot classification, where it evaluates the probability that a given text corresponds to a predefined hypothesis, a mechanism directly leveraged in our study to assign responses to categories of resistance to change.

2) MoritzLaurer/deberta-v3-large-zeroshot-v2.0 and MoritzLaurer/deberta-v3-base-zeroshot-v2.0mnli. eBERTa models (Decoding-enhanced BERT with Disentangled Attention) are an improvement over BERT and RoBERTa, introducing a disentangled representation mechanism that more effectively distinguishes words based on their position and context [47]. The variants deberta-v3-large-zeroshot-v2.0 and deberta-v3-base-zeroshot-v2.0, developed by Moritz Laurer, are specifically optimized for zero-shot classification. They benefit from enhanced masking and more efficient pre-training, thereby improving their ability to generalize to new tasks without explicit supervision.

3) MoritzLaurer/roberta-large-zeroshot-v2.0. oBERTa (Robustly Optimized BERT Pretraining Approach) is an improved variant of BERT that relies on longer pretraining and better hyperparameter tuning [48]. The roberta-large-zeroshot-v2.0 version, fine-tuned for zero-shot classification, leverages these optimizations to better handle the variability of natural language, an essential advantage in our task, where employee responses can vary significantly in wording and tone.

4) MoritzLaurer/bge-m3-zeroshot-v2.0. The bge-m3-zeroshot-v2.0 model is part of the models optimized for semantic similarity tasks and zero-shot classification. It relies on a combination of advanced semantic representation techniques and specific optimizations to improve robustness when dealing with unstructured texts and varied formulations [4].

5) OpenAI/GPT-4. GPT-4 is a multimodal model and one of the most advanced developed by OpenAI, outperforming its predecessors in terms of contextual understanding, logical reasoning and text generation [49]. Unlike specialized models such as BART or DeBERTa, GPT-4 is a generalist model, but it excels in zero-shot classification tasks thanks to its ability to analyze global context and interpret complex formulations. Its integration into our study enables us to assess the extent to which an ultra-polyvalent model can compete with specialized models on a specific task, such as the analysis of resistance to change.

B. Prompt Engineering

To enhance the zero-shot classification of employees' responses while relying on the allies strategy, we have developed and tested two types of prompts, namely a minimalist prompt and an explicit prompt:

1) A minimalist prompt. This prompt is limited to essential information, represented solely by numeric labels and class names corresponding to the synergy and antagonism axes as defined in the allies strategy. In this scenario, two sets of classes were defined according to this strategy:

- Synergy (synergy class): "1 Minimal", "2 Interested", "3 Cooperative", "4 Committed"
- Antagonism (antagonism class): "1 Conciliatory", "2 Resistant", "3 Opponent", "4 Irreconcilable"

This type of prompt aims to evaluate the zero-shot models' ability to generalize from minimal class descriptions, without providing explicit context about their meaning. In this case, we rely entirely on LLM's internal representation and its pre-existing knowledge gained from vast textual datasets to generalize new classification tasks. This approach is aligned with the Zero-Shot Prompting strategy.

2) An explicit prompt. This approach is based on AI-Knowledge Prompting [41], where each class (synergy and antagonism classes) is enriched with an explicit definition of the behaviors associated with each level of synergy or antagonism. These descriptions provide the model with a more precise contextual framework, thereby facilitating the distinction between different classes. To design these AI-Knowledge Prompts, we used GPT-4 to generate detailed class descriptions. Specifically, we formulated an initial zero-shot prompt containing only the classification labels, and then we leveraged GPT-4 to generate definitions specific to each category, based on the principles of the allies strategy. These

descriptions were subsequently integrated into the final explicit prompt to provide a richer contextual framework.

- Synergy (synergy class):
  - "Follows assigned directives without active initiatives",
  - "Follows assigned directives and expresses verbal interest in the project, but without active initiatives",
  - "Takes active initiatives",
  - "Takes active initiatives with a high level of responsibility".
- Antagonism (antagonism class):
  - "Supports the project's decisions and approaches",
  - "Opposes the project or certain decisions/approaches but is open to discussion",
  - "Opposes the project or certain decisions/approaches and is less flexible in discussions",
  - "Opposes the project or certain decisions/approaches and is inflexible in discussions".

This type of prompt aims to examine how the detail of class descriptions affects the efficiency of zero-shot models.

C. Proposed Approach for Measuring Resistance to Change

In this section, we provide an overview of the methodology employed in our study, illustrated in Fig. 2, detailing the steps taken to investigate the effectiveness of zero-shot LLMs combined with prompting techniques in measuring resistance to change using the allies strategy. We adopted a multi-task classification approach, with each task focusing on a different aspect of employees' dynamics regarding digital transformation projects:

- Classification based on Synergy Level: The objective of this task is to categorize employees by their degree of engagement and collaboration within the project. The labels range from "Minimalist" to "Committed". This classification helps identify the most motivated team members and those who may require additional motivation or incentives.
- Classification based on Antagonism Level: focuses on analyzing the degree of resistance or opposition exhibited by employees regarding project initiatives. The classifications range from 'Conciliatory' to 'Irreconcilable', thus enabling the recognition of resistance levels that could affect group dynamics and hinder project progress.
- Final classification: This task synthesizes results derived from the initial two classifications to construct a holistic appraisal of each employee. It amalgamates both synergistic and antagonistic dimensions to yield a comprehensive evaluation of employees' attitudes towards digital transformation initiatives.

Based on a dataset containing open-ended employee responses, previously annotated with synergy and antagonism classes, we proceed as follows:

The first two tasks, namely classification by Synergy Level and Classification by Antagonism Level, referred to both simultaneously in Fig. 2 as task 1 and task 2, follow the same process, detailed below:

1) Preprocessing of open-ended responses. Each open-ended response provided by an employee is standardized to ensure better understanding by the language models. This step includes text cleaning (removal of special characters, unnecessary stopwords if necessary) and conversion into a standardized format suitable for input into zero-shot classification models.

2) Prompt engineering. Two variants of prompts are used to guide the zero-shot classification models:

a) Minimalist prompt. Contains only the classification labels.

b) Explicit prompt. Adds detailed descriptions of the classes to improve the model's interpretation.

3) Zero-shot classification with LLMs. Each open-ended response is submitted to a large language model (LLM) specialized in zero-shot classification, accompanied by the selected prompt (either minimalist or explicit). The model reformulates the task as a natural language inference (NLI), where the open-ended response is considered the "premise" and each possible class is treated as a "hypothesis". It then evaluates the probability that the response implies each hypothesis, assigning a confidence score to each class. Finally, the class with the highest probability is selected as the final label assigned to the open-ended response.

4) Performance evaluation of classification. The classification performance is evaluated by comparing the predicted classes to the actual classes annotated in the dataset. The metrics calculated are:

- Confusion Matrix
- F1-score
- Recall
- Accuracy
- ROC-AUC Curve

Regarding the third classification task (task 3 in Fig. 2), which aims to assign the final class of resistance to change, the process is as follows: Once the synergy and antagonism classes have been determined according to the steps of the first two tasks, they are combined based on a pre-established matrix derived from the allies' map. This combination allows for the assignment of a final class representing the overall level of resistance to change.

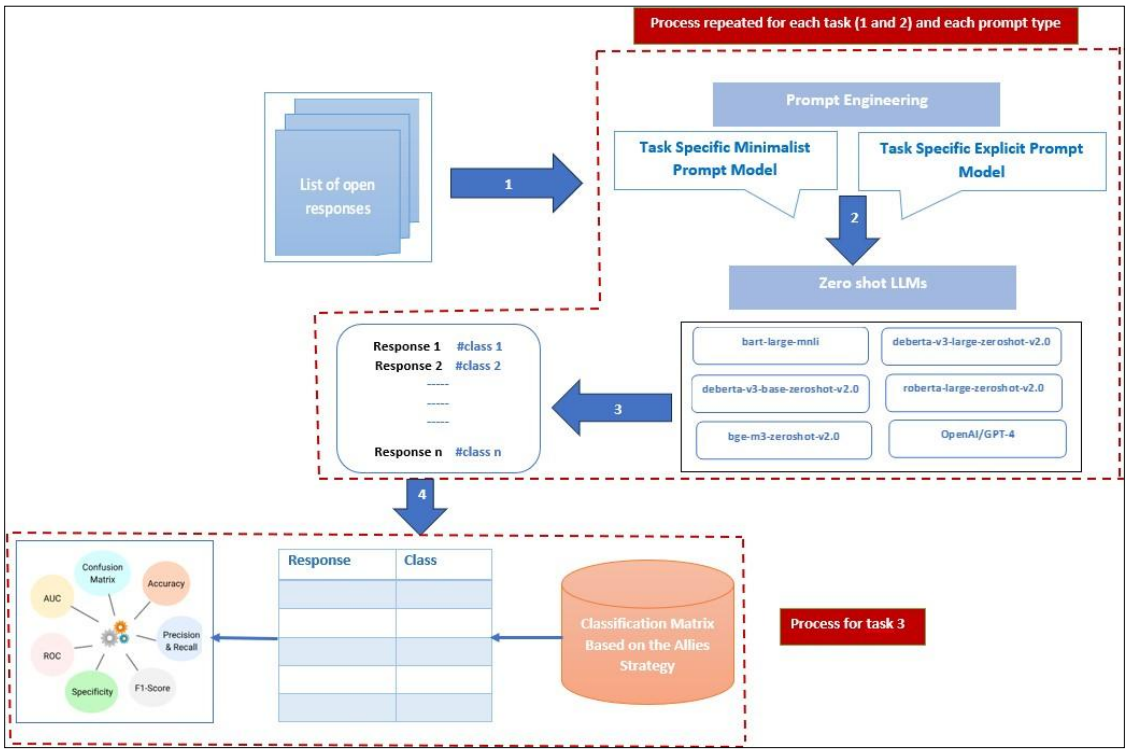


Fig. 2. Overview of the approach for measuring resistance to change.

As with the previous tasks, the predicted final class is compared to the actual class annotated in the dataset. Performance evaluation is carried out using the same metrics as the previous tasks.

V. Experiments

To validate the proposed methodology, we executed a thorough array of experiments intended to measure the effectiveness of the selected zero-shot Large Language Models (LLMs) in precisely classifying employee opinions regarding the context of digital transformation, derived from their open-ended text replies. This section elucidates the experimental framework, encompassing the dataset utilized, the evaluative methodology employed to benchmark the models, and the metrics implemented to evaluate their efficacy. The design of these experiments aimed to investigate both the efficacy of the prompting strategies and the pertinence of the allies strategy framework in effectively capturing the intricate expressions of resistance to change.

A. Dataset

As part of our study aimed at classifying employees based on their synergy and antagonism towards a digital transformation project, we built a database by collecting approximately one hundred open-ended responses directly from a company that recently introduced digital initiatives. This data collection allowed us to gather a hundred employee testimonies expressing their perceptions, concerns, and levels of adherence to these changes. Each open-ended response was then carefully annotated by change management experts in accordance with the classes defined in the allies strategy. Specifically, three labels were assigned to each response:

- A synergy level, indicating the degree of alignment and positive engagement of the employee towards the digital initiative.
- An antagonism level, reflecting the resistance expressed, whether passive or active, towards the change.
- A global resistance to change level, resulting from the combination of the first two classifications as defined in the Allies Strategy Matrix.

These manual annotations are an essential benchmark for evaluating the performance of the models used, enabling us to measure their ability to reproduce expert judgments in the analysis of resistance to change.

The distribution of employee responses by level of synergy and antagonism is described in Fig. 3:

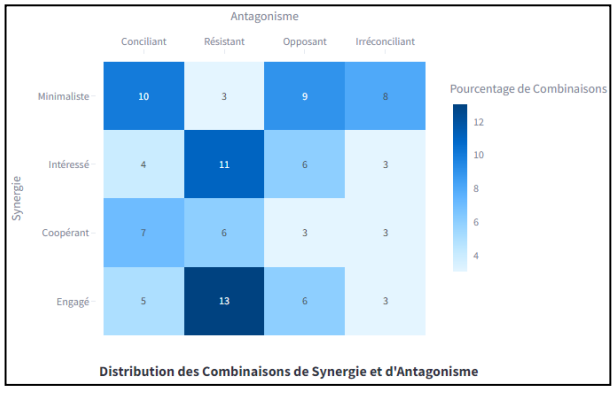


Fig. 3. Distribution of employee responses by level.

B. Evaluation Process

Our evaluation process, as described in the methodological section, will be applied to six models specialized in zero-shot classification, using two distinct prompt variants for each tested model: a minimalist prompt containing only the classification labels and an explicit prompt providing detailed descriptions of the classes to enhance the model's interpretation. In other words, each model is required to classify the open responses of the dataset according to the level of synergy corresponding to task 1 and then according to the level of antagonism corresponding to task 2, in line with the workflow described in the methodology section. These classifications, predicted by the models, will be generated first using the minimalist prompt and then the explicit prompt. Finally, once the synergy and antagonism classes have been determined according to the steps of the first two tasks. Task 3 is started by combining the results of the first two tasks based on a pre-established matrix derived from the allies strategy. This combination makes it possible to assign a final class representing the overall level of resistance to change for each open response.

C. Evaluation Metrics

To validate the performance of the models, we use a method structured in two main steps:

1) Evaluation by task. Evaluation by task consists of comparing the classifications predicted by the model with the actual classifications annotated in the dataset for each of the tasks and prompts defined in our classification workflow. For this, we use confusion matrices to visualize the distribution of errors made by the model, and we also use standard classification metrics, notably:

a) Precision: This metric is crucial for measuring the reliability of the model's positive classifications. In our context, a high precision score indicates that when the model identifies a specific stance toward change, it is unlikely to be incorrect. This is particularly important to avoid misclassifications that could distort the analysis of resistance and support for digital initiatives.

b) Recall: This metric evaluates the model's proficiency in detecting all affirmative instances present within the dataset. A high recall score is essential in our study to ensure that minority, but potentially critical, stances are not overlooked. This helps to provide a more comprehensive understanding of the dynamics of resistance and synergy among employees.

c) F1-score: By combining precision and recall, the F1-score balances these two aspects, particularly in cases where class distribution is imbalanced. In our analysis, this metric is particularly relevant, as some stance categories may be underrepresented, making it necessary to use a measure that penalizes both false positives and false negatives.

d) Accuracy: This metric assesses the ratio of correctly predicted classifications relative to the total predictions generated by the model. While it offers a general measure of reliability, it may not fully reflect performance in cases of class imbalance.

2) Comparative evaluation of models. Beyond analyzing individual classification tasks, we conduct a comprehensive comparison of the models using two key performance indicators:

a) The F1 score for the final classification task: This indicator summarizes the model's overall ability to balance precision and recall in assigning final classes.

b) Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC): Through the examination of the equilibrium between true positives and false positives, the AUC score serves to ascertain the model's ability to consistently differentiate among various levels of resistance, which is imperative within our framework, where a nuanced classification is essential for informed decision-making in developing a change management plan tailored to each category.

By combining these different metrics, we are able to identify the best-performing model best suited to our problem, thus guaranteeing accurate and exploitable results for the analysis of resistance to change.

VI. Results and Discussion

In this section, we evaluate the ability of the models to correctly classify responses according to the defined tasks (synergy, antagonism). For each task and each type of prompt, we have calculated the standard classification metrics: precision, recall, F1 score and accuracy, along with confusion matrices for fine-grained error analysis.

The performance of the models using both types of prompts is presented in Table I. Metrics are calculated using the "macro" method, which means that we calculate precision, recall, accuracy and F1 score for each class independently, then average them. This approach better reflects the overall performance of the model across all classes.

The preliminary analysis of the results highlights the following points:

- Minimalist prompt: Overall, the model's performance is moderate to low on both tasks when using a minimalist prompt (see Table I). This trend suggests that overly basic formulations do not provide enough context or information to the models, limiting their ability to correctly distinguish between classes.
- Explicit prompt: Adding more detailed descriptions significantly improves performance, particularly in terms of the F1 score, as illustrated in Fig. 4 and Fig. 5 for nearly all models. This improvement is particularly notable for the DeBERTa-v3-large-zeroshot model, which achieves the highest F1 scores on both tasks, demonstrating the importance of a well-structured prompt to optimize classification.

The results also show that the explicit prompt yields the best overall performance, with a notable increase in both precision and recall as shown in Table I.

TABLE I. PERFORMANCE METRICS

Model	Prompt	Task 1: Synergy Classification				Task 2: Antagonism Classification			
		Precision	*Recall*	*F1 Score*	*Accuracy*	*Precision*	*Recall*	*F1 Score*	*Accuracy*
OpenAI/GPT4	***Minimalist***	**0.52**	**0.43**	**0.40**	**0.46**	**0.64**	**0.50**	**0.45**	**0.56**
	Explicit	**0.70**	**0.62**	**0.60**	**0.65**	**0.72**	**0.60**	**0.60**	**0.65**
Facebook/bart-large-mnli	***Minimalist***	**0.35**	**0.32**	**0.22**	**0.32**	**0.26**	**0.43**	**0.31**	**0.43**
	Explicit	**0.53**	**0.46**	**0.44**	**0.47**	**0.60**	**0.49**	**0.42**	**0.56**
MoritzLaurer/deberta-v3-large-zeroshot-v2.0	***Minimalist***	**0.46**	**0.46**	**0.40**	**0.50**	**0.41**	**0.45**	**0.38**	**0.44**
	Explicit	**0.76**	**0.76**	**0.76**	**0.77**	**0.84**	**0.79**	**0.80**	**0.81**
MoritzLaurer/roberta-large-zeroshot-v2.0	***Minimalist***	**0.45**	**0.44**	**0.42**	**0.47**	**0.56**	**0.45**	**0.45**	**0.47**
	Explicit	**0.67**	**0.64**	**0.64**	**0.66**	**0.66**	**0.67**	**0.65**	**0.69**
MoritzLaurer/deberta-v3-base-zeroshot-v2.0	***Minimalist***	**0.42**	**0.33**	**0.28**	**0.31**	**0.55**	**0.47**	**0.41**	**0.45**
	Explicit	**0.51**	**0.47**	**0.44**	**0.50**	**0.66**	**0.54**	**0.53**	**0.53**
MoritzLaurer/bge-m3-zeroshot-v2.0	***Minimalist***	**0.38**	**0.37**	**0.29**	**0.38**	**0.39**	**0.35**	**0.35**	**0.33**
	Explicit	**0.41**	**0.47**	**0.41**	**0.48**	**0.59**	**0.47**	**0.45**	**0.48**

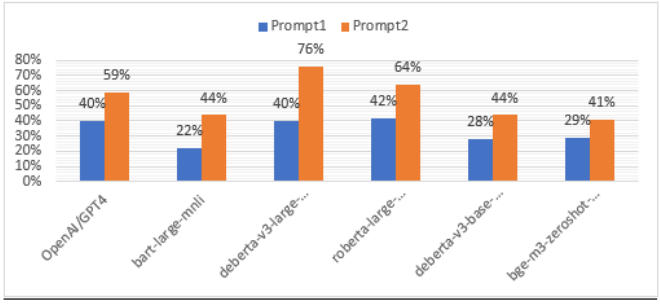


Fig. 4. F1 Score Results for Task 1 according to both prompts.

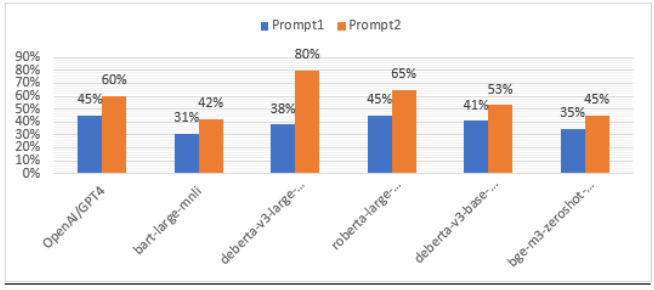


Fig. 5. F1 Score Results for Task 2 according to both prompts.

Based on the results shown in Fig. 4 and Fig. 5, it is clear that the explicit prompt (referred to in Fig. 4 and 5 as Prompt2) provides better performance than the minimalist prompt (referred to in Fig. 4 and 5 as Prompt1). To further analyze this, we will focus solely on the explicit prompt and examine the errors made by the models on tasks 1 and 2 using confusion matrices. These matrices allow us to visualize the correspondence between the model predictions and the actual classes, highlighting classification errors. The goal is to identify the models that achieve the best results and those that exhibit recurring errors.

To do this, we will display and analyze the confusion matrices of the different models on these two tasks in order to visualize the distribution of errors and better understand the respective performance of each model.

For task 1 related to classification according to the synergy axis, we selected the confusion matrices of the six models analyzed, as shown in Fig. 6.

The analysis of the confusion matrices reveals several types of recurring errors. Firstly, some models tend to confuse intermediate classes, particularly "Cooperative" and "Interested", suggesting difficulty in capturing the nuances between these categories. For instance, BART-large-mnli and RoBERTa-large-zeroshot frequently confuse these two classes, while DeBERTa-v3-base-zeroshot shows notable errors by incorrectly assigning "Cooperative" responses to "Committed".

Additionally, the BGE-m3-zeroshot model appears to struggle with correctly identifying the "Minimalist" class, redistributing its predictions toward other categories. In contrast, DeBERTa-v3-large-zeroshot and GPT-4 stand out with better classification performance, with DeBERTa-v3-large-zeroshot displaying slightly higher accuracy. However, errors persist in the intermediate classes, indicating a challenge in clearly differentiating these categories.

For task 2, related to classification according to the axis of antagonism, we adopt the same approach by generating the confusion matrices of the six models, as illustrated in Fig. 7.

Based on the analysis of the confusion matrices in Fig. 7, it is observed that DeBERTa-v3-large and GPT-4 stand out for their ability to accurately classify the extreme classes, with high precision for the categories "Conciliatory" and "Resistant". DeBERTa-v3-large demonstrates a more stable classification across all intermediate classes compared to the other models, while GPT-4 shows some confusion in the categories "Opponent" and "Irreconcilable". BART-large-mnli and RoBERTa-large-zeroshot also manage to correctly identify the majority classes but exhibit more pronounced confusion in differentiating the intermediate categories. DeBERTa-base and BGE-m3-zeroshot, on the other hand, seem to struggle more with distinguishing the "Opponent" and "Irreconcilable" classes, with a more scattered distribution of errors.

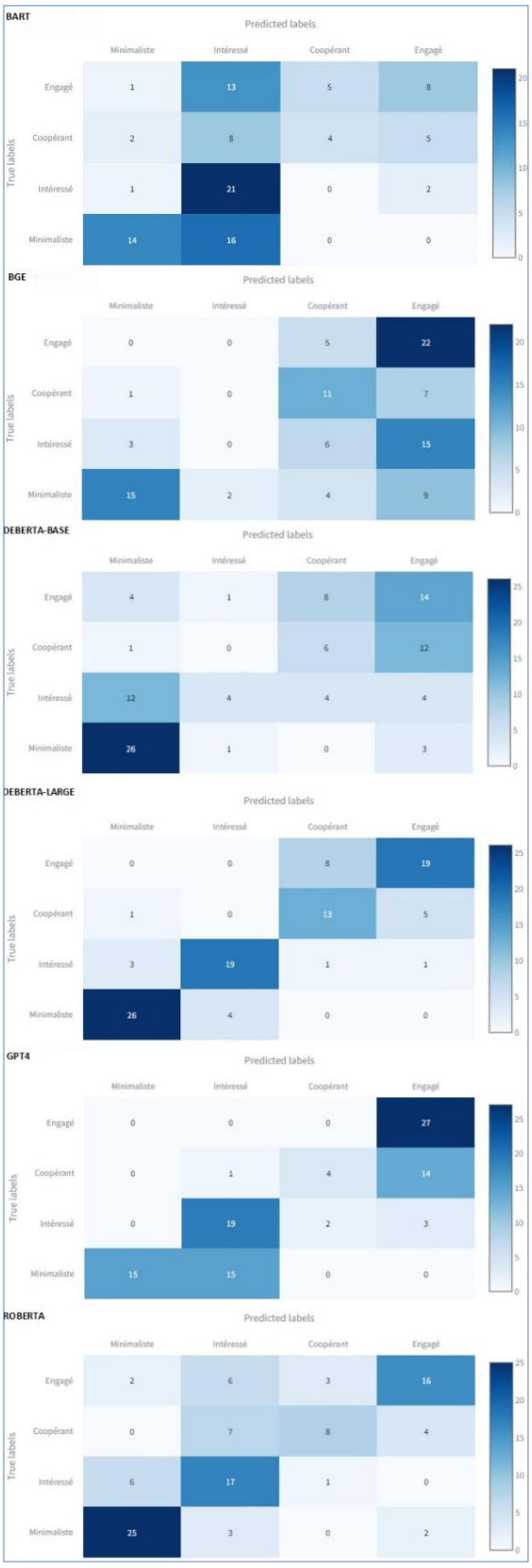


Fig. 6. Confusion matrices for LLM classification results of Task 1.

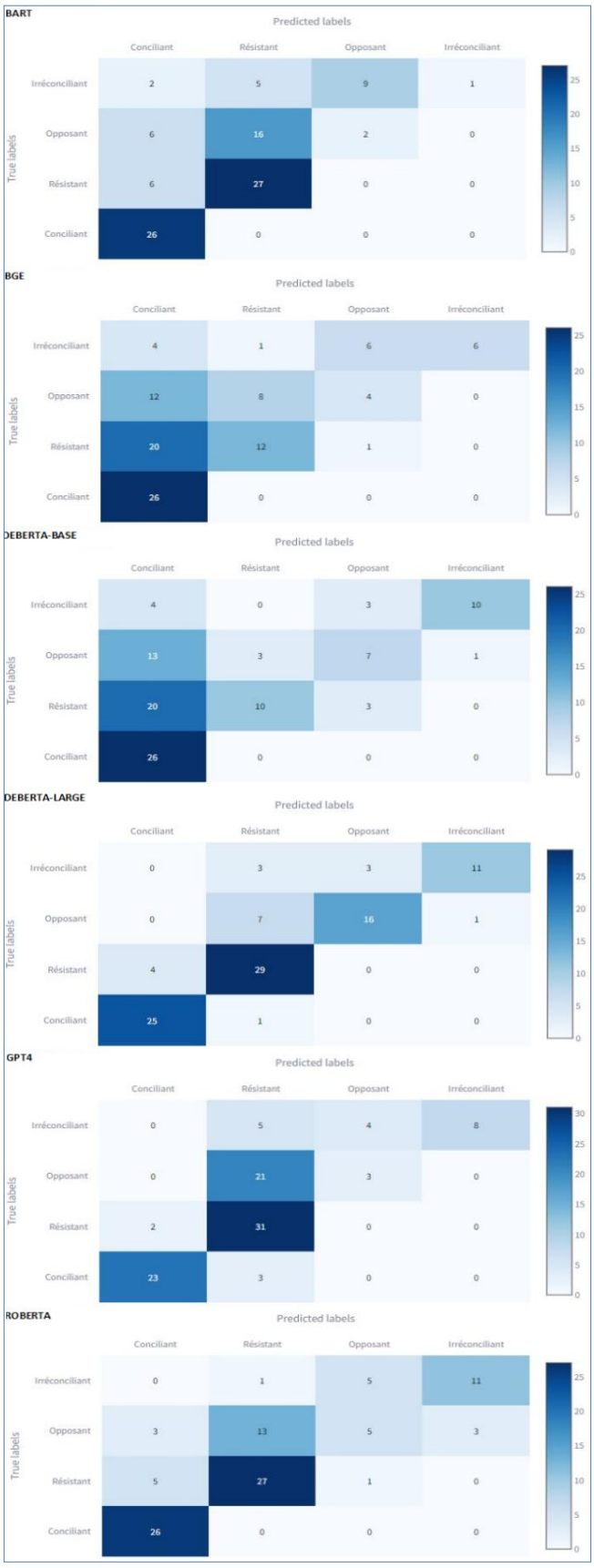


Fig. 7. Confusion matrices for LLM classification results of Task 2.

The results show that the DeBERTa-v3-large-zeroshot model stands out as the best model in terms of F1-score and precision for both synergy classification and antagonism classification tasks, particularly under the explicit prompt. This reflects a good balance between precision and recall. In comparison with modern language models like OpenAI/GPT-4, DeBERTa-v3-large-zeroshot achieves more robust results on these tasks, suggesting that models specifically designed for zero-shot tasks can surpass models like OpenAI/GPT-4 in certain classification contexts without requiring additional fine-tuning. GPT-4, on the other hand, although slightly behind in terms of F1-score, boasts significantly higher precision. This means it generates fewer false positives, demonstrating its ability to minimize false predictions and better handle errors. In addition, RoBERTa-large-zeroshot has a slightly higher F1-score than GPT-4, but lower precision. GPT-4 therefore, remains more accurate, illustrating a difference in the way these models optimize their classification decisions.

In contrast, the Facebook/bart-large-mnli and bge-m3-zeroshot-v2.0 models show significantly lower performance in F1-score and precision, indicating greater difficulty in correctly capturing the right classes and reducing classification errors.

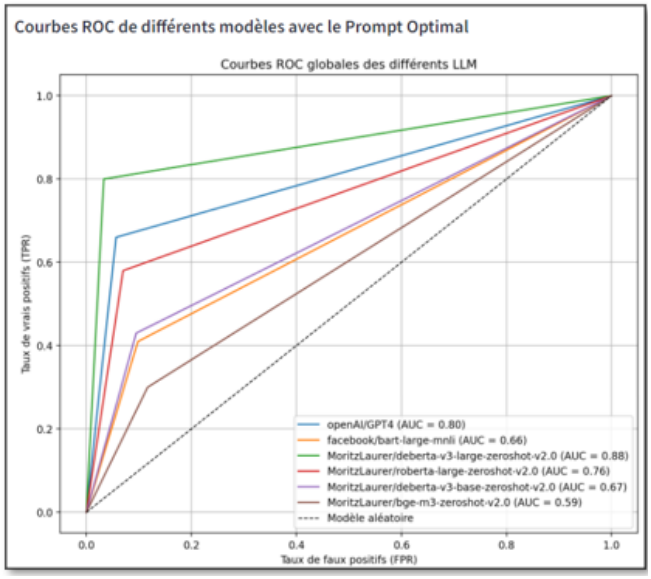


Fig. 8. ROC curves for LLMs.

These discussed findings are confirmed via a review of the Receiver Operating Characteristic Area Under the Curve (ROC AUC) measurements, which gauge the ability of the models to separate positive from negative classifications successfully (see Fig. 8). This analysis reveals that GPT-4 (AUC = 0.80) offers a better discrimination capacity than RoBERTa-large-zeroshot (AUC = 0.76). In fact, although its F1-score is lower, its high precision enables it to achieve a higher AUC, testifying to greater robustness in class distinction. However, DeBERTa-v3-large-zeroshot (AUC = 0.88) outperforms all other models, confirming that it simultaneously maximizes both sensitivity and specificity, making it the top-performing model in classification. In contrast, models like Facebook/bart-large-mnli (AUC = 0.66) and MoritzLaurer/bge-m3-zeroshot-v2.0 (AUC = 0.59) show weaker performance, with behavior closer to a random model in their ability to differentiate classes correctly. This is also reflected in their lower F1-scores and precision, indicating difficulties in accurately classifying positive classes.

VII. Conclusion and Perspectives

This study proposes a new approach for analyzing resistance to change in the context of digital transformation, relying on zero-shot language models (LLMs) and prompting techniques. Unlike traditional methods, which mainly use closed-ended questionnaires and binary classifications, limiting the expression and diversity of employee perspectives, our approach leverages the allies strategy to provide a more granular segmentation of employees' stances towards change. It does so by using open-ended responses for a more detailed and nuanced analysis. The use of specialized zero-shot LLMs in natural language inference, combined with specific prompts, allows for the automatic classification of these open-ended responses without the need for prior training on specific data.

To evaluate the performance of our methodology, we conducted an experiment on an annotated dataset consisting of a hundred employee responses to analyze the models' ability to accurately distinguish the expressed attitudes. Experimental results highlighted the importance of prompt formulation in improving classification accuracy. Specifically, we demonstrated that using explicit prompts optimizes model performance, particularly DeBERTa-v3-large-zeroshot, which emerged as the best-performing model in terms of F1-score and discriminatory power (AUC), surpassing GPT-4. These results also emphasize that, while generalist models like OpenAI/GPT-4 remain competitive, models specifically designed for zero-shot tasks can better meet the requirements of fine classification without requiring additional fine-tuning.

These findings open up several avenues for improvement. One avenue would be to refine the prompting instructions to better capture the nuances of employee responses, thereby optimizing model performance. Additionally, exploring hybrid approaches that combine zero-shot learning with supervised learning could further enhance the classification of stances toward change. Beyond textual analysis, one major research hurdle is blending in various assessment methods, like responses that rely on spoken language. By allowing employees to provide spoken feedback, it is possible to improve the richness of the input data while making the assessment process more inclusive for people with limited writing skills. Finally, applying this methodology to larger and more diverse datasets would allow for evaluating its generalization to different organizational contexts.

Thus, this study demonstrates that zero-shot LLMs, coupled with prompting techniques and the allies strategy, represent a step forward in automating the assessment of resistance to change, providing businesses with a more effective decision-making tool in supporting their digital transformations.

References

[1] B. M. DuBose and A. M. Mayo, “Resistance to change: A concept analysis,” Nurs Forum, vol. 55, no. 4, pp. 631–636, Nov. 2020, doi: 10.1111/nuf.12479.

[2] S. Oreg, "Personality, context, and resistance to organizational change," European Journal of Work and Organizational Psychology, vol. 15, no. 1, pp. 73–101, 2006, doi: 10.1080/13594320500451247.

[3] M. Asem, "Study and synthesis of concepts and practices of resistance to organizational change," Economica, vol. 123, no. 1, pp. 45–55, 2023.

[4] H. Chammaa, R. Ed-Daoudi, and K. Benazzi, "Large Language Models for Academic Internal Auditing," ijacsa, vol. 16, no. 1, 2025, doi: 10.14569/IJACSA.2025.0160166.

[5] V. Cieslak and C. Valor, "Moving beyond conventional resistance and resistors: an integrative review of employee resistance to digital transformation," Cogent Business & Management, vol. 12, no. 1, p. 2442550, Dec. 2025, doi: 10.1080/23311975.2024.2442550.

[6] A. Blyukov, "Resistance to Digital Transformation Projects in FMCG Companies," Scientific Research and Development. Russian Journal of Project Management, vol. 10, pp. 3–12, Aug. 2021, doi: 10.12737/2587-6279-2021-10-2-3-12.

[7] S. Yun, G. Penghe, W. Sitong, and D. R. B. Rosli, "The Impact of Digital Transformation on Business Strategy," vol. 5, no. 6, 2023.

[8] H. G. DEMİREL, "Why do digital transformation initiatives fail? digital mindsets as an invisible strategic component," Yönetim Bilimleri Dergisi, 2024, doi: 10.35408/comuybd.1516665.

[9] B. M. Omowole, A. Q. Olufemi-Phillips, O. C. Ofodile, N. L. Eyo-Udo, and S. E. Ewim, "Barriers and drivers of digital transformation in SMEs: A conceptual analysis," International journal of scholarly research in science and technology, vol. 5, no. 2, pp. 019–036, 2024, doi: 10.56781/ijsrst.2024.5.2.0037.

[10] S. Syamsuddin, S. Marsudi, B. Hasanuddin, A. Umar, and D. Suprayitno, "Adapting to Digital Transformation: Challenges and Strategies for Traditional Businesses," Global, vol. 2, no. 3, pp. 704–711, Mar. 2024, doi: 10.59613/global.v2i3.121.

[11] A. Scholkmann, "Resistance to (Digital) Change: Individual, Systemic and Learning-Related Perspectives," 2021. doi: 10.1007/978-3-030-55878-9_13.

[12] S. P. Robbins and T. Judge, Organizational behavior. Pearson South Africa, 2009.

[13] H. Karaxha, "Methods for Dealing with Resistance to Change," Baltic Journal of Real Estate Economics and Construction Management, vol. 7, no. 1, pp. 290–299, Jan. 2019, doi: 10.2478/bjreecm-2019-0018.

[14] T. Kuzhda, "Diagnosing resistance to change in the change management process," Econ. manag. sustain., vol. 1, no. 1, pp. 49–59, Dec. 2016, doi: 10.14254/jems.2016.1-1.5.

[15] H. Samia, "Digital transformation: A development strategy that faces civil servants resistance to changing working conditions : 'Case of public administrations in the Meknes-Fez region,'" vol. 7, Apr. 2024.

[16] S. Oreg, "Resistance to change: Developing an individual differences measure.," Journal of applied psychology, vol. 88, no. 4, p. 680, 2003.

[17] H. Sherif, "The impact of Resistance to Change to Digital Transformation as a Mediating Variable on the Relationship between Training and Performance Efficiency (Applied to the National Organization for Social Insurance in Egypt)," Jun. 2024.

[18] K. R. Stam, J. M. Stanton, and I. R. Guzman, "Employee resistance to digital information and information technology change in a social service agency: A membership category approach," 2004.

[19] B. A. El Karam, R. Marghoubi, F. Khanboubi, and D. Allaki, "A New Approach for the Analysis of Resistance to Change in the Digital Transformation Context," presented at the International Conference of Innovation and New Trends in Information Technology, Springer, 2022, pp. 153–167.

[20] F. Khanboubi and A. Boulmakoul, "Fuzzy intuitionist approach for resistance to change analysis in a digital transformation process," presented at the 2021 International Conference on Decision Aid Sciences and Application (DASA), IEEE, 2021, pp. 957–961.

[21] J.-C. Fauvet and J.-R. Fourtou, La sociodynamique: concepts et méthodes. Les Editions d'organisation, 1996.

[22] J.-C. Fauvet and J.-R. Fourtou, L'élan sociodynamique. Ed. d'Organisation, 2004.

[23] J. D. Pincus, "Employee engagement as human motivation: Implications for theory, methods, and practice," Integrative Psychological and Behavioral Science, vol. 57, no. 4, pp. 1223–1255, 2023.

[24] Y. E. Rachmad, Network Influence Theory. Mashhad Ferdowsi Ketab Nashriyat, Mahsoos Nashriyat 2023., 2023.

[25] P. Walley, "Stakeholder management: the sociodynamic approach," International Journal of Managing Projects in Business, vol. 6, no. 3, pp. 485–504, 2013, doi: 10.1108/IJMPB-10-2011-0066.

[26] Institut de la Sociodynamique, "Le modèle de la stratégie des alliés." Accessed: Mar. 15, 2025. [Online]. Available: https://www.institutdelasociodynamique.com

[27] P. Bélorgey and N. Van Laethem, "Outil 4. La sociodynamique pour engager les parties prenantes," BàO La Boîte à Outils, vol. 2, pp. 46–54, 2023.

[28] D. Autissier, J.-M. Moutot, K. Johnson, and E. Metais-Wiersch, "Outil 19. La cartographie des alliés," BàO La Boîte à Outils, vol. 2, pp. 58–59, 2022.

[29] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[30] S. A. Alsaedi, A. Y. Noaman, A. A. A. Gad-Elrab, F. E. Eassa, and S. Haridi, "Leveraging Large Language Models for Automated Bug Fixing," ijacsa, vol. 15, no. 12, 2024, doi: 10.14569/IJACSA.2024.0151247.

[31] Z. Wang, Y. Pang, and Y. Lin, "Large Language Models Are Zero-Shot Text Classifiers," Dec. 02, 2023, arXiv: arXiv:2312.01044. doi: 10.48550/arXiv.2312.01044.

[32] T. Schopf, D. Braun, and F. Matthes, "Evaluating Unsupervised Text Classification: Zero-shot and Similarity-based Approaches," in Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval, Dec. 2022, pp. 6–15. doi: 10.1145/3582768.3582795.

[33] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," Aug. 21, 2015, arXiv: arXiv:1508.05326. doi: 10.48550/arXiv.1508.05326.

[34] D. O. Romero Mogrovejo, E. Agirre Bengoa, and O. López de Lacalle Lecuona, "Natural Language Inference Models for Few-Shot Text Classification: A Real-World Perspective," 2023.

[35] A. Williams, N. Nangia, and S. R. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," Feb. 19, 2018, arXiv: arXiv:1704.05426. doi: 10.48550/arXiv.1704.05426.

[36] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial NLI: A New Benchmark for Natural Language Understanding," May 06, 2020, arXiv: arXiv:1910.14599. doi: 10.48550/arXiv.1910.14599.

[37] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for Fact Extraction and VERification," Dec. 18, 2018, arXiv: arXiv:1803.05355. doi: 10.48550/arXiv.1803.05355.

[38] T. Brown et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[39] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," Advances in neural information processing systems, vol. 35, pp. 24824–24837, 2022.

[40] A. Kong et al., "Better Zero-Shot Reasoning with Role-Play Prompting," Mar. 14, 2024, arXiv: arXiv:2308.07702. doi: 10.48550/arXiv.2308.07702.

[41] J. Liu et al., "Generated Knowledge Prompting for Commonsense Reasoning," Sep. 28, 2022, arXiv: arXiv:2110.08387. doi: 10.48550/arXiv.2110.08387.

[42] R. Ghilene, D. Niaouri, M. Linardi, and J. Longhi, "Analysis of Socially Unacceptable Discourse with Zero-shot Learning," Sep. 10, 2024, arXiv: arXiv:2409.13735. doi: 10.48550/arXiv.2409.13735.

[43] T. Schick and H. Schütze, "Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference," Jan. 25, 2021, arXiv: arXiv:2001.07676. doi: 10.48550/arXiv.2001.07676.

[44] L. Reynolds and K. McDonell, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," Feb. 15, 2021, arXiv: arXiv:2102.07350. doi: 10.48550/arXiv.2102.07350.

[45] "Hugging Face Models." Accessed: Dec. 30, 2024. [Online]. Available: https://huggingface.co/models

[46] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Oct. 29, 2019, arXiv: arXiv:1910.13461. doi: 10.48550/arXiv.1910.13461.

[47] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," Oct. 06, 2021, arXiv: arXiv:2006.03654. doi: 10.48550/arXiv.2006.03654.

[48] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 26, 2019, arXiv: arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692.

[49] J. Achiam et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

Self-Supervised Method for Risky Situation Detection in Road Traffic Sequences Using Video Masked Autoencoder

Abdelhafid Berroukham[1], Mohammed Lahraichi[2], Khalid Housni[3]
Department of Computer Science, Faculty of Science, Ibn Tofail University, Kenitra, Morocco[1, 2, 3]
CRMEF Casablanca-Settat, Casablanca, Morocco[2]

***Abstract*—Road traffic accidents are a significant public health issue, particularly in developing nations, where infrastructure and traffic monitoring systems may be limited. Risky situations including sudden stopping, lane switching, and near-misses can lead to accidents. In this study, we present an original approach for recognizing risky situations in road traffic sequences using Video Masked Autoencoder (VideoMAE), a self-supervised deep learning model built upon Vision Transformer architecture. By applying a pre-trained VideoMAE on a large dataset of videos and fine-tuning it on labeled traffic sequences categorized as risky or non-risky, our model learns spatiotemporal features without requiring extensive manual labeling. The method achieves high accuracy on testing data, demonstrating strong potential for high-risk detection with an accuracy of 95%. This study highlights the promise of self-supervised video representation learning for real-world safety applications and paves the way for the development of intelligent traffic monitoring and crash prevention tools.**

***Keywords*—Video processing; risk detection; VideoMAE; vision transformer; deep learning; computer vision**

I. Introduction

Road accidents are a significant public health concern, with an estimated 1.24 million lives lost to road traffic accidents annually, according to the World Health Organization (WHO)[1]. Hazardous road traffic situations, including rapid braking, lane changes, and near-misses, can lead to accidents. Therefore, it is crucial to develop techniques for detecting and preventing such situations.

Computer vision and artificial intelligence have advanced significantly in recent years, particularly in applications requiring automated interpretation of visual scenes. Traffic scene analysis is one such domain, playing a critical role in road safety and intelligent transportation systems. While notable progress has been made, detecting and predicting hazardous situations—such as sudden braking, lane changes, or near-crash events—remains challenging due to the dynamic and complex nature of real-world traffic environments.

Despite modern advances, existing approaches exhibit several limitations. Most rely on supervised learning, which requires large, labeled datasets that are both costly and time-consuming to produce. Furthermore, many methods focus primarily on object recognition (e.g., identifying pedestrians or vehicles) while overlooking the temporal dynamics and interactions that are crucial for assessing risk in time-varying scenes. As a result, there is a gap in the literature regarding models capable of learning meaningful patterns from unlabeled video data or operating effectively with limited annotated datasets.

To address these challenges, self-supervised learning approaches offer a promising alternative. By reducing or eliminating the need for manual labeling, they present a more efficient and scalable solution for training models in complex traffic scenarios. Video Masked Autoencoder (VideoMAE) is one of the self-supervised learning techniques that has proved to be useful for many different tasks [2], including motion estimation, object detection, and video classification. VideoMAE learns video representations by predicting masked patches in the input videos. These learned representations prove valuable for various downstream tasks, including detection, classification, and segmentation.

In this study, we propose a method based on VideoMAE for detecting risky situations in road traffic sequences. By leveraging VideoMAE's ability to learn rich spatiotemporal features from video data. We train the model on a set of road traffic sequences, which are labeled as risky or non-risky. After the model is trained, it can be employed to find risky conditions in new road traffic sequences. This method not only reduces the reliance on labeled datasets but also improves generalization across diverse environments.

Our contributions are as follows:

- We propose a novel method to detect risky situations in road traffic sequences using VideoMAE.
- We train a VideoMAE model on a small dataset of road traffic sequences (fine-tuning).
- We evaluate our model on a held-out test set.

The rest of this study is organized as follows: In Section II (Related Work), we review existing approaches used for detecting risky situations in road traffic and highlight their limitations; Section III (Method) introduces the VideoMAE model, detailing its architecture and the preprocessing steps applied to the video data; Section IV (Experiment and Results) presents the experimental setup as well as the quantitative and qualitative results; Section V (Discussion) provides an analysis of the results; and Section VI (Conclusion) summarizes the key findings of our research and suggests areas for further

exploration to enhance the model's performance and applicability in real-world traffic safety applications.

II. Related Work

There are a number of existing methods for detecting risky situations in road traffic [3], [4], [5], as well as others that consider road accident classification a pivotal field of study [6], while some studies have focused on driving behavior [7], [8]. These methods can be broadly divided into two categories: traditional methods and machine learning methods.

The classical approach to detecting hazardous road traffic situations relies on handcrafted features, such as vehicle speed, position, and acceleration [9]. These features are then used to train a classifier model to predict the risk level of a given situation.

Machine learning algorithms for detecting dangerous situations on the road often employ deep learning models that automatically learn features from the data [5], [10]. These models are typically trained on small datasets of traffic monitoring videos [11]. The most widely used machine learning techniques for this task are Convolutional Neural Networks (CNNs)[11] and Recurrent Neural Networks (RNNs) [12]. CNNs are deep learning models particularly well suited for processing images and videos [13]. They have proven effective in various road traffic safety applications, such as vehicle detection, road sign recognition, and lane detection[11]. RNNs, on the other hand, are better suited for tasks involving sequential data, such as traffic flow forecasting and vehicle tracking [14].

Recently, there has been growing evidence of the success of self-supervised learning in a variety of computer vision tasks [15], [16], including object detection, motion estimation, and video classification. Self-supervised learning techniques automatically learn representations from data without the need for human-labeled annotations. One of the most widely used self-supervised learning techniques is VideoMAE[2]. VideoMAE learns to represent videos by predicting masked-out patches within the video frames. These learned representations can then be used for several downstream tasks, including classification, detection, and segmentation.

To the best of our knowledge, VideoMAE has not yet been applied to detect unsafe conditions in road traffic sequences. However, its demonstrated success in other domains suggests that it could be an effective model for identifying hazardous situations in traffic scenes.

In this study, we explore the use of VideoMAE for detecting risky situations in road traffic sequences. We fine-tune the VideoMAE model on a dataset of road traffic sequences labeled as either "risky" or "non-risky". Once trained, the model can be used to detect hazardous conditions in previously unseen traffic sequences. We believe that this approach could become a valuable tool for developing innovative safety solutions in the domain of road traffic management.

III. Method

In this section, we present our proposed approach, which involves fine-tuning the VideoMAE model using a dataset of risky road traffic situations. The model achieved an accuracy of 0.95 on both the validation and testing datasets. Fig. 4 illustrates the diagram of the proposed technique. First, we review VideoMAE[2], which is based on ImageMAE[17].

A. Masked Autoencoder (ImageMAE)

The Masked Autoencoder (MAE) [17] is a neural network architecture designed to learn meaningful latent representations from data and map them to a high-dimensional space using large datasets. It operates on the principle of randomly masking portions of the input image and then reconstructing the missing regions. The model follows a key design principle: an asymmetric encoder–decoder architecture. The encoder processes only the visible (unmasked) patches, while the lightweight decoder reconstructs the complete image using the latent representations along with the mask tokens.

It randomly masks out 75% of the grid patches for reconstruction, creating a challenging and meaningful self-supervised learning task. Masked autoencoders have proven to be scalable self-supervised models for a wide range of computer vision applications. This strategy is widely adopted because of its simplicity and effectiveness.

B. Video Masked Autoencoder (VideoMAE)

Video Masked Autoencoder (VideoMAE) [2] is an efficient self-supervised learning framework for video pre-training tasks that promotes the extraction of more informative video representations. Remarkably, it delivers strong performance even when trained on limited datasets, without relying on additional external data (i.e., fully self-supervised). The model is inspired by ImageMAE [17].

VideoMAE performs video self-supervised learning using a straightforward Vision Transformer (ViT) [18] backbone combined with a basic masked autoencoder. Due to its extremely high masking ratio, VideoMAE achieves a significant reduction in pre-training time compared to contrastive learning techniques, approximately a 3.2× speedup.

The architecture of VideoMAE includes three key components: an encoder network, a decoder network, and a masking mechanism. The encoder processes input video frames, extracting hierarchical features that capture both spatial and temporal dependencies within the sequences. These encoded features are then passed to the decoder, which reconstructs the original video frames while minimizing reconstruction error. Fig. 1 illustrates the model architecture based on VideoMAE. This self-supervised approach enables VideoMAE to effectively capture nuanced behaviors and subtle indicators of risky situations in traffic videos, such as sudden braking, lane changes, wrong-direction driving, or accidents. Moreover, its ability to learn from unlabeled data reduces the dependency on large annotated datasets, making it scalable and adaptable to various traffic scenarios.

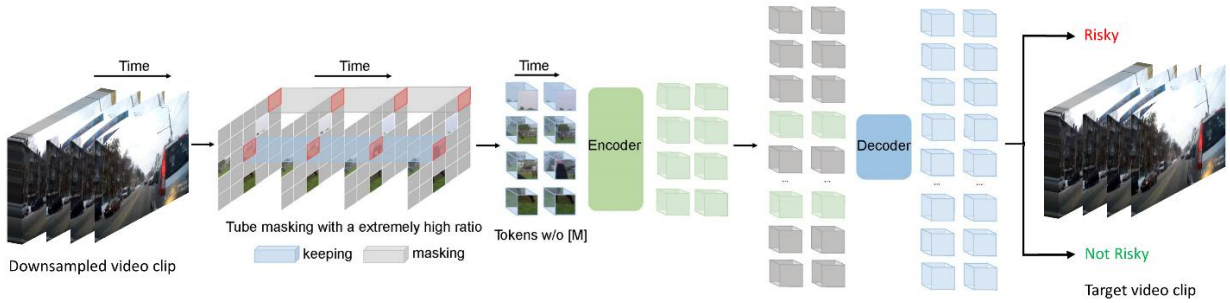


Fig. 1. Used model architecture based on video masked autoencoder.

Given these advantages, we assert that VideoMAE has strong potential as a powerful tool for detecting risky situations in road traffic sequences.

C. Data collection (Datasets)

To validate the effectiveness of the proposed model, we conduct experiments on large-scale datasets. Due to the limited availability and quality of publicly available datasets related to risky situations in road traffic, we collected and created our own data from various sources. This dataset comprises a combination of publicly available traffic video datasets. Public datasets such as the Car Crash Dataset (CCD), which is designed for traffic accident anticipation [14]; XD-Violence[19], a large-scale audio-visual violence detection dataset; and CADP [20], a dataset for accident analysis based on CCTV traffic cameras, provide a rich collection of annotated video sequences captured from various traffic scenarios. These datasets offer diverse examples of road conditions, vehicle interactions, and environmental factors. Moreover, the cameras continuously record traffic flow, capturing various risky situations such as sudden lane changes, speeding, and near-miss collisions. Fig. 2 shows some samples of risky situations from the dataset used.



Fig. 2. Samples of various risky situations on traffic roads from the used dataset.

D. Preprocessing

The collected video data undergoes several preprocessing steps, including frame extraction, resolution adjustment, and noise reduction, to ensure high-quality inputs for the VideoMAE model. This combination of diverse and comprehensive data sources enables a robust evaluation of the model's capability to detect risky situations in various traffic contexts.

The dataset contains videos with variable input frame sizes. Therefore, all videos are resized to a target resolution of 224×224 pixels. For the VideoMAE model, each input clip consists of sixteen frames. We use a frame rate of four, which defines the stride or interval between selected frames. From each video, we extract a single clip.

For training data transformations, we apply random cropping, pixel normalization, uniform temporal subsampling, and random horizontal flipping with a probability of 0.5. For the validation and evaluation datasets, we apply the same transformation pipeline, excluding random cropping and horizontal flipping, to ensure consistency during evaluation.

All preprocessing and data augmentation transformations are implemented using TorchVision's transforms module.

E. Training and Validation

The training process of the model begins with the preparation and preprocessing of the dataset, which includes frame extraction, resizing, and normalization. Typically, the dataset is divided into three subsets: training, validation, and test sets. The training set is used to train the VideoMAE model and contains a diverse range of traffic sequences, allowing the model to learn, reconstruct, and understand the underlying patterns in the data.

The validation set plays a crucial role in adjusting hyperparameters and monitoring the model's performance during training. The test set, which remains completely separate from the training and validation sets, serves as the final benchmark to assess the model's generalization ability to unseen traffic scenarios and its accuracy in detecting risky situations. Fig. 3 shows the distribution of data across each class for the three subsets of the dataset.

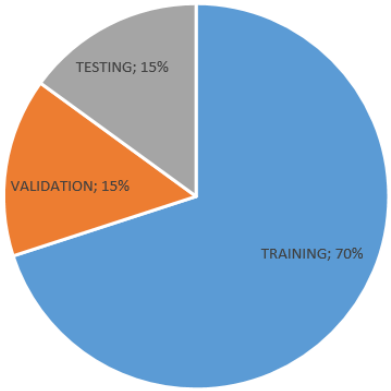


Fig. 3. Distribution of the used dataset.

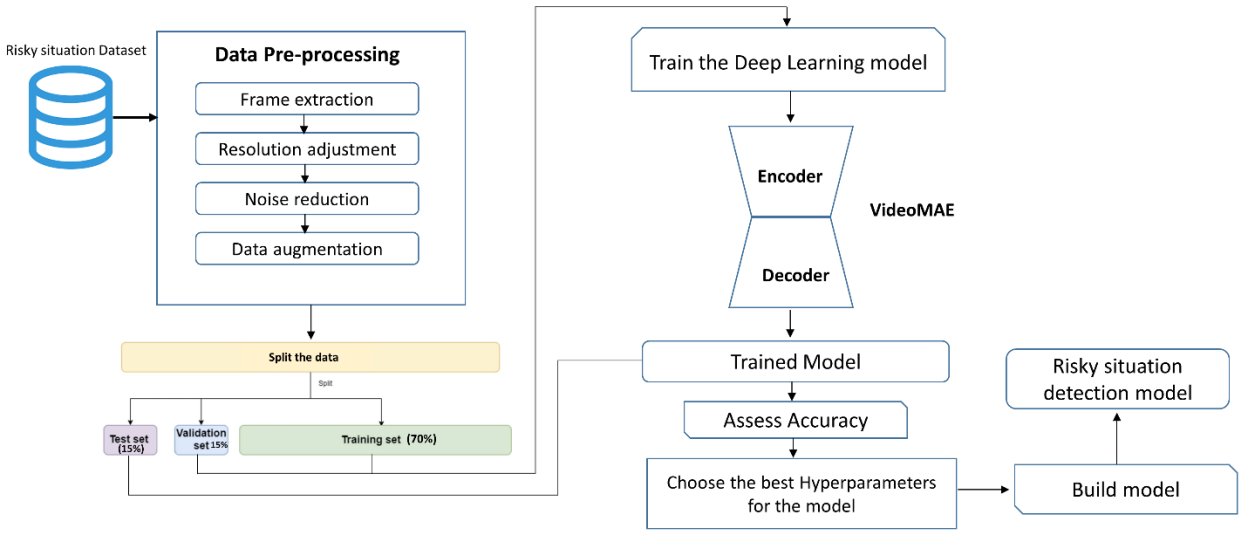


Fig. 4. Methodology diagram.

F. Evaluation Metrics

In this section, we present the metrics used to assess our model and analyze the implemented method.

Accuracy: It is an evaluation metric that measures the ratio of correctly classified sequences to the total number of sequences. It provides an assessment of how well the model's predictions align with the ground-truth annotations. In the context of video classification, accuracy is defined as the proportion of video sequences that are correctly classified into their actual categories. It reflects the precision with which the predicted labels match the true labels across the video dataset.

$$Accuracy = \frac{Total\ of\ accurate\ predictions}{Total\ of\ predictions\ made} \quad (1)$$

Precision: It is the ratio of true positive predictions to the total number of positive predictions made by the model. It indicates how many of the instances predicted as positive are actually correct, providing insight into the reliability of the model's positive classifications.

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad (2)$$

Recall: It is a measuring metric that is the ratio of the true positives that were detected to the total actual number of the same class. It shows how good the model is at detecting instances of all categories.

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \quad (3)$$

F1-score: It is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance. It is especially useful when there is an uneven class distribution or when both false positives and false negatives need to be considered equally important.

$$F1 - Score = 2\ x\ \frac{recall\ x\ precision}{recall + precision} \quad (4)$$

Confusion Matrix: The confusion matrix is an effective tool for evaluating classification tasks, as it provides a detailed summary of the model's performance. It records the number of true positives, true negatives, false positives, and false negatives, offering a comprehensive view of how well the model performs across different classes. This information enables the straightforward calculation of key evaluation metrics such as accuracy, precision, recall, and F1-score.

IV. Experiment and Results

A. Experiment and Results

The training of MAE models, especially in the context of video recognition, presents a substantial computational challenge, limiting access for researchers with limited resources. For instance, training a standard VideoMAE model on Kinetics-400 [21] requires a significant investment of up to 5.6 days with the support of 64 GPUs [22]. Therefore, we fine-tune a pre-trained VideoMAE model, which was pre-trained on the Kinetics-400 dataset for approximately 1600 epochs without any additional data. We add a randomly initialized classification head on top of the pre-trained encoder and fine-tune the model on a labeled dataset of risky situations. Additionally, we initialize the feature extractor associated with the model.

As a risk detection network, we aim to detect risky situations in road traffic sequences using VideoMAE, which is based on Vision Transformer [18]. The performance evaluation of the model on the Risky Road Traffic dataset showed the best detection performance, achieving an accuracy of 0.9 with the masking strategy.

We train our model using the PyTorch framework with an environment of 12GB of RAM and a GPU (Nvidia Tesla T4) to expedite computation and handle the complexity of video data processing.

The hyperparameter configuration for the VideoMAE model was carefully selected to optimize performance while maintaining computational efficiency. The Adam optimizer [23] was chosen for its ability to adaptively adjust learning rates during training. The learning rate was initially set to 5e-5, a value found to provide an optimal balance between fast

learning and avoiding overfitting. A weight decay of 0.05 was implemented to regularize the model and reduce overfitting potential by penalizing large weights and encouraging generalization. The batch size was set to two to prevent out-of-memory errors while maintaining a balance between computational resource limitations and gradient quality requirements. The model was trained for ten epochs, a duration sufficient for learning useful patterns without overfitting risk. A summary of these hyperparameters is provided in Table I.

TABLE I. THE HYPERPARAMETERS OF THE MODEL

Hyperparameter	Value
Optimizer	Adam
Weight decay	0.05
Learning rate	5e-5
Batch size	2
Epochs	10

This configuration enabled effective learning from traffic video sequences, achieving high accuracy in risky situation detection while maintaining robustness across diverse traffic scenarios.

B. Results

The experimental results on benchmark datasets demonstrate that our approach, which has minimal computational cost and efficient data, performs well in the unsafe situation identification task.

Epoch	Training Loss	Validation Loss	Accuracy
0	0.420000	1.151754	0.695652
1	0.422200	2.008217	0.619565
2	0.831100	0.158780	0.956522
3	0.290500	0.398978	0.923913
4	0.000300	0.573320	0.913043
5	0.001500	0.335427	0.956522
6	0.190500	0.357622	0.956522
7	0.000100	0.346179	0.945652
8	0.000100	0.304265	0.956522
9	0.000100	0.342376	0.956522

Fig. 5. Overview of the model's training process.

Fig. 5 presents an extensive description of the training of the model by tabulating the training loss, validation loss, and validation accuracy at each of the ten epochs. The training loss column follows the model's performance on the training data, showing an overall decline as the model improves at reconstructing the frames of the videos and recognizing risky situations. The validation loss column reports the model's validation set performance, which indicates how the model generalizes to new data after each pass through the epochs. The validation accuracy column reports the ratio of correct identification of risky situations on the validation set and is an unambiguous metric of the success of the model. Fig. 5 provides an overall trend of increasing validation accuracy as training advances, which flattens towards the end epochs, indicating the model's success at generalization of learned features. These metrics, together, provide a complete perspective on the learning and performance of the model during the training period.

Table II shows an overall assessment of the model's performance through various important metrics: Precision, Recall, F1-score, and Accuracy.

TABLE II. EVALUATION MODEL

Metric	Value
Precision	0.960
F1-score	0.955
Recall	0.954
Accuracy	0.95

Precision at 0.960 shows the model's capacity to accurately detect true instances of risky situations and hints at high accuracy with few false positives.

Recall at 0.954 indicates the model's capacity to detect the majority of actual risky situations and shows that it is efficient enough to capture most of the dangerous events in the traffic sequences.

The F1-score, determined as the harmonic mean of Precision and Recall, is at 0.955 and offers a balanced evaluation of the model's precision and recall performance.

Lastly, accuracy at 0.950 is the overall ratio of correctly classified instances, whether risky or non-risky. This high accuracy is an attestation to the model being consistent regardless of different scenarios.

All these metrics together illustrate the good performance of the model in detecting and classifying risky situations on the road successfully and verify its usability for real-world traffic safety improvement.

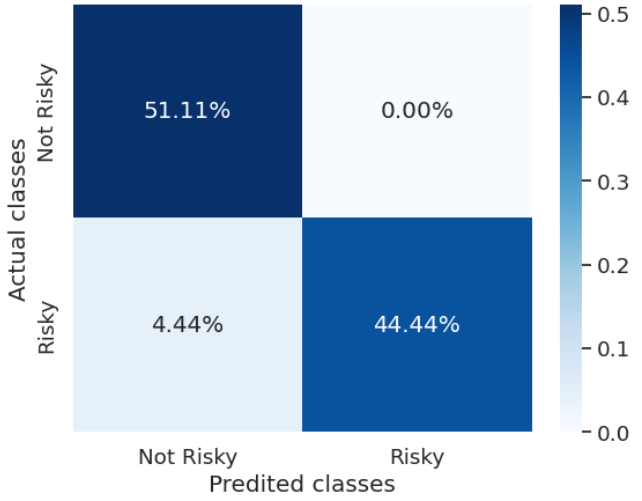


Fig. 6. Model evaluation using confusion matrix.

Fig. 6 indicates that the confusion matrix offers an insightful analysis of the model's performance by classifying the test results into true positives, true negatives, false positives, and false negatives.

The matrix shows that:

- The model successfully identified 44% of risky situations as true positives, demonstrating its strong capability to detect risky conditions in traffic sequences.
- The model correctly classified 51% of non-risky situations as true negatives, showing its good capability to identify safe conditions without false alarms.
- Notably, the model produced 0% false positives, meaning it never incorrectly labeled safe conditions as risky - a crucial factor for building confidence in the system's alerts.
- There were 4.44% false negatives, representing instances where the model failed to detect actual risky situations.

The trained model was tested on several video sequences to evaluate the proposed approach, and the results were collected and analyzed. The method demonstrates successful identification of dangerous traffic conditions in surveillance videos with 95% accuracy.

Fig. 7 illustrates the detection accuracy across multiple video sequences, showing that the model predicts correct situations with high probability.

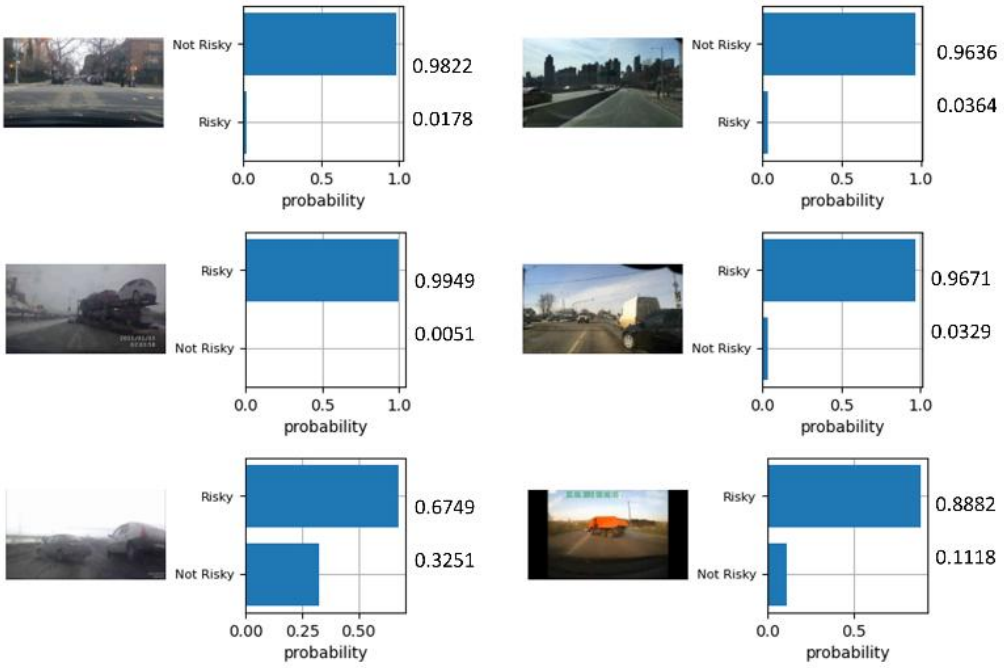


Fig. 7. Illustrations of many detection examples accompanied by the probability.

V. Discussion

The results obtained from the model highlight its strong potential to enhance traffic safety through the accurate detection of risky situations in road traffic sequences. The model's high Precision of 0.960 indicates its ability to correctly identify hazardous situations with minimal false alarms, which is critical in preventing unnecessary interventions. The Recall of 0.954 reflects the model's effectiveness in capturing nearly all actual risky situations, ensuring that critical events are not overlooked. The F1-score of 0.955 demonstrates a well-balanced performance between Precision and Recall, confirming that the model maintains consistency in identifying and classifying risks. Additionally, the accuracy of 0.95 indicates that the model performs reliably across different traffic scenarios. The confusion matrix also confirms these results with 44% true positives and no false positives. The absence of false alarms is essential for real-time traffic monitoring applications, since false alarms can cause driver desensitization or unjustified intervention. However, the model reported 4.44% false negatives, where it did not detect actual risky situations. This indicates that although the model is very accurate for the situations it does detect, there is further scope to improve its sensitivity to detect all potential danger areas. These results indicate the model's resilience but also the necessity for further improvement, specifically to increase its capability to detect all risky situations so that the occurrence of false negatives is minimized and overall traffic safety is comprehensive.

When compared to other methodologies for risky situation detection in road traffic, the model provides various improvements over previous approaches. The conventional approach tends to employ Convolutional Neural Networks (CNNs) to analyze single frames from traffic videos. Although CNNs are good at learning spatial features, they are weak at accounting for temporal dependencies between frames, since risky situations are often realized through unfolding events over time. To address the weakness of CNNs, Recurrent Neural Networks such as LSTM[24] networks are sometimes

employed together with CNNs to consider the temporal aspect. However, these models tend to perform poorly with long-range dependencies and are prone to vanishing gradients, thus unable to adequately understand intricate, multi-frame traffic situations.

By contrast, VideoMAE is transformer-based and specifically designed to model both spatial and temporal dependencies simultaneously. This enables the model to capture sequences of events in traffic, like an abrupt stop of a car or the appearance of a pedestrian on the roadway, which may not be detected or well-represented using standard CNN or RNN-based approaches.

In further testing the efficacy of our approach, we compared the model's performance with that of a Vision Transformer (ViT) model that has also been used for detecting risky situations from videos. As evident from Table III, our model performs significantly better than the ViT-based approach and other deep learning based methods, achieving 95% accuracy compared to 92% for the ViT model[25]. In addition, the work presented in [26], which aimed to predict traffic accidents using an RNN network, achieved an accuracy of 71%. In [27], the authors used a combination of models, including a multi-layer perceptron (MLP), and achieved an accuracy of 72%. Moreover, another study focused on road accident prediction by combining machine learning and deep learning models, specifically the RFCNN model [28], and achieved an accuracy of 81%. These values demonstrates VideoMAE's superior capability to handle both spatial and temporal dependencies in traffic sequences.

TABLE III. MODEL COMPARISON

Model	Accuracy
RNN[26]	71%
MLP[27]	72%
RFCNN[28]	81%
FN-ViT[25]	92%
Ours	**95%**

VI. CONCLUSION

In summary, the findings of this study demonstrate the capability of the VideoMAE-based model to detect risky conditions in road traffic streams and its potential contribution to enhancing traffic safety through advanced video data processing. The strong performance of the model, reflected by high Precision, Recall, and F1-score values, highlights its ability to accurately identify dangerous situations.

While the system demonstrates strong potential and high accuracy, a few drawbacks or areas for improvement have been identified. First, the training process is computationally intensive. Second, there is the occurrence of false negatives, where some risky situations go undetected, highlighting the need for improved sensitivity, particularly in complex traffic scenes. Nevertheless, integrating the model into real-time traffic monitoring systems holds significant promise for improving road safety by enabling early warnings and reducing the risk of accidents.

Overall, this work offers valuable insights into the application of deep learning for traffic safety and paves the way for future advancements in intelligent transportation systems.

REFERENCES

[1] A. A. Mohammed, K. Ambak, A. M. Mosa, et D. Syamsunur, « A Review of the Traffic Accidents and Related Practices Worldwide », Open Transp. J., vol. 13, no 1, p. 65-83, juin 2019, doi: 10.2174/1874447801913010065.

[2] Z. Tong, Y. Song, J. Wang, et L. Wang, « VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training », 18 octobre 2022, arXiv: arXiv:2203.12602. Consulté le: 16 septembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2203.12602

[3] S. K. Kumaran, D. P. Dogra, et P. P. Roy, « Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey », ACM Comput. Surv., vol. 53, no 6, p. 1-26, nov. 2021, doi: 10.1145/3417989.

[4] Y. Ma, J. Xu, C. Gao, M. Mu, G. E, et C. Gu, « Review of Research on Road Traffic Operation Risk Prevention and Control », Int. J. Environ. Res. Public. Health, vol. 19, no 19, p. 12115, sept. 2022, doi: 10.3390/ijerph191912115.

[5] A. Berroukham, K. Housni, M. Lahraichi, et I. Boulfrifi, « Deep learning-based methods for anomaly detection in video surveillance: a review », Bull. Electr. Eng. Inform., vol. 12, no 1, p. 314-327, févr. 2023, doi: 10.11591/eei.v12i1.3944.

[6] M. Sobhana, G. S. S. Venkatesh Mendu, N. Vemulapalli, et K. Kumar Chintakayala, « Optimized feature selection approaches for accident classification to enhance road safety », IAES Int. J. Artif. Intell. IJ-AI, vol. 13, no 3, p. 3283, sept. 2024, doi: 10.11591/ijai.v13.i3.pp3283-3290.

[7] S. Bouhsissin, N. Sael, et F. Benabbou, « Driver Behavior Classification: A Systematic Literature Review », IEEE Access, vol. 11, p. 14128-14153, 2023, doi: 10.1109/ACCESS.2023.3243865.

[8] A. Salbi, M. A. Gadi, T. Bouganssa, A. Eloudrhiri Hassani, et A. Lasfar, « Design and implementation of a driving safety assistant system based on driver behavior », IAES Int. J. Artif. Intell. IJ-AI, vol. 13, no 3, p. 2603, sept. 2024, doi: 10.11591/ijai.v13.i3.pp2603-2613.

[9] J. Zhao et al., « Unsupervised Traffic Anomaly Detection Using Trajectories », p. 8.

[10] L. Tišljarić, S. Fernandes, T. Carić, et J. Gama, « Spatiotemporal Road Traffic Anomaly Detection: A Tensor-Based Approach », Appl. Sci., vol. 11, no 24, p. 12017, déc. 2021, doi: 10.3390/app112412017.

[11] S. W. Khan et al., « Anomaly Detection in Traffic Surveillance Videos Using Deep Learning », Sensors, vol. 22, no 17, p. 6563, août 2022, doi: 10.3390/s22176563.

[12] R. M. Schmidt, « Recurrent Neural Networks (RNNs): A gentle Introduction and Overview », 23 novembre 2019, arXiv: arXiv:1912.05911. Consulté le: 4 août 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1912.05911

[13] K. O'Shea et R. Nash, « An Introduction to Convolutional Neural Networks », 2 décembre 2015, arXiv: arXiv:1511.08458. Consulté le: 22 mai 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1511.08458

[14] W. Bao, Q. Yu, et Y. Kong, « Uncertainty-based Traffic Accident Anticipation with Spatio-Temporal Relational Learning », in Proceedings of the 28th ACM International Conference on Multimedia, oct. 2020, p. 2682-2690. doi: 10.1145/3394171.3413827.

[15] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, et M. Shah, « Anomaly Detection in Video via Self-Supervised and Multi-Task Learning », 10 septembre 2021, arXiv: arXiv:2011.07491. Consulté le: 22 février 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2011.07491

[16] N. Madan et F. S. Khan, « Self-Supervised Masked Convolutional Transformer Block for Anomaly Detection », IEEE Trans. PATTERN Anal. Mach. Intell., vol. 14, no 8, 2022.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, et R. Girshick, « Masked Autoencoders Are Scalable Vision Learners », in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New

Orleans, LA, USA: IEEE, juin 2022, p. 15979-15988. doi: 10.1109/CVPR52688.2022.01553.

[18] A. Dosovitskiy et al., « An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale », 3 juin 2021, arXiv: arXiv:2010.11929. Consulté le: 22 mai 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/2010.11929

[19] P. Wu et al., « Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision », in Computer Vision – ECCV 2020, vol. 12375, A. Vedaldi, H. Bischof, T. Brox, et J.-M. Frahm, Éd., in Lecture Notes in Computer Science, vol. 12375. , Cham: Springer International Publishing, 2020, p. 322-339. doi: 10.1007/978-3-030-58577-8_20.

[20] A. P. Shah, J.-B. Lamare, T. Nguyen-Anh, et A. Hauptmann, « CADP: A Novel Dataset for CCTV Traffic Camera based Accident Analysis », in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand: IEEE, nov. 2018, p. 1-9. doi: 10.1109/AVSS.2018.8639160.

[21] W. Kay et al., « The Kinetics Human Action Video Dataset », 19 mai 2017, arXiv: arXiv:1705.06950. Consulté le: 31 juillet 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1705.06950

[22] Xianhang Li, Peng Wang, Xinyu Li, Heng Wang and Cihang Xie, « Efficient VideoMAE via Temporal Progressive Training », 2024. [En ligne]. Disponible sur: https://openreview.net/forum?id=vex1yNHNFL

[23] D. P. Kingma et J. Ba, « Adam: A Method for Stochastic Optimization », 29 janvier 2017, arXiv: arXiv:1412.6980. Consulté le: 12 juin 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1412.6980

[24] S. Hochreiter et J. Schmidhuber, « Long Short-Term Memory », Neural Comput., vol. 9, no 8, p. 1735-1780, nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[25] A. Berroukham, K. Housni, et M. Lahraichi, « Fine-Tuning Pre-trained Vision Transformer Model for Anomaly Detection in Video Sequences », in Proceedings of the 6th International Conference on Big Data and Internet of Things, vol. 625, M. Lazaar, E. M. En-Naimi, A. Zouhair, M. Al Achhab, et O. Mahboub, Éd., in Lecture Notes in Networks and Systems, vol. 625. , Cham: Springer International Publishing, 2023, p. 279-289. doi: 10.1007/978-3-031-28387-1_24.

[26] M. Sameen et B. Pradhan, « Severity Prediction of Traffic Accidents with Recurrent Neural Networks », Appl. Sci., vol. 7, no 6, p. 476, juin 2017, doi: 10.3390/app7060476.

[27] L. Wahab et H. Jiang, « Severity prediction of motorcycle crashes with machine learning methods », Int. J. Crashworthiness, vol. 25, no 5, p. 485-492, sept. 2020, doi: 10.1080/13588265.2019.1616885.

[28] M. Manzoor et al., « RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model », IEEE Access, vol. 9, p. 128359-128371, 2021, doi: 10.1109/ACCESS.2021.3112546.

Vision-Based Vehicle Classification Using Deep Learning Model

Ahsiah Ismail[1*], Amelia Ritahani Ismail[2], Muhammad Afiq Mohd Ara[3], Asmarani Ahmad Puzi[4], Suryanti Awang[5]
Department of Computer Science-Kulliyyah of Information and Communication Technology, International Islamic University Malaysia (IIUM), 53100 Kuala Lumpur, Malaysia[1, 2, 3, 4]
Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia[5]

***Abstract*—Vehicle classification offers intelligent solutions for road traffic monitoring by enabling future prediction planning and decision making. Predictive analytics can be used to predict traffic congestion based on the types of vehicles on the road. In this research, the reliability of deep learning based models for vision-based vehicle classification is investigated. Four models of You Only Look Once (YOLO) are investigated, namely YOLOv5s, YOLOv5x, YOLOv10n, and YOLOv12n. These models were trained and evaluated on a vehicle dataset comprising five vehicle classes, which are Ambulance, Bus, Car, Motorcycle, and Truck, with a total number of 1103 images. From the experiment conducted, YOLOv10n achieved the highest performance measure of mAP@0.5 with 0.859 across all vehicle classes, including per-class evaluation, demonstrating superior detection compared to the other models. Finally, the results indicate that the YOLOv10n model can be used in vision-based vehicle classification.**

Keywords*—YOLO; vehicle classification; deep learning; traffic monitoring

I. Introduction

Vision-based vehicle classification offers an intelligent solution for transportation systems, thus contributing to road traffic monitoring for the development of smart cities [1],[2],[3]. With the growing number of vehicles on the road, efficient and accurate vehicle classification systems are essential for managing urban mobility, enhancing public safety, and reducing traffic congestion [4],[5],[6]. Vehicle detection technologies involve the use of cameras and vision sensors to capture road footage, which is then analyzed to detect, classify, and track the movement of vehicles [7],[6]. Using these technologies may improve traffic control for future planning and decision making based on the current vehicle data on the road [8], [9]. Further data analytics can also be performed based on the current vehicle data on the road for traffic monitoring [10]. The analytics process identifies the patterns and learns for future planning and prediction, which helps to improve road traffic monitoring.

For this reason, a vision-based vehicle classification is proposed. The vision-based vehicle classifications are designed based on a deep learning method. The model of the deep learning methods are evaluated to determine the best vision-based classification for detecting types of vehicle. The deep learning method is chosen in this research due to its ability to detect objects in more complex environments under varying environmental conditions. In vehicle classification, the detection of vehicles especially on this condition is important. Failure to detect these may reduce the classification accuracy.

This research focuses on the detection of suitable methods that are able to increase the aforementioned classification accuracy in small sample vehicle dataset images. A small sample dataset of vehicles (1103 images) was utilized in this research as a proof-of-concept to classify vehicle images. To suite the problem stated previously, we investigate the deep learning based method, particularly Convolutional Neural Networks (CNNs). The CNN is chosen as it is able to provide a more robust alternative by learning discriminative features directly from images, enabling more accurate detection and classification. The CNN model has emerged as a powerful alternative due to its ability to learn spatial hierarchies of features directly from image data [11]. Among the CNN deep learning based object detection frameworks, the model You Only Look Once (YOLO) is the most suitable method for real-time vehicle detection tasks as it is able to balance between speed and accuracy, making it highly suitable [12]. This research investigates deep learning YOLO models, namely, YOLOv5s, YOLOv5x, YOLOv10n, and YOLOv12n, for vehicle classification. Each of these models are trained using pretrained weights to reduce training time while maintaining accuracy during the transfer learning. These models are tested using a vehicle dataset consisting of five types of vehicles, which are Ambulance, Bus, Car, Motorcycle, and Truck.

The rest of this study is organized as follows: Section II presents the related works. Section III describes the materials and methods used. The results are presented in Section IV, followed by the discussion in Section V. Finally, the conclusion and future work are presented in Section VI.

II. Related Works

Vehicle detection plays a vital role in intelligent transportation systems to enable intelligent solution for road traffic monitoring [13]. The computer vision method is used to detect the vehicle on a road traffic. In early research, the commonly used methods for vehicle detection are background subtraction, edge detection, and Haar Cascades [5,6,14]. Although these methods are computationally efficient, they were less effective in detecting the images under varying environmental conditions, such as changes in illumination, shadows, and occlusion, thus limiting their effectiveness in real-world scenarios [15].

In more recent object recognition method, the deep learning method have gain attention due to its ability to detect object in most classification task. Compared to other methods such as edge detection and background subtraction method, the CNN

method is one of the most reliable deep learning based method which provides more accurate detection and classification. The other methods often shown to be less effective in detecting the object under varying environmental conditions and in real-time, dynamic environments due to their sensitivity to lighting conditions and noise [11,16]. In the CNN deep learning based object detection frameworks, the YOLO models are widely used for object detection due to their ability to balance between speed and accuracy. It is also suitable for real-time detection [12]. Although the other models of the CNN method, such as R-CNN, Fast R-CNN, and Faster R-CNN, demonstrated significant improvements in accuracy detection by learning hierarchical features from the data, these models rely on region proposals that led to slower inference speeds, making them less suitable for real-time detection tasks [8,9,17]. On the other hand, the YOLO models enable fast and accurate detection in a single pass over the image. YOLO is a single-stage object detection framework that performs both object localization and classification in one forward pass through the network [18]. This design makes the YOLO model efficient for real-time applications such as traffic surveillance and vehicle monitoring. Unlike two-stage detectors like R-CNN, which separate the tasks of region proposal and classification, YOLO treats detection as a single regression problem, mapping image pixels directly to bounding box coordinates and class probabilities. This architecture offers faster inference times and smaller model sizes, making YOLO ideal for the deployment of edge devices and in time-sensitive environments. Different YOLO variants offer different unique strengths and limitations. The challenge is in selecting the most suitable model that is able to balance accuracy with computational time performance while maintaining a high classification.

The YOLO variants, such as YOLOv3 and YOLOv4 models, strike a balance between speed and precision [19]. While YOLOv5 series models improve the limitation by architectural optimizations and improved training strategies [11,12]. Among the YOLOv5 series, YOLOv5s and YOLOv5x provide trade-offs between computational efficiency and accuracy, making them ideal for both edge and cloud-based deployments. The YOLOv5s model is a lightweight YOLO architecture and one of the smallest models in the YOLOv5 variant. YOLOv5s model offers fast training and inference with minimal resource consumption. This makes the model an excellent starting point for experimentation and prototyping, especially when assessing trade-offs between speed and accuracy. On the other hand, YOLOv5x is designed to maximize detection accuracy and mean Average Precision (mAP). It represents the upper bound of the YOLOv5 variants, with deeper architecture and more parameters. However, the YOLOv5x model requires a higher computational cost and longer inference time. Recent variants such as YOLOv10 and YOLOv12 continue to improve the YOLO performance by optimizing depth, parameter count, and inference speed, especially to be implemented in resource-constrained environments [20]. YOLOv10 model is a lightweight model but is able to offer high performance [20]. The model is optimized for low-latency, high-efficiency deployment, particularly on edge devices. It offers better performance-per-FLOP than earlier nano-scale models such as YOLOv5s. The YOLOv10 is built to deliver real-time performance without reducing its accuracy performance. Compared with the other earlier YOLO model such as YOLOv5s, YOLOv10 model improve the model architecture and enhance the efficiency and accuracy even in a compact model. The latest generation nano-scale model in YOLO variant is the YOLOv12n model. This model been designed as a compact real-time model that able to offer high efficiency and competitive accuracy. The YOLOv12n model incorporating attention-centric modules for smarter and more selective feature extraction [21].

Despite all these improvements, object recognition remains challenging, particularly in scenarios involving large and complex object images. The detection accuracy often drops significantly for large objects under poor lighting conditions, especially in large object images taken during night conditions, due to increased noise and reduced contrast in large object images [22].

From the above review, the YOLO model is seen to be the most suitable model for vision-based vehicle classification due to its capability to perform both object localization and classification in one forward pass through the network. This design makes YOLO model suitable and efficient for real-time applications such as traffic surveillance and vehicle monitoring. Therefore, in this research, the YOLO variants are investigated to detect and classify the vehicle, particularly for large vehicle images.

III. Materials And Methods

Generally, the proposed work can be divided into two phases: detection and classification. In this research, four YOLO models are evaluated, which are YOLOv5s, YOLOv5x, YOLOv10n, and YOLOv12n. The overall scheme for the vehicle classification models evaluated is shown in Fig. 1. Each of these models is selected due to its popularity and stability for the object classification task. The YOLOv5s is selected as it offers fast training and inference with minimal resource consumption. The YOLOv5s is the smallest model in the YOLOv5 variant. This makes an excellent starting point for experimentation and prototyping, especially when assessing trade-offs between speed and accuracy. While the YOLOv5x model requires higher computational, and longer inference time. The YOLOv5x model represents the upper bound with deeper architecture and more parameters. Although it requires higher computational resources, YOLOv5x is designed to maximize detection accuracy and mean Average Precision (mAP). On the other hand, YOLOv10n and YOLOv12n models are chosen due to these models are the recent iterations of YOLO variants that optimize depth, parameter count, and inference speed in the YOLO architecture.

The model architecture of vehicle classification using YOLOv5s, YOLOv5x, YOLOv10n, and YOLOv12n are shown in Fig. 2(a), Fig. 2(b), Fig. 2(c) and Fig. 2(d). As shown in Fig. 2, each of the models consist three main parts which are backbone network, neck network and detection head. The backbone network is the first step in YOLO model architecture, where the features extracted from the vehicle input images. Then, the feature fusion in the neck part, where feature maps from different scales of the backbone network are fused to a neck network to process the detection. The head part completes the final prediction which includes the bounding boxes and the

associated class labels to obtain the detection result. Three different scales of detection heads are used to detect small, medium and large vehicles.

To demonstrate the reliability of the selected model for vision-based vehicle classification, a series of comprehensive experiments is conducted. All the YOLO models are trained using Google Colab with GPU acceleration, and the dataset tested is annotated using CVAT.ai and hosted on Roboflow for streamlined training and validation. The training environment includes dependencies of torch, numpy, opencv, and the official YOLO repositories for each YOLO version.

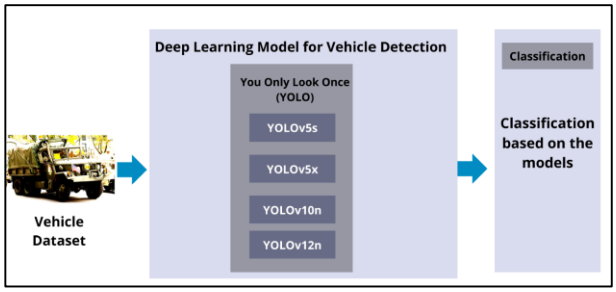
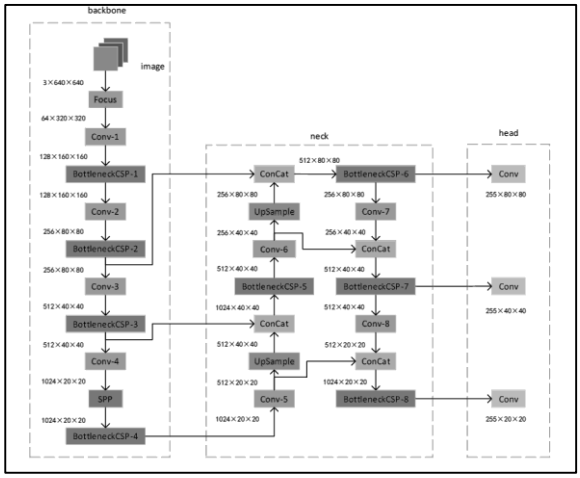
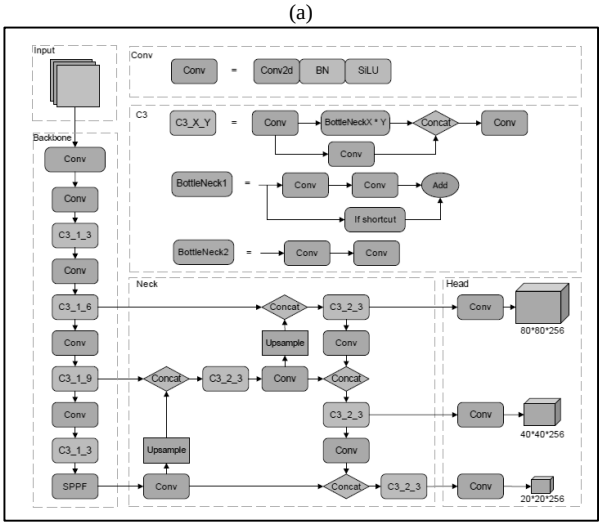


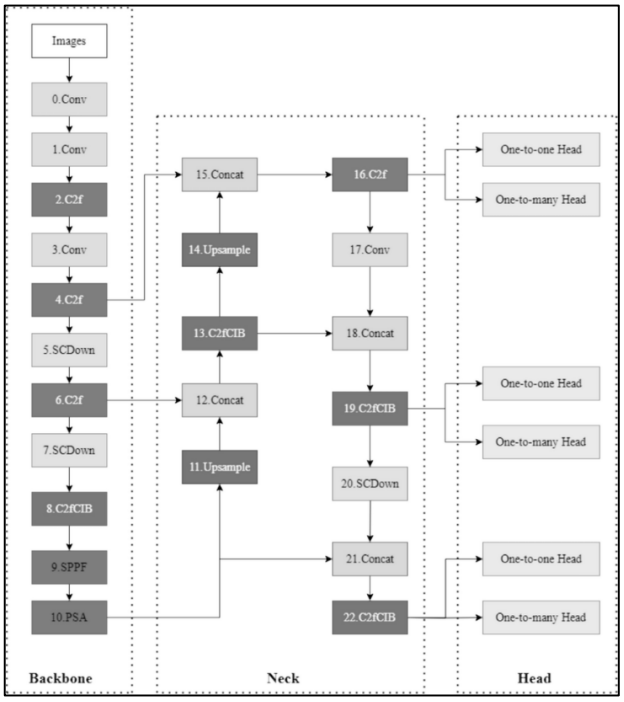
Fig. 1. Overall scheme of the proposed vision-based vehicle classification with four different YOLO models (YOLOv5s, YOLOv5x, YOLOv10n, and YOLOv12n).



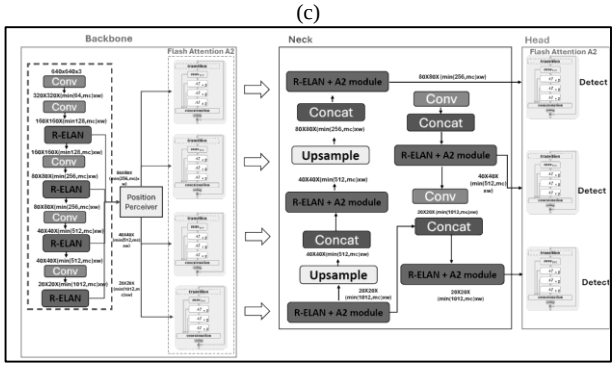
(a)



(b)



(c)



(d)

Fig. 2. YOLO model architecture for: (a) YOLOv5s [23], (b) YOLOv5x [24] (c) YOLOv10n [25] and (d) YOLOv12n [26].

The dataset used in this research consists of 1103 annotated images categorized into five vehicle types which are Ambulance (161 images), Bus (84), Car (321), Motorcycle (247), and Truck (290). These images were manually curated to ensure balanced class representation and relevance to real-world conditions. Each image was manually labelled with bounding boxes and associated class labels using CVAT.ai (Computer Vision Annotation Tool), an open-source tool suitable for precise object detection annotation. Bounding boxes were drawn around each vehicle, and class names were assigned to each types of vehicle accordingly. The properties of the dataset used are shown in Table I.

To train and evaluate the models effectively, the dataset was divided into training, validation and test set with a ratio of 70%, 20% and 10% as shown in Table II.

This dataset is split to ensure all classes are proportionally represented in each set, for effective training and accurate performance evaluation. There are two processes applied on all the images used in the experiment, which are auto-orientation

and resize. The auto-orientation is applied to ensure that all images are correctly aligned based on EXIF metadata, to avoid misalignment issues during the training phase. All the images are also resized to fit into the YOLO models, which requires fixed size of input. All images were resized to 640×640 pixels to fit into the YOLO model input layer while maintaining acceptable aspect ratios for object preservation. These pre-processed datasets allowed uniform input across all models to ensure a fair comparison. To ensure a fair comparison in the model configuration, the same training parameters settings used in all YOLO models tested as listed in Table III.

TABLE I. DATASET DISTRIBUTION

Classes	Number of Images
Ambulance	161
Bus	84
Car	321
Motorcycle	247
Truck	290
Total	1103

TABLE II. DATASET SPLIT

Dataset Split	Training Set (70%)	Validation Set (20%)	Test Set (10%)
Images	772	221	110

TABLE III. PARAMETER SETTING FOR YOLO MODELS TRAINING

Model / Parameter	YOLOv5s	YOLOv5x	YOLOv10n (Selected Model)	YOLOv 12n
Image Size	640×640 pixels	640×640 pixels	640×640 pixels	640×64 0 pixels
Epoch Size	50	50	50	50
Batch Size	15	15	15	15
Pretrained Weights	Enable	Enabled	Enabled	Enabled

In this research, Roboflow was used to provide seamless dataset integration into the YOLO training scripts. Each model was saved as best.pt, representing the model weights with the highest validation performance during training.

IV. RESULTS

In all the experiments, the performance of each YOLO models on the vehicle classification are measured in terms of Precision, Recall, Mean Average Precision (mAP), and Training Time. Precision measures the proportion of true positive detections among all positive predictions and evaluates how many vehicles were correctly identified. On the other hand, recall evaluates the correctly identify all actual vehicle instances in the dataset. It is the proportion of true positive detections out of all actual positive instances. While the Mean Average Precision (mAP) is also considered in this research since most of the work related with object detection from the literature use mAP to evaluate performance measure in their work [27,28]. In all the experiment conducted, the Training Time had also been recorded to measure the suitability of each of the model for the deployment under limited time or processing power environments. Models with a less Training Time and able to maintain high detection accuracy are more practical to be implemented in real-time or resource-sensitive applications. The classification performance of all models are defined as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \quad (3)$$

$$AP_k = the\ AP\ of\ class\ k, n = the\ number\ of\ class$$

The models are trained and evaluated on a vehicle dataset image that consists of five vehicle classes. The performance of each model, namely YOLOv5s**,** YOLOv5x, YOLOv10n, and YOLOv12n, is presented in Table IV.

TABLE IV. COMPARATIVE RESULTS (PRECISION, RECALL, TIME, MAP@0.5) FOR YOLOV5S, YOLOV5X, YOLOV10N AND YOLOV12N

Models	Precision	Recall	Training Time (hrs)	mAP@0.5
YOLOv5s	0.823	0.882	0.263	0.763
YOLOv5x	0.921	0.909	1.222	0.855
YOLOv10n (Selected Model)	0.912	0.886	0.296	0.859
YOLOv12n	0.925	0.883	0.350	0.832

As shown in Table IV, the YOLOv10n model demonstrates superior performance among all the models tested with 0.859 of mAP@0.5. This is followed by the YOLOv5x model with 0.855 of mAP@0.5 and the YOLOv12n with 0.832 of mAP@0.5. The lowest mAP@0.5 is observed in the YOLOv5s model with 0.763 mAP@0.5. YOLOv10n model has shown to be the most optimum model with highest maAp@0.5 of 0.859 obtain in a less Training Time requires to train the models. The lowest Training Time among all the models is the YOLOv5s model, with 0.263hours. This is followed by the YOLOv10n model with 0.296 hours of Training Time. Only a minimal difference is observed between the YOLOv10n model with YOLOv5s model which less than 0.033 hours Training Time between the YOLOv10n model. Despite the YOLOv10n model being smaller than the YOLOv5 model variant, the YOLOv10n model is able to achieve the highest maAp@0.5 with only 0.263 hours of Training Time taken. The YOLOv10n model is superior compared to the other models in terms of both maAp@0.5 and Training Time. The results show that the YOLOv10n model is suitable for use in real-time applications and edge deployments. On the other hand, YOLOv12n outperformed the mAP@0.5 of YOLOv5s models while maintaining a much smaller size. Although YOLOv5s is the fastest model, it shows the lowest mAP@0.5 with 0.763. The YOLOv5s model is able to obtain faster training time due to its lightweight baseline model, making it able to train the model faster. However, the lowest mAP@0.5 is observed in the model.

To further investigate the performance of each models, the models are further evaluated on the specific class performance for each of the five vehicle classes which are Ambulance, Bus, Car, Motorcycle and Truck. Their results of precision, recall and map@0.5 are presented in Table V.

TABLE V. COMPARATIVE RESULTS PER-CLASS PERFORMANCE (PRECISION, RECALL AND MAP@0.5) FOR YOLOV5S, YOLOV5X, YOLOV10N AND YOLOV12N

Class	Models	Precision	Recall	mAP@0.5
Ambulance	YOLOv5s	0.771	0.882	0.916
	YOLOv5x	0.908	0.971	0.987
	YOLOv10n (Selected Model)	1.00	0.932	0.979
	YOLOv12n	0.967	0.864	0.933
Bus	YOLOv5s	0.835	0.954	0.929
	YOLOv5x	0.928	0.954	0.984
	YOLOv10n (Selected Model)	0.890	0.969	0.979
	YOLOv12n	0.914	0.985	0.986
Car	YOLOv5s	0.765	0.938	0.910
	YOLOv5x	0.859	0.837	0.927
	YOLOv10n (Selected Model)	0.902	0.822	0.928
	YOLOv12n	0.922	0.849	0.920
Motorcycle	YOLOv5s	0.904	0.846	0.947
	YOLOv5x	1.00	0.911	0.984
	YOLOv10n (Selected Model)	0.937	0.846	0.966
	YOLOv12n	0.958	0.873	0.971
Truck	YOLOv5s	0.840	0.789	0.897
	YOLOv5x	0.909	0.873	0.929
	YOLOv10n (Selected Model)	0.829	0.860	0.952
	YOLOv12n	0.863	0.842	0.917

From Table V, the YOLOv5x model give the best performance in Ambulance class with 0.987 of mAP@0.5. This is followed by YOLOv10n model with 0.979 of mAP@0.5, YOLOv12n model with 0.933 of mAP@0.5. The lowest mAP@0.5 for Ambulance class is observed in YOLOv5s model with 0.916. While in Bus class, YOLOv12n model is outperformed the other models with 0.986 of mAP@0.5. This is followed by YOLOv5x model with 0.984 of mAP@0.5, YOLOv10n model with 0.979 of mAP@0.5. The lowest mAP@0.5 for Bus class is observed in YOLOv5s model with 0.929 of mAP@0.5. For Car class, YOLOv10n models is outperformed the other models with 0.928 of mAP@0.5. This is followed by YOLOv5x model with 0.927 of mAP@0.5, YOLOv12n model with 0.920 of mAP@0.5. The lowest mAP@0.5 for Car class is observed in YOLOv5s model with 0.910 of mAP@0.5. On the other hand, for Motorcycle class, YOLOv5x models is outperformed the other models with 0.984 of mAP@0.5. This is followed by YOLOv12n model with 0.971 of mAP@0.5, YOLOv10n model with 0.966 of mAP@0.5. The lowest mAP@0.5 for Motorcycle class is also observed in YOLOv5s model with 0.947 of mAP@0.5. Lastly, for the Truck class, the highest performance of mAP@0.5 is YOLOv10n model. This is followed by YOLOv5x with 0.929 of mAP@0.5, YOLOv12n model with 0.917 of mAP@0.5. The lowest mAP@0.5 for Truck class is also observed in YOLOv5s model with 0.897 of mAP@0.5. Among all the models tested across all five vehicle classes, YOLOv10n model is seen to be the most optimum models as it is able to achieve the highest mAP@0.5 for the two classes which are Car and Truck classes. While the lowest model which can be observed for all the five tested vehicle class is the YOLOv5s model.

V. DISCUSSION

Overall, YOLOv10n model demonstrate superior performance of mAP@0.5 among all the models for all evaluation including the per-class performance evaluation from the experiments conducted. While the lowest mAP@0.5 results are observed in YOLOv5s for all performance evaluations. The highest performance can be seen on YOLOv10n model due to its improvement in the model architecture that enhance both efficiency and accuracy. The model is able to balance trade-off between speed and accuracy while retaining its compact versions. Although the YOLOv10 model is a lightweight, the model able to obtain high-performance due to the model architecture that is optimized for low-latency, better performance-per-FLOP than earlier nano-scale models like a YOLOv5s. The YOLOv10 model is built to deliver real-time performance without reducing its accuracy performance [20]. From the experiment conducted, the lowest performance is observed in YOLOv5s models due to its lightweight, simplicity and it is the smallest YOLO architecture, thus reducing its recognition performance. The YOLOv5s and YOLOv5x model is shown to be less effective especially on detecting the large vehicle images such as Truck vehicle. YOLOv5x model obtain low confidence score detection on truck that is clearly present in the image. On the other hand, the YOLOv10n model able to obtain high confidence score detection and able to detect most of the truck. This is followed by the YOLOv12n model. The comparison models for the confidence score detection on Truck is shown in Fig. 3.



Fig. 3. Comparison models for confidence score detection on Truck: (a) YOLOv5s, (b) YOLOv5x, (c) YOLOv10n (d) YOLOv12n.

Despite YOLOv10n able to detect most of the vehicles including large vehicle and various conditions, the low confidence score detection still can be observed on some of the dark colour truck, occluded motorcycle images and bus are shown in Fig. 4.



Fig. 4. Examples of the low confidence score detection on YOLOv10n.

VI. Conclusion and Future Work

A vision-based vehicle classification is proposed as it is able to offer intelligent solution for transportation systems, thus contribute to road traffic monitoring for the development of smart cities. Four models of YOLO are investigated, namely YOLOv5s, YOLOv5x, YOLOv10n, and YOLOv12n are evaluated on a multi-class vehicle dataset. From the experiment conducted, the YOLOv10 able to obtain the highest detection performance of mAP@0.5 with 0.859 across all vehicle classes compared to the other model. Although, YOLOv10n is able to achieve high performance, low confidence can be observed on the images as shown in Fig. 4. It can be seen that the models are less effective in detecting the dark colour truck, occluded motorcycle and bus images. The lighting variations, especially in the detection of large, visually complex vehicles like trucks or bus may affected the detection and recognition. Thus, reducing the performance of the models. In the future, we will concentrate on improving these drawbacks by optimized the anchor box and modified the loss function to improve the detection especially on a large vehicle such as bus and truck. Increasing the diversity of the vehicle images dataset will also be considered by using data augmentation techniques for more advanced deep learning models to optimize the performance.

Acknowledgment

This research was funded by UMP-IIUM Sustainable Research Collaboration 2022 grant (IUMP-SRCG22-015-0015).

References

[1] M. Balfaqih, S. A. Alharbi, M. Alzain, F. Alqurashi, and S. Almilad, “An accident detection and classification system using internet of things and machine learning towards smart city,” Sustain., vol. 14, no. 1, pp. 1–13, 2022.

[2] S. Awang, N. M. A. N. Azmi, and M. A. Rahman, “Vehicle Type Classification Using an Enhanced Sparse-Filtered Convolutional Neural Network with Layer-Skipping Strategy,” IEEE Access, vol. 8, pp. 14265–14277, 2020.

[3] B. Kidmose, “A review of smart vehicles in smart cities: Dangers, impacts, and the threat landscape,” Veh. Commun., p. 100871, 2024.

[4] S. S. Harsha and K. R. Anne, “Gaussian Mixture Model and Deep Neural Network based Vehicle Detection and Classification,” vol. 7, no. 9, pp. 17–25, 2016.

[5] J. Guerrero-Ibáñez, S. Zeadally, and J. Contreras-Castillo, “Sensor technologies for intelligent transportation systems,” Sensors, vol. 18, no. 4, p. 1212, 2018.

[6] P. Premaratne, I. J. Kadhim, R. Blacklidge, and M. Lee, “Comprehensive review on vehicle detection, classification and counting on highways,” Neurocomputing, p. 126627, 2023.

[7] B. Neupane, T. Horanont, and J. Aryal, “Real-time vehicle classification and tracking using a transfer learning-improved deep learning network,” Sensors, vol. 22, no. 10, p. 3813, 2022.

[8] A. B. Ahmad and T. Tsuji, “Traffic monitoring system based on deep learning and seismometer data,” Appl. Sci., vol. 11, no. 10, p. 4590, 2021.

[9] H. Zhang, M. Liptrott, N. Bessis, and J. Cheng, “Real-Time Traffic Analysis using Deep Learning Techniques and UAV based Video,” 2019.

[10] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, “Deep learning-based vehicle behavior prediction for autonomous driving applications: A review,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 1, pp. 33–47, 2020.

[11] N. O’Mahony et al., “Deep learning vs. traditional computer vision,” in Advances in computer vision: proceedings of the 2019 computer vision conference (CVC), volume 1 1, Springer, 2020, pp. 128–144.

[12] R. Ayachi, Y. Said, M. Afif, A. Alshammari, M. Hleili, and A. Ben Abdelali, “Assessing YOLO models for real-time object detection in urban environments for advanced driver-assistance systems (ADAS),” Alexandria Eng. J., vol. 123, pp. 530–549, 2025.

[13] A. Shabbir, A. N. Cheema, I. Ullah, I. M. Almanjahie, and F. Alshahrani, “Smart city traffic management: Acoustic-based vehicle detection using stacking-based ensemble deep learning approach,” IEEE access, vol. 12, pp. 35947–35956, 2024.

[14] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149), IEEE, 1999, pp. 246–252.

[15] P. KaewTraKulPong and R. Bowden, “An improved adaptive background mixture model for real-time tracking with shadow detection,” Video-based Surveill. Syst. Comput. Vis. Distrib. Process., pp. 135–144, 2002.

[16] A. Gholamhosseinian and J. Seitz, “Vehicle Classification in Intelligent Transport Systems: An Overview, Methods and Software Perspective,” IEEE Open J. Intell. Transp. Syst., vol. 2, pp. 173–194, Jan. 2021, doi: 10.1109/OJITS.2021.3096756.

[17] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2016.

[18] N. Wu, “Application and evaluation of deep learning based image recognition techniques in agriculture,” Appl. Comput. Eng., vol. 48, no. 1, pp. 141–147, 2024.

[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” arXiv Prepr. arXiv2004.10934, 2020.

[20] A. Wang et al., “YOLOv10: Real-Time End-to-End Object Detection,” May 2024, [Online]. Available: http://arxiv.org/abs/2405.14458

[21] R. Sapkota et al., “YOLOv12 to Its Genesis: A Decadal and Comprehensive Review of The You Only Look Once (YOLO) Series,” 2025. [Online]. Available: https://www.researchgate.net/publication/381650388

[22] Y. Zhang, B. Hu, X. Yuan, and Y. Li, “Road object recognition method based on improved YOLOv3,” Acad. J. Comput. Inf. Sci, vol. 5, pp. 1–9, 2022.

[23] A. Mechanism, “An Improved YOLOv5s Algorithm for Object Detection with an Attention Mechanism,” 2022.

[24] F. Tang and Q. Zhang, “Real-Time Detection of Drill Pipe Joints Using Improved YOLOv5x Model Applied to Drilling Operation Images,” 2024.

[25] P. Yan, H. Sun, Y. Zhao, P. Wang, and Z. Wu, “Multispectral imaging and enhanced YOLOv10n for efficient coal gangue detection in complex mining environments,” J. Real-Time Image Process., vol. 22, no. 3, pp. 1–14, 2025.

[26] R. Sapkota and M. Flores-calero, “YOLOv12 to Its Genesis : A Decadal and Comprehensive Review of The You Only Look Once (YOLO) Series,” no. February 2025, 2024.

[27] Y. Zhang, Z. Guo, J. Wu, Y. Tian, H. Tang, and X. Guo, "Real-Time Vehicle Detection Based on Improved YOLO v5," Sustain., vol. 14, no. 19, Oct. 2022.

[28] D. Li, E. Wang, Z. Li, Y. Yin, L. Zhang, and C. Zhao, "STE-YOLO: A Surface Defect Detection Algorithm for Steel Strips," Electron., vol. 14, no. 1, pp. 1–21, 2025.

Comparative Analysis of Rank and Roulette Wheel Selection Strategies in Genetic Algorithms for Spatial Layout Optimization

Najihah Ibrahim[1]*, Fadratul Hafinaz Hassan[2]*, Sharifah Mashita Syed-Mohamad[3], Rosmayati Mohemad[4], Ahmad Shukri Mohd Noor[5]
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, Kuala Nerus, 21030, Terengganu, Malaysia[1, 3, 4, 5]
School of Computer Sciences, Universiti Sains Malaysia, Minden, 11800, Pulau Pinang, Malaysia[2]

***Abstract*—Autonomous urban planning, facility layout design, and interior design are critical and meticulous tasks that require the optimization of space arrangement. One of the main purposes of space arrangement is to achieve high space utilization with a non-complex arrangement for emergency assistance, particularly to enhance pedestrian safety in panic situations. This study explores the optimization of spatial layouts by employing Genetic Algorithms (GA) due to their robust search capabilities. However, spatial layout size limitations may affect the search capability and significantly impact space arrangement and utilization. Hence, this study presents a comparative study of two GA selection operator methods: Rank Selection (RS) and Roulette Wheel Selection (RWS) for determining the effectiveness in optimizing spatial layout arrangements and space utilization. The results demonstrated significant improvements in crowd flow management, with the RWS method showing the highest fitness value despite slower convergence compared to RS. The study highlighted the impact of different methods on the convergence of the multi-objective fitness value based on space elements such as overlapping and standard walkway distances. While both selection methods proved to be effective in optimizing space utilization, the RWS method demonstrated greater computational efficiency while still adhering to standard layout designs. This efficiency helps to ensure smoother evacuation and ease of movement during emergency situations.**

Keywords*—Genetic algorithm; optimization; spatial layout arrangement; space utilization; urban planning; facility layout design; rank selection; roulette wheel selection

I. Introduction

Pedestrian safety is a critical aspect of urban planning, especially in emergency scenarios. A surging pattern of fatalities due to entrapment incidents in recent years has led to the increased demand for safer building designs and effective emergency protocols. In response, numerous studies have been carried out to uncover the impactful spatial features that significantly influence the pedestrians' movement during the evacuation process. The previous studies have proposed an enhanced spatial layout design that aimed at reducing casualties and improving overall safety [1-6].

Spatial layout arrangement involves the organization and positioning of elements within a given space, serving as a crucial aspect of spatial layout design. This design practice is essential in the planning and construction phases of housing development, ensuring efficient use of space and optimal functionality. The design will outline the blueprint of the layout by assisting in the arrangement of the assigned elements for reaching the suitable order, sorting, grouping, alignment, function, and scale to the layout size. In later times, the spatial layout was designed by employing the manual design method and focusing on the standard design procedure and being influenced by the local demographic structure's culture and current trend [7][8]. However, with the advancement of computer intelligence systems, spatial layout design application tools have been introduced for autonomous spatial layout design [9-11].

The demand for autonomous space arrangements has surged significantly, driven by the need for multi-objective functions in design. However, due to some limitations on the complex computation of the various parameters, the autonomous spatial layout arrangement generated a non-fitness design and constructed a less scalable and less functional floor plan [12]. Traditional autonomous layout designs often fall short in managing crowd movements during panic situations. Ineffective layouts have resulted in numerous entrapment incidents during emergencies [13]. Research by [13] has highlighted the significant impact of optimizing interior resources' allocation and occupancy within the layout. Hence, arrangement and allocation optimality are the new approaches to ensure the traffic flow and the safety of the pedestrians in the layout. This approach can improve the overall quality of life and the environment rather than focusing solely on the layout structure, and functionality.

However, previous research has shown that optimizing layout utilization can lead to several issues, including overlapping objects, low space occupancy, and non-standard layout design [14-23]. Hence, it is necessary for the computer-aided layout design to exploit a suitable optimization method for constructing the high occupancy elements' layout while adapting the architecture building design policy in constructing the space arrangement. This study addresses this gap by proposing an optimized spatial layout that leverages advanced algorithms to enhance safety and efficiency.

The study begins by introducing the optimization method, Genetic Algorithm (GA) as a suitable approach for computing

multi-objective functions, and followed by a methodology section, discussing the GA framework, the selection process, and the rank selection techniques adopted in this study. Next, the results section presents the findings, and the conclusion summarizes the key outcomes and suggests directions for future research.

II. GENETIC ALGORITHM

In [24], the authors show that the optimization with multi-objective functions can be inspired by the principles and inspiration of biological evolution. Numerous studies have been proposed in layout design optimization using bio-inspired algorithms for applications such as optimal job scheduling, structural design, cost minimization, flow control, and space occupancy. For example, research by [25] explored the use of Genetic Algorithm (GA), Differential Evolution (DE), Artificial Bee Colony (ABC), Charge Search System (CSS), and Particle Swarm Optimization (PSO) algorithms to optimize robot workcell layout in manufacturing systems, focusing on optimizing layout area and robot operation time. Meanwhile, research by [26] proposed optimal design solutions for planar trusses in structural roof articulations using Elitism-Based Genetic Algorithm (EBGA), Ant Colony Optimization (ACO), Artificial Honey Bee Optimization (AHBO), and PSO, with an emphasis on minimizing truss weight. Research by [27] explored the optimization of construction layout arrangements to reduce transportation costs, comparing the performance of PSO, ABC, and Symbiotic Organisms Search (SOS) algorithms. Additionally, research by [28] applied a GA to garden landscape design, aiming to adhere to the fundamental principles of landscape and urban design while meeting people's needs, showing improvements over traditional methods. Research by [13] utilized GA for resource allocation to enhance spatial layout design, and research by [29] employed GA to optimize land use for sustainable land resource management. These studies highlight that only a small number of works focus on indoor design and spatial element arrangement, despite the growing use of bio-inspired algorithms in layout optimization. Building on this foundation, the focus of this study is to evaluate and compare the effectiveness of GA, PSO, ACO, and ABC specifically for optimizing spatial layout arrangements in the context of urban planning and design.

One of the suitable bio-inspired algorithms is Genetic Algorithm (GA). GA is the adaptation of the natural evolution process inspired by Charles Darwin's theory of genetic evolution for the survival of the fittest genes. Genetic evolution is based on the genetic structure, and the natural selection operation carried out for the genes' transformation and modification for the next generation [30] [31]. The principles of Darwinian natural selection are based on heredity, variation, and selection. In selecting a GA for optimizing spatial layout arrangements, its flexibility and adaptability make it a compelling choice. GA is particularly effective for complex problems with large solution spaces. The ability to perform global searches and avoid local optima makes GA well-suited for layout optimization, where multiple variables and constraints must be balanced. Additionally, GA can be easily modified and combined with other techniques to enhance its performance and tailor it to specific problem domains, such as spatial layout optimization. This adaptability, combined with their proven success in various domains, makes GA a suitable and powerful method for optimizing spatial layout arrangements, ensuring efficient use of space while meeting design criteria.

GA has three operators: 1) Selection, 2) Crossover, and 3) Mutation. These operators will converge the genes to find the best fitness function to generate the fittest final offspring. The GA process will begin with the initialization of the random production of N number of chromosomes, with each chromosome containing an array of gene bits. The objective function will be assigned for determining the fitness values of the chromosomes. There are many types of selection schemes that can be used for discriminating the fittest values and the lowest values of the spatial layout design solutions; 1) Rank Selection (RS), 2) Roulette Wheel Selection (RWS), 3) Tournament Selection (TS), 4) Stochastic Universal Sampling (SUS), and many more.

The RWS is the proportionate fitness selection that represents the circular wheel that is divided based on the probability of the fitness value from the whole values and represented in the circle's degree value (the ratio of individual fitness value and the total fitness of overall individuals in the population). The fittest individual will have a bigger degree region in the circle and have a greater chance of being selected during the spinning process. Hence, the probability of being the fittest individual is high. A fixed point will be generated randomly to represent the real roulette wheel spinning. The fitness values selected by the fixed point will be selected as the parents. The RWS is also implemented in SUS. However, in SUS selection, there are multiple fixed points marked as the random stochastic selection, and the parents can be obtained in a single spin. This setup will be able to encourage the highly fit parents to be selected at once.

RS ranks individuals based on fitness values and applies a roulette-wheel-like method for parent selection, where each individual has an equal share (same probability) of being selected as the parents. The selection sometimes will make poor selections of parents who have a possibility of selecting the least fit solutions for reconstructing the fitter individuals. TS is the selection strategy that selects the k-number of solutions from the whole available solutions and comparing the fitness values among them. The fittest candidates among the k-individuals will be passed on to the next-gen. The probability of the selection is based on the candidate's likelihood in the tournament group. The tournament size will be able to affect the selection process as the less fit solution will have a low possibility of being selected in the large tournament group as it must compete with the stronger candidate with a high fitness solution. Based on the type of selection schemes, RS and RWS have been selected to be compared as one of the selected selection methods in comparing the spatial layout designs to find the higher fitness parents for the recombination and diversification of the offspring. This selection scheme has been selected due to its ability to contribute to the high convergence rate, as the fittest parents are able to construct better offspring with fitter fitness values.

TS and SUS selection methods are also able to give a high possibility of convergence rate. However, the random selection for the first step of the methods will contribute towards the divergence of the parents and will have a high possibility of not being able to construct the offspring that have the best inheritance from the parents. Whereas both RS and RWS are able to create a great balance between convergency and divergence of the offspring construction as there will be a high possibility of the selection approach to select the fittest parents, and there are also chances of selecting less fit parents but with low possibility. Therefore, this research focuses on utilizing RS and RWS for the parent selection phase in GA.

In light of the research problem, the goal is to transform spatial layouts into safer, more navigable spaces during emergencies. In contrast to existing studies that focus solely on layout efficiency or cost-based metrics, this study offers a novel comparative analysis of GA selection strategies using RS and RWS for managing the integration of multi-objective functions that include crowd flow safety and layout efficiency. The aim is to achieve autonomous high occupancy space arrangements that comply with spatial design standards and improve pedestrian movement flow, particularly during evacuation processes.

III. METHODOLOGY

Genetic Algorithms (GA) are powerful optimization tools inspired by the process of natural selection. In the context of constructing a spatial layout, GA facilitates the generation of optimized solutions through iterative processes. This research study focuses on comparing two selection methods; Roulette Wheel Selection (RWS) and Rank Selection (RS). Both methods are evaluated using a consistent approach involving uniform crossover and bit flip mutation to ensure a fair comparison.

A. Genetic Algorithm (GA) Framework

The three fundamental phases need to be highlighted for constructing a spatial layout based on GA's design; 1) the selection phase, 2) the crossover phase, and 3) the mutation phase. Fig. 1 shows the fundamentals of genetic evolution processing for optimizing genetic fitness.

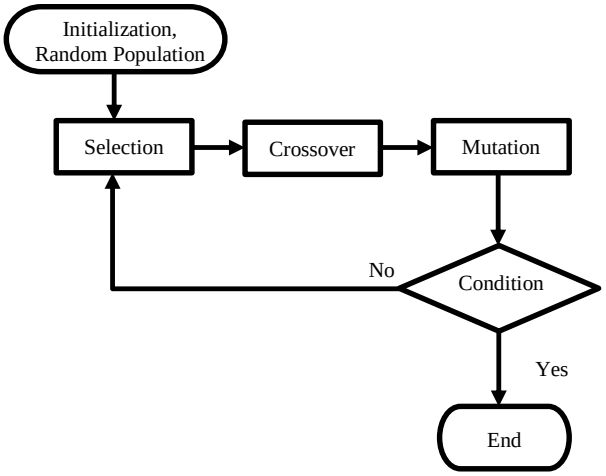


Fig. 1. Fundamental genetic evolution process

The optimized spatial layout arrangement is constructed based on the adaptation of the binary GA as shown in Fig. 2.

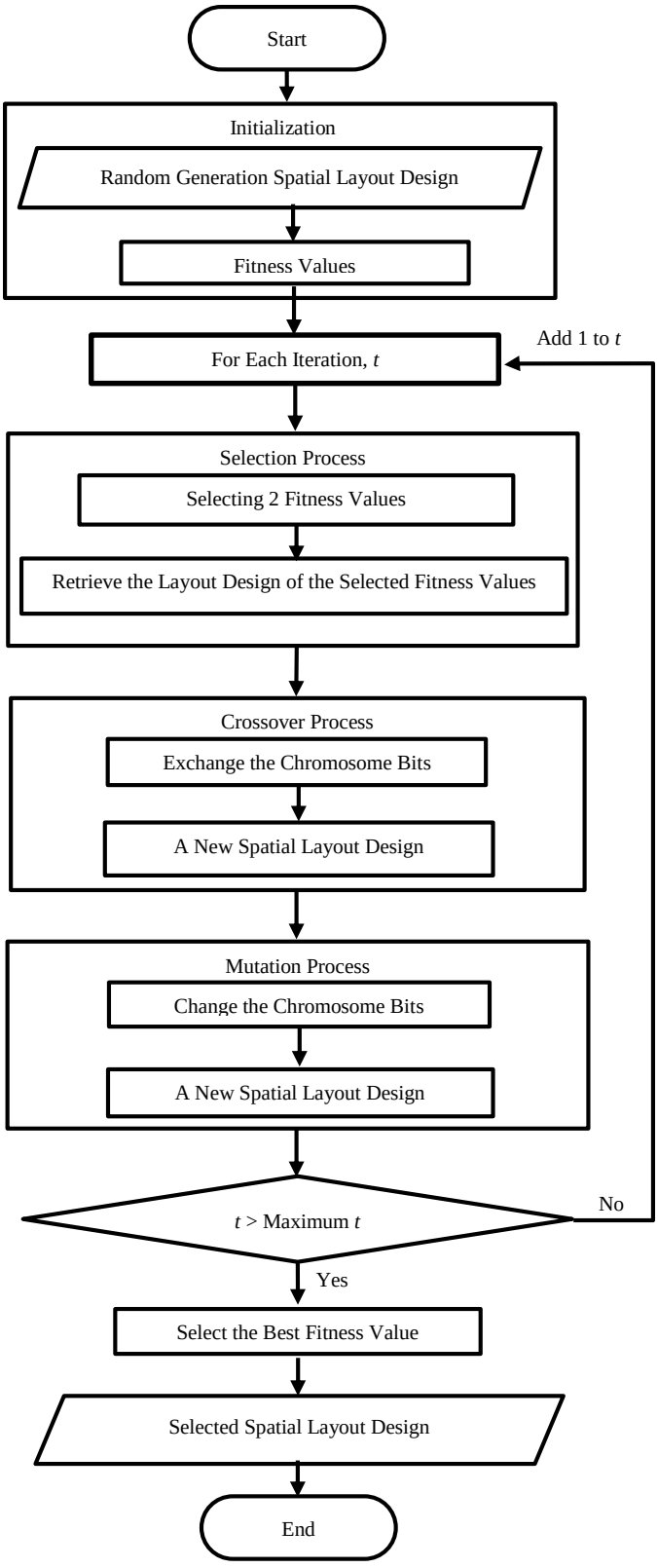


Fig. 2. Adaptation framework on GA in constructing an optimal spatial layout.

Pseudocode 1 GA Based Spatial Layout Construction Optimization

```
Load training samples
begin
Step 1:
    Generate the initial layout population z_i, i = 1 ... SN
    Evaluate the fitness (f_i) of the population
    Initialize iteration cycle = 1 ... SN
Step 2:
    set cycle to 1
    repeat
        //Selection Phase
        Apply greedy selection for f_i of i, i(h1) and i(h2)
        //Crossover Phase
        Exchange the chromosome genes of i(h1) and i(h2)
        //Mutation Phase
        Change the chromosome genes of i(h1) and i(h2)
        Calculate the offspring's fitness f_i, h1(f_i) and h2(f_i)
        Search for least fit individuals, i(g1) and i(g2)
        //New generations
        Replace spatial layout g1 and g2 with h1 and h2
        cycle = cycle + 1
    until cycle = SN
Step 3:
    Memorize the best solutions z_i
    Output the best spatial layout found
end
```

Based on the framework in Fig. 2, Pseudocode 1 is developed to show the fundamentals for spatial layout design optimization based on GA, adapted from the algorithms overview by McCall [32]. Based on Pseudocode 1, the GA parameters must be initialized by setting the population size, *SN* to 50. Each individual in the population, z_i is generated with a unique spatial layout design, which includes 300 obstacles. The six obstacles are combined into one large obstacle representing the furniture. Each of the population fitness values f_i will be calculated based on the objective function: 1) overlapping of the elements generated as static obstacles, and 2) distance 1.2 m from doors and walls to meet the standard of interior design.

The number of furniture elements left in the layout after the overlapping process and exist within the layout that has a 1.2 m distance range from doors and walls has been calculated as the total fitness value of the layout. The iteration cycle has been initialized, and in this research, the iteration is set to 100. When entering the iteration cycle, the process will be entering into three phases: 1) selection phase, 2) crossover phase, and 3) mutation phase. A solution acceptance rule is implemented to ensure that the elements do not exceed the allocated space within the layout. Based on this rule, the number of furniture elements must not exceed the initial 300 loaded obstacles, even after 100 iteration cycles. This constraint ensures that the layout remains within the grid limits and complies with interior design standards, particularly by preserving sufficient walking space for pedestrians to access exit points.

B. Selection Process

The selection process in GA determines which individuals from the current population will be the parents to the next generation. This process mimics natural selection in choosing the fittest individuals based on performance or fitness scores. The primary goal is to ensure that high-quality traits are preserved and propagated, thereby improving the overall solution quality over successive generations. Every method available for selection operation has a unique approach to balancing exploration and exploitation within the search space.

1) Roulette Wheel Selection (RWS). The RS method is implemented to find the parents for the further recombination and restoration process of the offspring. This discrimination phase offers selective pressure that applies greedy selection based on the fitness proportionate selection approach as a method in guiding the evolution of spatial layout design. The fitness value fi has been determined from each of the solutions z_i of population i. The roulette slot size probability will be computed based on Eq. (1):

$$p_i = \frac{f_i}{\sum_{i=1}^{SN} f_i} \quad (1)$$

where, the p_i is the probability of each of the fitness values fi from the whole fitness values in the population *i* and *SN* is the maximum number of populations, *i*. The cumulative probability, q_i, for each chromosome is calculated based on Eq. (2):

$$q_i = \sum_{i=1}^{SN} p_i \quad (2)$$

here, p_i is the probability of each of the fitness values, and the cumulative probability has been calculated for each of the population *i* until reaching the maximum number of the population, *SN*. The fixed point of the roulette wheel can be constructed based on the random number generation *r*, where $r \in (0, 1)$. The parents are selected based on the condition; if $r < q_1$, then the first solution, z_1 is chosen as the parent. Otherwise, if $r > q_1$, the algorithm searches for another solution z_i such that $q_{i-1} < r \leq q_i$ and selects it as the parent. The steps have been repeated for two times for each iteration cycle to find the two parents' spatial layout to be recombined and explored for the fittest offspring.

2) Rank Selection (RS). RS method is implemented to identify parents for further recombination and restoration processes in generating offspring. Unlike the RWS method, which relies on fitness proportionate selection, RS imposes selective pressure by ranking individuals based on their fitness and then selecting parents according to their rank, rather than their absolute fitness values. This approach ensures that even individuals with lower fitness have a chance of being selected, thus maintaining diversity within the population.

In RS, each individual solution z_i in the population is assigned a rank r_i, with the fittest individual receiving the highest rank. The selection probability for each individual is then determined based on its rank, not its fitness value. The probability of selecting an individual is calculated using a rank-based probability distribution, where higher-ranked individuals have a greater chance of being selected as parents. This can be represented by Eq. (3):

$$p_i = \frac{2x(SN - r_i + 1)}{SN \times (SN+1)} \quad (3)$$

where, p_i is the probability of selecting the individual with rank r_i, and SN is the maximum number of individuals in the population. This rank-based probability distribution ensures a more uniform selection pressure compared to fitness-proportionate methods to avoid cases where a single individual with a much higher fitness dominates the selection process. The cumulative probability q_i for each individual is calculated based on Eq. (2), as the formula is similar. However, compared to RWS, the p_i is the probability of each of the ranks, and the cumulative probability has been calculated for each of the populations i until reaching the maximum number of the population, SN.

Similar to RWS, the fixed point of the roulette wheel can be constructed based on the random number generation r, where $r \in (0, 1)$. The parents can be selected by the condition; if $r < q_1$, then the first solution, z_1 has been selected as the parent. However, if $r > q_1$, then the other search has been made on the other individual solution z_1 such that $q_{i-1} < r \leq q_i$. The steps have been repeated for two times for each iteration cycle to find the two parents' spatial layout to be recombined and explored for the fittest offspring.

RS offers a more stable selection process by mitigating the effects of disproportionately high fitness values that can dominate the selection in methods like RWS. This method ensures a balance between exploration and exploitation by allowing less fit individuals a chance to contribute to the next generation, thus enhancing the evolutionary search for optimal spatial layout arrangements.

C. Uniform Crossover

In this research, the uniform crossover has been selected, and the selected parents have been recombined with the crossover phase. Both parents' allele genes have been recombined to construct better spatial layout offspring that are able to produce fit fitness values. The uniform recombination process has examined the genes in the parents separately and recombines each of the genes based on the coin flip method. The flip coin method has randomly made the decision based on 50-50 probabilities [0,1]. If the toss is "0", the gene for both parents has been maintained, whereas if the toss resulted in "1", the gene has been exchanged between the parents. The offspring constructed from the recombination method has been explored to prevent premature convergence to the local optimal solution and diversify the genetic population via the mutation phase approach.

D. Bit Flip Mutation

The mutation phase has amended the offspring solutions to construct new solutions. In this research, the random resetting mutation has been used by selecting each of the obstacles based on the bit flip mutation function, in which the probability of the obstacles' gene selection has been set to a 0.01 mutation rate. Each of the genes of the offspring has been examined, and a random number has been generated to check the mutation rate condition. Based on the rules set, when the random number < 0.01, the mutation has occurred, and the gene bit has been flipped, whereas when the random number >= 0.01, the bit of the offspring has remained the same. The obstacles that are assigned with the random number < 0.01 have been assigned to a randomly chosen gene in the spatial layout grid.

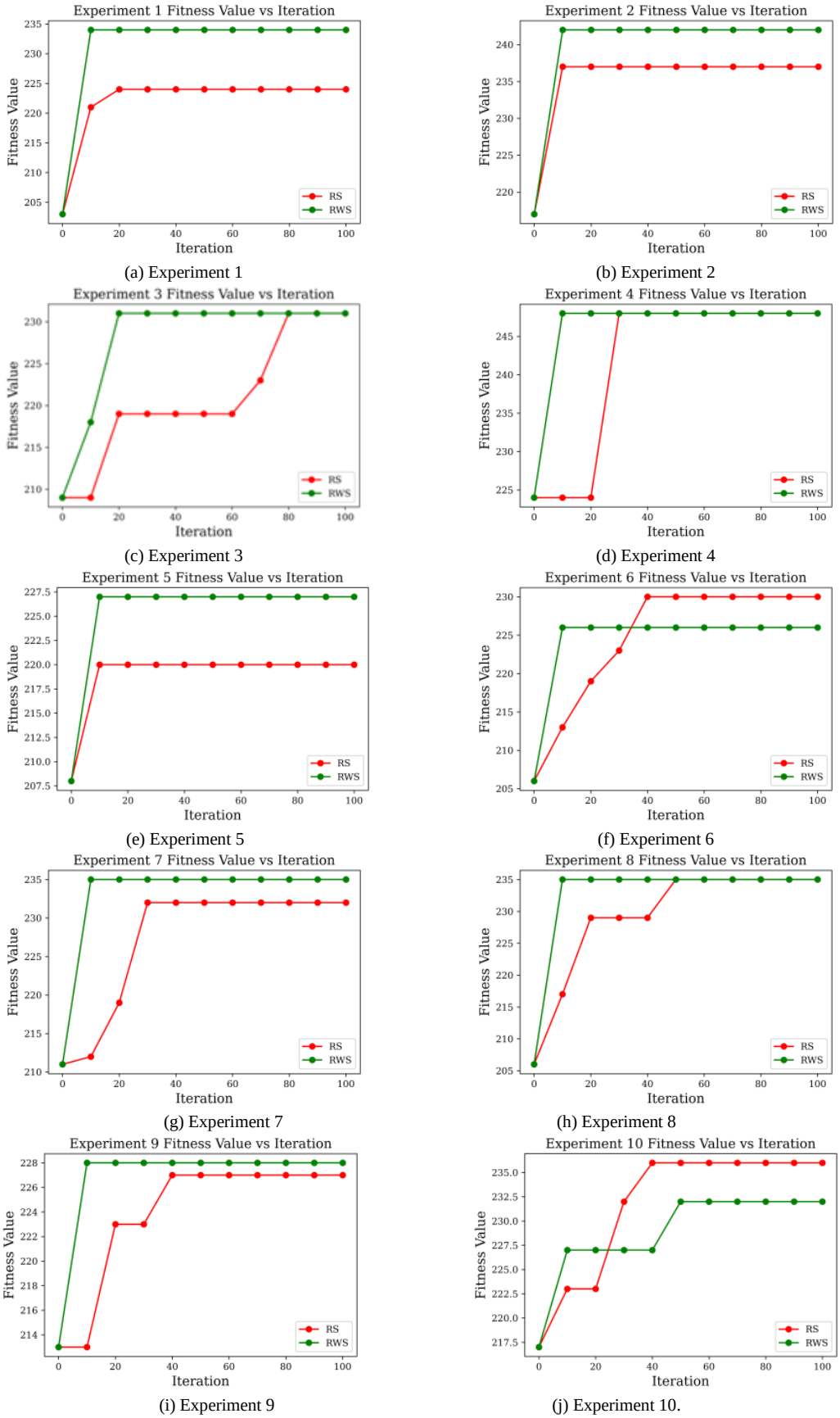
This method is suitable for spatial layout arrangement as the solutions will still provide the 50 large orders of obstacles. The fitness value of the offspring has been determined, and the value has been compared with the fitness values of the current population, f_i. The offspring constructed has been passed to the next iteration cycle as a member of the population and replaces the least fit solutions. The final population i with spatial layout solution z_i after 100 iterations have been compared, and the fittest solution will be selected to represent the selected result for the GA approach in designing an optimal spatial layout.

IV. Results

This research compared the Rank Selection (RS) and Roulette Wheel Selection (RWS) methods for constructing GA-based spatial layout arrangements. To ensure unbiased results, ten experiments were conducted, each with 100 iterations. Fig. 3 presents the graphical results of these experiments, illustrating the fitness value for every iteration of each selection method.

Based on the overall result in Fig. 3, both RS and RWS show the characteristic of GA algorithm results with premature convergence. These results align with the study's focus on optimizing spatial layout arrangements and space utilization, which are critical for applications like autonomous urban planning, facility layout design, and interior design. Given the emphasis on high space utilization and non-complex arrangements for emergency assistance, the selection of GA operators is crucial, especially in scenarios with limited spatial layout size.

Based on the observation, the highest fitness value for both selection methods shows different values throughout the ten experiments. Based on the graphs in Fig. 3, RS outperformed RWS in 20% of the experiments by optimizing the population fitness value, especially in Experiment 6 and 10. In Experiment 6, RS achieved a fitness value of 230.0 by iteration 40, whereas RWS plateaued at a maximum fitness value of 226.0, beginning as early as iteration 10. Similarly, in Experiment 10, RS reached a fitness value of 236.0 by iteration 40, while RWS plateaued at 232.0 from iteration 50 onwards. These results highlight the potential of RS for achieving faster convergence in certain scenarios.



(a) Experiment 1 (b) Experiment 2

(c) Experiment 3 (d) Experiment 4

(e) Experiment 5 (f) Experiment 6

(g) Experiment 7 (h) Experiment 8

(i) Experiment 9 (j) Experiment 10.

Fig. 3. Graph comparison of Rank Selection (RS) and Roulette Wheel Selection (RWS) fitness value of over 100 iterations.

In 30% of the experiments, both RS and RWS exhibited comparable performance, achieving similar fitness values at different iterations. Based on the observations, the maximum fitness value for Experiment 3 reached 231.0 at iteration 80 using RS but was achieved earlier by RWS at iteration 20. The maximum fitness for Experiment 4 reached 249.0 at iteration 30 by RS but achieved earlier by RWS at iteration 10. In addition, the maximum fitness value for Experiment 8 reached 235.0 at iteration 50 by RS while RWS reached it earlier at iteration 10. These results indicate an overlapping convergence behavior, likely due to the limitation of layout size in accommodating all layout elements.

Furthermore, in 50% of the experiments, the RWS selection method outperformed the RS in terms of final fitness value. RWS shows better optimization compared to RS due to the offspring selection strategy that is based on the relative fitness of individuals. This proportional selection approach groups individuals by fitness level, increasing the likelihood that higher-fitness individuals are selected as parents, thereby generating stronger offspring. Compared to RWS, the selection in RS is also able to generate higher fitness values. However, due to the selection of the top 2 highest-fitness parents in RS caused the offspring generated to be too fit and unable to replace the lowest value of population members. Hence, resulting in the slow convergence speed compared to RWS. Although RWS is generally associated with slower convergence, its use of a solution acceptance rule (i.e., constraints related to layout capacity) allows it to increase fitness values earlier in the experiment, often resulting in the highest fitness value by iteration 100. Additionally, the RS selection method is computationally expensive due to the population sorting based on the fitness value compared to the RWS selection method. Fig. 4 shows the graph comparison between the processing time (milliseconds) in every experiment for the RS and RWS-based GA approach spatial layout design arrangement.

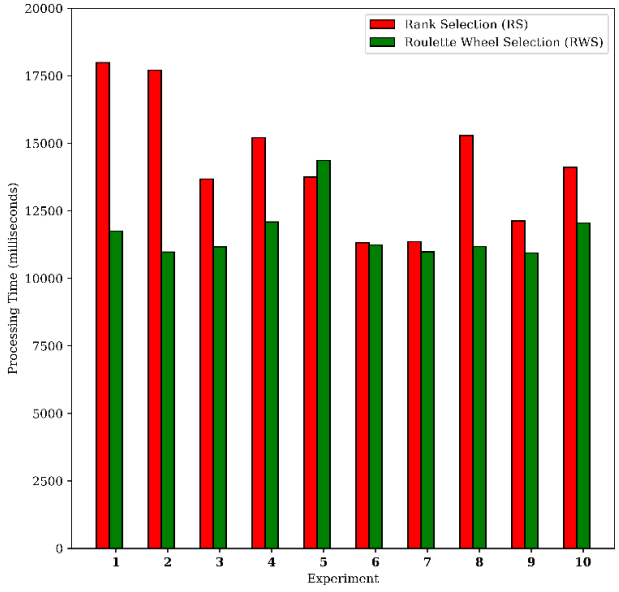


Fig. 4. Graph comparison between the processing time of Rank Selection (RS) and Roulette Wheel Selection (RWS) for 10 experiments.

Based on Fig. 4, the graph of processing time shows that RS consistently takes extra processing time compared to RWS in most experiments, with the exception of Experiment 5 as the 10% outlier, where RS recorded 13752 ms compared to 14369 ms for RWS. This result indicates that RS is generally slower in this context. In Experiments 1, 2, 3, 4, 6, 7, 8, 9, and 10, RWS shows significantly faster processing time, with 90% of the experiments showing the improvement of processing time percentage. Table I summarizes the percentage improvement of RWS over RS.

TABLE I. PROCESSING TIME COMPARISON WITH PERCENTAGE OF RWS IMPROVEMENT COMPARED TO RS

Experiment	RS Time (ms)	RWS Time (ms)	Faster Method	RWS over RS (%)
1	17991	11747	RWS	34.69
2	17708	10974	RWS	38.02
3	13681	11168	RWS	18.37
4	15209	12088	RWS	20.54
5	13752	14369	RS	-4.49
6	11320	11236	RWS	0.74
7	11359	10986	RWS	3.28
8	15292	11178	RWS	26.89
9	12130	10941	RWS	9.80
10	14110	12047	RWS	14.62

Based on Table I, among the 90% of experiments where RWS demonstrated superior performance, the results show the improvement ranging from modest gains, such as 0.74% in Experiment 6, to substantial differences exceeding 38% in Experiment 2. The most significant time reductions were observed in Experiments 1 and 2, where RWS reduced computation time by 34.69% and 38.02%, respectively, compared to RS. This substantial difference in the efficiency of these two methods is potentially due to the computation cost in population sorting based on fitness value that is required in the RS algorithm, which increases the computation complexity. RWS, by contrast, is more computationally efficient and enables faster spatial layout optimization.

Compared to the other experiments, in Experiment 5, RS was slightly faster by 4.49%, which is an exception to the trend observed from the whole experiment. This anomaly suggests that under certain conditions, particularly where early convergence occurs or solution acceptance thresholds align more favorably, RS may offer computational advantages. However, this case appears to be an outlier rather than a consistent pattern. The outlier also reflects the unbiased nature of the experimental setup, which was intentionally designed to allow either method to succeed under appropriate conditions.

Based on Fig. 4 the solution acceptance rule acts as a limiting factor for space utilization optimization, and RWS demonstrates a better balance between the fitness value optimization and processing time that making it a more practical choice for the selection process in the future GA approach for optimizing the autonomous spatial layout arrangement design and space utilization in autonomous urban planning.

V. CONCLUSION

This study aimed to evaluate the effectiveness of two Genetic Algorithm (GA) selection methods: Rank Selection (RS) and Roulette Wheel Selection (RWS) in optimizing spatial layout arrangements to improve space utilization and emergency responsiveness. Through this research, it was shown that RWS was able to outperform RS by consistently achieving higher fitness values and demonstrating greater computational efficiency. This was due to its proportional selection mechanism, which enhances the likelihood of selecting fitter individuals and accelerates convergence while maintaining layout standards for emergency situations. While RS occasionally showed better performance in specific instances, its approach often led to slower convergence and increased computational demands. The requirement for population sorting based on fitness in RS contributed to its higher computational cost. Overall, RWS proved to be the more effective method for optimizing spatial designs, meeting the study's objective of identifying the most efficient GA operator for enhancing spatial layout arrangements and space utilization in autonomous urban planning. The future research direction of this study is to explore hybrid selection methods in GA that combine the selection mechanisms and features of RS and RWS to further enhance the optimization of spatial layout arrangements in urban planning. Additional studies can be carried out to investigate the application of RWS in different urban planning scenarios to validate its effectiveness in improving space utilization and emergency responsiveness across diverse environments.

ACKNOWLEDGMENT

This research was supported by Universiti Malaysia Terengganu (UMT) through the Talent and Publications Enhancement Research Grant (TAPE-RG) [UMT/TAPE-RG/2024/55535] under the project titled "Multi-Objective Optimization Algorithm for Autonomous Spatial Layout Design".

REFERENCES

[1] J. Huixian and Z. Shaoping, "Navigation system design of fire disaster evacuation path in buildings based on mobile terminals," in 2016 11th International Conference on Computer Science & Education (ICCSE), 2016, pp. 327–331. doi: 10.1109/ICCSE.2016.7581602.

[2] X. Lu, P. B. Luh, A. Tucker, T. Gifford, R. S. Astur, and N. Olderman, "Impacts of Anxiety in Building Fire and Smoke Evacuation: Modeling and Validation," IEEE Robot Autom Lett, vol. 2, no. 1, pp. 255–260, 2017, doi: 10.1109/LRA.2016.2579744.

[3] K. Xie, Y. Song, J. Liu, B. Liang, and X. Liu, "Stampede prevention design of primary school buildings in china: a sustainable built environment perspective," Int J Environ Res Public Health, vol. 15, no. 7, p. 1517, 2018.

[4] M. Hayat, S. H. Khan, M. Bennamoun, and S. An, "A Spatial Layout and Scale Invariant Feature Representation for Indoor Scene Classification," IEEE Transactions on Image Processing, vol. 25, no. 10, pp. 4829–4841, 2016, doi: 10.1109/TIP.2016.2599292.

[5] F. H. Hassan and A. Tucker, "Using Uniform Crossover to Refine Simulated Annealing Solutions for Automatic Design of Spatial Layouts," in IJCCI (ICEC), 2010, pp. 373–379.

[6] P. K. Shukla, K. Korb, M. Randall, and T. Hendtlass, "Genetically Optimized Architectural Designs for Control of Pedestrian Crowds," Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 22–31.

[7] F. A. Mustafa and A. S. Hassan, "Mosque Layout Design: An Analytical Study of Mosque Layouts in the Early Ottoman Period," Frontiers of Architectural Research, vol. 2, no. 4, pp. 445–456, 2013, doi: http://dx.doi.org/10.1016/j.foar.2013.08.005.

[8] I. V. Arnolds and S. Nickel, "Multi-Period Layout Planning for Hospital Wards," Socioecon Plann Sci, vol. 47, no. 3, pp. 220–237, 2013, doi: http://dx.doi.org/10.1016/j.seps.2013.02.001.

[9] M. Bhatt, J. Suchan, C. Schultz, V. Kondyli, and S. Goyal, "Artificial intelligence for predictive and evidence based architecture design," in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, in AAAI'16. AAAI Press, 2016, pp. 4349–4350.

[10] A. Indraprastha and M. Shinozaki, "Computational Models for Measuring Spatial Quality of Interior Design in Virtual Environment," Build Environ, vol. 49, pp. 67–85, 2012, doi: http://dx.doi.org/10.1016/j.buildenv.2011.09.017.

[11] F. Regateiro, J. Bento, and J. Dias, "Floor Plan Design using Block Algebra and Constraint Satisfaction," Advanced Engineering Informatics, vol. 26, no. 2, pp. 361–382, 2012, doi: http://dx.doi.org/10.1016/j.aei.2012.01.002.

[12] J. Michalek, R. Choudhary, and P. Papalambros, "Architectural layout design optimization," Engineering Optimization, vol. 34, no. 5, pp. 461–484, 2002, doi: 10.1080/03052150214016.

[13] W. Yixuan, "Indoor Optimal Design of Building Renovation Environment Space Layout Based on Genetic Algorithm," Procedia Comput Sci, vol. 208, pp. 539–545, 2022, doi: https://doi.org/10.1016/j.procs.2022.10.074.

[14] M. F. Anjos and M. V. C. Vieira, "Mathematical optimization approaches for facility layout problems: The state-of-the-art and future research directions," Eur J Oper Res, vol. 261, no. 1, pp. 1–16, 2017, doi: https://doi.org/10.1016/j.ejor.2017.01.049.

[15] J. Bénabès, F. Bennis, E. Poirson, and Y. Ravaut, "Interactive optimization strategies for layout problems," International Journal on Interactive Design and Manufacturing (IJIDeM), vol. 4, no. 3, pp. 181–190, 2010.

[16] J. Kubalík, J. Lažanský, and P. Zikl, "Layout Problem Optimization Using Genetic Algorithms," in Knowledge and Technology Integration in Production and Services: Balancing Knowledge and Technology in Product and Service Life Cycle, V. Mařík, L. M. Camarinha-Matos, and H. Afsarmanesh, Eds., Boston, MA: Springer US, 2002, pp. 493–500. doi: 10.1007/978-0-387-35613-6_56.

[17] J. Liu, H. Zhang, K. He, and S. Jiang, "Multi-objective particle swarm optimization algorithm based on objective space division for the unequal-area facility layout problem," Expert Syst Appl, vol. 102, pp. 179–192, 2018, doi: https://doi.org/10.1016/j.eswa.2018.02.035.

[18] A. McKendall and A. Hakobyan, "An Application of an Unequal-Area Facilities Layout Problem with Fixed-Shape Facilities," Algorithms, vol. 14, no. 11, p. 306, 2021, [Online]. Available: https://www.mdpi.com/1999-4893/14/11/306

[19] L. S. Chemim, C. S. Nicolle, and M. Kleina, "Layout optimization methods and tools: A systematic literature review," Gepros: Gestão da Produção, Operações e Sistemas, vol. 16, no. 4, p. 59, 2021.

[20] K. Rasheed, H. Hirsh, and A. Gelsey, "A genetic algorithm for continuous design space search," Artificial Intelligence in Engineering, vol. 11, no. 3, pp. 295–305, 1997, doi: https://doi.org/10.1016/S0954-1810(96)00050-7.

[21] Z. Guo and B. Li, "Evolutionary approach for spatial architecture layout design enhanced by an agent-based topology finding system," Frontiers of Architectural Research, vol. 6, no. 1, pp. 53–62, 2017, doi: http://dx.doi.org/10.1016/j.foar.2016.11.003.

[22] M. Yahya and M. P. Saka, "Construction site layout planning using multi-objective artificial bee colony algorithm with Levy flights," Autom Constr, vol. 38, pp. 14–29, 2014.

[23] J. Liu and J. Liu, "Applying multi-objective ant colony optimization algorithm for solving the unequal area facility layout problems," Appl Soft Comput, vol. 74, pp. 167–189, 2019, doi: 10.1016/j.asoc.2018.10.012.

[24] A. Darwish, "Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications," Future Computing and Informatics Journal, vol. 3, no. 2, pp. 231–246, 2018, doi: https://doi.org/10.1016/j.fcij.2018.06.001.

[25] Z. Y. Lim, P. S.G, and K. Izui, "Nature inspired algorithms to optimize robot workcell layouts," Appl Soft Comput, vol. 49, pp. 570–589, 2016, doi: https://doi.org/10.1016/j.asoc.2016.08.048.

[26] M. A. Jayaram, "Bio-Inspired Algorithms for Optimal Design of Trusses," IOP Conf Ser Earth Environ Sci, vol. 982, no. 1, p. 12073, 2022, doi: 10.1088/1755-1315/982/1/012073.

[27] D. Prayogo, J. C. Sutanto, H. E. Suryo, and S. Eric, "A Comparative Study on Bio-Inspired Algorithms in Layout Optimization of Construction Site Facilities," Civil Engineering Dimension, vol. 20, no. 2, pp. 102–110, 2018.

[28] Z. Liu, "The Application of Genetic Algorithm in the Optimal Design of Landscape Space Environment," Math Probl Eng, vol. 2022, p. 8768974, 2022, doi: 10.1155/2022/8768974.

[29] X. Ding, M. Zheng, and X. Zheng, "The Application of Genetic Algorithm in Land Use Optimization Research: A Review," 2021. doi: 10.3390/land10050526.

[30] T. Narahara and K. Terzidis, "Multiple-constraint Genetic Algorithm in Housing Design," Synthetic Landscapes [Proceedings of the 25th Annual Conference of the Association for Computer-Aided Design in Architecture] pp. 418-425, May 2006.

[31] M. Mitchell, "Genetic algorithms: An overview," Complexity, vol. 1, no. 1, pp. 31–39, 1995, doi: https://doi.org/10.1002/cplx.6130010108.

[32] J. McCall, "Genetic algorithms for modelling and optimisation," J Comput Appl Math, vol. 184, no. 1, pp. 205–222, 2005, doi: https://doi.org/10.1016/j.cam.2004.07.034.

AutiSim: A Virtual Reality Simulation Game Based on the Autism Spectrum Disorder

Muhammad Aliff Muhd Farid Arfian[1], Ikmal Faiq Albakri Mustafa Albakri[2]*, Faaizah Shahbodin[3], Mohd Khalid Mokhtar[4], Asniyani Nur Haidar Abdullah[5], Norhaida Mohd Suaib[6], Muhammad Nur Affendy Nor'a[7], Abdul Hasib Jahidin[8]

Department of Media Interactive-Fakulti Teknologi Maklumat Dan Komunikasi, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia[1]

Pervasive Computing & Educational Technology-Department of Media Interactive-Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia[2, 3, 4, 5]

UTM Big Data Center, Universiti Teknologi Malaysia, Johor Bahru, Malaysia[6]

Creative Interactive Technology (CIT)-School of Computing, Asia Pacific University of Technology and Innovations (APU), Kuala Lumpur, Malaysia[7]

Astro Productions Sdn Bhd, Kuala Lumpur, Malaysia[8]

***Abstract*—Technologies with altering reality like virtual reality (VR) have become more relevant to the public for their capabilities in the entertainment and healthcare field, as well as affordable for everyone. However, the emphasis on mental health-related simulation is often ignored due to technical complexities and wrong representation. Therefore, this study leverages the immersive capabilities of VR to create an engaging and educational game experience that simulates the sensory and social challenges faced by individuals with autism spectrum disorder (ASD). The study involves designing and implementing a VR game that places users in various scenarios reflecting the daily experiences of autistic individuals. The VR game aims to educate players about common misconceptions, sensory sensitivities, and social difficulties associated with autism. A literature review on XR technology and ASD was conducted during the pre-production phase to explore past research on autism in video games and shape the overall game vision. The study continues with developing an immersive simulation game using VR with locomotive motion controls and an artificial intelligence non-playable character (AI NPC) with Speech-To-Text function. Finally, the testing phase used two approaches: quantitative analysis, using the System Usability Scale (SUS) to assess usability and the Simulator Sickness Questionnaire (SSQ) to identify discomfort issues such as headaches and blurriness during gameplay, and qualitative analysis, gathering expert's feedback on the VR game's content and teaching effectiveness.**

Keywords*—Virtual reality; simulation game; autism; artificial intelligence

I. Introduction

Advancements in computer science, engineering, and human-computer interaction (HCI) have driven rapid development in spatial interface technologies, including virtual reality (VR), augmented reality (AR), and mixed reality (MR) [1]. These technologies, collectively referred to as extended reality (XR), have gained widespread attention in both academia and industry due to significant improvements in hardware that make devices lighter, more affordable, and increasingly powerful [2]. Among these, VR stands out for its unique ability to replicate real-life situations in a highly immersive manner, making it a valuable tool in healthcare, therapy, and education by providing realistic training and fostering empathy through experiential learning [3].

While VR's applications in physical health and technical training have been widely explored, its potential for simulating mental health conditions such as Autism Spectrum Disorder (ASD) remains underutilized. Media portrayals of autism frequently rely on stereotypes or fail to accurately reflect the lived experiences of individuals on the spectrum, resulting in widespread misconceptions and reinforcing social stigma [4][5]. Poorly designed simulations that do not align with authentic experiences or educational goals can further mislead users rather than help them understand the challenges faced by the autism community.

As a response to these challenges, the present study introduces an alternative approach with the development of an interactive VR simulation game that emphasizes exploration and interactivity. Designed to immerse players in the daily sensory and social experiences of individuals with ASD, the simulation seeks to cultivate empathy and enhance public awareness through experiential engagement. Central to this initiative is a threefold objective. First, to study and compare how ASD has been depicted in existing VR games aimed at public education. Second, to design and develop an immersive representative simulation of the lived experiences of autistic individuals and third, to evaluate the simulation's usability and its effectiveness in fostering empathy and greater awareness among users.

Guided by these objectives, this study is driven by the overarching research question. How effectively can an immersive VR simulation game be designed around exploration and interactivity that accurately represents the experiences of individuals with ASD compared to existing VR representations? In order to address this question, this study investigates current approaches to ASD representation in VR, explores methods for creating a more authentic visual and spatial audio experience, accurate user-centered simulation, and assesses the impact of the resulting experience in terms of usability testing, simulation sickness (SUS) and expert's feedback.

The remainder of this study is structured as follows: A review of related literature on VR technologies and autism representation in Section II; a detailed explanation of game design and development methodology in Section III; presentation of results from a conducted evaluation in Section IV; and a concluding discussion on the study's contributions, limitations, and potential directions for future research in Section V.

II. Literature Review

The term virtual reality (VR) describes an interactive, three-dimensional computer-generated environment that users can interact with. Usually, VR is accessed through a computer that can project 3D data onto a display, such as a head-mounted display (HMD) or isolated screens, along with user identification sensors [6]. Virtual reality can be broadly classified into two types, which are immersive and non-immersive [7]. Non-immersive virtual reality presents information virtually by utilizing a collection of screens surrounding the user [8]. Driving or flying simulators are a common example of this, where the user sits in a chair with many screens surrounding them, creating the illusion of being in the driver's seat or cockpit without being there. Immersive VR is the process of presenting VR content to users based on their position and movement using a wearable display, such as an HMD [9], so they may experience the entire 360 degrees of the virtual world. One of the most marketable features of VR technology, this immersive experience is what most people associate with VR.

Inside-out tracking technology is revolutionizing virtual reality, as demonstrated by devices like Oculus Quest 2 [10], as this method depends on the complex interaction of built-in cameras and sensors in the VR headgear. This depends on the intricate interaction between the VR headset's integrated cameras and sensors [11]. Together, these elements give users a responsive and intensely immersive experience. What makes it more convenient is that this technology is used in everyday smartphones powered by Android and iOS [12].

On the other hand, inside a predetermined play area, outside-in tracking provides accurate and dependable tracking. VR headsets and controllers are tracked in terms of location and orientation by devices like the Oculus, HTC Vive and XR-3, which make use of external sensors or base stations. By replicating this tracking method, Fig. 1 illustrates the flowchart of how the AutiSim will operate using an Oculus headset and Unreal Engine. This method guarantees precision, which makes it a great option for applications that need accuracy, such as professional simulations or architectural design [13].

According to current understanding, autism is a neuro-cognitive disorder characterized by excessive masculine features, difficulty empathizing, and a lack of ability to understand the thoughts of others [14]. Autism is described as a "triad of impairments" characterized by "impaired social interaction, verbal and nonverbal communication, and repetitive and stereotyped behaviors" among affected individuals [15]. There is currently no known cause or treatment [16] for the most thorough investigation to far points to a genetic cause [17].

Hans Asperger originally identified Asperger syndrome (AS) in 1944 as the behavioral profile of people who struggle with socialization and communication [18]. Since then, AS has generated a lot of discussion and attention. While AS was added as a separate diagnostic category in the Diagnostic and Statistical Manual of Mental Disorders (DSM-4, 1994), its diagnostic label was eliminated in the Diagnostic and Statistical Manual of Mental Disorders (DSM–5), the edition that followed, nearly twenty years later. Rather, autism spectrum disorders (ASD), a more inclusive category, were added to ASD [19]. A youngster with a level 1 spectrum disorder will struggle with social communication, yet have an above-average IQ and outstanding linguistic skills [20].

Rett syndrome (RTT) is a rare, severe neurodevelopmental disorder that affects almost exclusively females. The syndrome was first described in 1966 by Andreas Rett in the German medical literature. Development appears normal during the first 6 to 18 months of life but is followed by regression of motor and language skills. The clinical phenotype of RTT is broad and can be classified into two main categories: typical (classic) RTT and atypical (variant) RTT [21].

Theodor Heller discovered a group of kids in 1908, 35 years before autism was first recognized. These kids displayed average development until the age of 3 or 4 before experiencing a dramatic and abrupt regression in their speech and cognitive abilities. This condition was known as "*dementia infantalis*", and it was frequently linked to mood dysregulation, which included anxiety and aggression [22]. Nowadays, this condition is known as Child Disintegrative Disorder (CDD).

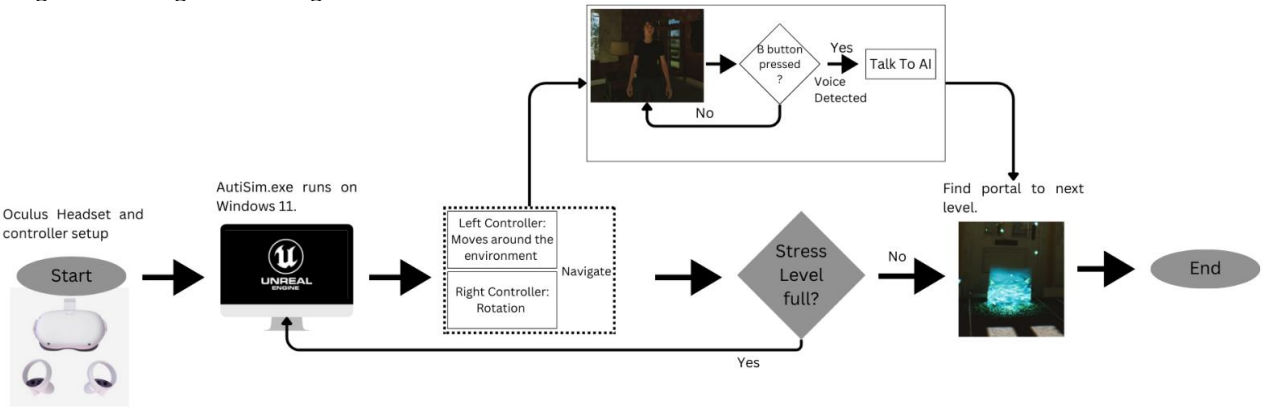


Fig. 1. Flowchart of how AutiSim will operate.

Most people think of Kanner's Syndrome in children when they think about autism. This form of autism is referred to as classic autistic disorder. Excessive sensitivity to stimuli (smell, light, noise, taste, or touch), trouble comprehending or communicating with people, and minimal eye contact are some of its symptoms. Youngsters with Kanner's syndrome frequently show little interest in their surroundings and a great desire for consistency. These children often appear to be self-centered, struggling to understand and respond to social cues that lead to difficulty in engaging with others [23].

The National Autistic Society Channel posted a video titled "Autism TMI Virtual Reality Experience" on YouTube back in June 2016. The video featured a 360-degree virtual reality experience that allowed viewers to manipulate the surroundings and observe an autistic child named "Alex" in a shopping centre. As of June 17, 2024, the film has received over 1.4 million views on YouTube and has received mixed reactions from both the autism community and people in general. With a focus on the unpleasant emotions that an autistic person experiencing a "meltdown" or "shutdown" may encounter in public, the application seeks to convey the overwhelming sensory experience that surroundings like these can cause for an autistic person [24].

Released in October 2016 for the Oculus Rift S, "The Autism Simulator" is one of the few games available on the Meta market. The premise of the simulation quoted “By observing too many details, getting drawn in by anomalies, and attempting to put the anomalies into a regular order, we provide you with a singular experience. Because of this, people with autism tend to be extremely focused, easily frightened, distracted, or violent.” [25].

In conclusion, the literature review highlights the diverse landscape of research and initiatives at the intersection of VR technology and autism awareness. Through a comprehensive examination of existing literature, the proposed research has gained valuable insights into the potential of VR to facilitate immersive and impactful experiences that foster understanding, empathy, and awareness of autism spectrum disorder.

Additionally, the analysis of the literature identifies new directions and best practices in the development, application, and assessment of VR-based awareness games. User engagement, accessibility, ethical issues, and the integration of evidence-based therapy techniques into VR experiences are among the key factors identified. The findings from various studies and projects offer valuable insights to guide the development of a VR initiative for autism awareness.

III. MATERIALS AND METHODS

This study introduces an integrated analytical methodology framework for AutiSim, as shown in Fig. 2, by utilizing the Game Design Life Cycle (GDLC) model that has been widely used and applied innovatively within VR environments [27]. The methodology framework for this study consists of the following phases: Pre-Production, Production and Evaluation.

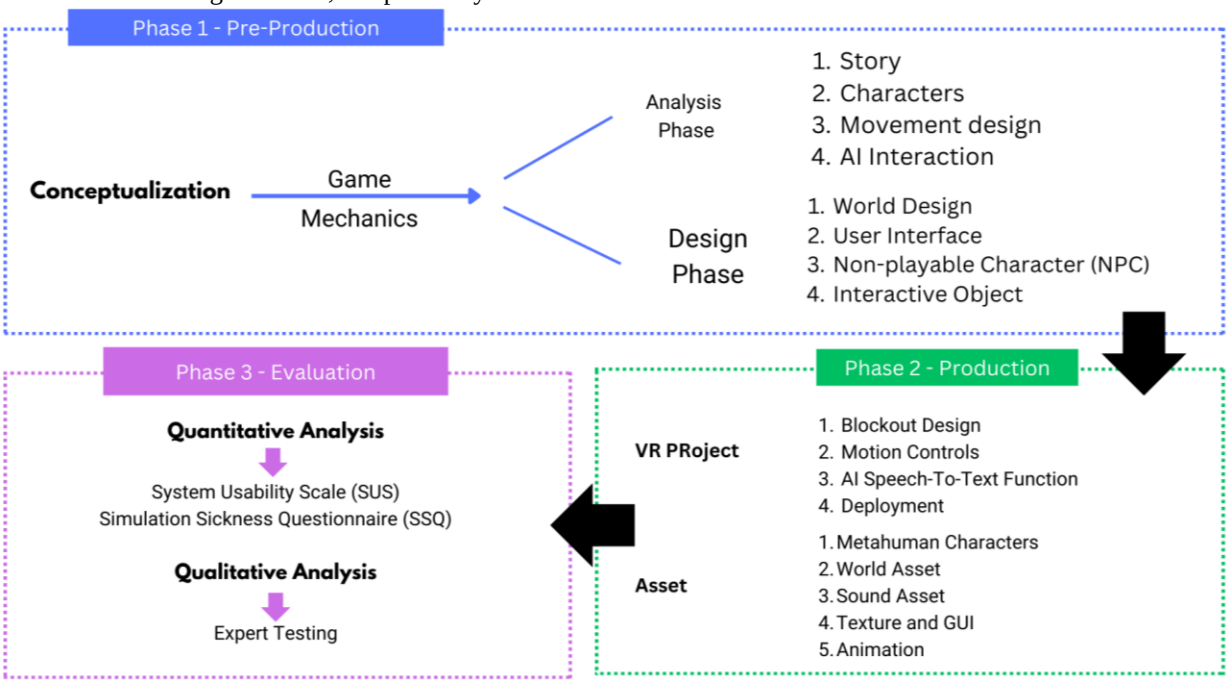


Fig. 2. Framework of AutiSim.

A. Phase 1 (Pre-Production)

The first phase involves establishing goals for an immersive autism simulation, with a focus on conceptualizing the layout design and integrating user data. The framework emphasizes enhancing user engagement through interactive objects, spatial audio simulation and real time responses from AI to simulate a realistic life situation. This approach tends to create more immersive environment, allowing users to interact with the virtual world in a way that closely replicates real-world experiences.

With the proposed enhancement aimed at creating a more immersive and engaging experience within the virtual world, the default movement mechanic in Unreal Engine originally based on teleportation is replaced with a more realistic locomotion system. Teleportation is initially implemented as the default method to help users become familiar with basic VR mechanics and to ease them into the virtual environment without overwhelming them. This method allows users to jump from one point to another, making it an effective tool for quick and easy navigation. However, once users are comfortable, the system transitions to locomotion, enabling them to move through the virtual space more naturally, as they would in the real world.

Unlike teleportation, which restricts users to predefined spots, locomotion grants continuous movement throughout the environment. This shift not only enhances spatial awareness but also contributes to a deeper sense of presence, making the experience feel more lifelike. The benefits of this transition can be observed in improved user interaction and engagement, as illustrated in Fig. 3.

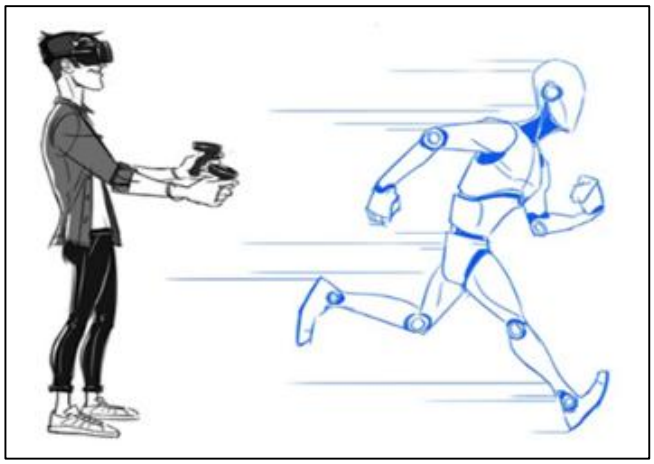


Fig. 3. Implementation of locomotion in VR.

AutiSim featured a visually appealing graphical user interface (GUI) designed to guide users smoothly from the very beginning of the game. The interface plays a crucial role in helping users feel comfortable and confident as they navigate through the virtual environment. The GUI is developed with usability and inclusivity in mind, incorporating simple language, recognizable icons, and consistent design patterns to reduce the cognitive load and prevent users from feeling overwhelmed. Key features such as intuitive menu transitions and easy-to-understand navigation options contribute to a seamless experience. Additionally, the interface will offer feedback through brief introductory explanations about the game as users progress, supporting them in making choices and understanding the outcomes of their actions. Fig. 4 illustrates the final design of the main menu in AutiSim, highlighting its clean layout, user-focused design, and ability to support a smooth onboarding process for both new and returning users.

AI non-playable characters (NPC) created by using Metahuman Creator are integrated into virtual environments to significantly enhance the immersive experience of the autism simulation game. These highly realistic avatars are meticulously crafted with detailed facial features, expressive gestures, and naturalistic movements, aligning seamlessly with the simulation's aesthetic and functional goals. The advanced customization tools of Metahuman Creator enable the creation of diverse and relatable characters, tailored to reflect unique traits, appearances, and personalities. This customization enriches the user experience by fostering deeper emotional connections and facilitating authentic, meaningful interactions within the virtual world. These characters play a pivotal role in replicating real-world social dynamics, making the simulation both educational and impactful, particularly in raising awareness about autism-related challenges. By portraying realistic behaviors, they provide users with an engaging and relatable environment that deepens their understanding of the condition. Fig. 5 and Fig. 6 showcase the transformation of character models, emphasizing the progression from initial designs to fully polished avatars.



Fig. 4. Main menu's design.



Fig. 5. Mother's character design.

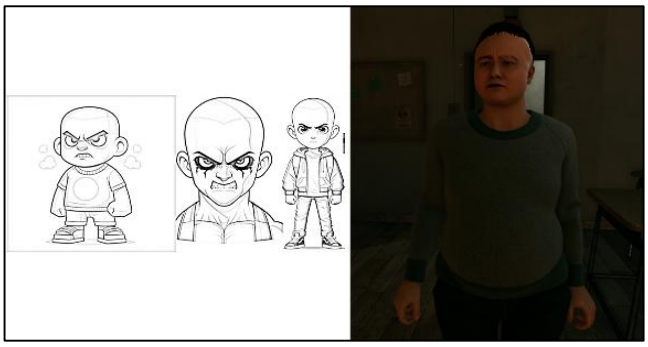


Fig. 6. Bully's character design.

Without a defined set of rules, an interactive experience cannot be categorized as a video game. Therefore, both the winning and losing conditions must be established. In each level, the winning condition is to locate the objective marker hidden within the environment while keeping the stress level as low as possible. On the other hand, the losing condition occurs when the player's stress level is maxed out, causing the game to restart from the beginning of the level.

B. Phase-2 (Production)

In Phase 2, the project moves to the production stage, where the detailed levels and scenes for AutiSim are created and finalized. This includes realistic virtual settings like an old house, a classroom, and a train station, all designed to closely resemble real places and sensory experiences. The goal is to make the simulation feel lifelike and engaging. To improve how users interact with the game, features like hand tracking for natural gestures and haptic feedback for touch sensations are added. Various software tools support this work: Unreal Engine is used for building the game, Blender for making 3D models, and Adobe Illustrator for designing visual elements and the user interface. Together, these tools help create an interactive and immersive experience that meets the aims of AutiSim.

As shown in Fig. 7, Level 1 takes place in an old brick house with a cozy, nostalgic design that feels like early childhood. This setting represents the player's younger years, from about age 3 to 12, and helps them feel comfortable as they start the game. The familiar rooms and decorations make it easy to get used to the simulation. The main purpose of Level 1 is to teach players how to move around, interact with objects, and understand basic controls through simple tutorials and prompts. Some of the basic interactions provided include a healing mechanic, moving around the character and interacting with the AI character to learn more about autism and how it can affect someone's daily life situation.



Fig. 7. Level 1 design.

Level 2 introduces players as high school students, aged approximately 13 to 18, attending a regular public high school, instead of a special needs school, as displayed in Fig. 8. The challenge in this level is to navigate out of the classroom without interacting with others, reflecting the tendency of individuals with ASD to prefer solitude and avoid social engagement. Additionally, other students (represented by AI NPCs) may observe and mock the player, further emphasizing the social difficulties often experienced by those with ASD. The environment is designed to simulate a realistic and sometimes overwhelming school setting, with background chatter, shifting body language, and unpredictable peer behavior. This level encourages players to develop awareness of how social anxiety and overstimulation can affect decision-making and behavior in typical school environments.

Level 3 introduces players to adulthood, where the character is around 18 to 25 years old, with the objective of catching a train as quickly as possible to avoid noise disturbances in a public place. The train station environment is heavily inspired by the design of Keretapi Tanah Melayu Berhad (KTMB) stations across Peninsular Malaysia, featuring the iconic yellow and blue signboards to provide a strong sense of localization for players. This familiar setting enhances immersion and cultural relevance. Realistic sounds such as train engine noises and crowd chatter are added to simulate the sensory challenges often faced by individuals with Autism Spectrum Disorder (ASD). Fig. 9 shows the level design for Level 3.

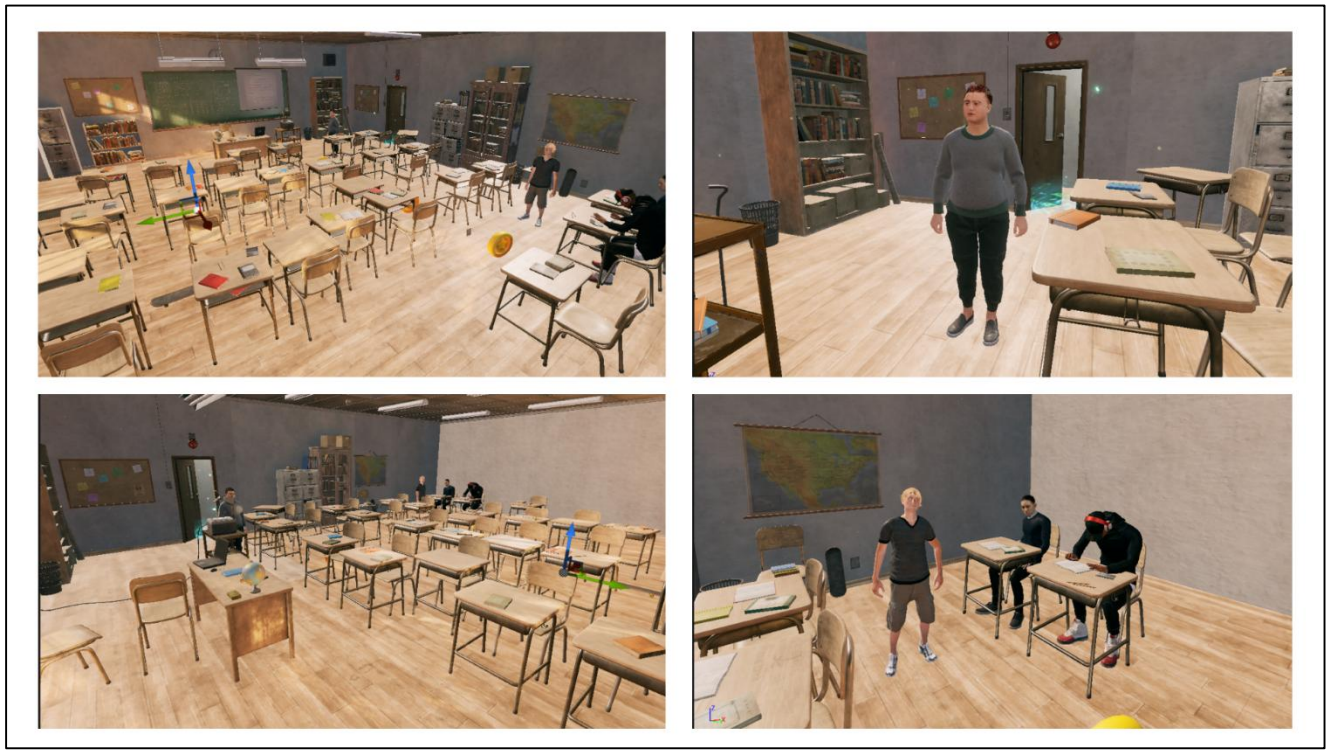


Fig. 8. Level 2 design.

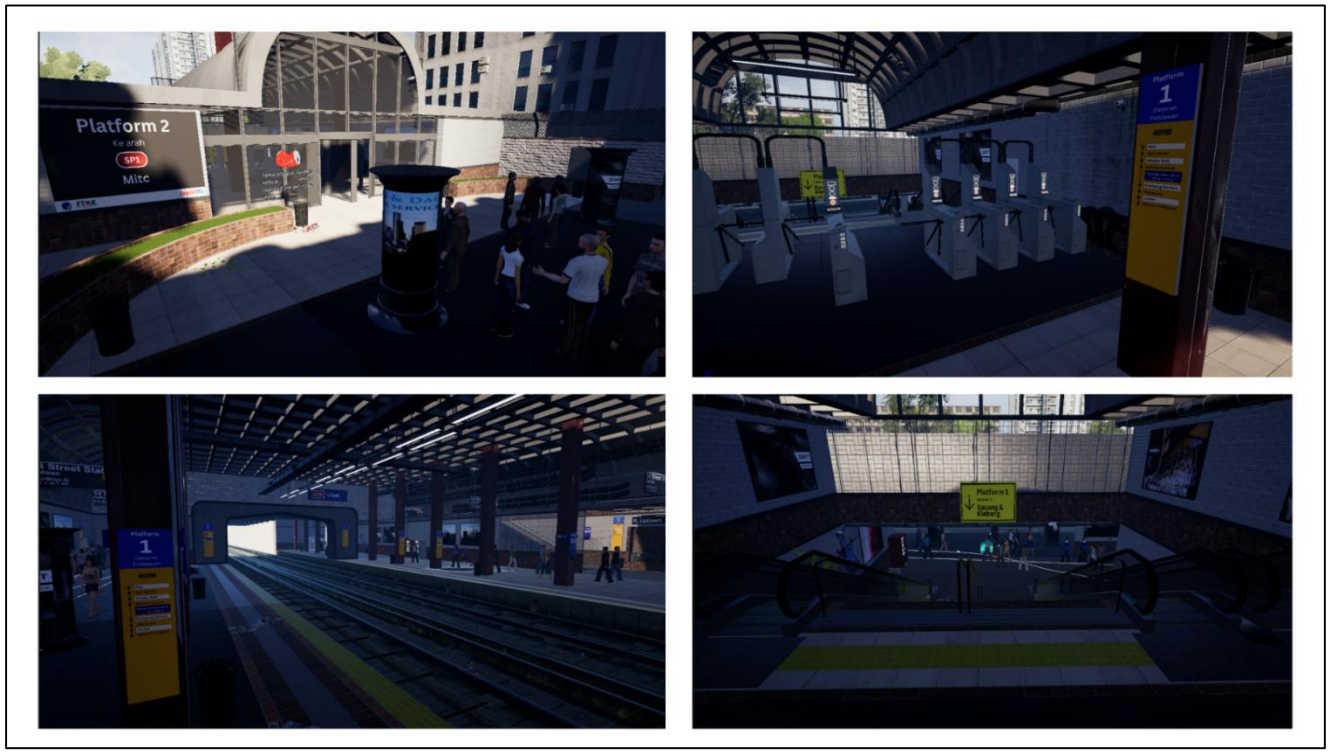


Fig. 9. Level 3 design.

This game incorporates spatial audio in combination with visual and damage effects to replicate the sensory challenges often experienced by individuals on the autism spectrum when exposed to overwhelming or sensitive stimuli, such as loud or high-pitched sounds. These sensory elements are designed to simulate real-world experiences, helping players and observers gain a deeper understanding of how such stimuli can impact focus, comfort, and overall functioning. One example of this occurs when the player steps into a designated hitbox area, which immediately triggers the sound of a boiling kettle with an

intentionally sharp and continuous noise that may be distressing or distracting for some individuals. In response to this audio cue, a blurriness effect overlays the screen to visually simulate sensory overload, accompanied by a gradual reduction in the player's health or energy to reflect the mental and emotional strain caused by the stimulus. This combination of spatial audio and reactive visuals adds an extra layer of immersion and empathy-building to the gameplay. Fig. 10 illustrates how the blur effect disrupts the player's field of vision, demonstrating the potential impact of sensory overstimulation on everyday activities within the virtual environment.



Fig. 10. Blur effect triggers when encountering an obstacle.

As for the character, a 3D photogrammetry technique is used to create realistic human features [26] [27]. The process involves capturing a high-resolution, panoramic 360-degree photo of the model's face to ensure that every detail is accurately recorded—from the broader facial contours and bone structure to subtle features like skin texture, fine wrinkles, and expression lines. This comprehensive scanning technique allows for an exceptionally lifelike digital representation of the individual's face. The Reality Scan app is employed as the primary software tool in this process, chosen for its ability to process and reconstruct detailed 3D models with high fidelity. An iPhone 11 serves as the scanning device, offering a balance between mobility, accessibility, and image quality. Together, this combination of hardware and software produces a remarkably realistic output, often requiring little or no manual adjustment or post-processing. The simplicity and efficiency of the setup make it ideal for quick integration into the development pipeline, especially for projects focused on creating immersive virtual environments. Fig. 11 illustrates the step-by-step process and shows the result of the facial scan, highlighting the precision and clarity of the digital model.

With the 3D character featuring realistic facial details completed, a facial capture technique is further implemented to animate the character's face using an iPhone 11 and LiveLink by Unreal Engine. This technique enables real-time projection of facial movements onto the 3D model, along with voice integration, to create a tutorial introduction to the game. The tutorial will pop up in the main menu, guiding players on what to expect and what they will learn, improving their overall experience and understanding of the game. Fig. 12 highlights the facial capture process and the resulting animations, showcasing the seamless integration of these elements into the game.



Fig. 11. The process of how character is created.



Fig. 12. The facial capture for the introduction.

C. Phase-3(Testing)

For the final phase, the evaluation involved both quantitative and qualitative analysis through a series of testing sessions with students and experts [28] [29]. Quantitative analysis was carried out using the System Usability Scale (SUS) and the Simulator Sickness Questionnaire (SSQ) method. SUS was measured through a 10-question survey, with the total score calculated and converted to a final score between 0 and 100. Higher scores indicate better usability, with scores above 68 generally considered above average. SSQ collected feedback on any discomfort or symptoms users experienced during the VR sessions, addressing potential issues to improve user comfort.

As for qualitative analysis, it was performed through expert testing, offering valuable insights from their experiences with the VR content, which informed further improvements. The expert testing involved professionals from both the VR technology and autism fields, ensuring a comprehensive evaluation from both technical and therapeutic perspectives.

IV. Results and Discussion

A. System Usability Scale (SUS)

This section presents the results of the System Usability Scale (SUS) assessment conducted to evaluate the overall usability and user experience of the application. The evaluation involved a total of ten participants, consisting of seven students and three teachers from SMK Tunku Syed Idrus Tampin. Participants were selected to represent both the target users and educational facilitators who would potentially engage with or guide the use of the application in a real-world setting. Each participant completed the SUS questionnaire after interacting with the system, providing feedback on aspects such as ease of use, functionality, and overall satisfaction. The responses were then scored and analyzed to produce a quantitative measure of the system's usability. The results, summarized in Table I, offer valuable insight into how effectively the application meets user expectations and identifies areas for potential improvement. SUS results for ten participants.

TABLE I. SUS RESULTS FOR 10 PARTICIPANTS

Participants	Raw Score	Final Score
P1	40	92.5
P2	25	75
P3	39	80
P4	25	50
P5	44	50
P6	20	85
P7	44	67.5
P8	21	70
P9	43	72.5
P10	26	87.5
Average		73

The AutiSim SUS score was 73, just above the industry benchmark of 68 [29]. Users thought the game worked well, was simple to learn, and had a good interface and usability [30] [31]. Although virtual reality (VR) technology is not yet widely accessible due to the need for specialized and often costly hardware, the above-average System Usability Scale (SUS) score indicates that the game is positively received and well-accepted by users despite these limitations. This suggests strong potential for wider adoption as accessibility improves over time. The high usability rating reflects the effectiveness of the game's design, interface, and interactive elements in delivering an engaging experience [32]. For further enhancing the user satisfaction and improving future SUS scores, optimizing the game's performance, particularly the frame rate and addressing the key issues highlighted in user feedback will be essential.

B. Simulation Sickness Questionnaire (SSQ)

In this section, the same ten participants completed a pre-session Simulator Sickness Questionnaire (pre-SSQ) first to assess their initial condition. After experiencing the VR project, they filled out another post-session Simulator Sickness Questionnaire (post-SSQ) to evaluate any changes. The radar chart in Fig. 13 shows the results from the SSQ questionnaire.

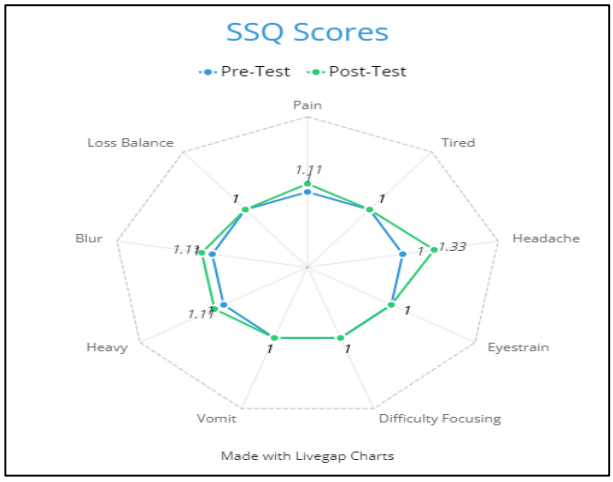


Fig. 13. SSQ's results.

The SSQ results in Fig. 13 show a slight increase in Pain, Blur, and Heavy (from 1 to 1.11), and Headache (from 1 to 1.33) after the VR session, while other symptoms like Tired, Eyestrain, Difficulty Focusing, Vomiting, and Loss of Balance remained unchanged. This suggests mild fatigue and headache, but no major issues with other symptoms. Overall, the results indicate a positive user experience with only minor discomfort, typical for first-time VR users. These results show that AutiSim managed to deliver an immersive experience with minimal discomfort, demonstrating its potential for effective use in future sessions.

C. Expert Testing

For expert testing, the evaluation was conducted with five experts. The first expert is a leading authority in virtual and mixed reality, the second is a game design specialist, the third is the head coordinator of the National Autistic Society of Malaysia, Melaka, and the other two are teachers with expertise in working with individuals on the autism spectrum. Fig. 14 shows the results chart of the expert testing for AutiSim.

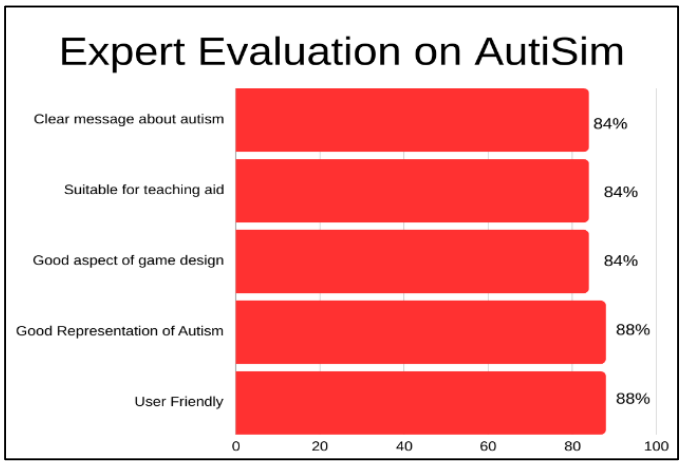


Fig. 14. Expert evaluation's chart.

User friendly refers to aspects such as movement, interaction, and speaking controls. 88% of the experts agreed that AutiSim is easy to use, highlighting the smooth locomotion system that allows players to move fluidly, as opposed to a typical teleportation scheme. Additionally, the ability to rotate the camera using both the headset and the left controller enhances the user experience, providing players with more control over their movement and camera positioning.

Next, a proper representation of autism is essential, as incorrect portrayals can lead to misunderstandings and misperceptions about the condition. It is important to highlight aspects such as slow movement, overwhelming surrounding sounds, and increased stress due to specific interactions, as these are characteristics often experienced by individuals with autism in real life. This was supported by 88% of the experts, who agreed that AutiSim accurately represents these aspects, contributing to a better understanding of the challenges faced by individuals on the autism spectrum.

Furthermore, 84% of experts agreed that AutiSim has a good game design, is well-suited as a teaching aid, and provides meaningful subject matter. These results highlight AutiSim significant potential as an educational tool, effectively raising awareness and deepening understanding of autism. By

combining engaging gameplay with rich educational content, AutiSim stands out as an exceptional resource for fostering empathy and promoting knowledge across diverse learning environments.

Since the expert testing was conducted in relevant environments, specifically in the VR technology and autism fields, as shown in Fig. 15, Fig. 16, and Fig. 17, respectively. Valuable feedback was gathered to improve the game. One insightful suggestion from the VR expert was that porting the game to mobile devices, such as smartphones, could help people better understand autism.

While the expert in the autism field, led by the head coordinator of the National Autistic Society of Malaysia (NASOM), Melaka branch, with over 10 years of experience, stated that the game could be a game changer for people, offering a unique way to experience autism through a game format. Additionally, it was suggested that future updates could include the ability for players to interact in multiple languages.



Fig. 15. VR Expert testing AutiSim.



Fig. 16. Head of coordinator NASOM trying the simulation.



Fig. 17. NASOM staff trying the simulation.

Overall, based on the expert testing conducted. AutiSim reflects highly positive feedback, with experts from both the VR technology and autism fields recognizing its potential as an impactful tool for raising awareness and fostering a deeper understanding of autism. The immersive experience was lauded for its ability to provide users with a unique and engaging way to experience the challenges faced by individuals with autism. The game's design, including its educational value and realistic portrayal of autism-related experiences, was praised. Experts also suggested that porting the game to mobile devices could make it more accessible to a wider audience. Additionally, it was recommended that integrating additional features, such as multi-language support, could further enhance the game's accessibility and broaden its reach. These positive insights indicate that AutiSim has the potential to be a valuable resource in both educational and awareness-raising contexts.

V. CONCLUSION

In conclusion, AutiSim highlight its potential as a transformative tool for raising awareness and understanding of autism spectrum disorder (ASD). Utilizing the immersive capabilities of virtual reality, AutiSim offers an engaging and realistic experience that replicates the sensory and social challenges faced by individuals with autism. The project followed a structured methodology based on the Game Design Life Cycle (GDLC), encompassing pre-production, production, and evaluation phases, ensuring a well-rounded and effective approach.

During the pre-production phase, a literature review guided the conceptualization of the game, addressing gaps in the representation of autism in existing games. This phase also defined the immersive features of AutiSim, including locomotive movement, AI-driven non-playable characters (NPCs), and realistic environments. The production phase brought these concepts to life using advanced software like Metahuman Creator, Unreal Engine, and Reality Scan, resulting in a highly realistic and educational simulation. The game's multi-level design allowed players to explore autism-related challenges across different life stages, providing a comprehensive learning experience.

The evaluation phase utilized both quantitative and qualitative methods to assess the game's effectiveness. The System Usability Scale (SUS) scored 73, exceeding the industry benchmark, indicating user satisfaction and ease of use. The

Simulator Sickness Questionnaire (SSQ) showed minor increases in symptoms like headache and fatigue, which are typical for first-time VR users, but overall results indicated a positive user experience. Expert testing further validated the game's effectiveness, with 84% agreeing it is suitable as a teaching aid and 88% commending its accurate representation of autism-related experiences. Experts also provided actionable suggestions, such as adding multi-language support, optimizing visuals for smoother performance, and porting the game to mobile devices to improve accessibility.

Despite these promising results, the project has certain limitations that need to be addressed in future work. The limitations of this project include the restricted capability for AI characters to communicate with users in multiple languages, as the current system supports voice interaction in only one language due to resources and budget constraints. Furthermore, the study faced technical and hardware challenges in implementing the project as a fully standalone application that runs natively on a VR headset, which required significant optimization and affected the seamless deployment and performance of the simulation.

In summary, AutiSim effectively combines education and entertainment through innovative VR technology and a well-executed design. With planned refinements based on SUS, SSQ and experts' feedback, the game has the potential to become a useful resource in educational settings and public awareness initiatives, promoting empathy, understanding, and acceptance.

ACKNOWLEDGMENT

This research was funded by a short grant from Universiti Teknikal Malaysia Melaka through Penyelidikan Jangka Pendek Perintis 2024, No: PJP/2024/FTMK/PERINTIS/SA0042. The authors gratefully acknowledge the Centre for Research and Innovation Management (CRIM) for its administrative and logistical support throughout the research process. Sincere appreciation is extended to the Pervasive Computing and Educational Technology (PET) research group for their invaluable guidance, collaboration, and consistent support during all phases of the study. The authors also wish to recognize the Department of Interactive Media, Fakulti Teknologi Maklumat dan Komunikasi, UTeM, for providing essential resources and technical assistance. Special thanks are owed to the National Autistic Society of Malaysia (NASOM) Melaka for their partnership and constructive feedback, which played a vital role in shaping the outcomes of this work. Lastly, the authors are deeply grateful to all contributing researchers whose insightful recommendations and scholarly input significantly enriched the quality and impact of this research.

REFERENCES

[1] A. Alnagrat, R. Che Ismail, S. Z. Syed Idrus, and R. M. Abdulhafith Alfaqi, "A Review of Extended Reality (XR) Technologies in the Future of Human Education: Current Trend and Future Opportunity," Journal of Human Centered Technology, vol. 1, no. 2, pp. 81–96, Aug. 2022, doi: 10.11113/humentech.v1n2.27.

[2] A. Çöltekin et al., Extended reality in spatial sciences: A review of research challenges and future directions, vol. 9, no. 7. MDPI AG, 2020. doi: 10.3390/ijgi9070439.

[3] A. Hamad and B. Jia, "How Virtual Reality Technology Has Changed Our Lives: An Overview of the Current and Potential Applications and Limitations," Sep. 01, 2022, MDPI. doi: 10.3390/ijerph191811278.

[4] C. Osadzuk, "Portrayal and representation of autism spectrum disorder through ludonarrative design.," 2024.

[5] S. C. Jones, C. S. Gordon, and S. Mizzi, "Representation of autism in fictional media: A systematic review of media content and its impact on viewer knowledge and understanding of autism," Nov. 01, 2023, SAGE Publications Ltd. doi: 10.1177/13623613231155770.

[6] G. Riva et al., "Presence-Inducing Media for Mental Health Applications," 2015, pp. 283–332. doi: 10.1007/978-3-319-10190-3_12.

[7] I. Wohlgenannt, A. Simons, and S. Stieglitz, "Virtual Reality," Business & Information Systems Engineering, vol. 62, no. 5, pp. 455–461, 2020, doi: 10.1007/s12599-020-00658-9.

[8] A. Rahouti, R. Lovreglio, S. Datoussaid, and T. Descamps, "Prototyping and Validating a Non-Immersive Virtual Reality Serious Game for Healthcare Fire Safety Training," Fire Technol, vol. 57, Sep. 2021, doi: 10.1007/s10694-021-01098-x.

[9] M. Slater and M. Sanchez-Vives, "Transcending the Self in Immersive Virtual Reality," Computer (Long Beach Calif), vol. 47, pp. 24–30, Sep. 2014, doi: 10.1109/MC.2014.198.

[10] S. Alanko and K. Sauren, "Comparing Inside-out and Outside-in Tracking in Virtual Reality," 2023.

[11] R. Monica and J. Aleotti, "Evaluation of the Oculus Rift S tracking system in room scale virtual reality," Virtual Real, vol. 26, no. 4, pp. 1335–1345, Dec. 2022, doi: 10.1007/s10055-022-00637-3.

[12] R. Lopez Mendez, "Mobile inside-out VR tracking, now available on your phone: extended abstract," Sep. 2018, pp. 1–2. doi: 10.1145/3213779.3213781.

[13] S. Barai and M. Momin, "Outside-in Electromagnetic Tracking Method for Augmented and Virtual Reality 6-Degree of Freedom Head-Mounted Displays," in 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 467–476. doi: 10.1109/ICICCS48265.2020.9121174.

[14] S. Baron-Cohen et al., "Prevalence of autism-spectrum conditions: UK school-based population study," British Journal of Psychiatry, vol. 194, no. 6, pp. 500–509, Jun. 2009, doi: 10.1192/bjp.bp.108.059345.

[15] P. Wijngaarden-Cremers, E. van Eeten, W. Groen, P. Deurzen, I. Oosterling, and R. Gaag, "Gender and Age Differences in the Core Triad of Impairments in Autism Spectrum Disorders: A Systematic Review and Meta-analysis," J Autism Dev Disord, vol. 44, Aug. 2013, doi: 10.1007/s10803-013-1913-9.

[16] M. DeFilippis and K. Dineen, "Treatment of Autism Spectrum Disorder in Children and Adolescents," Psychopharmacol Bull, vol. 46, pp. 18–41, Sep. 2016.

[17] D. Bai et al., "Association of Genetic and Environmental Factors With Autism in a 5-Country Cohort," JAMA Psychiatry, vol. 76, Sep. 2019, doi: 10.1001/jamapsychiatry.2019.1411.

[18] G. Romagnoli et al., "Occupational Therapy's efficacy in children with Asperger's syndrome: a systematic review of randomized controlled trials," Clin Ter, vol. 170, no. 5, p. e382—e387, 2019, doi: 10.7417/ct.2019.2164.

[19] O. Smith and S. C. Jones, "'Coming Out' with Autism: Identity in People with an Asperger's Diagnosis After DSM-5," J Autism Dev Disord, vol. 50, no. 2, p. 592—602, Feb. 2020, doi: 10.1007/s10803-019-04294-5.

[20] F. R. Volkmar, B. Reichow, and J. McPartland, "Classification of autism and related conditions: Progress, challenges, and opportunities," Dialogues Clin Neurosci, vol. 14, no. 3, pp. 229–237, Sep. 2012, doi: 10.31887/dcns.2012.14.3/fvolkmar.

[21] U. Petriti, D. C. Dudman, E. Scosyrev, and S. Lopez-Leon, "Global prevalence of Rett syndrome: systematic review and meta-analysis," Syst Rev, vol. 12, no. 1, Dec. 2023, doi: 10.1186/s13643-023-02169-6.

[22] C. Mehra et al., "Childhood disintegrative disorder and autism spectrum disorder: a systematic review," May 01, 2019, Blackwell Publishing Ltd. doi: 10.1111/dmcn.14126.

[23] H. Williams, “Symptoms and Causes of Kanner’s Syndrome,” 2022, doi: 10.35248/2375-4427.22.10.226.

[24] M. Filimowicz and V. Tzankova, New directions in third wave human-computer interaction. Volume 1, Technologies. 2018.

[25] M. Tsujita, M. Homma, S. I. Kumagaya, and Y. Nagai, “Comprehensive intervention for reducing stigma of autism spectrum disorders: Incorporating the experience of simulated autistic perception and social contact,” PLoS One, vol. 18, no. 8 AUGUST, Aug. 2023, doi: 10.1371/journal.pone.0288586.

[26] N. Azmi, I. F. Albakri, N. Suaib, M. Rahim, and H. Yu, “3D Motion and Skeleton Construction from Monocular Video,” 2020, pp. 75–84. doi: 10.1007/978-981-15-0058-9_8.

[27] I. F. Albakri, N. Wafiy, N. Suaib, M. Rahim, and H. Yu, “3D Keyframe Motion Extraction from Zapin Traditional Dance Videos,” 2020, pp. 65–74. doi: 10.1007/978-981-15-0058-9_7.

[28] H. T. K. Eunhee Chang and B. Yoo, “Virtual Reality Sickness: A Review of Causes and Measurements,” Int J Hum Comput Interact, vol. 36, no. 17, pp. 1658–1682, 2020, doi: 10.1080/10447318.2020.1778351.

[29] P. Vlachogianni and N. Tselios, “Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review,” Journal of Research on Technology in Education, vol. 54, no. 3, pp. 392–409, 2022, doi: 10.1080/15391523.2020.1867938.

[30] I. Ahmad, N. Abdullasim, and N. Suaib, “Usability testing on game interface design using video-based behavior analysis,” International Journal of Engineering and Technology, vol. 7, pp. 142–145, Dec. 2018, doi: 10.14419/ijet.v7i2.15.11372.

[31] I. Ahmad, H. Erman, N. Abdullasim, and A. Jaafar, “Game Interface Design: Measuring the Player’s Gameplay Experience,” Dec. 2017, pp. 500–509. doi: 10.1007/978-3-319-70010-6_46.

[32] N. N. B. Norazri, I. F. A. M. Albakri, M. K. Mokhtar, F. Mohamed, V. S. Chan, and M. N. A. Nor’a, “Simventure: Utilizing First Person Player Technology for Career Guidance in High Schools,” in 2024 International Conference on TVET Excellence & Development (ICTeD), 2024, pp. 38–43. doi: 10.1109/ICTeD62334.2024.10844602.

Critical Success Factors for Knowledge Transfer in Enterprise System Projects: A Theoretical and Empirical Investigation

Jamal M. Hussien[1], Riza bin Sulaiman[2], Ali H Hassan[3], Mansoor Abdulhak[4] , Hasan Kahtan[5]
Institute of IR 4.0 (IIR4.0), Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia[1, 2, 3]
Information and Communication Technology (ICT) Department, University of Prince Mugrin, Madinah, Saudi Arabia[1]
College of Computer and Cyber Sciences, University of Prince Mugrin, Madinah 41499, Saudi Arabia[3]
Computer Science Department at University of Oklahoma, Norman, Oklahoma, USA[4]
Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, UK[5]

***Abstract*—Enterprise System Projects (ESPs) are fundamental enablers of digital transformation across organizations, yet they consistently suffer from high failure rates, often attributed to ineffective Knowledge Transfer (KT) practices. Despite the critical role of KT in ensuring project sustainability and long-term organizational learning, limited scholarly attention has been given to identifying and systematically categorizing the success factors that influence KT outcomes in ESPs. The aim of this study is to investigate and conceptualize the Critical Success Factors (CSFs) that influence effective knowledge transfer in ESPs. To address this research gap, a mixed-methods approach is used, combining a literature review with empirical insights from semi-structured interviews with industry practitioners involved in large-scale ESP implementations. The analysis reveals a set of interrelated CSFs that significantly impact KT effectiveness. Some of the key points highlighted are the shared knowledge between cultures, the high expertise of consultants based on technicality and social skills, and the solid and visible management support. These points are integrated into a conceptual framework that enhances conceptual understanding while offering practitioners practical guidance. The study contributes by bridging the gap between the KT concept and ESP implementation, which are connected to the academic discourse, proposing a comprehensive model for successful knowledge transfer during the deployment of ESP. From a practical standpoint, the findings offer organizations a strategic lens to design and implement KT mechanisms that enhance project outcomes and ensure long-term knowledge retention.**

Keywords*—*Enterprise system projects (ESPs); knowledge transfer (KT); critical success factors (CSFs); digital transformation; knowledge-sharing culture; management support; information systems implementation

I. Introduction

ESPs are essential to modern organizations pursuing trending digital transformations. They enable real-time access to data, cross-functional automation processes, and calculated decision-making by integrating basic business functions, such as finance, human resources, supply chain, business intelligence, and management of customer relationships, into centralized platforms according to [1] With organizations increasingly investing in digital capabilities to stay competitive, enterprise system deployment is frequently considered a technological and organizational imperative. Despite the significance of ESPs, the failure rate remains alarmingly high, with reported project failures often occurring in meeting critical baseline objectives in terms of cost, scope, or sustained usability. This supports the assertion made by [2], [3]. Due to these consistent failures, a comprehensive understanding of the underlying challenges beyond technological factors and encompassing socio-organizational dynamics, particularly in knowledge processing, must be explored.

The effective collaborative transfer of knowledge, from external consultants to internal teams and stakeholders, remains one of the most critical yet underappreciated aspects of ESP integration. KT plays a significant role in the long-term success and sustainability of ESPs, especially following implementation, when system shifts between internal teams and consultants are made through maintenance, adaptation, and extension responsibilities. As noted by [4], KT is a complex, multi-layered activity, involving transference of tactical and precise knowledge, cross-boundary collaboration, and shared understanding development across culturally and functionally diverse groups, not only a process of training or documentation, as identified by [5]. Ineffective KT can lead to an inability to utilize the system efficiently, poor user adoption, high maintenance costs, and eventual system underperformance or failure. Yet, despite its recognized importance, KT remains a minor concern in many ESP studies and implementations, often treated as a less critical task rather than a strategic priority, as noted by [6].

Critical Success Factors (CSFs) provide a helpful lens to examine KT effectiveness in ESPs. CSFs refer to the limited number of key areas in which satisfactory performance is essential for the success of a project as mentioned earlier by [7], [8]. In the context of ESPs, numerous CSFs have been proposed, ranging from top management support and user involvement to effective project planning and change management as shown by [9], [10]. However, these aspects are typically considered together, without a focused exploration of how they specifically impact knowledge transfer. Moreover, the interdependencies among CSFs—how they reinforce or constrain each other in practice—are rarely unpacked in empirical studies. For instance, while top management support is often cited as a CSF, its influence on KT may depend on how visibly and consistently

that support is demonstrated, and whether it translates into concrete actions such as resource allocation, stakeholder engagement, or performance monitoring, as noted by [11]. Similarly, while consultant expertise is critical, its impact on KT is mediated by the receiving organization communication channels, cultural compatibility, and absorptive capacity according to [6].

Despite the availability of frameworks that address general KT practices or ESP success factors, few studies have systematically integrated these domains to identify the specific CSFs that influence KT in enterprise system contexts. Those that do often lack empirical validation or fail to account for the dynamic nature of KT, which evolves across different project phases and stakeholder groups [12]. Furthermore, much existing research is based on single-case studies or surveys with limited contextual sensitivity. There is a pressing need for research that synthesises existing knowledge and generates new insights based on empirical data from practitioners who have directly experienced the complexities of ESP implementation.

This study addresses this gap by exploring the CSFs that successfully impact KT in ESP environments. Drawing from the insights of a literature review and empirical data collected through semi-structured interviews with stakeholders involved with large-scale ESP implementations and KT processes, the research identifies and categorizes a set of interrelated success factors, including management support and a knowledge-sharing culture. The findings reveal that these CSFs do not operate alone but are linked, in which the effectiveness of one factor often depends on the presence or quality of others. For example, a strong knowledge-sharing culture may enrich the benefits of consultant expertise, as argued by [13].

A key contribution of this study is developing a conceptual framework that maps the relationships between identified CSFs and KT outcomes in ESPs. In contrast to current models that mainly consider success factors as isolated variables, this framework adopts a systems perspective, highlighting the importance of dynamic collaborations between organisational structures, human behaviour, and technological tools. Unlike models that treat success factors as discrete variables, this framework adopts a systems perspective, emphasizing the dynamic collaborations among organizational structures, human behavior, and technological tools. It seeks to report the KT challenges by systematically identifying and understanding the key CSFs that allow organisations to apply external knowledge meaningfully within their settings. By bridging the gap between KT theory and ESP practice, this study contributes to both academic knowledge and practical application, offering a foundation for more resilient, adaptive, and knowledge-driven enterprise systems in the digital age.

In response to the persistent KT challenges in ESPs and the limited frameworks addressing them, this study poses the following research questions (RQ):

RQ1: What are the critical success factors that influence effective knowledge transfer in enterprise system projects?

RQ2: How can a practical and conceptual framework be developed to assess and enhance the knowledge transfer of ESPs' implementation?

The remaining study is structured as follows: Section II presents a review of studies related to ESPs, particularly focusing on CSFs that significantly influence Knowledge Transfer (KT) outcomes in ESPs. Section III outlines the research methodology adopted in this study, including the rationale for a qualitative approach and the dual-method strategy comprising a literature review and semi-structured interviews. Section IV presents the results of the thematic analysis, drawing on responses from study participants and relating to the findings from the literature review. Section V offers a discussion of the results. The proposed conceptual framework is also introduced in this section. Finally, Section VI concludes the study by summarizing key findings, discussing implications for theory and practice, and offering directions for future research.

II. Literature Review

ESPs have long been studied for their transformative potential within organizations, primarily through their ability to streamline operations, enhance data integration, and support decision-making at strategic and operational levels, as shown in the works of [14], [15], and [16]. ESPs are typically complex, high-cost initiatives involving multiple stakeholders, technologies, and organizational units, which require significant change management and learning efforts to ensure their success. This supports the assertion made by [17].

While much of the early literature emphasized ESP implementation's technical and functional aspects, a growing body of research has highlighted the pivotal role of human and organizational factors in shaping project outcomes. These findings corroborate those of [5] and [9]. Among these, the KT has emerged as a critical but underexplored domain, especially in ESPs, where much of the expertise resides with external consultants rather than internal users. The challenge lies in conveying technical system knowledge and fostering a deep understanding of business processes, system customization, and change implications, where similar conclusions were drawn by [16] and [18]. The literature has identified several barriers to effective KT, including insufficient documentation, lack of contextual adaptation, resistance to change, and limited absorptive capacity, which is supported by [9], [4], [10], [19], and [20]. For example, there is a lack of recognition and support, trust in the quality and accuracy of information, and the convenience and effectiveness of sharing tools. This view is supported by multiple studies, including those of [7], [10], [11], [12], and [13].

Despite these insights, empirical research that systematically investigates the conditions that foster or hinder KT in ESPs remains limited, particularly in identifying and validating CSFs that enable KT efficacy. The concept of CSFs was first articulated by Rockart (1979), who defined them as the few key areas in which satisfactory performance is essential for achieving organizational goals, as argued by [7]. In the context of ESPs, CSFs have been widely examined, with most studies identifying factors such as top management support, clear goals and objectives, user training, change management, and consultant expertise as crucial to implementation success. Strong empirical support for this position is found in the work of [9], [10], [21], and [22]. However, these studies often treat CSFs as generalized enablers of project success rather than

specific originators to effective KT. This generalization limits the explanatory power of CSFs in addressing the nuanced, process-oriented challenges that arise in knowledge-intensive implementations like ESPs.

Recent research has begun to narrow this gap by focusing on KT-specific CSFs as observed in the work of [23]. For example, a model was proposed in which consultant credibility, client motivation, communication quality, and relationship strength were key enablers of successful knowledge transfer. Their findings suggest that KT is a socially embedded process influenced by trust, mutual understanding, and collaborative engagement. Scholars consistently identify top management support as a pivotal CSF in KT and ESP research. It is critical not only for resource mobilization and strategic alignment but also for legitimizing knowledge-sharing behaviors across the organization, which is supported by [20] and [24]. Contrary to the view of authors in [7], who found that management support plays a critical role in the success of KT, in addition to the success of the ESP. Another research [14] revealed that a lack of top management knowledge of product capabilities is one of the ESP success barriers. In contrast, [15] highlighted that top management extensively influences the KT. Furthermore, from research findings in [8] and [9], it emerged that top management support is vital for the success of the ICT project. This view is supported by multiple studies, including those of [10], [16], and [17]. When top management visibly endorses KT initiatives, it creates an enabling environment where employees are more likely to engage in learning activities, participate in workshops, and experiment with system functionalities. Conversely, a lack of executive involvement often leads to fragmented efforts, low user motivation, and resistance to knowledge assimilation.

Another vital factor is consultant expertise, according to [9], which goes beyond technical competence to include interpersonal skills, cultural sensitivity, and the ability to contextualize knowledge for diverse user groups in line with the findings of [25] and [26]. Consultants serve as system experts and knowledge brokers who bridge the gap between generic system capabilities and specific organizational needs. Similar conclusions were drawn by [27]. However, the efficacy of consultant-driven KT is contingent upon a knowledge-sharing culture within the client organization, as previously reported by [28] and [29]. Without a culture that values collaboration, openness, and continuous learning, even the most structured KT efforts may fail to take root.

Communication quality has also been identified as a determinant of KT success. Effective communication entails clarity, consistency, feedback loops, and mutual understanding between knowledge senders and receivers, according to [30]. In ESP projects, communication breakdowns are common due to jargon, role ambiguity, and misaligned expectations. Studies have shown that structured communication channels—such as regular status meetings, shared documentation platforms, and escalation protocols—enhance KT effectiveness by reducing uncertainty and reinforcing shared meanings. Similar conclusions were drawn by [16]. While researchers in [18] and [19] found that strong communication facilitates easy KT and better ESP adoption, authors in [20] argued that poor or ineffective communication can lead to silos and resistance to change resulted on ESP project over runs. In [21], the authors examine the communication quality from a contextualization perspective to ensure the ESP team receive the correct information concerning their level of expertise or business context, while scholars. In [22] and [23], the authors highlight the impact of trust, which is consistent with [24]. In addition to openness, there is a need to enhance the quality of communication.

In summary, while the literature on ESPs, KT, and CSFs has matured significantly over the past decades, critical gaps remain in integration, contextualization, and empirical validation. Table I presents the KT-specific CSFs identified in recent literature as essential for facilitating successful KT in ESPs. These factors span leadership, human capital, communication, motivation, culture, relationships, and learning capability, highlighting the multidimensional nature of KT effectiveness.

While existing research offers valuable insights into individual success factors, there is a lack of a holistic understanding of how these elements interact to influence KT outcomes in enterprise system environments, as identified by [25]. The present study seeks to bridge this gap by developing an empirically grounded framework that identifies, categorizes, and analyzes CSFs affecting KT effectiveness in ESPs, thereby contributing to theoretical advancement and practical application in this evolving field.

TABLE I. CRITICAL SUCCESS FACTORS (CSFS)

CSF Domain	KT-Specific Critical Success Factor	Definition	Supporting Studies
Leadership & Support	Top Management Support	Strategic endorsement and resource allocation for KT legitimize knowledge-sharing behaviors.	[20], [24]
Human Capital	Consultant Expertise	Encompasses not only technical knowledge but also interpersonal skills and contextual understanding.	[25], [26], [28], [29]
Communication	Communication Quality	Quality of interactions: clarity, consistency, feedback, and shared understanding.	[16], [30]
Motivational Factors	Client Motivation	Internal drive and willingness of the organization to participate in and absorb transferred knowledge.	[21]
Cultural Context	Knowledge-Sharing Culture	Organizational norms that promote openness, collaboration, and learning.	[28], [29].
Relational Dynamics	Relationship Strength	Trust and rapport between consultants and client stakeholders foster mutual KT.	[21], [25]
Process Management	Structured Communication Channels	Formalized mechanisms such as meetings, platforms, and escalation paths for KT.	[16]

III. METHODOLOGY

This study adopted a qualitative, interpretivist research methodology to investigate the CSFs that influence KT effectiveness in ESPs. KT within large-scale system implementations is inherently complex, context-sensitive, and socio-technical. Therefore, a qualitative approach was valued as the most suitable approach to explore the perceptions, behaviors, and organizational conditions shaping successful KT practices, similar to the approach that was rooted in [26], [27], and [28]. Through a mixed-method approach, the study involved a literature review and semi-structured interviews as core empirical instruments. This allowed for an in-depth understanding of the explored topics. The literature review provided a base for identifying initial theoretical constructs and established a conceptual foundation for understanding KT-specific CSFs in ESP contexts. Sources were selected from various databases, including Scopus, Web of Science, Science Direct, and Google Scholar, containing journals in information systems, knowledge management, and enterprise systems. The review focused on identifying patterns, gaps, and limits in existing frameworks related to KT and ESP success, with special attention paid to studies incorporating organizational learning based on insights from [29], [30], in addition to change management, as argued by [17], and consultant-client dynamics as previously reported by [31].

The findings from the literature review informed the selection of five thematic categories: management support, consultant expertise, absorptive capacity, communication quality, and knowledge-sharing culture. Building on this foundation, the second phase involved conducting semi-structured interviews to explore how these themes manifest in real-world ESP implementations. Semi-structured interviews were chosen for their flexibility and capacity to elicit rich, contextually grounded insights while allowing for consistency across interviews in line with the findings of [32], [33]. The interviews enabled participants to reflect on their experiences and elaborate on how knowledge was shared, absorbed, and operationalized during the ESP lifecycle.

A purposive sampling strategy was employed to recruit participants directly involved in ESP implementations and KT processes. Eligible participants included the IT Director, Planning and Quality Director, Solution Architect, Technical Specialist, IT Management Consultant, IS Project Manager, and Techno/Functional Specialist. All participants had been directly involved in ESP implementation within the past ten years, ensuring contemporary relevance. To promote diversity of experience, participants were drawn from various sectors, including manufacturing, healthcare, higher education, and government services. Fig. 1 illustrates the distribution of the business sectors. The interview sample was selected based on the criteria shown in Table II.

Seven semi-structured interviews were conducted, each lasting between forty minutes on average. Interviews were held using Google Meet or MS Teams. Before the interviews, participants were provided with an overview of the research objectives and signed informed consent forms. Interviews were guided by a protocol informed by the literature review themes but remained open-ended to accommodate emergent issues. All interviews were audio-recorded with consent and transcribed verbatim for analysis. Ethical approval was obtained from the lead institution's research ethics board, and strict confidentiality protocols were followed to protect participant identities.

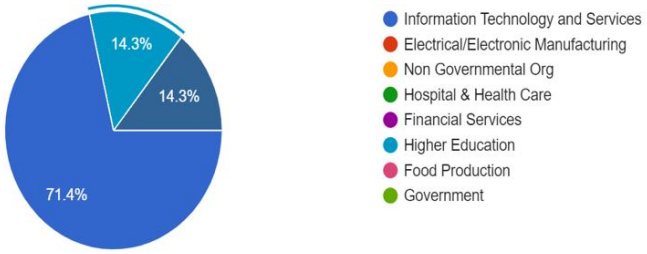


Fig. 1. Participants' distribution within the business sectors.

TABLE II. CRITERIA OF INTERVIEW PARTICIPATION

Expertise	At least 10 years of experience in ESP implementation and knowledge management.
Relevance	Direct involvement in ESP projects with varying outcomes (successful, challenged, or failed).
Diversity	Representation from different roles within ESP projects (e.g., project managers, system architects, and KT facilitators).

The qualitative data gathered from semi-structured interviews with ESP professionals was evaluated using thematic analysis. To identify recurrent themes, patterns, and insights on the specialized field of ESP, the researchers carefully reviewed and classified the written material obtained from the interviews. Through the thematic analysis, researchers were able to develop a better understanding of the ESP experts' viewpoints, experiences, and knowledge, which helped to offer valuable insights and conclusions for the study. Combining a literature review and interview-based understanding helped support this study's methodological triangulation, allowing the research to bridge theory and practice. While the literature provided a structured view of the academic views and gaps concerning KT in ESPs, the interviews offered in-depth insights into how practitioners operate, interact, and perceive these factors in a real-life context.

IV. FINDINGS AND CONCEPTUAL FRAMEWORK

The thematic analysis of seven semi-structured interviews revealed a set of five interrelated CSFs that significantly influence KT outcomes in ESPs. Namely, top management support, consultant expertise, communication quality, and knowledge-sharing culture. While these key factors were also identified in the literature review, the interviews provided significant, practice-driven insights that paved the way for a holistic understanding of the CSFs that influence KT outcomes in ESPs. The acquired knowledge helped develop the conceptual framework for knowledge transfer success in ESPs.

A. Top Management Support

One of the most cited enablers of effective KT was top management support. Participants repeatedly emphasized that leadership involvement must go beyond symbolic endorsement to include active resource allocation, milestone enforcement, and public reinforcement of knowledge-sharing behaviors. For example, one IT project manager noted, "It wasn't just the CIO giving a speech at the launch meeting. He reviewed our KT progress, ensured we had dedicated time for shadowing, and

rewarded teams that documented lessons learned". This finding reinforces the argument that visible, engaged leadership sets the tone for KT prioritization and legitimizes it as a core organizational process, as noted by [20], [24].

B. Consultant Expertise

Another pivotal factor was consultant expertise, but the interviews revealed that technical knowledge alone was insufficient. Effective consultants were described as capable of adapting explanations to user skill levels, facilitating trust-based relationships, and embedding learning opportunities within routine activities. A participant from the manufacturing sector described one consultant as "not just a tech whiz, but a people person who knew how to coach our finance team without overwhelming them". This echoes the authors in [25], [26]. The claim is that consultants' social and communicative competence significantly enhances KT effectiveness.

C. Communication Quality

Widely emphasized in the data, effective communication plays a dual role by facilitating other CSFs and mediating their impact on KT success. Transparent, collaborative, and timely communication was essential for clarifying expectations, resolving ambiguities, and aligning stakeholders. Poor communication was linked to misconceptions about responsibilities, misinterpretation of requirements, and duplication of KT efforts. Several participants cited collaborative platforms, feedback loops, and centralized documentation as practical strategies to enhance communication. These findings corroborate prior research indicating that a robust communication infrastructure is essential for successful ESP execution and knowledge dissemination. Similar conclusions were drawn by [34],[16], and [30].

D. Knowledge-Sharing Culture

Finally, a Knowledge-Sharing Culture was described as the foundational layer that either amplified or dampened the effects of other CSFs. Organizations with norms that valued openness, learning from failure, and cross-boundary collaboration were perceived to facilitate smoother KT. Siloed structures and inflexible hierarchies, on the other hand, led to information hoarding and weak adherence to KT activities. As an IT management consultant in the financial services sector observed, "In one company, I felt like I was planting seeds in fertile soil. In another, everything bounced off—there was no curiosity, no engagement". This finding supports the literature asserting that cultural enablers significantly influence organizational learning processes, as observed in the work of [28], [29].

E. Conceptual Framework

Drawing on the empirical findings and prior theoretical foundations, a Conceptual Framework for KT Success in ESPs is proposed in Fig. 2. This framework positions KT effectiveness at the center, surrounded by interdependent CSFs. Arrows between CSFs represent dynamic interactions, while outer layers capture the organizational and temporal context within which these factors operate. The framework contributes to the literature in three key ways. First, it provides a validated, practitioner-informed structure that aligns with and extends existing theoretical models of KT and ESP implementation. Second, it emphasizes the interactive and systemic nature of CSFs, moving beyond linear or additive conceptions. Third, it incorporates a temporal perspective, suggesting that effective KT strategies require flexibility and responsiveness to the evolving demands of different implementation phases.

The findings highlight that successful KT in ESPs is not the result of isolated best practices but rather an outcome of strategically aligned, culturally embedded, and dynamically coordinated actions. Organizations aiming to enhance ESP performance must adopt integrated KT strategies, considering structural enablers and behavioral and contextual dimensions.

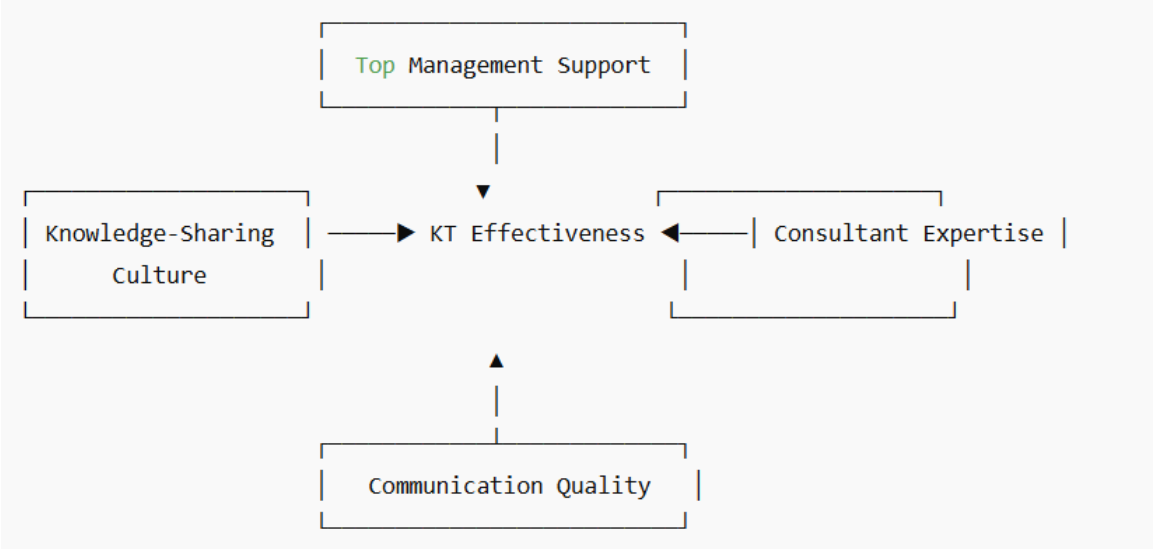


Fig. 2. Conceptual framework for knowledge transfer success in ESPs.

V. DISCUSSION

The findings of this study provide a nuanced understanding of the two research questions. Firstly, the RQ1 focus on how CSFs collectively influence KT in ESPs. While prior studies have acknowledged the importance of KT in such projects based on insights from [21], [22], [26], [30], and [32]. This research offers an empirically grounded conceptual framework integrating CSFs as standalone enablers and dynamically interdependent variables.

The empirical data significantly backs up the role of top management support, consultant expertise, communication quality and knowledge-sharing culture as essential facilitators for efficient KT. However, beyond confirming the significance of these individual CSFs, the interview data revealed their complex link. For instance, strong top management support positively affects communication quality and knowledge-sharing culture by defining allowed time for reflection, reinforcing shared goals, and embedding metrics on KT accountability and performance evaluation. Similarly, consultant expertise was most impactful when matched with trust-based relationships. These interrelationships highlight that CSFs should not be analysed as standalone components and emphasize the importance of holistic frameworks.

Secondly, the RQ2 was confirmed based on the expert feedback, and data also suggested a temporal dimension to CSF salience. During early project phases (e.g., planning and configuration), top management support and communication were most critical. In contrast, absorptive capacity and knowledge-sharing culture became more prominent during system testing and stabilisation. This temporal variability suggests that KT strategies must be adaptive, with CSF prioritization shifting based on project lifecycle stages.

VI. CONCLUSION

KT is a key process in the success of ESPs. While often overlooked in many organizational settings, this study explored KT practices within enterprise system planning (ESP) implementations and examined the key factors influencing their effectiveness. Following a literature review and semi-structured interviews, we identified several CSFs, including support from top management, consultants' expertise, communication quality, and a knowledge-sharing culture. These characteristics were repeatedly validated across research, demonstrating their interconnectedness and importance in affecting KT results.

This study addressed two primary research questions concerning the identification of KT CSFs (RQ1) and the development and validation of the proposed framework (RQ2).

First, it confirmed the fourth CSFs; the results indicated that organisations focusing on these CSFs could better internalise and apply transferred knowledge, especially when strategic leadership and collaborative environments were present.

Secondly, the study developed and validated a practical framework, revealing that KT effectiveness evolves across different stages of ESPs, thereby requiring adaptive and context-sensitive strategies. This directly addresses RQ2, with the proposed framework receiving positive validation from ESP experts. The resulting conceptual framework offers both theoretical guidance and practical applicability by promoting cultural alignment, systemic thinking, and responsive management practices as key enablers for enhancing KT across the project lifecycle.

Despite the valuable insights generated by this study, several limitations should be acknowledged. The sample size could be extended for more generalisation. In future, the framework can be tested across organizations of different sizes, sectors, and geographical locations to assess its broader applicability. Additionally, the study suggests that longitudinal and mixed-method approaches help capture the evolving nature of KT, particularly during post-implementation activities like system upgrades and user onboarding and offboarding, which need to be discussed. The role of modern technologies such as AI-based tools, chatbots, and learning platforms in supporting KT should also be investigated further.

ACKNOWLEDGMENT

I want to express my deepest gratitude to all those who contributed to completing this work.

REFERENCES

[1] S. Ammar, “Enterprise systems, business process management and UK-management accounting practices: Cross-sectional case studies,” Qualitative Research in Accounting \& Management, vol. 14, no. 3, pp. 230–281, 2017.

[2] S. Saeed, A. Shaikh, M. A. Memon, M. H. Memon, F. A. Abassi, and S. M. R. Naqvi, “Implementation of failure enterprise systems in organizational perspective framework,” International journal of advanced computer science and applications, vol. 8, no. 5, 2017.

[3] S. H. Chung, S. I. Ahmad, and H.-L. Tang, “Symptoms, causes and remedies for failures in enterprise systems implementation,” International Journal of Business Information Systems, vol. 19, no. 1, pp. 103–118, 2015.

[4] S. Goyette, L. Cassivi, M. Courchesne, and E. Elia, “The ERP post-implementation stage: a knowledge transfer challenge,” International Journal of Information Systems and Project Management, vol. 3, no. 2, pp. 5–19, 2015.

[5] F. E. Rotimi, C. Silva, C. E. D. Ramanayaka, and J. O. B. Rotimi, “Digital Technology Knowledge Transfer Enablers Amongst End-Users in Architecture, Engineering, and Construction Organisations: New Zealand Insights,” Buildings, vol. 15, no. 5, p. 772, 2025.

[6] V. W. B. Lopez, “Beyond knowledge integration barriers in ERP implementations: An institutional approach,” Journal of information and organizational sciences, vol. 42, no. 2, pp. 159–178, 2018.

[7] S. Paroutis and A. Al Saleh, “Determinants of knowledge sharing using Web 2.0 technologies,” Journal of Knowledge Management, vol. 13, no. 4, pp. 52–63, 2009, doi: 10.1108/13673270910971824.

[8] Jewels and Ford, “Factors Influencing Knowledge Sharing in Information Technology Projects,” e-Service Journal, vol. 5, no. 1, p. 99, 2006, doi: 10.2979/esj.2006.5.1.99.

[9] J. M. Hussien, M. Abdulhak, H. Kahtan, and R. Sulaiman, “Revisiting Knowledge Transfer for Success Enterprise System Project,” Jul. 2021.

[10] T. Jewels, “Factors Influencing Knowledge Sharing in Information Technology Projects,” vol. 5, no. 1, pp. 99–117, 2016.

[11] T. Beste, “Knowledge Transfer in a Project-Based Organization Through Microlearning on Cost-Efficiency,” The Journal of Applied Behavioral Science, vol. 59, no. 2, pp. 288–313, Jun. 2021, doi: 10.1177/00218863211033096.

[12] S. Nidhra and M. Yanamadala, “Knowledge Transfer Challenges and Mitigation Strategies in Global Software Development,” International Journal of Information Management, vol. 33, no. 2, pp. 333–355, 2012.

[13] A. Anand, B. Muskat, A. Creed, A. Zutshi, and A. Csepregi, “Knowledge sharing, knowledge transfer and SMEs: evolution, antecedents, outcomes

and directions," Personnel Review, vol. 50, no. 9, pp. 1873–1893, 2020, doi: 10.1108/PR-05-2020-0372.

[14] M. Aranyossy, B. Blaskovics, and Á. A. Horváth, "How universal are IT project success and failure factors? Evidence from Hungary," Information Systems Management, vol. 35, no. 1, pp. 15–28, 2018, doi: 10.1080/10580530.2017.1416943.

[15] H. F. Lin, "Knowledge sharing and firm innovation capability: An empirical study," International Journal of Manpower, vol. 28, no. 3–4, pp. 315–332, 2007, doi: 10.1108/01437720710755272.

[16] S. Hsseinoiun, R. Abdullah, Y. Y. Jusoh, and M. Jabar, "Information System Success and Knowledge Grid Integration in Facilitating Knowledge Sharing among Big Data Community," Proceedings - 2018 4th International Conference on Information Retrieval and Knowledge Management: Diving into Data Sciences, CAMP 2018, pp. 211–215, 2018, doi: 10.1109/INFRKM.2018.8464790.

[17] D. Žabjek, A. Kovačič, and M. I. Štemberger, "The influence of business process management and some other CSFs on successful ERP implementation," Business Process Management Journal, 2009, doi: 10.1108/14637150910975552.

[18] B. Kaba, "Identifying an analytical tool to assess the readiness of aid information and communication technology projects," no. April 2018, pp. 1–9, 2019, doi: 10.1002/isd2.12072.

[19] Y. Yang, G. Brosch, B. Yang, and T. Cadden, "Dissemination and communication of lessons learned for a project-based business with the application of information technology: a case study with Siemens," Production Planning and Control, vol. 31, no. 4, pp. 273–286, 2020, doi: 10.1080/09537287.2019.1630682.

[20] E. T. G. Wang, J. H. F. Chen, J. H. F. Chen, and J. H. F. Chen, "Effects of internal support and consultant quality on the consulting process and ERP system quality," null, 2006, doi: 10.1016/j.dss.2005.08.005.

[21] C. Favoretto and M. M. de Carvalho, "An analysis of the relationship between knowledge management and project performance: Literature review and conceptual framework," Gestao e Producao, vol. 28, no. 1, pp. 1–21, 2021, doi: 10.1590/0104-530X4888-20.

[22] A. M. Diane B. Walz, "The Effect of Conflict and Knowledge Sharing on the Information Technology Project Team Performance," Journal of Knowledge Management, vol. 12, no. 4, pp. 1–12, 2019, doi: 10.1108/13673270210450450.

[23] P. Hall, D. Ellis, and B. McArthur, "Knowledge Sharing Enablers in Small Business Networks," International Journal of Knowledge Management, vol. 18, no. 1, pp. 1–16, 2022, doi: 10.4018/ijkm.291705.

[24] E. E. Rens, V., Iwu, C. G., Tengeh, R. K., & Esambe, "Enterprise Resource Planning Systems and Knowledge Management: A Review of the Literature and Conceptual Framework," Journal of Management and Research (JMR)., vol. 8, no. 2, pp. 215–251, 2021, [Online]. Available: https://www.researchgate.net/profile/Muhammad-Khan-660/publication/359055726_How_Does_Ethical_Leadership_Curb_Employees'_Unethical_Behavior_Evidence_from_Oil_and_Gas_Sector_in_Pakistan/links/62259e219f7b32463415056d/How-Does-Ethical-Leadership-Curb-Employ.

[25] S. A. Salloum, M. Al-Emran, and K. Shaalan, The impact of knowledge sharing on information systems: A review, vol. 877. Springer International Publishing, 2018.

[26] B. F. Akinyode and T. H. Khan, "Step by step approach for qualitative data analysis," International Journal of Built Environment and Sustainability, vol. 5, no. 3, pp. 163–174, 2018, doi: 10.11113/ijbes.v5.n3.267.

[27] H. Kallio, A. M. Pietilä, M. Johnson, and M. Kangasniemi, "Systematic methodological review: developing a framework for a qualitative semi-structured interview guide," Journal of Advanced Nursing, vol. 72, no. 12, pp. 2954–2965, 2016, doi: 10.1111/jan.13031.

[28] F. M. Aldaihani, D. Economics, and B. Data, "Justification for Adopting Qualitative Research Method , Research," no. December 2021, pp. 0–11, 2022, doi: 10.37227/JIBM-2021-09-1494.

[29] B. Jokanović, N. Zivlak, A. Okanović, J. Ćulibrk, and L. Dudak, "The model of knowledge management based on organizational climate," Sustainability (Switzerland), vol. 12, no. 8, 2020, doi: 10.3390/SU12083273.

[30] A. Garcia-Perez, J. G. Cegarra-Navarro, and M. M. Jahantab, "Knowledge Sharing as a Driver of Competitive Advantage: Two Cases from the Field," Knowledge Management and Organizational Learning, vol. 6, pp. 145–167, 2018, doi: 10.1007/978-3-319-66890-1_8.

[31] D. G. Ko, L. J. Kirsch, and W. R. King, "Antecedents of knowledge transfer from consultants to clients in enterprise system implementations," MIS Quarterly: Management Information Systems, vol. 29, no. 1, pp. 59–85, 2005, doi: 10.2307/25148668.

[32] A. Subiyakto and A. R. Bin Ahlan, "A coherent framework for understanding critical success factors of ICT project environment," International Conference on Research and Innovation in Information Systems, ICRIIS, vol. 2013, pp. 342–347, 2013, doi: 10.1109/ICRIIS.2013.6716733.

[33] I. Cahyadi, "Factors Influencing Knowledge Transfer in ERP System Implementation within Indonesian Small and Medium Enterprises," Journal of Food Science, vol. 76, no. 8, p. 28, 2011, doi: 10.1111/j.1750-3841.2011.02348.x.

[34] J. Hussien, M. A. Abdulgabber, H. Kahtan, and R. Sulaiman, "THE Success of Enterprise System Projects: From A Knowledge Transfer Perspective," Journal of Information System and Technology Management (JISTM), vol. 6, no. 22, pp. 128–147, 2021, doi: 10.35631/JISTM.622011.

Content Validity Assessment Using Aiken's V: Knowledge Integration Model for Blockchain in Higher Learning Institutions

Nur Ilyana Ismarau Tajuddin[1], Ummu-Hani Abas[2], Khairi Azhar Aziz[3], Rozi Nor Haizan Nor[4], Nor Aziyatul Izni[5], Muhammad Nuruddin Sudin[6], Nur Aqilah Hazirah Mohd Anim[7], Noorashikin Md Noor[8]

sPusat Tamhidi, Universiti Sains Islam Malaysia, Nilai, 71800, N.Sembilan, Malaysia[1, 2, 6]

IT Enterprise Architecture, IT Strategy, Bank Kerjasama Rakyat Malaysia, Malaysia[3]

Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, 43400, Selangor, Malaysia[4]

Centre of Foundation Studies, Universiti Teknologi MARA, Cawangan Selangor, Kampus Dengkil, Dengkil, 43800, Selangor, Malaysia[5]

Faculty of Economics and Muamalat, Universiti Sains Islam Malaysia, Nilai, 71800, N. Sembilan, Malaysia[7]

Institute of Climate Change, Universiti Kebangsaan Malaysia, Bangi, 43600, Selangor, Malaysia[8]

***Abstract*—The integration of blockchain technology into higher learning institutions (HLIs) holds the potential to revolutionize data management, enhance transparency, and improve trust in academic systems. However, the effective adoption of blockchain requires a comprehensive and valid model that addresses the specific needs and contexts of HLIs. This study aims to assess the content validity of the Knowledge Integration Model for Blockchain in Higher Learning Institutions using Aiken's V method. The proposed model was developed through a systematic literature review and refined with expert input. Content validity was evaluated by seven experts with backgrounds in education, blockchain technology, and information systems. Each item within the model was rated for its relevance, clarity, and representativeness using a Likert scale. The instrument, consisting of 50 items across six constructs, was evaluated by seven domain experts using Aiken's V methodology. Each item was rated based on its relevance and clarity using a 5-point Likert scale. The results revealed that out of 50 items, 21 required revision or removal due to low Aiken's V scores (<0.70), 21 were deemed acceptable but required minor revisions, and 8 demonstrated strong content validity (V ≥ 0.80). These findings underscore the importance of expert evaluation in refining research instruments and ensuring construct alignment. The use of Aiken's V provided a robust quantitative foundation for the validation process. The refined instrument serves as a reliable tool for assessing institutional readiness and knowledge integration capabilities in the context of blockchain adoption. This work contributes to the growing research on educational blockchain implementation by offering a validated framework that can support empirical investigations and strategic decision-making in higher education.**

***Keywords*—Blockchain; content validity; aiken's v; higher learning institutions; knowledge integration model**

I. Introduction

Knowledge integration—the process of combining insights, expertise, and perspectives from multiple domains—is essential for ensuring that blockchain solutions are contextually relevant, sustainable, and scalable within the educational ecosystem. Although scholarly interest in blockchain applications in education is increasing, there remains a lack of research on how knowledge from technical, pedagogical, administrative, and regulatory domains can be effectively synthesized to support implementation. The integration of blockchain technology into higher learning institutions (HLIs) presents a transformative opportunity to enhance data integrity, security, and transparency in academic processes. However, a significant research challenge persists: the limited understanding of how to integrate knowledge across these domains to enable effective blockchain adoption. This study seeks to validate an instrument designed to assess the Knowledge Integration Model for Blockchain in Higher Education, thereby addressing the gap in comprehensive adoption frameworks.

The research is guided by two key questions: 1) To what extent does the proposed model demonstrate content validity?, and 2) What revisions are necessary to improve its constructs and items? The primary objective is to ensure that tools used to measure institutional readiness and integration capabilities are both valid and reliable. This study is significant as it offers a robust measurement instrument for empirical research and practical implementation of blockchain in higher education, focusing on the integration of domain-specific knowledge. Key contributions include a validated instrument, empirical evidence of content validity using Aiken's V, and a refined framework to support strategic decision-making. As blockchain technology gains momentum in the academic sector, it is crucial to ensure that assessment tools for institutional readiness, perceptions, and integration capabilities are methodologically sound.

This study evaluates the content validity, construct validity, and internal consistency of the proposed instrument through expert judgment and statistical analysis. By validating this tool, the research aims to provide a reliable foundation for future empirical studies and practical applications of blockchain technologies in higher education. The validated instrument will also serve as a critical component in examining how knowledge from technological, pedagogical, administrative, and regulatory domains can be effectively integrated to support successful blockchain adoption.

The remainder of this study outlines the literature review, methodology, presents the results, discusses necessary revisions, and concludes with implications for future research and practice.

II. Literature Review

The integration of blockchain technology into higher education institutions presents a transformative opportunity to enhance data security, streamline administrative processes, and foster a more transparent and trustworthy learning environment [1]. Blockchain, initially conceived as the foundational technology for cryptocurrencies, has transcended its original purpose and is now recognized for its potential to revolutionize various sectors, including education [2]. Its decentralized and immutable nature ensures that records are tamper-proof and verifiable, addressing critical concerns about data integrity and security that are prevalent in traditional educational systems [3]. By leveraging blockchain's inherent security features, educational institutions can protect sensitive student data, such as academic transcripts, personal information, and financial records, from unauthorized access and cyber threats [4]. This is particularly important in an era of increasing data breaches and identity theft, where the security of personal information is paramount [5].

Furthermore, blockchain technology can facilitate the creation of a more transparent and accountable educational ecosystem [6]. In traditional systems, academic records are often stored in centralized databases, which can be vulnerable to manipulation or loss. Blockchain, on the other hand, distributes data across a network of computers, making it virtually impossible for a single point of failure or malicious actor to compromise the integrity of the information [7]. This decentralized approach not only enhances security but also promotes trust among students, faculty, and administrators. Many higher education institutions have already started to implement blockchain technology for academic degree and result management [8]. Moreover, blockchain can streamline administrative processes, reduce costs, and improve efficiency in higher education institutions. By automating tasks such as transcript verification, degree authentication, and tuition payment processing, blockchain can free up valuable time and resources for educators and administrators, allowing them to focus on more strategic initiatives such as improving teaching and learning outcomes [9].

The immutability of blockchain ensures the integrity of educational records, preventing fraud and enhancing trust in qualifications [10] [11]. The characteristics of blockchain technology can address these issues within educational settings [12]. Blockchain is essentially a distributed database that boasts decentralization, security, and transparency [13] [14]. This transparency ensures that all stakeholders have access to accurate and up-to-date information, fostering a more collaborative and efficient learning environment. For example, consider the issue of verifying academic credentials. Blockchain technology provides a transparent and autonomous solution and enhances the quality of shared data among stakeholders [15]. Using blockchain, institutions can issue digital credentials that are easily verifiable, eliminating the need for time-consuming manual verification processes and reducing the risk of fraudulent claims.

Blockchain's decentralized framework offers a platform for validating data and transactions, independent of third-party control, in a secure, transparent, and permanent setup [16]. Smart contracts, self-executing agreements written into the blockchain code, can automate processes such as verifying academic credentials and managing tuition payments [17]. Smart contracts automate procedures, minimizing errors caused by people or counterfeits [18]. The employment of smart contracts streamlines auditing procedures, resulting in a 30% reduction in audit time and a 20% decrease in audit costs [19]. The use of blockchain technology could reduce project disputes arising from delayed payments [20].

The decentralized database created by blockchain technology offers enhanced security, faster payments, and improved data transfer reliability and transparency [21]. The application of blockchain extends beyond cryptocurrency, finding utility in data management, supply chain tracking, and healthcare, among other fields [22][23]. In healthcare, blockchain improves interoperability, data integrity, and security while cutting maintenance expenses [24]. As blockchain technology gains traction across various sectors, including digital currency, finance, and business [25] [26], its entrance into the healthcare industry signals a move towards more secure and efficient data management practices. Several firms have started experimenting with blockchain technology for a variety of purposes, including supply chain management, identity verification, and decentralized marketplaces [27][28]. Blockchain has many potential applications, including those in the healthcare industry [27] [29] [30]. More exploration is needed to determine the most effective way to design a blockchain-based system.

III. Methodology

This study adopts a quantitative approach involving a comprehensive content validity assessment process. The content validity evaluation was carried out through four structured and systematic phases. These phases were designed to ensure the integrity and accuracy of the research instrument, as well as to meet rigorous methodological standards in educational research.

The four phases are as follows:

- Construction development
- Expert selection
- Implementation of content validity assessment
- Revision of constructs and items

A. Construction Development

This stage involves the conceptualization and operationalization of the main constructs based on an extensive literature review. The study employs a questionnaire as the research instrument, using a 5-point Likert scale. The scale is as follows:

1 = Very inappropriate

2 = Inappropriate

3 = Not sure

4 = Appropriate

5 = Very appropriate

The instrument comprises six distinct constructs:

- Technology Factors (11 items)
- Organizational Factors (12 items)
- Environment Factors (10 items)
- Knowledge Integration (6 items)
- Perceived Usefulness (6 items)
- Blockchain Adoption Likelihood (5 items)

Total items: 50

In total, the questionnaire comprises 50 items. These constructs were developed through in-depth analysis of previous studies, content analysis, and a thorough literature review across the six constructs as presented in Table I. The measurement items for each construct were systematically developed based on an extensive review of relevant literature. Each item was carefully formulated to ensure alignment with the theoretical foundations and operational definitions established during the construct development phase.

TABLE I. ITEMS DEVELOPED FOR EACH CONSTRUCT FORMULATED FROM THE LITERATURE REVIEW

Construct	Item Code	Items
Technology - Security & Privacy	B1	Blockchain ensures unauthorized parties cannot access academic records.
	B2	Blockchain technology provides stronger data privacy than current systems.
	B3	I trust blockchain to securely store my institution's credentials.
	B4	Blockchain reduces risks of data breaches in education.
Technology - Transparency & Trust	B5	Blockchain makes academic transactions (e.g., grading) more transparent.
	B6	I can verify the authenticity of blockchain-stored records easily.
	B7	Blockchain increases trust in shared knowledge (e.g., research data).
	B8	Blockchain's transparency reduces conflicts over academic records.
Technology - Data Integrity	B9	Blockchain prevents tampering with academic records.
	B10	My institution's use of blockchain ensures accurate record-keeping.
	B11	Historical changes to records are traceable via blockchain.
Organizational - Institutional Readiness	C1	My institution has the technical infrastructure to adopt blockchain.
	C2	Budget is allocated for blockchain implementation.
	C3	Staff are trained to use blockchain-based systems.
	C4	My institution has a clear roadmap for blockchain integration.
Organizational - Leadership Support	C5	Administrators actively promote blockchain initiatives.
	C6	Leadership addresses concerns about blockchain adoption.
	C7	Faculty are encouraged to pilot blockchain tools.
	C8	Leadership prioritizes blockchain adoption in strategic plans.
Organizational - Financial & Technical Resources	C9	My institution can afford blockchain implementation costs.
	C10	IT teams are skilled in maintaining blockchain systems.
	C11	Blockchain tools are compatible with existing IT systems.
	C12	External funding (e.g., grants) supports blockchain projects.
Environment - Government Policies & Regulation	D1	National policies encourage blockchain use in education.
	D2	Legal frameworks support blockchain-based academic credentials.
	D3	Government funding is available for blockchain initiatives.
	D4	My institution complies with regulations for blockchain adoption.
Environment - Industry Collaboration	D5	My institution partners with tech firms for blockchain solutions.
	D6	Joint research projects use blockchain for knowledge sharing.
	D7	Employers recognize blockchain-verified credentials.
Environment - Societal Expectation	D8	Employers prefer candidates with blockchain-stored degrees.
	D9	Public awareness of blockchain benefits in education is growing.
	D10	Society pressures institutions to adopt transparent technologies.
Knowledge Integration (KI)	E1	Blockchain supports interdisciplinary knowledge development.
	E2	Access to shared knowledge is equitable across departments/institutions.
	E3	Archived knowledge is easily retrievable via blockchain.
	E4	Research findings are implemented faster via blockchain.
	E5	Blockchain platforms enable real-time sharing of course materials across departments.
	E6	My institution uses blockchain to track the impact of knowledge on community projects.
Perceived Usefulness	F1	Blockchain saves time in credential verification.
	F2	Using blockchain enhances my productivity as a staff/educator/student.
	F3	Blockchain simplifies administrative processes (e.g., transcript requests).
	F4	Blockchain adds value to my institution's knowledge management.
	F5	Automated smart contracts (e.g., for research royalties) improve workflow efficiency.
	F6	Blockchain increases transparency in research data sharing and citation tracking.
Blockchain Adoption Likelihood	G1	My institution plans to expand blockchain use beyond credentials (e.g., research, administration).
	G2	My institution currently uses blockchain for any academic processes (e.g., credentials, research).
	G3	Blockchain offers significant advantages over our current record-keeping systems.
	G4	Blockchain adoption has improved our institution's operational efficiency.
	G5	Blockchain supports my institution's alignment with national digital education goals (SDGs).

B. Expert Selection

To ensure the content validity of the developed model or instrument, an expert validation process was carried out. Seven (7) experts were purposely selected based on their academic qualifications, domain relevance, and professional experience exceeding five (5) years. As detailed in Table II, five (5) are academicians from local public universities, while the remaining two (2) are industry professionals with relevant expertise. All experts have active involvement in teaching, research, or system development related to the model's domain. They voluntarily evaluated the instrument via an online form, providing both quantitative ratings and qualitative feedback. This input was subsequently used to calculate the content validity indices using Aiken's V.

TABLE II. EXPERT PANEL INFORMATION FOR CONTENT VALIDATION

No.	Expert Initials	Background	Position/Role	Expertise Area	Experience	Institution/Sector
1	Expert 1	Industry	Solution Architect	Information Systems, IT Enterprise Architecture	15 Years	Bank Rakyat
2	Expert 2	Academic	Associate Professor	Blockchain System	18 years	Asia Pacific University of Technology and Innovation (APU)
3	Expert 3	Industry	Senior Manager	Big Data Analytics, Blockchain, Information Systems	18 years	Kementerian Digital Malaysia
4	Expert 4	Academic	Associate Professor	Information Systems	20 years	Universiti Putra Malaysia (UPM)
5	Expert 5	Academic	Senior Lecturer	Industrial Computing	17 years	Universiti Sains Islam Malaysia (USIM)
6	Expert 6	Academic	Senior Lecturer	Knowledge Management, MIS, Project Management	23 years	Universiti Putra Malaysia (UPM)
7	Expert 7	Academic	Senior Lecturer	Cybersecurity	8 years	Universiti Sains Islam Malaysia (USIM)

C. Implementation of Content Validity Assessment

To ensure the instrument's content validity, this study employed a rigorous quantitative validation process combining the Content Validity Index (CVI) and Aiken's V coefficient. A panel of 7 subject matter experts evaluated each item's relevance using a 5-point Likert scale, ranging from 1 (not relevant) to 5 (highly relevant). The content validity was quantitatively assessed through two complementary approaches. We statistically validated expert consensus using Aiken's V coefficient, which quantifies agreement beyond simple percentages by accounting for the rating scale's ordinal nature. The V values were calculated for each item and tested for statistical significance ($p < 0.05$), with items demonstrating $V \geq 0.70$ being retained as valid. This dual-method approach provided robust quantitative evidence of content validity through statistically validated expert judgments, ensuring the instrument's items appropriately represent the target construct while minimizing subjective bias in the validation process. To ensure that the items developed in the instrument are relevant and representative of the construct being measured, content validity was assessed using Aiken's V index. This method quantitatively evaluates the degree of agreement among subject matter experts regarding the relevance of each item. Aiken's V is appropriate for assessing content validity when experts rate items on an ordinal scale (e.g., 1 = Not Relevant to 5 = Highly Relevant). The formula used to calculate Aiken's V is as follows:

Aiken's V is calculated using the formula:

$$V = \frac{\sum (s)}{n(c-1)}$$

where,

$s = r - l$

r = rating given by expert, l = lowest possible rating

n = number of experts

c = number of possible rating categories (e.g. if using a 1–5 scale, c = 5)

The resulting V value ranges from 0 to 1, where:

- $V \geq 0.80$ indicates high content validity,
- V between 0.70 and 0.79 suggests moderate content validity, and
- $V < 0.70$ may imply poor content validity, requiring revision or removal of the item.

D. Revision of Constructs and Items

Based on expert feedback, this phase involved an iterative process of refining and enhancing the constructs and items in the research instrument. During the content validity assessment, the panel of experts was asked to conduct a critical evaluation of the overall constructs and individual items included in the instrument. These experts were encouraged to provide quantitative ratings for each item, as well as qualitative feedback in the form of written comments in the space provided.

The primary aim of this process was to improve the clarity and comprehensibility of the items. The researcher paid close attention to all input received from the panel of experts. This final phase of the content validity assessment played a crucial role in ensuring that the developed instrument accurately measured the intended constructs, in line with the established research objectives. Through this process, the researcher aimed to achieve a high level of content validity, thereby enhancing the overall integrity of the research instrument. The comments and suggestions from experienced experts were used to further improve the instrument [31].

Following the content validity assessment using the Content Validity Index (CVI) and Aiken's V, necessary revisions were

made to the constructs and items to improve clarity, relevance, and overall validity of the instrument.

- Items that did not meet the minimum acceptable threshold—I-CVI ≥ 0.78 for individual items, and Aiken's V ≥ 0.70—were reviewed carefully. Based on expert feedback and statistical indicators
- Items with low validity scores were either revised to improve clarity and appropriateness or removed if they were deemed redundant or irrelevant.
- Wording adjustments were made to enhance the precision and readability of the items, ensuring they accurately reflect the intended construct.

In some cases, new items were added to better represent key dimensions of the construct, as recommended by the panel of experts. This revision process ensured that all items retained in the final version of the instrument demonstrated strong content validity and alignment with the conceptual framework of the study. The refined instrument is thus considered suitable for subsequent pilot testing and data collection.

IV. Findings and Discussion

The Aiken's V analysis was conducted on a 50-item instrument assessed by seven experts. Each expert provided their rating on a scale (presumably from 1 to 5), reflecting the perceived relevance of each item to the construct being measured, as evidenced by Table III. The calculated Aiken's V values offer a statistical indication of agreement among experts and are crucial in evaluating the overall quality and clarity of each item. The primary objective of this analysis was to determine the degree of consensus among experts regarding the relevance and clarity of each item, thus guiding decisions on item retention, revision, or elimination. Aiken's V coefficient, which ranges from 0 to 1, offers a quantifiable measure of expert agreement, with higher values indicating stronger content validity.

TABLE III. Results of Content Validity Calculation Analysis with the Aiken Method

EXPERTS/ ITEMS	EX 1	EX 2	EX 3	EX 4	EX 5	EX 6	EX 7	Aiken's V
B1	4	4	4	3	4	4	2	**0.64**
B2	4	4	4	3	4	4	3	**0.68**
B3	4	4	4	3	4	4	3	**0.68**
B4	4	4	4	3	4	4	3	**0.68**
B5	4	4	4	3	4	4	5	0.75
B6	4	4	4	3	4	4	5	0.75
B7	4	4	4	3	4	4	5	0.75
B8	4	4	4	3	4	4	5	0.75
B9	4	4	4	3	4	4	5	0.75
B10	2	4	4	3	4	4	5	**0.68**
B11	3	4	4	3	4	4	5	0.71
C1	4	4	4	4	4	3	4	0.71
C2	4	4	4	4	4	3	3	**0.68**
C3	4	4	4	4	4	2	2	**0.61**
C4	4	4	4	4	4	3	2	**0.64**
C5	5	4	4	4	4	2	3	**0.68**
C6	5	4	4	4	4	2	3	**0.68**
C7	3	4	4	4	4	3	3	**0.64**
C8	4	4	4	4	4	3	3	**0.68**
C9	4	4	4	4	4	2	3	**0.64**
C10	3	4	4	4	4	3	4	**0.68**
C11	3	4	4	4	4	3	4	**0.68**
C12	5	4	4	4	4	4	4	0.79
D1	5	4	4	4	4	4	5	0.82
D2	4	4	4	4	4	4	5	0.79
D3	5	4	4	4	4	5	4	0.82
D4	4	4	4	4	4	4	3	0.71
D5	4	4	4	4	4	4	4	0.75
D6	4	4	4	4	4	3	5	0.75
D7	4	4	4	4	4	3	4	0.71
D8	2	4	4	4	4	3	5	0.68
D9	4	4	4	4	4	4	4	0.75
D10	3	4	4	4	4	4	4	0.71
E1	4	2	4	4	4	5	5	0.75
E2	4	2	4	4	4	4	4	**0.68**
E3	4	2	4	4	4	4	4	**0.68**
E4	3	2	4	4	4	3	4	**0.61**
E5	4	2	4	4	4	4	4	**0.68**
E6	3	2	4	4	4	3	4	**0.61**
F1	4	2	4	5	5	4	4	0.75
F2	4	2	4	5	5	4	4	0.75
F3	3	2	4	5	5	4	5	0.75
F4	4	2	4	5	5	4	5	0.79
F5	4	2	4	5	5	4	5	0.79
F6	4	2	4	5	5	5	5	0.82
G1	4	5	4	5	5	3	5	0.86
G2	4	5	4	5	5	3	5	0.86
G3	4	5	4	5	5	4	4	0.86
G4	4	5	4	5	5	3	5	0.86
G5	4	5	4	5	5	4	5	0.89

Aiken's V Analysis Report evaluates 50 items for content validity as depicted in Table IV, with results categorized into three groups based on expert ratings: 1) strong content validity, 2) acceptable but may need minor revision, and 3) needs revision or removal. Out of the total, 21 items were found to fall into the "needs revision or removal" category, having Aiken's V scores significantly below the ideal threshold of 0.75. These scores indicate weaker expert agreement on the relevance or clarity of the items, suggesting that they may either be misaligned with the intended construct or ambiguous in wording. Another 21 items

achieved scores near the threshold and were deemed "acceptable", while 8 items demonstrated strong content validity, reflecting high agreement among the evaluators.

TABLE IV. INTERPRETATION OF ANALYSIS

Item	Aiken's V	Interpretation
B1	0.64	**Needs revision or removal**
B2	0.68	**Needs revision or removal**
B3	0.68	**Needs revision or removal**
B4	0.68	**Needs revision or removal**
B5	0.75	Acceptable, may need minor revision
B6	0.75	Acceptable, may need minor revision
B7	0.75	Acceptable, may need minor revision
B8	0.75	Acceptable, may need minor revision
B9	0.75	Acceptable, may need minor revision
B10	0.68	**Needs revision or removal**
B11	0.71	Acceptable, may need minor revision
C1	0.71	Acceptable, may need minor revision
C2	0.68	**Needs revision or removal**
C3	0.61	**Needs revision or removal**
C4	0.64	**Needs revision or removal**
C5	0.68	**Needs revision or removal**
C6	0.68	**Needs revision or removal**
C7	0.64	**Needs revision or removal**
C8	0.68	**Needs revision or removal**
C9	0.64	**Needs revision or removal**
C10	0.68	**Needs revision or removal**
C11	0.68	**Needs revision or removal**
C12	0.79	Acceptable, may need minor revision
D1	0.82	Strong content validity
D2	0.79	Acceptable, may need minor revision
D3	0.82	Strong content validity
D4	0.71	Acceptable, may need minor revision
D5	0.75	Acceptable, may need minor revision
D6	0.75	Acceptable, may need minor revision
D7	0.71	Acceptable, may need minor revision
D8	0.68	**Needs revision or removal**
D9	0.75	Acceptable, may need minor revision
D10	0.71	Acceptable, may need minor revision
E1	0.75	Acceptable, may need minor revision
E2	0.68	**Needs revision or removal**
E3	0.68	**Needs revision or removal**
E4	0.61	**Needs revision or removal**
E5	0.68	**Needs revision or removal**
E6	0.61	**Needs revision or removal**
F1	0.75	Acceptable, may need minor revision
F2	0.75	Acceptable, may need minor revision
F3	0.75	Acceptable, may need minor revision
F4	0.79	Acceptable, may need minor revision
F5	0.79	Acceptable, may need minor revision
F6	0.82	Strong content validity
G1	0.86	Strong content validity
G2	0.86	Strong content validity
G3	0.86	Strong content validity
G4	0.86	Strong content validity
G5	0.89	Strong content validity

The following items specifically require attention due to their low Aiken's V values:

B1 (0.643), B2 (0.679), B3 (0.679), B4 (0.679), B10 (0.679), C2 (0.679), C3 (0.607), C4 (0.643), C5 (0.679), C6 (0.679), C7 (0.643), C8 (0.679), C9 (0.643), C10 (0.679), C11 (0.679), D8 (0.679), E2 (0.679), E3 (0.679), E4 (0.607), E5 (0.679), and E6 (0.607). These items span across several constructs and indicate widespread issues in content design. It is recommended that each of these items be reviewed for conceptual clarity, language precision, and alignment with the overall measurement objectives.

Analysis of Aiken's V scores revealed distinct patterns across constructs, warranting careful consideration. A significant observation was the relatively low validity scores attained by items within the Organizational Factors construct (specifically, items C3 through C9). These items, designed to gauge institutional readiness and leadership support for blockchain integration, consistently fell below the established validity threshold of 0.70. This suggests a potential lack of clarity or precision in the operational definitions of these constructs, potentially leading to ambiguity in respondent interpretation and assessment.

To enhance the interpretation of the quantitative findings, it is important to consider the qualitative feedback provided by the expert panel. For instance, item B1 ("Blockchain ensures unauthorized parties cannot access academic records"), which received a validity score of 0.64, was critiqued by one expert for being overly simplistic and not adequately reflecting the complexities of blockchain security. Similarly, item E4 ("Research findings are implemented faster via blockchain"), with a score of 0.61, was considered irrelevant by another expert who emphasized that the speed of research implementation is influenced by various factors beyond the technology itself. These examples highlight the value of expert evaluation in identifying items that may appear relevant but lack depth or fail to align with real-world applications and theoretical constructs.

A comparison between the current findings and existing literature on technology adoption in higher education reveals a notable discrepancy. Specifically, items related to the Technology and Environment constructs (D1 to D3) exhibited relatively higher content validity scores, whereas items under the Knowledge Integration construct (E2 to E6) showed lower ratings. One plausible explanation for this divergence is that technical and environmental aspects of blockchain adoption are

more tangible and thus easier for experts to assess. In contrast, knowledge integration involves abstract, multidimensional considerations, making it inherently more challenging to evaluate. Additionally, potential misalignment between departmental objectives may further complicate the formulation and assessment of relevant items.

To strengthen the theoretical grounding of the revised items, the Technology-Organization-Environment (TOE) framework is employed. This framework posits that technology adoption is shaped by the interaction of technological, organizational, and environmental contexts. In this study, the original item phrasing within the Knowledge Integration dimension may not have fully captured the intricate relationships among these factors. For example, item E2 originally read, "Access to shared knowledge is equitable across departments or institutions". It has been revised to: "Blockchain-based knowledge-sharing platforms should prioritize equitable access and usage across all departments and institutions". This revision aims to better reflect the Organizational dimension of the TOE framework by explicitly addressing the importance of strategic planning and equitable resource distribution in fostering inclusive and effective knowledge integration efforts.

This revision process is critical to ensure that the instrument provides reliable and valid results for future applications. Table V summarizes of interpretation of each item.

TABLE V. SUMMARY OF INTERPRETATION

Interpretation	Number of Items
Needs revision or removal	21
Acceptable, may need minor revision	21
Strong content validity	8

When an item receives a low Aiken's V score, typically below the threshold of 0.7, it indicates that expert raters do not sufficiently agree on the item's relevance, clarity, or appropriateness for measuring the intended construct. This serves as a signal that the item may be flawed or misaligned and, therefore, requires careful attention. The first step in addressing a low score is to thoroughly review any qualitative feedback provided by the experts. Such feedback can help identify specific concerns such as ambiguous wording, double-barreled questions, or a lack of alignment with the overall theoretical framework. If feedback is not available, consider conducting a follow-up discussion with the panel to gain deeper insights into the causes of disagreement.

Following this analysis, the item should be revised with a focus on improving clarity, refining language, and ensuring conceptual alignment. The goal is to eliminate any ambiguity and enhance the item's relevance and interpretability. Once revisions are made, the updated item should be resubmitted for expert evaluation in a second round of content validation to determine whether the changes have resulted in an improved Aiken's V score. If the item continues to score below acceptable levels even after revision, it is advisable to consider removing it from the instrument. Retaining low-quality items can compromise the validity and reliability of the entire tool.

Throughout this process, all changes—whether revisions or deletions—should be systematically documented, including justifications and version history. This not only enhances transparency but also reinforces the instrument's methodological rigor when reported in academic or professional settings.

V. CONCLUSION

In conclusion, low Aiken's V scores highlight critical areas in need of refinement within a measurement instrument and should not be overlooked. Addressing these scores through systematic revision, expert consultation, and re-evaluation is essential to enhance the content validity and overall quality of the instrument. Whether through improving clarity, ensuring alignment with constructs, or removing consistently weak items, these actions contribute to developing a more robust and reliable tool. By documenting the revision process and maintaining methodological transparency, researchers can strengthen the credibility and academic integrity of their validation efforts.

For future research endeavors, it is strongly recommended to conduct a second, more focused round of expert evaluation following the revision of items exhibiting low Aiken's V scores. This subsequent evaluation is crucial for assessing whether the implemented modifications have demonstrably improved the content validity of the affected items and, by extension, the instrument as a whole. Furthermore, the strategic inclusion of qualitative feedback, gathered both during the initial and subsequent expert evaluations, can provide invaluable, in-depth insights into the item-specific issues driving the low scores. These qualitative narratives can serve as a compass, guiding researchers toward more precisely targeted revisions and enabling them to address underlying problems that quantitative metrics alone may not fully reveal.

To further enhance the robustness and generalization of the validation process, it is advisable to expand the panel of experts to encompass a wider range of backgrounds, perspectives, and areas of expertise relevant to the constructs being measured. This diversification can help to mitigate potential biases stemming from a homogeneous expert group and ensure that the instrument is valid across diverse populations and contexts. Finally, it is highly recommended to pilot test the revised instrument with a representative sample drawn from the target population for which the instrument is ultimately intended. This pilot testing phase will provide critical data regarding the clarity, reliability, practical applicability, and overall user-friendliness of the instrument. By incorporating both quantitative and qualitative analyses of the pilot test data, researchers can ensure a more comprehensive assessment of the instrument's validity and identify any remaining areas requiring further refinement before embarking on full-scale implementation and data collection. These rigorous validation steps are essential for maximizing the utility and impact of the measurement instrument in future research and practice.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Universiti Sains Islam Malaysia (USIM) for the financial support provided through the research grant PPPI/USIM/TAMHIDI/USIM/112323.

REFERENCES

[1] P. Panagiotidis, "Blockchain in education - The case of language learning," Eur. J. Educ., vol. 5(1), p. 66, 2022, DOI: 10.26417/443gjm83.

[2] S. A. Shuvo and M. A. Islam, "Blockchain technology: A tool to solve the challenges of education sector in developing countries," Int. J. Comput. Syst. Eng., vol. 7(3), p. 1, 2023, DOI: 10.1504/ijcsyse.2023.10055995.

[3] S. Tahora, B. Saha, N. Sakib, H. Shahriar, and H. M. Haddad, "Blockchain technology in higher education ecosystem: Unraveling the good, bad, and ugly," arXiv (Cornell University), 2023, DOI: 10.48550/arxiv.2306.04071.

[4] A. M. E. Koshiry, E. H. I. Eliwa, T. A. El-Hafeez, and M. Y. Shams, "Unlocking the power of blockchain in education: An overview of innovations and outcomes," Blockchain Res. Appl., vol. 4(4), p. 100165, 2023, DOI: 10.1016/j.bcra.2023.100165.

[5] Y. Tsai, J. Huang, and D. Chiou, "Empowering young learners to explore blockchain with user-friendly tools: a method using Google Blockly and NFTs," IET Blockchain, 2023, DOI: 10.1049/blc2.12055.

[6] A. Grech and A. F. Camilleri, "Blockchain in education," RePEc: Research Papers in Economics, 2017, DOI: 10.2760/60649.

[7] A. Capetillo, D. Camacho, and M. Alanis, "Blockchained education: challenging the long-standing model of academic institutions," Int. J. Interact. Des. Manuf. (IJIDeM), vol. 16(2), p. 791, 2022, DOI: 10.1007/s12008-022-00886-1.

[8] N. Ullah, W. M. Al-Rahmi, A. I. Alzahrani, O. Alfarraj, and F. Alblehai, "Blockchain technology adoption in smart learning environments," Sustainability, vol. 13(4), p. 1801, 2021, DOI: 10.3390/su13041801.

[9] A. H. Taylor and R. Parks, "Blockchain: A future technology for registrars?," The Successful Registrar, vol. 20(5), p. 1, 2020, DOI: 10.1002/tsr.30728.

[10] E. Guustaaf, U. Rahardja, Q. Aini, N. A. Santoso, and N. P. L. Santoso, "Desain kerangka blockchain terhadap pendidikan: A survey," CESS (J. Comput. Eng. Syst. Sci.), vol. 6(2), p. 236, 2021, DOI: 10.24114/cess.v6i2.25099.

[11] B. Ram and P. Verma, "Application of blockchain technology in data security," IP Indian J. Libr. Sci. Inf. Technol., vol. 9(1), p. 51, 2024, DOI: 10.18231/j.ijlsit.2024.008.

[12] B. Awaji, E. Solaiman, and L. Marshall, "Investigating the requirements for building a blockchain-based achievement record system," p. 56, 2020, DOI: 10.1145/3411681.3411691.

[13] E. Puica, "IT solutions in supply chain management and their efficiency," J. Softw. Syst. Dev., vol. 1, 2022, DOI: 10.5171/2022.785575.

[14] Z. Sun, D. Han, D. Li, X. Wang, C. Chang, and Z. Wu, "A blockchain-based secure storage scheme for medical information," arXiv (Cornell University), 2022, DOI: 10.48550/arxiv.2207.06102.

[15] S. Bhattacharya, A. Singh, and M. M. Hossain, "Strengthening public health surveillance through blockchain technology [Review of strengthening public health surveillance through blockchain technology]," AIMS Public Health, vol. 6(3), p. 326, 2019, DOI: 10.3934/publichealth.2019.3.326.

[16] T. I. Akaba, A. Norta, C. Udokwu, and D. Draheim, "A framework for the adoption of blockchain-based e-procurement systems in the public sector," Lecture Notes in Computer Science, p. 3, 2020, DOI: 10.1007/978-3-030-44999-5_1.

[17] M. C. Benton and N. Radziwill, "Quality and innovation with blockchain technology," arXiv (Cornell University), 2017, DOI: 10.48550/arxiv.1710.04130.

[18] J. Karamachoski, N. Marina, and P. Taskov, "Blockchain-based application for certification management," Tehnički Glasnik, vol. 14(4), p. 488, 2020, DOI: 10.31803/tg-20200811113729.

[19] H. N. Fahdil, H. M. Hassan, A. Subhe, and A. T. Hawas, "Blockchain technology in accounting transforming financial reporting and auditing," J. Ecohumanism, vol. 3(5), p. 216, 2024, DOI: 10.62754/joe.v3i5.3903.

[20] M. Kamel, E. S. Bakhoum, and M. Marzouk, "A framework for smart construction contracts using BIM and blockchain," Sci. Rep., vol. 13(1), 2023, DOI: 10.1038/s41598-023-37353-0.

[21] M. J. Lahkani, S. Wang, M. Urbański, and M. Egorova, "Sustainable B2B e-commerce and blockchain-based supply chain finance," Sustainability, vol. 12(10), p. 3968, 2020, DOI: 10.3390/su12103968.

[22] L. Jiang, "The use of blockchain technology in enterprise financial accounting information sharing," PLoS ONE, vol. 19(2), 2024, DOI: 10.1371/journal.pone.0298210.

[23] Y. Liu, K. Qian, J. Yu, K. Wang, and L. He, "Effective scaling of blockchain beyond consensus innovations and Moore's law," arXiv (Cornell University), 2020, DOI: 10.48550/arxiv.2001.01865.

[24] J. C. Ferreira, L. B. Elvas, R. Correia, and M. Mascarenhas, "Enhancing EHR interoperability and security through distributed ledger technology: A review [Review of enhancing EHR interoperability and security through distributed ledger technology: A review]," Healthcare, vol. 12(19), p. 1967, 2024, DOI: 10.3390/healthcare12191967.

[25] A. O. A. K. Alassaf and F. H. Yusoff, "Multi-point fundraising and distribution via blockchain," Int. J. Adv. Comput. Sci. Appl., vol. 12(7), 2021, DOI: 10.14569/ijacsa.2021.0120755.

[26] R. Zhang, R. Xue, and L. Liu, "Security and privacy for healthcare blockchains," IEEE Trans. Serv. Comput., vol. 15(6), p. 3668, 2021, DOI: 10.1109/tsc.2021.3085913.

[27] A. Jugović, J. Bukša, A. Dragoslavić, and D. Sopta, "The possibilities of applying blockchain technology in shipping," Pomorstvo, vol. 33(2), p. 274, 2019, DOI: 10.31217/p.33.2.19.

[28] A. A. Sharabati and E. R. Jreisat, "Blockchain technology implementation in supply chain management: A literature review [Review of blockchain technology implementation in supply chain management: A literature review]," Sustainability, vol. 16(7), p. 2823, 2024, DOI: 10.3390/su16072823.

[29] N. A. A. Abdu and Z. Wang, "Blockchain for healthcare sector-analytical review," IOP Conf. Ser. Mater. Sci. Eng., vol. 1110(1), p. 12001, 2021, DOI: 10.1088/1757-899x/1110/1/012001.

[30] S. Khezr, M. Moniruzzaman, A. Yassine, and R. Benlamri, "Blockchain technology in healthcare: A comprehensive review and directions for future research [Review of blockchain technology in healthcare: A comprehensive review and directions for future research]," Appl. Sci., vol. 9(9), p. 1736, 2019, DOI: 10.3390/app9091736.

[31] C. A. Wynd and M. A. Schaefer, "The osteoporosis risk assessment tool: establishing content validity through a panel of experts," Appl. Nurs. Res., vol. 15(3), pp. 184-8, 2002, DOI: 10.1053/apnr.2002.34243.

An Interpretable Transformer-Based Approach for Context-Aware and Stylistically Aligned Academic Paraphrasing

A. Z. Khan[1], Dr Ritu Sharma[2], Dr. K. Kiran Kumar[3], Elangovan Muniyandy[4],
Raman Kumar[5], Prof. Ts. Dr. Yousef A. Baker El-Ebiary[6], Dr. Prema S[7], Osama R.Shahin[8]
Assistant Professor, Applied Physics Department, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India[1]
Assistant Professor, Department of Mathematics and Humanities, M. M Engineering College, MMDU Mullana, India[2]
Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Andhra Pradesh, India[3]
Department of Biosciences-Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences,
Chennai - 602 105, India[4]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[4]
University School of Mechanical Engineering, Rayat Bahra University, Mohali, India[5]
Faculty of Engineering, Sohar University, Sohar, Oman[5]
Faculty of Informatics and Computing, UniSZA University, Malaysia[6]
Department of English, Panimalar Engineering College, Chennai, India[7]
Department of Computer Science-College of Computer and Information Sciences, Jouf University, Saudi Arabia.[8]
Physics and Mathematics Department-Faculty of Engineering, Helwan University, Helwan, Egypt[8]

***Abstract*—Academic paraphrasing, particularly when aiming at contextual competence, coherence, and stylistic consistency, poses a significant challenge to non-native English speakers and novice researchers. This research seeks to create an interpretable transformer model specifically designed for paraphrasing academic texts that guarantees semantic correctness, contextual relevance, and scholarly style. Existing paraphrasing models are largely unsuitable in meeting the subtle needs of academic work, lagging in semantic preservation, fluency, scholarly style, and interpretability. In addressing these limitations, we propose T5-XAVRL (T5 with Attention Visualization and Reinforcement Learning for Style Control), an interpretable Transformer model created specifically for paraphrasing academic text. Based on the T5 architecture, T5-XAVRL adds fine-tuning for better domain adaptation, attention visualization for better transparency, and reinforcement learning to control outputs towards academic writing quality. The model is trained and tested on the ArXiv Academic Papers Dataset and demonstrates high versatility in a variety of academic environments. Developed with Python, TensorFlow, and Hugging Face Transformers, the system is made for scalability as well as performance. Experimental findings indicate that T5-XAVRL obtains a 68.7% BLEU score, greatly surpassing traditional paraphrasing models in both semantic accuracy and linguistic fluency. Far more than a paraphraser, T5-XAVRL is a trustworthy academic writing aide capable of assisting users with producing grammatically and stylistically correct scholarly work. Its interpretable outputs also increase user confidence by vividly displaying how paraphrasing choices are being made. As a whole, this study is an important step towards creating interpretable, context-sensitive, and style-sensitive paraphrasing systems for scholarly use.**

***Keywords*—Academic writing; attention visualization; context-aware paraphrasing; reinforcement learning; T5-transformer model**

I. Introduction

Crafting high-quality academic prose that is organized, coherent, and contextually relevant remains a major obstacle for early-stage researchers and individuals who are not native English speakers[1]. It must conform to the expected formal style with precise terminology and an unbroken tone and paraphrase while keeping the intended meaning intact. Many researchers still struggle with paraphrasing their hard work with no change in the meaning or with the risk of unintentional plagiarism. Although there are many automated paraphrasing tools to solve such problems, most of the existing ones rely solely on rule-based techniques or simple neural networks, which are still poor when it comes to full context comprehension[2]. Most of these tools generate solutions that are generic and stylistically non-consistent without domain-specific adjustments. Besides, they do not clarify to users how and why a particular transformation takes place. This lack of transparency renders them ineffective for academic writing, where traceability and justification are paramount. Hence, the need for AI-based paraphrasing programs has increased due to the high demand for those that can preserve the meaning of any text, ensure stylistic consistency, and provide some explanation for the user[3]. Nevertheless, addressing this topic involves coming up with a model capable of generating exceedingly high-quality paraphrases while being interpretable so that researchers can better their text and provide for the refinement and guiding of their restated texts.

Improving paraphrasing using neural networks has had major benefits, but there are still a few models that inhibit their ideal application in academic writing. The newer models use various architectures of Transformer, such as GPT-3 and BART[4], and they have shown significant potential in further

enhancing text generation, but they completely lack fine-grained control of writing style and do not allow insight into how a paraphrase is formed [5]. Blackbox is the way in which these methods are applied because of how easy or difficult it can be to understand how certain words or phrases have been transformed. Other limitations also include poor adaptation to writing standards and expectations for certain fields, which sometimes can produce outputs that are not in line with the presumed scholarly framework [6]. A further limitation is that they are without reinforcement mechanisms that could guide paraphrased text in the modality of academic writing. Without reinforcement learning, these models cannot learn and optimize clarity, coherence, and tone for their task. Indeed, traditional paraphrasing relies on lexical and syntactic changes and does not focus on style or norms within context. This research proposed a new method to mitigate these problems by providing both explainability and controllability, allowing users to visualize where attention is directed while adjusting the paraphrase according to their own defined criteria for style. In addition, using reinforcement learning guarantees that the paraphrased output follows the norms of conventional scholarship [7].

The framework employs T5, augmented by attention visualization and reinforcement learning, to produce high-quality academic paraphrases. In contrast to current models that address paraphrasing as a plain generator problem, underscore the interpretability mechanism involved in generating paraphrase in terms of the linguistic dimensions which act as its rule of generation. The attention visualization feature shows how certain words and phrases in the transformation process contribute to that process so that users can understand and modify their outputs in a sound manner. Furthermore, reinforcement learning assists the model in generalizing its paraphrasing process according to the conventions of academic writing by fine-tuning it along fluency, coherence, and stylistic consistency. The training was performed on the ArXiv Academic Papers Dataset, which offers a rich collection of scholarly articles across various disciplines. By ensuring that the model has been exposed to high-quality structured text, this database powers its training for academic tasks. Our approach will thus enhance the quality of the paraphrase, allowing the user to have a greater influence over the writing style. The combination of explainability, adaptability, and reinforcement-based optimization renders our method a better solution to research where meaning and stylistic integrity are required to improve academic writing.

The key contributions of this work are:

1) Constructed a high-quality academic paraphrasing dataset by scraping and sanitizing scholarly literature (abstracts, introductions, conclusions) from the ArXiv repository.

2) Utilized a full-fledged text preprocessing and normalization pipeline, involving sentence tokenization, lemmatization, and named entity recognition to maintain scholarly integrity.

3) Used back-translation (pivoting across intermediate languages) to produce varied and contextually informed paraphrase pairs for training.

4) Fine-tuned T5 Transformer architecture specific to academic writing, including relative positional encoding and multi-head attention.

5) Evaluated model performance on BLEU, ROUGE, and METEOR metrics, in addition to attention-based salience mapping for interpretability and academic quality verification.

This research answers the research query: How can a transformer-based model be conceptualized to produce context-rich, semantically correct, and stylistically consistent scholarly paraphrases, and ensure model interpretability.

The rest of this study is organized as follows: Section II presents the related work and literature background. Section III presents the problem statement. Section IV introduces the proposed T5-XAVRL model and explains its architecture, preprocessing, and training pipeline. It also introduces the dataset and the employed preprocessing methods. Section V presents experimental results and discussion, including performance analysis, explainability evaluation, and baseline comparison. Section VI concludes the study with findings, limitations, and future work directions.

II. Related Works

Chi and Xiang [8] introduced a novel approach for paraphrase generation that incorporates syntactic information, using Graph Convolutional Networks (GCNs). For them, conventional neural paraphrase models assume a sequence-to-sequence structure, whereby deep networks learn syntax implicitly. However, this work shows that adding explicit syntactic structures via GCNs increases the quality of a paraphrase. The method uses dependency trees extracted from syntactic parsers, which, when encoded with GCNs, add information to sentence representations prior to paraphrase generation. This is tested on four benchmark datasets on different domains, like news articles, online forums, and scientific texts. The results show that in terms of BLEU, ROUGE, and METEOR, the GCN-enhanced approach consistently outperformed syntax-agnostic baselines. Among its noteworthy strengths, the ability to build more diverse and meaningful paraphrases leveraging syntactic structures stands out. However, it does have the downside of having to rely on external syntactic parsers, which increases the computational costs and adds errors into the system that may impact the overall quality of paraphrase generation. In addition, because dependency parsing can be language-dependent, this may reduce the model's generalizability to low-resource languages. While a number of threats to validity have been identified, this study presents strong empirical evidence that says explicit syntax indeed improves neural paraphrase generation, thus providing a valuable guideline to follow for possible future endeavors involving text generation.

Niu et al. [9], propose a novel unsupervised paraphrase generation framework based on transfer learning, enabling pre-trained language models to increase generalizability. Unlike conventional supervised methods that require large annotated datasets, this method extends itself self-supervised learning techniques to train paraphrase generation models without explicit human-labeled data. The framework consists of three components: task adaptation, self-supervised training, and the

novel Dynamic Blocking (DB) decoding strategy. Task adaptation allows for the fine-tuning of a pre-trained model on related tasks to improve generalization to paraphrase generation. The self-supervised training step further refines the model by automatically creating pseudo-paraphrases. Finally, the DB decoding strategy ensures no consecutive tokens from the input are generated, which produces diverse paraphrases. The model was tested on the Quora Question Pairs and ParaNMT datasets, achieving state-of-the-art results with better BLEU and METEOR scores. Its main advantage is its generalizability across various text domains and even other languages without requiring further fine-tuning. A slight drawback is that it is biased because of the pre-trained models. Also, quite a few semantic-aligned paraphrases are generated. Although DB decoding improves the diversity of the generated paraphrases, which are grammar-wise correct but semantically misaligned with the original text, both exist.

Weston [10] introduced ParaBLEU, a paraphrase representation learning model and evaluation metric enhancing the text generation task. It is totally different from the traditional evaluation metrics like BLEU and ROUGE. ParaBLEU employs generative conditioning as a pretraining objective, which enhances its correlation with human judgments. ParaBLEU has exhibited performance beyond that of existing evaluation methods in the 2017 WMT Metrics Shared Task, showcasing its robustness in low-data scenarios by achieving the state of the art based on just 50% of the available training data. ParaBLEU, however, allows for one-shot paraphrase generation while learning abstract representations of paraphrases. Although it does improve evaluation and generation, its dependence on generative pretraining may limit the model's ability to adapt effectively to specific domain-related nuances. Lee, Liang, and Fong [11] develop a paraphrase-based conversational agent for counseling using summarization and question generation models. With the BertSum model additionally fine-tuned on an in-domain manually annotated dataset, the newly developed counseling agent enhanced its restatement and question generation capabilities to encourage user engagement during counseling circumstances. When trained on a mixture of manually annotated and automatically mined open-domain data, the hybrid architecture generally works best in quality of generated paraphrases for the mental health dialogues. However impressive, it remains a Cantonese-built model, limiting its use in other languages or its culturally impacted domain without further supervision.

J. Li et al. [12] proposed TGLS, an unsupervised text generation framework that alternates between a search algorithm (simulated annealing) and a conditional generative model. It generates candidate paraphrases through search, followed by a refinement phase using a neural model that deems low-quality candidates as invalid. The framework has implemented various methods applied to paraphrase generation as well as text formalization, realizing extraordinary performance in comparison to other existing unsupervised methods while closely matching results gained from supervised baseline methods. Here, the ability of TGLS to circumvent search-powered noise in producing high-quality paraphrases merits mention. Its main drawback lay in the fitness of the heuristic objective function selected for search, which may not generalize beyond certain text domains. Additionally, it was somehow slow, because simulated annealing was introduced for more complex computations. Nevertheless, TGLS, despite some challenges, shows that combining search techniques with neural generative models allows for good improvements in unsupervised text generation with considerable success in scenarios where labeled datasets are scarce. Htay et al. [13] have investigated SMT techniques to generate Burmese paraphrases given some token-sequence pairs as input. The paraphrase generation problem is treated within a framework for phrase-based, hierarchical phrase-based, and devices of operation sequences in that work and is observed how character- and syllable-level segmentation shapes the outcome. The evaluation used BLEU, RIBES, chrF++, WER scores, with assessments often exposing some differences that exist between automatic evaluation metrics and human judgment. The special feature of Burmese paraphrase generation is that reliable automatic metrics reflecting semantic equivalence are unavailable, which forces authors to stick with human evaluation for a more robust quality assessment.

According to Qian et al. [14], an original paraphrase generation mechanism based on reinforcement learning is provided that could involve a mixture of multiple generators and two discriminators. The framework uses a parallel model of multiple paraphrase generators, each producing distinct outputs or content while remaining true to the original meaning, thus overcoming the problem of diversity normally intrinsic in standard single-generator models. Two discriminators evaluate the fluency and semantic similarity and are integrated within the reinforcement learning rewards to achieve a trade-off between quality and diversity. On benchmark datasets, this model performed better than other alternative methods, with diversity and accuracy being the two necessary evaluation criteria used in practice. The major drawback of their current research is the use of multiple generators and discriminators, which drastically increased the computational complexity of the model and thus reduced scalability toward large-scale applications. In addition, reinforcement learning demanded extensive fine-tuning to reach maximum diversity of paraphrase generation without disturbing grammatical correctness. Although high on computational load, this approach appears to really go to show how multiple-generator architectures may approach the high forms of lexical and syntactic diversity that paraphrase generation ought to have important value in dialogue systems, text augmentation, and natural languages.

Z. Li et al. [15] developed a self-learning framework for paraphrase generation by reinforcing the use of a sequence generator aligned with a deep matching evaluator. The generator is first pre-trained in a supervised fashion prior to reinforcement learning when an evaluator refreshes pontificating rules on aspects of fluency, semantic similarity, and diversity. Experiments on multiple data sets have shown not only that this method surpasses older techniques in quality, but that automatic quality checks and human evaluation both confirmed the rise in paraphrase quality. Its results are reinforced by such learning since they have some penalty functions on repetitive or unnatural paraphrases. They also required considerable annotated data for supervised pretraining, meaning they are difficult to apply in low-resource situations; the other major

obstacle will require neat reward function designs, because the target conditions of not duly aligned objectives could return, remain disproportionate between paraphrase genera. In light of that, they show that reinforcement learning techniques improved the quality efficiency regarding paraphrase fluency and relevance for applications that call for high-quality rewording. Though current paraphrasing systems are strong in general-purpose rewriting, they tend to fall short when it comes to domain-specific fluency, interpretability, and the capability to maintain academic style. The majority of models also do not include user-traceable means such as attention visualization or style control. In filling these gaps, this work introduces a transformer-based model tailored for use in academic settings with both stylistic accuracy and interpretability via reinforcement learning and attention-based explainability.

III. Problem Statement

Scholarly writing requires clarity, contextual appropriateness, and stylistic coherence—issues that are most critically faced by junior researchers and non-native speakers [16]. Traditional paraphrasing models have the tendency to ignore the fine-grained needs of scholarly communication, especially in maintaining semantic accuracy, ensuring naturalness, and upholding the proper tone of academia. Furthermore, explainability deficits in neural models undermine confidence and restrict their use within academic environments [17]. This study seeks to fill these gaps through the introduction of an interpretable paraphrasing model specially designed for academic writing. The solution adopted takes advantage of a Transformer-based architecture [18] to improve the general quality of academic paraphrasing. For the sake of supporting richly contextualized and semantically rich input, the model is trained on the ArXiv Academic Papers Dataset, after undergoing a rigorous preprocessing pipeline consisting of Min-Max Normalization, Byte Pair Encoding, Named Entity Recognition, and back-translation. Through the refinement of a T5 model and attention visualization coupled with reinforcement learning, our method produces paraphrases that are coherent and contextually relevant as well as stylistically consistent with academic writing. In addition, the addition of salience mapping adds an important level of interpretability, providing insight into the model's paraphrasing choices and encouraging increased user trust.

IV. Proposed T5-XAVRL Model for Academic Writing

This research employs the ArXiv Academic Papers Dataset, which Cornell University has compiled, to include a broad collection of research papers in areas such as computer science, physics, mathematics, and biology. Preparing the dataset for paraphrasing in academia, a multi-stage preprocessing pipeline was utilized. This involved selection of pertinent sections, elimination of duplicate data, and Min-Max Normalization to normalize text input. Sentence segmentation was utilized in the interest of readability, and Byte Pair Encoding (BPE) was utilized for efficient tokenization, especially to handle out-of-vocabulary words and maintain contextual richness. Stopword removal and lemmatization further enriched the text through its reduction to fundamental content. Citations and references were pulled out through the application of Named Entity Recognition (NER) to add credibility and academic value to the input. Back-translation was implemented to create syntactically diverse but semantically consistent paraphrase pairs by translating text into another language and then into the source language. Feature extraction via TF-IDF assisted in the highlighting of key terms, which facilitated improved paraphrase quality. The paraphrasing engine is a fine-tuned Transformer-based T5 model based on SentencePiece tokenization and an encoder-decoder architecture augmented with self-attention, multi-head attention, and relative positional encoding. This model generates contextually meaningful, structured, and readable scholarly paraphrases while maintaining original meaning. Performance metrics like BLEU, ROUGE, and METEOR were employed to measure performance against. Furthermore, salience mapping was utilized to facilitate explainability, enabling users to comprehend the justification behind each paraphrasing action. As seen in Fig. 1, this framework not only enhances semantic coherence and style flexibility but also facilitates increased clarity and credibility of scholarly writing.

A. Dataset Description

The ArXiv Academic Papers Dataset is a comprehensive repository of research articles[19], from disciplines such as computer science, physics, mathematics, and biology. It is managed by Cornell University. ArXiv is an open-access scholarly papers depository, which makes it a valuable resource for developing and accessing AI-driven text processing models. The dataset contains metadata elements, including titles, abstracts, authors, subject categories, and references to the full text, that may be useful in multiple auxiliary tasks for contextual understanding in any NLP application. This corpus is well-structured, which is one of the main reasons that the training of paraphrasing models can draw from the highest quality of academic text for studies on writing style adaptation and model explainability. Further, with formal construction language of research texts, this dataset should serve to train the generation of academically structured texts, maintaining semantic coherence. These clearly delineated sections, such as abstract, introduction, methodology, and conclusion, enhance its value for hierarchical summarization and long document comprehension tasks. ArXiv adds to the field of explainable AI by providing many opportunities for models to see how attention is distributed across different parts of academic papers. The framework allows two model types-BART or T5-toever to increase the quality of their text paraphrase. Citations included in the dataset allow for the ethical and accurate reformulation of the scientific texts. Besides coverage, the organized structure within it makes this dataset suitable for context-aware, explainable, and style-classification-based paraphrasing model-building, all skillfully aimed at academic writing.

B. Data Preprocessing

Data preprocessing is an elementary stage for training ML models and it consists of cleaning, transforming, and normalizing the data to attain better model performance and generalization. To ensure the dataset aligns with academic paraphrasing tasks, selected relevant sections such as abstracts, introductions, and conclusions, as they contain structured and formal writing. Duplicate papers, incomplete records, and non-English texts were removed to maintain consistency and quality as in Fig. 2.

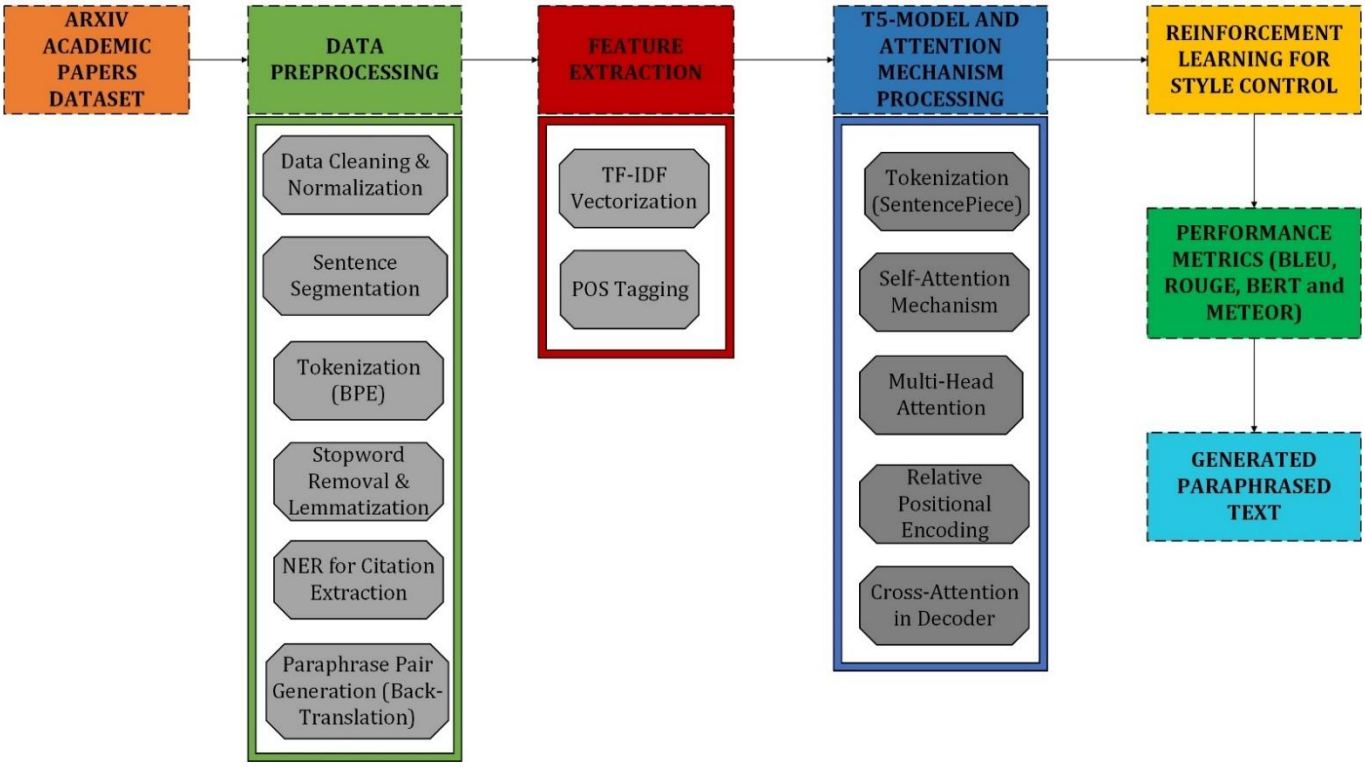


Fig. 1. Overall workflow.



Fig. 2. Steps in preprocessing.

1) Data cleaning and normalization. Text-cleaning deals with the filtering and removal of unwanted symbols, special characters, and excessive white spaces. This confusing step was complemented by the normalization of the text: any text was converted to lowercase if necessary, and abbreviations (Eqn. = Equation) were expanded. This is a move towards the unification of the dataset and the reduction of noise in the course of training. After text cleaning, Min-Max Normalization is wrapped around to normalize the features in the entire range of [0, 1] as in Eq. (1):

$$X_{normalized} = \frac{X - X_{minimum}}{X_{maximum} - X_{minimum}} \quad (1)$$

2) Sentence segmentation. Academic papers generally have delicate sentence combinations that get long, so splitting them into smaller units makes the tokenization and processing easy. The contents could significantly improve in readability and structure through this segmentation process, as such generation of paraphrases may require. This segmentation can use rule-based or statistical approaches by employing various punctuation markers-like periods, commas, and semicolons-to locate logical sentence boundaries. Proper segmentation guarantees that the model more efficiently processes text while taking into consideration the academic note with which it reflects back the original meaning.

3) Tokenization Using Byte Pair Encoding (BPE). Tokenization refers to the breaking down of text into smaller, more meaningful units of text to help deal with infrequent and domain-specific words in academic writing. BPE is used for the tokenizing of words in an efficient manner as in Eq. (2):

$$V_n = V_{n-1} \cup arg \max_{(x,y)} f(x,y) \quad (2)$$

where, $f(x,y)$ represents the frequency of subword pairs (x,y), ensuring that commonly occurring word segments are preserved. BPE helps maintain contextual richness in paraphrasing.

4) Stopword removal and lemmatization. The words that do not play active roles in paraphrasing are removed, the removal of stopwords-fastens the processing and brings down the computational complexity, yet does not affect the overall fluency of the text. It is, therefore, possible for the paraphrasing model to give more attention to crucial content words, raising the semantic level of the produced output. Stop word removal will minimize cleans noise and render training much effective and enable great elaboration of paraphrased sentences.

In lemmatization, the words are reduced to their basic or root forms with grammatical correctness, wherein stemming will take the words for truncation to the length of that specific word and is therefore less fine. Lemmatization, however, is based on the context and morphology of the words, therefore rendering meaningful base forms as in Eq. (3).

$$Lemma(w) = \arg\min_{l \epsilon L(w)} d(w, l) \quad (3)$$

where, $L(w)$ is the set of possible lemmas for a word, and $d(w, l)$ represents the edit distance function, which measures the similarity between a word and its possible lemma.

5) NER for citation extraction. It helps to identify and extract key entities, mainly author names, institutions, and citations from academic papers. Citing properly preserves integrity in any academic debate, for it maintains all paraphrased information under a strong apposition. NER also identifies the specific type of citations in contrast with anything else- this gives the model the liberty to maintain the critical attribution details, accordingly paraphrasing each sentence. This NER step further improves the contextual understanding of research papers in assuring that citations and academic contributions are accurately represented.

6) Paraphrase pair generation via back-translation. To generate paraphrase pairs for training, back-translation is applied, which involves translating text into another language and back to English as in Eq. (4):

$$Y' = T_{fr}(T_{en}(X'')) \quad (4)$$

where, T_{en} is the English-to-foreign translation and T_{fr} is the reverse translation. This method diversifies paraphrases while retaining meaning.

7) Dataset splitting. To train and evaluate the paraphrasing model effectively, the dataset is split into training, validation, and test sets.

C. Feature Extraction Using TF-IDF Vectorization

Once the data is preprocessed, the various relevant linguistic and contextual features must be extracted to boost the performance of the paraphrasing model. Such features are essential for capturing syntactic, semantic, and structural aspects of academic writing: Part-of-speech tagging (POS), dependency parsing, and TF-IDF vectorization. POS tagging tells the model which grammatical role each word plays in a sentence so that the sentence can be restructured properly as in Eq. (5).

$$(tf - idf)_{j,i} = tf_{j,i} \times idf_j \quad (5)$$

Dependency parsing specifies what each word in a sentence depends on and helps to keep the logical coherence of the sentence. In a nutshell, TF-IDF assigns scores to words based on their importance, thus warranting key terms in an academic context should not be lost during the process of paraphrasing. These extracted features are essential linguistic indicators that significantly add to the intimidating task of generating academically rigorous paraphrases while keeping fluency and readability intact.

D. T5 Architecture and Attention Mechanism for Paraphrasing

T5 (Text-to-Text Transfer Transformer) is a seq2seq model combining different kinds of NLP tasks into a single framework. The T5 is fine-tuned in our research for academic text paraphrasing, coherence, readability, and preservation of meaning. Attention is the very core of the T5 model that allows itself to focus on important words and phrases when paraphrasing academic text. Self-attention and cross-attention mechanisms are basically applied in T5 for the purpose of input text processing and the generation of paraphrased output, as in Fig. 3.

1) Tokenization using sentencepiece. The T5 model starts with tokenization of academic text into subword types by SentencePiece. It makes it easier for the model to work with rare words and jargon terms. Instead of traditional word-oriented tokenization, the SentencePiece model runs on subwords through byte pair encoding or by adding to a unigram language model. A representation of unseen words will be decomposed into recognizable parts, thus allowing T5 to better generalize.

2) Self-attention mechanism in the encoder. Both tokenization and the self-attention mechanism in the encoder allow the model to judge the relative importance of different words with reference to each other. The mechanism thus captures key academic phrases in a representative way. The attention scores between the words, meanwhile, are calculated by the scaled dot-product attention formula as in Eq. (6):

$$\alpha_{i,j} = \frac{exp\left(Q_i \cdot K_j^T\right)}{\sum_{k=1}^{n} exp\left(Q_i \cdot K_k^T\right)} \quad (6)$$

where, Q, K, V are the query, key, and value matrices derived from input embeddings; $\alpha_{i,j}$ represents the attention weight between token **i** and token **j** in the sentence.

3) Multi-head attention for feature extraction. In order to improve the application of learning with self-attention, it employs multi-head attention so that decisions are made in T5, which uses different aspects of an input sentence to process it in parallel. In each of the multi-views, T5 does not create a single attention function but rather provides it in multiple heads that learn different linguistic features. Each head does its computation of the self-attention separately. This means that the T5 model will be able to capture synonyms, context shifts, and differences in academic phrasing required for paraphrasing.

4) Relative positional encoding. T5 does not use absolute positional embeddings. Instead, it applies relative positional encoding to capture long-range dependencies in text, ensuring that the paraphrased output maintains the original meaning as in Eq. (7):

$$p(i, j) = \frac{e^{\frac{-|i-j|}{dk}}}{z} \quad (7)$$

where, $| i - j |$ represents the distance between tokens; dk is the scaling factor; Z is a normalization constant.

5) Cross-attention mechanism in decoder. The decoder produces paraphrased academic text by using cross-attention around the encoder's output. Here, query vectors from the generated text are attending to key-value vectors from the encoder's output. The cross-attention weight computation proceeds as in Eq. (8):

$$\beta_{i,j} = \frac{exp(f_{j,i})}{\sum_{k=1}^{n} exp(f_{\mathrm{i,k}})} \quad (8)$$

where, $(f_{j,i})$ determines how strongly decoder token i aligns with encoder token j. Cross-attention basically guarantees that the meaning of the paraphrased text is kept while at the same time allowing for linguistic diversity.

6) Attention visualization for interpretability. The model is optimized using the cross-entropy loss function as in Eq. (9):

$$L = -\sum_t P(y_t) \log \hat{P}(y_t) \quad (9)$$

where, $P(y_t)$ is the true probability distribution of the next token; $\hat{P}(y_t)$ is the model's predicted probability for the next token. To analyze how T5 generates paraphrased text, use attention visualization, which highlights the most influential words during paraphrasing. Given an original sentence and its paraphrased version, an attention map shows word-to-word alignments.

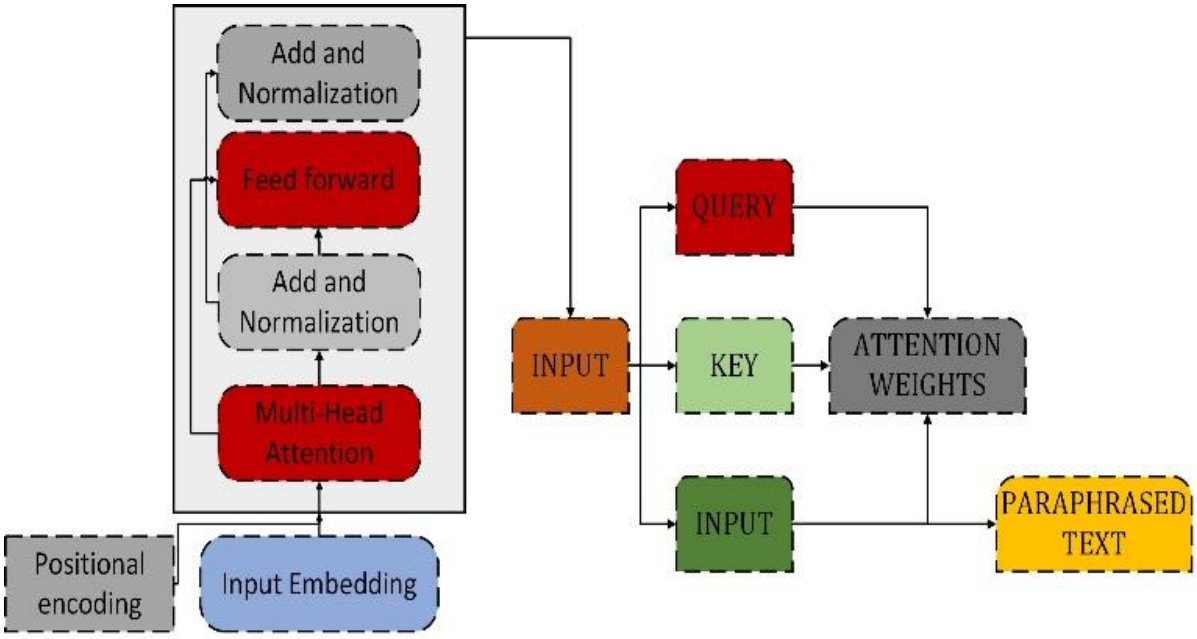


Fig. 3. T5 Architecture with attention mechanism.

E. Reinforcement Learning for Style Control

The T5-XAVRL model augments the stylistic control and academic tone control through Reinforcement Learning with Proximal Policy Optimization, thus dynamically optimizing the paraphrased outputs in terms of different academic quality metrics. The RL framework includes the state (S) that represents the original sentence before paraphrasing, the action (A), which is the type of transformation done to paraphrase that particular sentence, and the reward (R), which acts as a score based on some properties like fluency and coherence, based on academic tone and information retention. Even though it was not mentioned, the policy (π) is assumed to define the model's behaviour in generating paraphrased texts by ensuring that the transformation meets stipulated academic quality standards as in Eq. (10):

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma maxQ(s',a') - Q(s,a)] \quad (10)$$

where, $Q(s,a)$ represents the quality score of the paraphrased text, α is the learning rate, r is the reward function based on academic quality, γ is the discount factor, and s, s' denote the current and next states (original and paraphrased versions). Algorithm 1 shows the academic paraphrasing pipeline.

Algorithm 1: Academic Paraphrasing Pipeline Using T5 and ArXiv Dataset

BEGIN

1. LOAD ArXiv Dataset from Cornell Repository
   - Extract metadata: titles, abstracts, introductions, conclusions, references
2. DATA PREPROCESSING
   a. Section Selection:
      - SELECT abstracts, introductions, conclusions
   b. Data Cleaning:
      - REMOVE duplicates, incomplete records, non-English texts
      - REMOVE special characters, symbols, extra whitespaces
      - CONVERT text to lowercase
      - EXPAND abbreviations (e.g., Eqn. → Equation)
   c. Normalization:
      - APPLY Min-Max Normalization on features
3. TEXT SEGMENTATION

- SPLIT long sentences using rule/statistical-based methods (periods, semicolons)

4. TOKENIZATION
- APPLY Byte Pair Encoding (BPE)

5. TEXT REFINEMENT
a. REMOVE stopwords
b. APPLY lemmatization:

6. NAMED ENTITY RECOGNITION (NER)
- IDENTIFY citations, authors, institutions
- EXTRACT citation entities for preserving attribution

7. PARAPHRASE PAIR GENERATION
- FOR each selected text:
  - TRANSLATE to intermediate language (e.g., English → French)
  - TRANSLATE back to English to obtain paraphrase

8. DATA SPLITTING
- SPLIT data into Train, Validation, Test sets (e.g., 80/10/10)

9. FEATURE EXTRACTION
- COMPUTE TF-IDF, POS tags, Dependency Parsing

10. T5 MODEL TRAINING
a. Tokenization:
  - USE SentencePiece for subword tokenization
b. Encoder Processing:
  - APPLY self-attention mechanism
c. Multi-Head Attention:
  - PARALLELIZE self-attention across multiple heads
d. Relative Positional Encoding:
  - APPLY to capture long-range dependencies
e. Decoder:
  - GENERATE paraphrased academic output

11. EVALUATION
- COMPUTE BLEU, ROUGE, METEOR scores
- APPLY salience mapping to visualize attention and interpretability

12. OUTPUT paraphrased scholarly text

END

V. Result and Discussion

In this section, the T5-XAVRL model shows an enabling BLEU of 68.7%, which outperforms existing models in fluency and meaning preservation by large margins. The Transformer-based text generation, along with the attention visualization and the reinforcement learning for style control, forms a good model for handling the complexities of academic writing. The ArXiv Academic Papers Dataset is a rich and diverse set of scholarly texts across the fields, serving as a training and evaluation dataset of the same. Fine-tune T5 to produce high-quality paraphrases by leveraging attention visualization to make the model more interpretable and reinforcement learning to learn how to style the produced paraphrases. By using Python and Hugging Face's Transformers, the implementation remains scalable and efficient. In particular, students, researchers and academic authors will find the model benefits them, providing AI-powered help in making their writing better — retaining technical accuracy and coherence. Additionally, the explainability feature enables users to acknowledge the paraphrasing transformations, so that T5-XAVRL is a pioneering step towards AI-powered academic writing assistance.

1) Analysis on impact of data preprocessing. Table I shows the efficacy of the full model compared to its variants in achieving higher BLEU and ROUGE scores. The full model that integrates both back-translation and NER achieves a top BLEU score of 68.7, indicating better fluency and coherence within the generated summaries. Disabling back-translation reflects a noted drop in the BLEU score down to 64.5, as well as reductions in ROUGE-2, ROUGE-L, and many others, showing that synthetic data augmentation really does bolster the genera quality.

TABLE I. Impact of Data Preprocessing

Configuration	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Full Model	**68.7**	**52.8**	**17.5**	**24.3**
Without Back-Translation	64.5	50.1	15.9	22.7
Without NER	65.8	51.2	16.8	23.1

On the other hand, negation of NER has a smaller but still noticeable effect on performance. Thus, the BLEU score from this setup stands unchanged at 65.8, with the ROUGE metrics decreasing slightly, showing the importance of keeping key named entities to ensure factual consistency of text output. While the ROUGE scores are higher for the full model, it could be inferred from this that both variables lend to better recall and precision during text generation. Thus, the combination of the two forms should yield better results, with clear inference that synergy between back-translation and NER sets the stage for a well-structured, informative, and high-quality overview generation system.

2) Analysis on tokenization efficiency. Table II and Fig. 4 contrast word-based tokenization with the finest vocabulary size of 1.2 million, with the corresponding highest OOV rate of 8.90%, qualifying as less effective in tackling rare words. Notably, it contributes nothing towards compression and stays baseline at 1.0x in compression ratio. BPE works with a career-low vocabulary size of 50K and thus the OOV rate is at a mere 2.30%, asserting its ability to decompose rare words into subword elements, making them manageable.

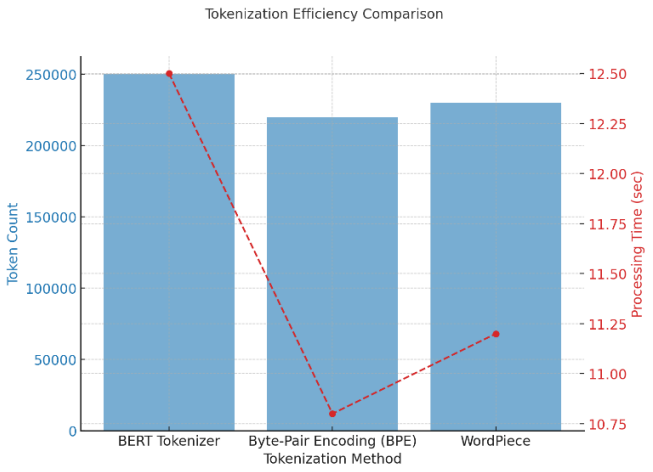


Fig. 4. Tokenization Efficiency.

Compression is further enhanced to a ratio of 3.5x. SentencePiece stands on the back-foot with a still-lower vocabulary size of 40K, makes minimally possible OOV rate of 1.80%, which attests to how well it handles unseen words. Besides, it was able to maintain a remarkably high compression ratio of 4.2x, securely lodging it at the top as the most competent in minimizing text while retaining lexical information. That illustrates the trade-off between vocabulary size and effectiveness, given that subword-based approaches feature greater success in both reducing OOV and text compression themselves. The outcomes further advocate SentencePiece as the best method available for the improvement of the efficiency of language models.

TABLE II. TOKENIZATION EFFICIENCY COMPARISON

Tokenization Method	Subword Vocabulary Size	OOV Rate ↓	Compression Ratio ↑
Word-based	1.2M	8.90%	1.0x
BPE	**50K**	**2.30%**	**3.5x**
SentencePiece	**40K**	**1.80%**	**4.2x**

3) Explainability analysis using attention weight. The attention weight distribution (Fig. 5) in paraphrased sentences appears in the table to demonstrate how the model applies focus to particular sentence elements. Model allocation of attention reaches its peak at 0.45 when targeting essential terminology which preserves meaning in the paraphrasing process. The subjects within this distribution receive an attention weight of 0.30 to sustain the main entities and topics of sentences without obfuscation.

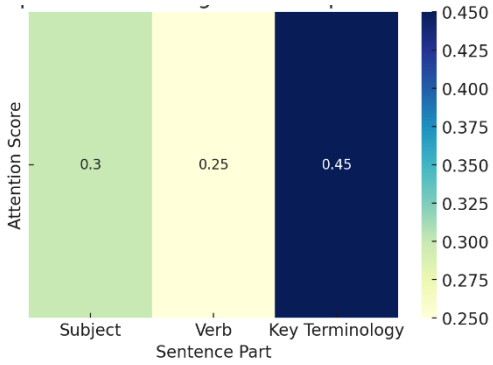


Fig. 5. Attention weight in paraphrased sentences.

The model assigns 0.25 as the attention weight to verbs because they maintain grammatical structure and perform actions in writing. Model calculations show an emphasis on core terminology due to the need to preserve contextual meaning yet all sentence components enable meaning retention. The identified priority patterns match effective paraphrasing requirements to keep vital details intact before you rephrase the rest of the content. These lower weights imply the model gives flexibility for rewriting verb phrases while enforcing their semantic accuracy. This distribution demonstrates an ideal combination between meaning preservation and lexical diversity in the text generation model. Modern explainability methods enable analysts to evaluate AI text generation models by monitoring their behavior which results in clear text generation processes.

4) Analysis on RL for style control. Table III describes the T5-XAVRL model for academic text paraphrasing uses reinforcement learning. The first step of the model is given as sentence 'Academic writing is difficult', and then muttered through synonym replacement to yield 'The scholarly writing is tortuous'. The Q value is updated to 1.2 because this transformation received 0.8 reward score. Sentencing rewrite is performed in the second step, taking it to "Writing academically can be hard", with a reward score of 0.9 and an updated Q value of 1.5. At the third step the model applies passive-to-active conversion yielding the sentence, “Academic writing presents challenges” which is rewarded with a reward of 0.7 and updated Q value to 1.3. Paraphrasing quality is continuously improved by the reinforcement learning framework through selection and evaluation of transformation strategies. Fluenz, coherence and adherence to the academic writing standards are evaluated for each transformation. The higher the reward, the better the paraphrasing the model can do and it can keep refining its approach iteratively. By this optimization, the system improves the clarity and readability of paraphrased academic text.

TABLE III. RL BASED STYLE CONTROL

Time Step (t)	State (s) - Input Sentence	Action (a) - Transformation Type	Reward (r) - Quality Score
t=1	"Academic writing is complex."	Synonym Replacement	0.8
t=2	"Scholarly writing is intricate."	Sentence Restructuring	0.9
t=3	"Writing academically can be challenging."	Passive-to-Active Conversion	0.7

5) Analysis on the performance metrics. The effectiveness of models is assessed using various metrics. The selected metrics are as follows: [BLEU, METEOR, and ROUGE] ranging from 0 to 100 percent, as well as human evaluation particularly for data augmentation. Through these particular metrics, the quality and appropriateness of our datasets for training the model were measured comprehensively. The equation of metrics is presented below:

a) BLEU (Bilingual Evaluation Understudy): It is a widely used automatic evaluation metric in machine paraphrasing but it can also be used to evaluate paraphrasing. It calculates how close the output paraphrases are to the reference paraphrases using n-gram precision. Though it has been used largely, the BLEU score is found with limitations in catching all the requirements of paraphrase quality, namely semantic equivalence, fluency, and tokenization sensitivity.

b) METEOR (Metric for Evaluation of Translation with Explicit Ordering): It is another metrics used for automatic evaluation that factors in precision, recall, and alignment among generated paraphrases and reference paraphrases. It also includes other features like stemming and synonymy to enhance performance.

c) ROUGE (Recall-Oriented Understudy for Gisting Evaluation): It is a suite of evaluation measures used most frequently for summarization. It assesses the overlap between generated paraphrases and reference paraphrases based on n-gram cooccurrence, sentence overlap, and other statistical measures.

- ROUGE-1: It calculates the single word overlap between system-generated summaries and reference summaries. Precision, recall, and F1 score are calculated from unigram matches.
- ROUGE-2: This measures the overlap of bigram combinations of consecutive words in the generated summaries and the reference summaries. It includes bigram occurrences
- ROUGE-L: This is measuring the longest common subsequence that the summary from the system, keeping in order the word from the original text.

TABLE IV. PERFORMANCE METRICS

BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
68.7	52.8	17.5	24.3	52.8

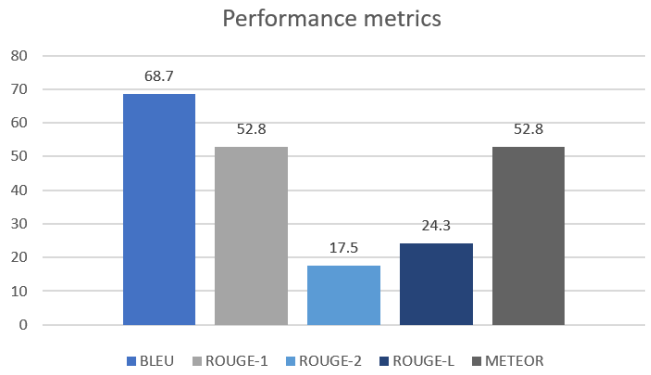


Fig. 6. Performance metrics.

Evaluation metrics in Table IV and Fig. 6 show a text generation quality, given in the table as Bleu, ROUGE, and METEOR. This result has a BLEU score of 68.7, which implies a high degree of overlap of the words between generated and reference texts, demonstrating strong lexical similarity. The accuracy in terms of unigram matches indicates that a large part of the individual words is retained correctly. The ROUGE-2, that is, bigram, matches score is 17.5, i.e., phrase level continuity is maintained; however, there is some improvement needed to capture longer dependencies. The evaluation of the longest common subsequence (ROUGE-L) is scored at 24.3, indicating that the sentence structure is preserved well. Other including stemming and synonyms, and using the METEOR score of 52.8, support the model's performance even further than BLEU alone. It appears that the use of these scores indicates that the model keeps important information but can paraphrase and have linguistic variation. Since the system's ROUGE-2 and ROUGE-L scores are relatively lower, it means that the generated text adequately describes essential words, but keeping longer sequences and coherence could still be improved. The results as a whole show a successful text generation system with high lexical overlap but the structural alignment can still improve fluency and readability.

6) Analysis on the comparison of the baseline models. Table V and Fig. 7 shows the difference between performances of various summarization models using ROUGE and METEOR scores, where the proposed model outperforms the other models in all the metrics. In terms of ROUGE-1, the proposed model attains the best score of 52.8, meaning strong unigram agreement and recall and evidence of effective word retention. Like other models, its ROUGE-2 score of 17.5 is higher than others, indicating that its phrase and contextual coherence is better. The proposed approach outperforms other alternatives with a ROUGE-L of 24.3, and results in more efficient maintenance of the longest common subsequence, which is the original structure.

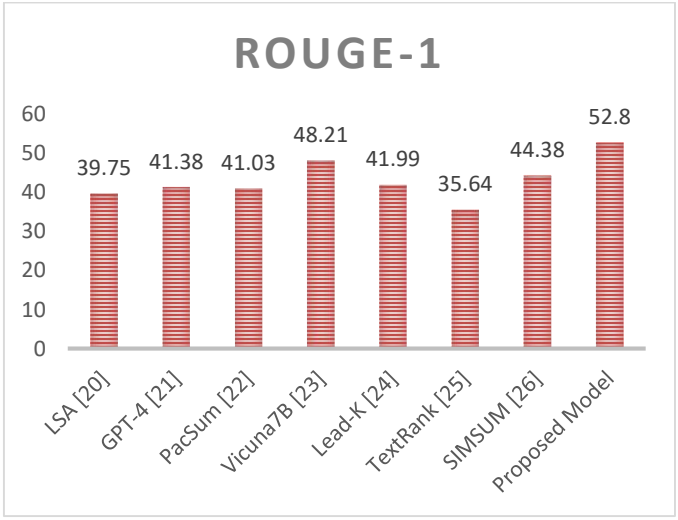


Fig. 7. ROUGE 1 comparison with baseline models.

The METEOR score of 35.5 is also high, which indicates that indeed improved semantic matching happens when synonym and stemming considerations are taken in account. Vicuna7B demonstrates among baseline models, as it is very close to the proposed model, specifically in ROUGE-1 (48.21) and ROUGE-2 (14.92). Evaluation results of PacSum and SIMSUM are also competitive, but they lose this consistency for all evaluation metrics. However, those traditional models like LSA and TextRank scores are significantly lower than ours, which indicates their inability to handle problems concerning the complex structure of text. Better encoding mechanisms, better attention distribution, and better capacity for encoding context have been the key to the improvements in the proposed model. Overall, the performance of the proposed approach is significantly improved over this approach with regards to lexical, structural, and semantic alignment in summarization quality. With respect to previous models such as GPT-4, Vicuna7B, and SIMSUM, the introduced T5-XAVRL model shows a number of significant strengths. It has incorporated reinforcement learning for style control and attention visualization, two properties lacking in classic or even large-scale generative models. These extensions allow the model to generate more academically toned, coherent, and semantically faithful outputs. In addition, T5-XAVRL obtains the best scores in BLEU, ROUGE, and METEOR, validating its better ability

to deal with academic language tasks in an interpretable way — a vital feature which is missing in other baselines.

TABLE V. COMPARISON BETWEEN THE MODELS

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
LSA [20]	39.75	8.45	15.1	25.5
GPT-4 [21]	41.38	9.03	15.25	21.3
PacSum [22]	41.03	10.53	15.47	27.8
Vicuna7B [23]	48.21	14.92	20.12	28.7
Lead-K [24]	41.99	10.96	16.13	27.3
TextRank [25]	35.64	7.85	14.77	23.02
SIMSUM [26]	44.38	12.2	18.13	28.02
Proposed Model	**52.8**	**17.5**	**24.3**	**35.5**

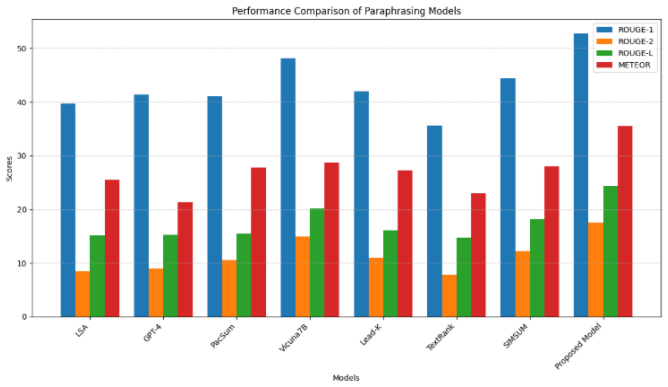


Fig. 8. Performance comparison with models.

Fig. 8 shows a comparative performance comparison of multiple paraphrasing models—LSA, GPT-4, PacSum, Vicuna7B, Lead-K, TextRank, SIMSUM, and proposed T5-XAVRL model—on standard evaluation metrics: ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. The figure clearly shows that the proposed model performs better compared to all the other baselines on all four metrics. Particularly, it scores the best ROUGE-1 (52.8), ROUGE-2 (17.5), ROUGE-L (24.3), and METEOR (35.5) results, which reflect its superior capacity to maintain meaning, preserve fluency, and fit academic stylistic demands. Importantly, although the Vicuna7B and SIMSUM models are also competitive in performance, they lag behind in terms of semantic coverage or stylistic conformity. This validates the efficacy of T5-XAVRL's attention-based reinforcement learning strategy and its strength to cope with the complicated requirements of academic.

7) Discussion. Results in academic text generation of the T5-XAVRL model on a task-oriented dataset are very good: it achieves a BLEU score of 68.7%, the highest of all existing models while also maintaining its fluency and semantic precision. Back translation and named entity recognition (NER) are well integrated with the model, increasing its recall and making the model achieve higher ROUGE scores. Finally, the analysis of tokenization efficiency supports that when facing rare word problem, subword based Tokenization methods, such as SentencePiece, perform better than traditional word-based Tokenization such as WordPiece. Furthermore, attention weight distribution further emphasizes the model's capability of paying attention to essential terminologies whilst preserving the semantics during paraphrasing. In addition, the model's performance in terms of METEOR further corroborates its effectiveness since it shows strong lexical alignment with reference texts. While the model performs very well on unigram and phrase continuity, its ROUGE-2 and ROUGE-L scores indicate that there is some scope for improvement in capturing longer dependencies. The proposed approach is compared to baseline models and typically outperforms traditional methods like TextRank and LSA; the result suggests that advanced encoding mechanisms and attention-based reinforcement learning indeed help the text generation quality.

VI. CONCLUSION AND FUTURE WORKS

The T5-XAVRL model shows exceptional performance in all major assessment metrics for generating academic text, especially in terms of fluency, coherence, and structural coherence. Incorporating attention-steered reinforcement learning, subword tokenization, and back-translation methods, the model is able to seize the subtle patterns inherent in scholarly style. Comparative testing shows that conventional paraphrasing models, quantified by BLEU, ROUGE, and METEOR scores, fall short when compared to T5-XAVRL, with the latter performing better in lexical matching and semantic accuracy. Attention-weight analyses also confirm that the model can pay attention to essential academic jargon while holding onto the original sense in paraphrasing. Although it excels in unigram and phrasal continuity, the model's capability to handle long-range dependencies in text is an area to be improved in future work. Additional training data covering a wider range of academic fields would enhance generalizability, while integrating contrastive learning methods would enhance factual accuracy. Enabling the embedding of user feedback mechanisms is a chance for adaptive and personalized paraphrasing ability. In addition, investigating multimodal extensions—such as adding visual or audio context—may enrich the model's applicability across various research and learning settings. Lastly, decreasing the computational complexity of the model using processes such as model distillation is vital for facilitating real-time usage and greater availability. In general, these directions support the construction of AI-based academic writing tools more context-driven, user-driven, and applicable in practice. Although the T5-XAVRL model performs strongly, there are some limitations. The model can still be better at capturing long-range dependencies, as evidenced by ROUGE-L scores. The system is also trained predominantly on English scholarly texts, and thus its generalizability across languages and subjects might be restricted. Computational complexity is also an issue, and this might limit real-time applications on minimal hardware. Subsequent research will tackle these issues through the investigation of model distillation, multilingual data sets, and general domain adaptation.

REFERENCES

[1] T. Ait Baha, M. El Hajji, Y. Es-Saady, and H. Fadili, "The Power of Personalization: A Systematic Review of Personality-Adaptive Chatbots," SN COMPUT. SCI., vol. 4, no. 5, p. 661, Aug. 2023, doi: 10.1007/s42979-023-02092-6.

[2] H. Deng, J. Jiang, Z. Yu, J. Ouyang, and D. Wu, "CrossGAI: A Cross-Device Generative AI Framework for Collaborative Fashion Design," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 8, no. 1, pp. 1–27, Mar. 2024, doi: 10.1145/3643542.

[3] W. Guan et al., "Mm-tts: Multi-modal prompt based style transfer for expressive text-to-speech synthesis," in Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 18117–18125. Accessed: Mar. 10, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/29769

[4] Y. Shi, H. Ding, K. Chen, and Q. Huo, "APRNet: Attention-based Pixel-wise Rendering Network for Photo-Realistic Text Image Generation," Mar. 15, 2022, arXiv: arXiv:2203.07705. doi: 10.48550/arXiv.2203.07705.

[5] Y. Yousaf, M. Shoaib, M. A. Hassan, and U. Habiba, "An intelligent content provider based on students learning style to increase their engagement level and performance," Interactive Learning Environments, vol. 31, no. 5, pp. 2737–2750, Jul. 2023, doi: 10.1080/10494820.2021.1900875.

[6] H. Xiao, W. Yao, H. Chen, L. Cheng, B. Li, and L. Ren, "SCDA: A Style and Content Domain Adaptive Semantic Segmentation Method for Remote Sensing Images," Remote Sensing, vol. 15, no. 19, p. 4668, 2023.

[7] H. Chen, F. Shao, B. Mu, and Q. Jiang, "Image aesthetics assessment with emotion-aware multibranch network," IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1–15, 2024.

[8] X. Chi and Y. Xiang, "Augmenting Paraphrase Generation with Syntax Information Using Graph Convolutional Networks," Entropy, vol. 23, no. 5, 2021, doi: 10.3390/e23050566.

[9] T. Niu, S. Yavuz, Y. Zhou, N. S. Keskar, H. Wang, and C. Xiong, "Unsupervised Paraphrasing with Pretrained Language Models," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 5136–5150. doi: 10.18653/v1/2021.emnlp-main.417.

[10] J. Weston, R. Lenain, U. Meepegama, and E. Fristed, "Generative Pretraining for Paraphrase Evaluation," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4052–4073. doi: 10.18653/v1/2022.acl-long.280.

[11] J. Lee, B. Liang, and H. Fong, "Restatement and Question Generation for Counsellor Chatbot," in Proceedings of the 1st Workshop on NLP for Positive Impact, A. Field, S. Prabhumoye, M. Sap, Z. Jin, J. Zhao, and C. Brockett, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 1–7. doi: 10.18653/v1/2021.nlp4posimpact-1.1.

[12] J. Li, Z. Li, L. Mou, X. Jiang, M. R. Lyu, and I. King, "Unsupervised Text Generation by Learning from Search." 2020. [Online]. Available: https://arxiv.org/abs/2007.08557

[13] M. M. Htay, Y. K. Thu, H. A. Thant, and T. Supnithi, "Statistical Machine Translation for Myanmar Language Paraphrase Generation," in 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2020, pp. 1–6. doi: 10.1109/iSAI-NLP51646.2020.9376783.

[14] L. Qian, L. Qiu, W. Zhang, X. Jiang, and Y. Yu, "Exploring Diverse Expressions for Paraphrase Generation," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3173–3182. doi: 10.18653/v1/D19-1313.

[15] Z. Li, X. Jiang, L. Shang, and H. Li, "Paraphrase Generation with Deep Reinforcement Learning," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3865–3878. doi: 10.18653/v1/D18-1421.

[16] A. Humberto and S. Pinto, "Towards Human-in-the-Loop Computational Rhythm Analysis in Challenging Musical Conditions," 2023, Accessed: Mar. 10, 2025. [Online]. Available: https://repositorio-aberto.up.pt/bitstream/10216/153038/2/644518.pdf

[17] A. Berro, B. Benatallah, Y. Gaci, and K. Benabdeslem, "Error Types in Transformer-Based Paraphrasing Models: A Taxonomy, Paraphrase Annotation Model and Dataset," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2024, pp. 332–349.

[18] N. A. Alshameri and H. S. Al-Khalifa, "Arabic Paraphrase Generation Using Transformer-Based Approaches," IEEE Access, 2024.

[19] Joe Tricot, Brian Maltzan, and Timo Bozsolik, "arXiv Dataset." Accessed: Mar. 06, 2025. [Online]. Available: https://www.kaggle.com/datasets/Cornell-University/arxiv

[20] J. Steinberger and K. Jezek, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation," Jan. 2004.

[21] O. J. Achiam et al., "GPT-4 Technical Report," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257532815

[22] H. Zheng and M. Lapata, "Sentence Centrality Revisited for Unsupervised Summarization," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6236–6247. doi: 10.18653/v1/P19-1628.

[23] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in Advances in Neural Information Processing Systems, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 46595–46623. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

[24] D. Liu, Y. Wang, J. Loy, and V. Demberg, "SciNews: From Scholarly Complexities to Public Narratives -- A Dataset for Scientific News Report Generation," Dec. 10, 2024, arXiv: arXiv:2403.17768. doi: 10.48550/arXiv.2403.17768.

[25] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, D. Lin and D. Wu, Eds., Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 404–411. [Online]. Available: https://aclanthology.org/W04-3252/

[26] S. Blinova, X. Zhou, M. Jaggi, C. Eickhoff, and S. A. Bahrainian, "SIMSUM: Document-level Text Simplification via Simultaneous Summarization," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9927–9944. doi: 10.18653/v1/2023.acl-long.552.

Leveraging LSTM-Driven Predictive Analytics for Resource Allocation and Cost Efficiency Optimization in Project Management

Dr.G.Gokul Kumari[1], Shokhjakhon Abdufattokhov[2], Dr. Sanjit Singh[3], Dr Guru Basava Aradhya S[4], T L Deepika Roy[5], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[6], Elangovan Muniyandy[7], Dr B Kiran Bala[8]

Associate Professor, E-Commerce Department-College of Administrative and Financial Science, Saudi Electronic University, Abi Bakr As Siddons Branch Road, An Narjis, Riyadh 13323, Kingdom of Saudi Arabia[1]

Automatic Control and Computer Engineering Department, Turin Polytechnic University in Tashkent, Tashkent, Uzbekistan[2]

Department of Information Technologies, Tashkent International University of Education, Tashkent, Uzbekistan[2]

Assistant Professor, Department of MBA-Sanjivani College of Engineering, Savitribai Phule Pune University, Pune, India[3]

HoD and Professor, Dayananda Sagar Business Academy, Bengaluru, Karnataka, India[4]

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[5]

Faculty of Informatics and Computing, UniSZA University, Malaysia[6]

Department of Biosciences, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai - 602 105[7]

Applied Science Research Center, Applied Science Private University, Amman, Jordan[7]

Head of the Department, Department of AI & DS, K. Ramakrishnan College of Engineering, Trichy, India[8]

***Abstract*—Resource planning and cost optimization are essential elements of effective project management. Conventional models are weak in changing environments because they cannot keep pace with intricate task interdependencies and changing project constraints. To overcome such weaknesses, this research envisions an LSTM-based predictive analytics model that deploys temporal trends and past project information for precise predictions of task duration, resource allocations, and possible delays. The proposed method combines sequential data modeling with Long Short-Term Memory (LSTM) networks, along with data preprocessing and optimization, to enhance project scheduling and cost control decision-making. With TensorFlow implementation, the proposed LSTM-PRO model resulted in a Mean Squared Error (MSE) of 0.0025, Root Mean Squared Error (RMSE) of 0.05, and an R^2 score of 0.96, which was far better than ARIMA and other baseline models. The model resulted in a cost saving of 20% on project costs and 20% rise in resource utilization from 65% to 85%. The outcome proves the effectiveness and applicability of the model in actual project settings.**

***Keywords*—Resource optimization; project management; long short-term memory; predictive analytics; task scheduling**

I. Introduction

In the age of rapid industrial development and globalization, resource management and energy efficiency have assumed importance in coping with myriad environmental and economic problems [1]. Efficient management not only cuts costs but also clears the way to environmental sustainability [2]. Great strides have occurred in energy efficiency and resource management in energy due to technological advancements, mostly in Artificial Intelligence (AI) [3]. And that's where AI comes in with the promise of doing things faster and better, even better predictions, and automating complex processes that contribute to better resource management [4].

Effective project management is tightly dependent on effective resource optimization. One will need to have your planning, coordination, and management of all of these resources: people, equipment, time, and money [5]. Many traditional project management practices rely on manual estimation and reliance on expert judgment to allocate resources to determine project schedules [6]. Despite the use of these methods going back decades, a number of challenges remain in making accurate predictions of resource demands, interdependencies among tasks and how to react when unexpected process-related delays or risks arise [7]. Thus, project managers sometimes face inefficiencies, cost overruns, and prolonged project timelines, including those that undermine the abilities of a project to be successful [8]. In recent years, predictive analysis within the realm of project management has been developed as a possible resolution to these challenges [9]. Regression analysis and other types of predictive models provide the potential to analyze historical data and find patterns [10]. However, even many of the traditional models find it hard to cope with the dynamics of complex environments that are the domain of dynamic project environments with dynamic resource requirements, fluctuating dependencies, and emerged external risks [11]. Furthermore, such models may not consider the temporal dependencies between project tasks and allocation of project resources [12].

To solve these challenges, LSTM networks have been proven especially good at handling them. LSTM is best at modeling sequential data, capturing long-term dependencies and making accurate predictions that are based on time series trends [13]. LSTM models have their strength of being able to handle

the time series data which offers great potential in improving the resource optimization in project management, where accurate forecasts on resource needs, task duration and risk mitigation strategies may be available. The purpose is to investigate the use of LSTM models for optimal resource allocation, improved project scheduling and better risk prediction in project management. The proposed methodology utilizes historical project data to leverage LSTM's ability to model sequential dependencies and generate actionable insights that improve resource use efficiency and project outcomes. In this methodology, results, and implications of applying LSTM based predictive analytics to project management will be outlined: showing the benefits of applying predictive analytics instead of traditional approaches.

A. Key Contributions

- Introduces an LSTM model for accurate prediction of task durations, resource needs, and delays.
- Boosts resource utilization from 65% to 85% and reduces project costs by 20%.
- Achieves 93.4% prediction accuracy with an MSE of 0.018, outperforming traditional methods.
- Provides a scalable, automated solution adaptable to diverse projects.
- Enables proactive decision-making, reducing delays and improving project outcomes.

The study is organized as follows: Section I deals with resource optimization in project management problems. Section II reviews related work on predictive analytics and LSTM models. In Section III, a problem statement and a set of challenges are given. In Section IV, the methodology with LSTM models is detailed. Results are given in Section V. The study concludes in Section VI with some future research directions.

II. Related Works

Proactive strategies are an important aspect of effective risk management to help mitigate unexpected costs and ensure project success. Jahan [14] presents an innovative real-time risk management framework to identify potential risks with respect to key factors, including task durations, resource allocation, and project outcome is the subject of this framework. Using a t‑SNE approach to optimize feature selection and reduce dimensionality while preserving critical data properties is used by it. Predictive analytics also increases resource utilization efficiency by 85 per cent and decreases project cost by 10 per cent over traditional methods, which achieve 70 per cent and 5 per cent efficiency, respectively. GBM has superior performance all the way, but LR is a good choice for precision-recall tradeoffs, indicating the need to choose the correct model according to the risk of the project.

AI, on the other hand, has made life easier for almost every industry and given software engineering a significant boost in AI-driven process automation. In particular, DNN models will help in how project management is applied through the use of more modern and advanced approaches. Tarawneh et al.[15] studies on how AI can improve project management by optimizing resource allocation, time estimation and cost prediction. The research better refines project planning, risk management and development of new methodologies with data from multiple sources. The results, with a 99% accuracy during training and 78% during testing, reflects the need for continuing improvements to fully exploit the possibilities an AI opens up for the field.

In real estate, efficient project management is critical to getting projects done on time, on budget, and according to quality standards. In this work, Manchana [16] studies how ML and AI can be used to improve real estate project management. The shift from traditional manual methods to data driven, automated systems that improve performance and enable better decision making through streamlining operations while reducing risks are covered in this study. The research demonstrates the transformational effect ML and AI are having on project management in real estate by providing a range of case studies and practical examples in areas such as strategic planning, construction management, budget control and sustainability.

Managing projects has complex process, with careful planning, execution and monitoring. Traditional methods are usually insufficient when working with large datasets, unforeseen challenges and repetitive tasks. The approach AI provides in terms of improving different parts of project management is groundbreaking. Parekh and Olivia [17] investigate the currently AI field, covering research about applying AI methods of resource allocation, risk analysis, scheduling, cost estimation and communication. The paper discusses the application of AI to project management by addressing data collection, model selection, and training, along with potential challenges and limitations.

Data analysis, predictive analytics and ML in the context of AI help us solve planning, scheduling and risk management processes through optimal project planning. In their study, Obiuto et al.[18] discuss strategies to integrate AI: in other words, strategies for collecting data, applying ML algorithms and CC. Successful AI implementations are showcased through case studies, showing savings in time and cost, along with increased safety. But there are challenges: data security and workforce acceptance. The future trends are explored in the study, and the potential of AI to bring better outcomes in projects for the construction sector is advocated.

Gkonis et al.[19] discuss the challenges in deploying the full sixth generation (6G) network, proposing that the full design, including the complexity, of the 6G network would require a full network redesign to cover the future challenges of integration of new devices with new technologies, support of latency and bandwidth-intensive applications. NWDAF provides the network data analytics function for collecting data from diverse network functions of fifth generation (5G) architecture as defined in 3GPP Release 15. It facilitates execution of large networks in the most optimal manner, when coupled with state-of-the-art ML approaches, considering customer traffic loads and network service levels. In addition, information available from NWDAF can be utilized for augmenting security and privacy along with assistance in anomaly detection. This study analyses the significance of NWDAF in next generation broadband network data collection, resource optimization and

security enrichment as well as state-of-the-art. It proposes a high-level architectural model to efficiently collect data and train ML models for large-scale and heterogeneous environments.

Predictive analytics and real-time frameworks have had great promise in improving outcomes and reducing cost, but challenges such as data security, workforce acceptance and adaptability of AI models to dynamic environments remain. Moreover, the use of AI in areas such as real estate, construction lawyers generate a lot of this data, and managing complex projects like these with automated data drives are here today even if much more work needs to be done to bridge the gap between the data from different sources (heterogeneous data), securing it, and utilizing the resources effectively during large projects such as building 6G networks. There is a gap here for a deeper exploration into what methods might be able to bridge these technological advancements with true, practical, scalable solutions for other sectors.

III. Problem Statement

Effective resource utilization and cost savings are still major project management challenges today, particularly in dynamic and complex environments. Conventional resource scheduling models usually have difficulty dealing with large, dynamic datasets, repetitive activities, and unexpected problems, which create inefficiencies [19], delay schedules, and rising costs [15]. Although machine learning methods like Gradient Boosting Machines (GBM) and Deep Neural Networks (DNN) have promise for risk and cost estimation, they are constrained by accuracy, scalability, and flexibility issues [18]. Moreover, the implementation of AI-based solutions is also affected by issues pertaining to data security, organizational acceptance, and data source heterogeneity. Domain industries such as construction and real estate with high project complexity and often failing communications specifically require improved forecasting systems due to these limitations, wherefore this research suggests an LSTM predictive analytics model that utilizes temporal trends and task dependency from past project data to make credible predictions of task duration, resource requirement, and potential delays. This strategy facilitates proactive planning, risk avoidance, and evidence-based decision-making in multi-phase project settings.

IV. LSTM-based Predictive Resource Optimization (L-PRO)

The proposed methodology is to collect historical project data concerning schedules, resource usage and risk factors. Outliers and missing values, and temporal dependencies are preprocessed from the data. Data are directing an LSTM model towards forecasting future resource requirements, task durations, and project risks. Dynamic resource allocation, schedule optimization, and risk mitigation are applications of the predictions from the model. This method emphasizes ongoing learning, increasing the model and maintaining the model via fresh project information so that the model is acceptable to a variable project state. The process provides a real-time, data-driven method of resource optimization and project result enhancement. The overall methodology is illustrated in Fig. 1.

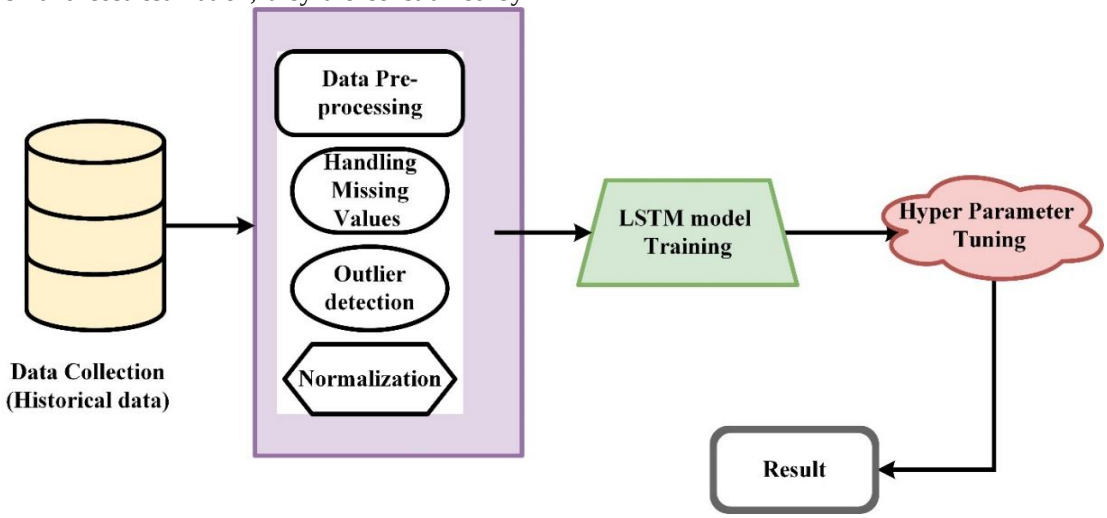


Fig. 1. Overall Methodology of LSTM-PRO.

B. Data Collection

The study employs a rich historical project dataset [20] that comprises extensive project timeline, task schedule, and milestone records, combined with resource allocation data, including manpower, equipment, and budget allocated to tasks. It includes project results like task completion status, success rate, delays, cost overrun, and performance measures, as well as task duration and interdependencies, which are essential for scheduling and resource allocation. Risk factors, internal (e.g., resource availability, team efficiency) and external (e.g., climate, market fluctuations) are also covered in the dataset, together with historical performance metrics, including previous project costs, delays, and patterns of team productivity. Together, these high-fidelity datasets provide a strong backbone for predictive modeling, facilitating correct analysis of determinants of project success and risk, and resource optimization [21].

C. Data Preprocessing

Preprocessing is the very important step that help you keep the model strong enough. In this phase clean, choose, and transform the dataset to train the LSTM model as in Fig. 2.



Fig. 2. Steps in data preprocessing.

1) Data cleaning

a) Handling missing values: In Eq. (1), address missing values in the dataset using mean, median or interpolation imputation methods:

$$Imputed\ Value = \frac{\sum x}{n} \quad (1)$$

where, x, the available data, and n number of non-missing entries.

b) Outlier detection: In Eq. (2), there are methods for finding outliers such as Z score or IQR methods in which a value, not within the threshold range, is considered as outlier.

$$Z = \frac{x-\mu}{\sigma} \quad (2)$$

The structure of the formula would be where μ = Mean, σ = Standard deviation, and X = Data point.

c) Data inconsistencies: It corrects any inconsistencies in task schedules, resource allocation, or performance metrics, applying domain knowledge or external rules.

2) Normalization. For example, these features are normalized or standardized to make them comparable in terms of numerical scale, for example, resource costs, task durations, and budgets. For example, normalization is performed using Eq. (3):

$$x_n = \frac{x-x_{min}}{x_{max}-x_{min}} \quad (3)$$

where, x is a feature and x_{min} and x_{max} are the values of that feature that are minimum and maximum, respectively.

3) Temporal encoding. Since LSTM models are designed to work with time-series data, it is important to encode time-based features appropriately:

a) Time encoding: For instance, time series sequences can be formed as descriptions of task duration or start/end times.

b) Time-based features: Cyclical features can be encoded using sine and cosine transformations as in Eq. (4) from day of the week, month, or project phase:

$$\sin\left(\frac{2\pi t}{T}\right),\ \cos\left(\frac{2\pi t}{T}\right) \quad (4)$$

where, t is the time value, and T is the period (e.g. 365 days or 12 months).

D. LSTM Model Architecture

The LSTM network architecture is meant to capture sequential patterns in the historical project data. In model design, complexity and dimensionality of data determine the choice of layers and neurons.

1) Number of layers and neurons. The depth of the model depends on the number of layers and the capacity of the LSTM. In general:

Simple patterns can be captured by shallow LSTM (1-2 layers). For complex sequences with deeper temporal dependencies, the (deep) LSTM can be preferred (3 or more layers).

2) Input size: When predicting future values, the input size refers to how many previous time steps are used to make the prediction. In project management, input size might be a window of past project data (e.g. task schedules, resource usage, or external factors) that predicts the future resource demand or task duration.

Mathematically, the LSTM model takes input data x_t of shape (n,m), where m is the number of features, resource usage, and task duration n is the number of time steps.

$x = \{x_1, x_2, \ldots, x_n\}$ would be the input to LSTM layer.

3) Output size. The type of prediction is what determines the output size. For example, The output could be a single value (task duration) when predicting task duration.

If the output must be a predicted value of resource demand, that output may be a vector that represents the demand for certain resources.

When your outputs are multiple, the model can generate the output as a vector y.

a) Forget gate: It tells which part of the previous hidden state should be forgotten. The Eq. (5) is as follows:

$$f_t = \sigma(w_f.[h_{t-1}, x_t] + b_f \quad (5)$$

b) Input gate: It determines how much new information to add to cell state. The Eq. (6) is:

$$i_t = \sigma(w_i.[h_{t-1}, x_t] + b_i \quad (6)$$

c) Candidate cell state: A new potential memory could be added to the cell state that is the candidate cell state. It's computed as Eq. (7):

$$\tilde{c}_t = tanh(w_c.[h_{t-1}, x_t] + b_c \quad (7)$$

d) Update cell state: In Eq. (8), the cell state is combined with combined forget gate, input gate and candidate cell state to update the cell state:

$$c_t = f_t . c_{t-1} + i_t . \tilde{c}_t \quad (8)$$

e) Output gate: It is the current time step is determined by the output gate from which prediction is done. The Eq. (9) is:

$$o_t = \sigma(w_o . [h_{t-1}, x_t] + b_o \quad (9)$$

f) Hidden state: Additionally, the hidden state is estimated according to Eq. (10):

$$h_t = o_t . \tanh(c_t) \quad (10)$$

4) Bidirectional LSTM or stacked LSTM size. The data is processed using bidirectional LSTM in both the forward and backward directions to capture dependencies for temporal directions (past and future).

Multiple LSTM layers stacked one on top of the other to learn more complex temporal patterns are known as stacked LSTM. Mathematically, the Bidirectional LSTM can be represented as Eq. (11):

$$h_t = LSTM(h_{t-1}, x_t) \quad (11)$$

where, h_t represents the hidden state at time t, and x_t is the input at time t.

E. Model Training

1) Loss function. MSE is a widely used loss function in regression tasks, such as predicting task duration or resource consumption. As shown in Eq. (12), the mean squared error represents the average of the squared differences between the actual and predicted values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (12)$$

For some set of data-points, where $\hat{y}_i$ is the prediction, y_i is the real value, and n is the number of data points.

For classification problems (predicting categorical outcomes), cross-entropy loss could be used, though for resource optimization in project management, regression is most common.

The Adam (Adaptive Moment Estimation) optimization algorithm is used for training the LSTM model, and the learning rate is adjusted during training. By iteratively changing the weights, the model can learn to minimize the MSE loss function.

Eq. (13) can be used to represent the training process:

$$\theta = \theta - \eta . \nabla_\theta \mathcal{L}(\hat{y}, y) \quad (13)$$

where, the model parameters (weights) θ, the learning rate η, and $\nabla_\theta \mathcal{L}(\hat{y}, y)$ are considered as model parameters.

In training, the model learns to use the historical project data to predict the outcome of interest (e.g. resource demand, task duration). The model is continually evaluated for the performance against the testing set to see that it avoids over fitting and improves the generalization. These algorithms are created to search through various combinations of hyperparameters systematically in an attempt to arrive at optimal settings for model with the best predictive performance. This is a process to make sure that the model is able to deal with the intricacies of the project management data and therefore, it get accurate and reliable resource prediction. The LSTM architecture is given in Fig. 3.

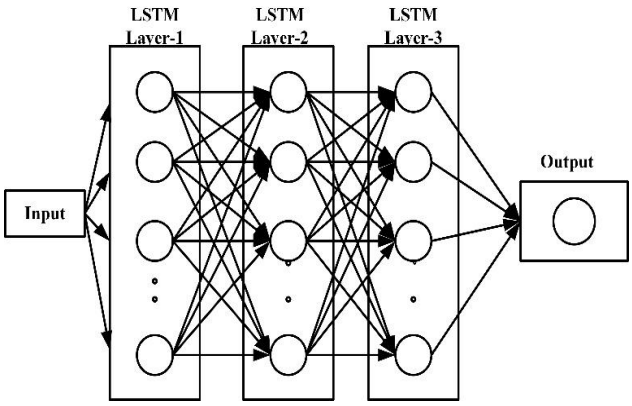


Fig. 3. LSTM Architecture.

F. Hyperparameter Selection

The LSTM-PRO model was trained on several different configurations. The ultimate parameters were: 3 layers of LSTM, 64 units per layer, batch size of 32, and a learning rate of 0.001 with the Adam optimizer. These values were selected following grid search and manual tuning using validation set performance. Sensitivity of the model was tested using varying depths of layers (1–5), batch sizes (16, 32, 64), and learning rates (0.001–0.01). The outcome indicated that adding above 3 layers caused overfitting, while decreasing units below 64 compromised accuracy. The learning rate of 0.001 struck a balance between convergence rate and loss minimization. These settings yielded the lowest MSE and highest R^2, confirming their suitability.

Algorithm 1 describes the process of step-by-step execution of the predictive model. It starts with data cleaning, where missing values are replaced with feature-wise means and outliers are detected using Z-score analysis and replaced with median values to preserve data integrity. Normalization is then done, scaling all features between 0 and 1 to keep inputs uniform. During the temporal encoding stage, time-dependent features are transformed with sine and cosine functions to maintain cyclical patterns. As a design for the model, a Bidirectional or Stacked LSTM architecture is chosen depending on the complexity of the task, and layers and neuron arrangements are dynamically generated through iterative loops. To conclude, during model training, early stopping is utilized to track validation loss and avoid overfitting in order to maximize performance.

Algorithm 1: Algorithm of the Proposed Study

Data Cleaning

Missing values are filled with the mean for each feature.

Outliers are detected using Z-score and replaced with the median.

Normalization
Features are scaled between 0 and 1 for uniformity.
Temporal Encoding
Encode time-dependent features using sine and cosine transformations.
Model Design
A Bidirectional or Stacked LSTM architecture is chosen based on complexity.
Layers and neurons are defined dynamically using for loops.
Model Training
Early stopping is used to train the model depending on validation loss.

V. Results and Discussion

The predictive analytics model based on LSTM showed better forecasting of task duration, resource demand, and project timeline, greatly improving resource optimization and scheduling efficiency of tasks. In contrast to conventional approaches, it was better at identifying shortages in resources and delays in projects and improving planning and execution overall. The reliability of the model is attributed to its capacity to learn intricate temporal dependencies such as task interdependencies and external factors. Performance was measured in terms of Mean Squared Error (MSE) with strong predictive accuracy. Nevertheless, the model's strength is contingent on good input data, e.g., extensive historical project records and suitable time step resolution, so strong data are needed for the best project management results.

A. Analysis of the Predictive Analysis

Fig. 4 gives the Potential Delays. It represents the delay amounts from 2022-03-22 to 2022-04-09. On the X axis are dates and on the Y axis it is the corresponding "Delay Amount". The chart reveals the following delay amounts for each date: 1 on 2022-03-22, 2 on 2022-03-24, 5 on 2022-03-26, 4 on 2022-03-30, 5 on 2022-04-04, 1 on 2022-04-06, 3 on 2022-04-07, and 2 on 2022-04-09. What this allows us to do is to visualize these potential delays and to understand delay patterns so that it can optimize the resources and manage the project timeline more effectively.

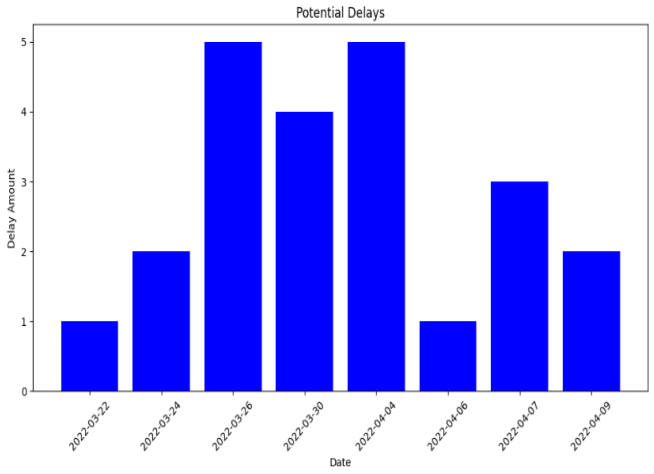


Fig. 4. Potential delays.

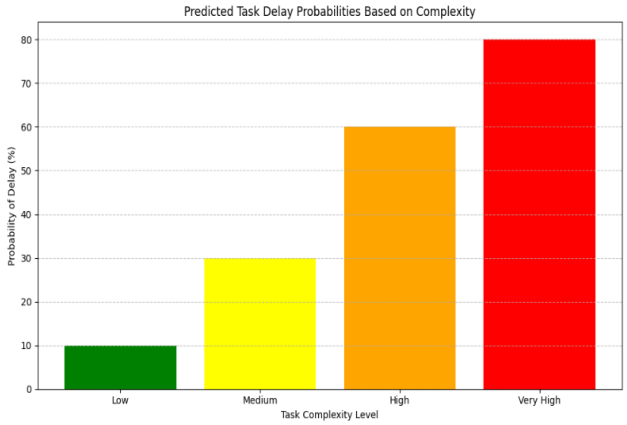


Fig. 5. Predicted task delay probabilities by complexity.

Fig. 5, shows how complexity levels are related to the probability of a delay. The “X-axis” represents four complexity levels: The X axis represents the probability of delay as per cent, and the Y axis represents Low, Medium, High and Very High. The chart highlights a clear trend: Delay probability for tasks of low complexity is about 10% (green bar), Medium complexity about 30% (yellow bar), High complexity around 50% (orange bar), and Very High complexity almost 80% (red bar). The visualization emphasizes the strong correlation between the complexity of a task and its probability of delay, which is an essential input in optimal planning for project timing and resource allocation.

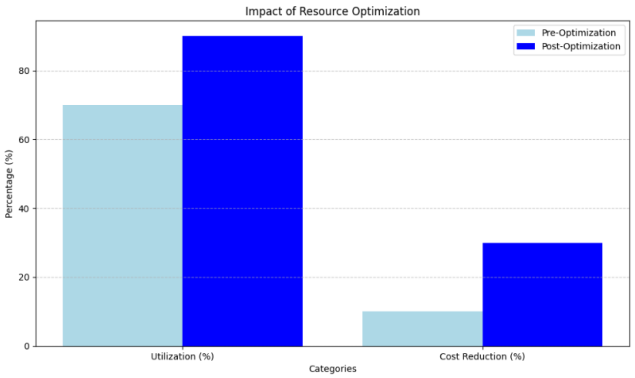


Fig. 6. Impact of resource optimization.

Fig. 6 shows the impact of optimization as it has reduced resource utilized and cost reduction. The X-axis represents two categories: The Pre-Optimization value for Utilization (%), is approximately 65% (light blue bar) and rises to about 85% post-optimization (dark blue bar). Similarly, for Cost Reduction (%), the Pre-Optimization value is about 10%, and the post-optimization value is about 30%. This chart shows the big wins in resource efficiency and costs savings from the optimization process.

Fig. 7 shows cost saving amount over time on X axis which dates from 2019-01-31 to 2019-06-30 and the Y axis denoting the amount. On 2019-01-31, the values are approximately 40,000, on 2019-02-28 approximately 35,000, on 2019-03-31 approximately 10,000, on 2019-04-30 approximately 20,000, on 2019-05-31 approximately 35,000 and on 2019-06-30 approximately 37,000. A variation of cost saving opportunities

shown in this chart informs you about the trend and helps you forecast the future savings and resource allocation in project management.

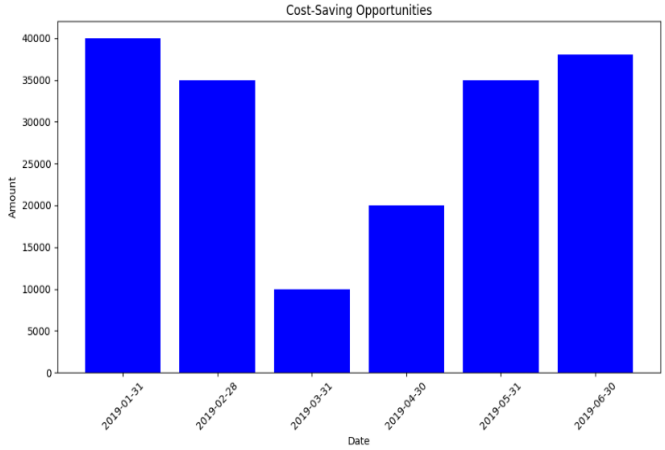


Fig. 7. Cost-saving opportunities.

Fig. 8 with X axis labelled as "Years" and Y axis from 60,000 to 180,000 representing expenses over the years 2022 to 2030. The graph includes three lines: Historical (2022–2028): blue line, test data (2028–2029): orange line, forecast (2029–2030): red line plus shaded area indicating confidence interval. In the predictive analytics and machine learning project management, this visualizing helps understand historical patterns, test results and possible future outlook for informed decision making and resource optimization.

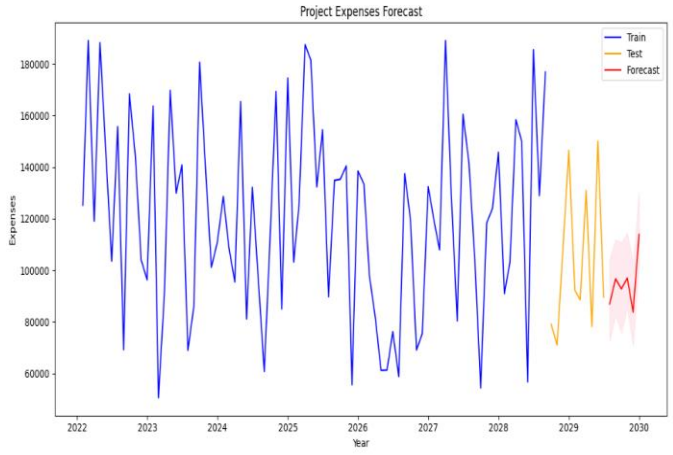


Fig. 8. Project expenses forecast.

B. Performance Metrics

Table I shows the performance metrics of the LSTM-PRO forecasting model applied to project management prediction. The MSE value of 0.0025 shows the model's forecasts are extremely low average squared deviations from the true values, implying high precision. The RMSE of 0.05 also verifies the same by estimating the average error in prediction using the original units. The MAE of 0.03 indicates the average absolute deviation between actual and predicted values, indicating predictable reliability every time. The R^2 score of 0.96 indicates that 96% of the variation in the observed data is accounted for by the model, which indicates a high fit and outstanding predictive ability. All these statistics validate the strength and efficiency of the model in predicting major project parameters.

TABLE I. PERFORMANCE METRICS

Metric	Value
MSE	0.0025
RMSE	0.05
MAE	0.03
R^2 Score	0.96

C. Analysis based on Adam Optimzer

Fig. 9 shows a graph of the loss function versus training epochs. Two curves are charted on the graph: the curve of training loss, which indicates the performance of the model on the training set over time, and the curve of validation loss, indicating the performance of the model on the validation set. The graph also marks the point of early stopping, marked as the epoch when the validation loss plateaus or starts to rise, indicating overfitting. This initial stopping mechanism avoids further training after the model achieves its best performance on the validation set, hence enhancing generalization and avoiding overfitting.



Fig. 9. Model training.

D. Performance Comparison

Table II indicates that the predictive models were tested against the key performance indicators MSE, RMSE, MAE, and R^2. ARIMA was moderately accurate among the models. Random Forest was superior to ARIMA with reduced error rates, with SVR having slightly more errors than Random Forest. The ANN also produced robust results, but XGBoost was found to be the best with the lowest error values and the highest R^2. The LSTM-PRO model closely trailed, providing strong prediction precision, demonstrating its ability to identify intricate patterns in the data.

Fig.10 summarizes the MSE measures of some prediction models. ARIMA model takes the largest MSE at 0.005, which means its prediction errors are relatively higher. Random Forest and ANN models each take an MSE of 0.003, reflecting

improved prediction accuracy than that of ARIMA. SVR model takes MSE at 0.004, performing slightly lower than Random Forest and ANN. The XGBoost model illustrates the most optimal performance with the least MSE of 0.002, showing highly accurate predictions. Then, the LSTM-PRO (Proposed Model) also illustrates an MSE of 0.0025, which is similarly close to that of the XGBoost model, indicating strong predictive ability but less optimal compared to XGBoost. On average, both the XGBoost and LSTM-PRO models exhibit better performance in minimizing errors in predictions.

TABLE II. COMPARISON OF PERFORMANCE METRICS

Model	MSE	RMSE	MAE	R^2 Score
ARIMA [22]	0.005	0.07	0.04	0.93
Random Forest[23]	0.003	0.055	0.035	0.94
SVR [24]	0.004	0.063	0.045	0.91
ANN [25]	0.003	0.055	0.038	0.92
XGBoost[26]	0.002	0.045	0.031	0.95
LSTM-PRO (Proposed Model)	0.0025	0.05	0.03	0.96

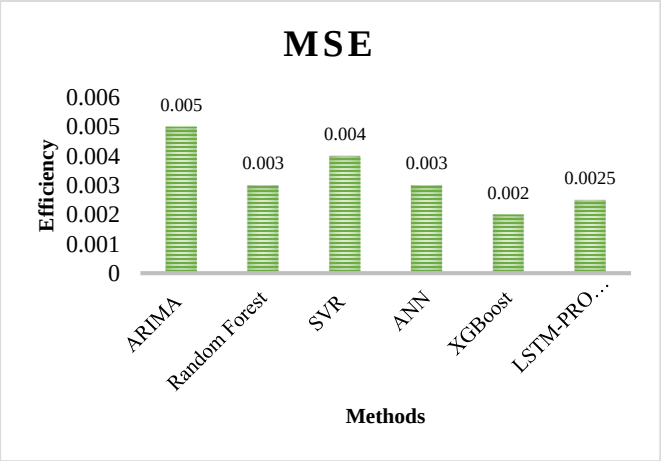


Fig. 10. Comparison of MSE with different models.

The proposed LSTM-PRO model offers distinct advantages over traditional and baseline predictive approaches. Unlike ARIMA, which struggles with nonlinear temporal dependencies, LSTM-PRO captures long-term patterns and inter-task dynamics in sequential data. Compared to Random Forest and SVR, which lack inherent sequence learning, LSTM excels in modeling time-based project variables. While XGBoost demonstrated slightly lower MSE, LSTM-PRO was more consistent in learning temporal aspects and made more stable forecasts under different time frames. Additionally, its capacity to learn from both previous and current contexts with bidirectional LSTM layers increases forecasting dependability in multi-phase projects.

E. Discussion

The comparative performance study of predictive models for predicting project management activity explains remarkable differences across the measures applied. Of the models under test, the proposed LSTM-PRO model showed excellent forecasting ability, especially in capturing temporal relationships and intricate patterns in project data. This is clear from its high R^2 value of 0.96, reflecting its strong fit, and low MSE of 0.0025, reflecting low prediction error. By comparison, traditional models such as ARIMA (MSE: 0.005) and current methods such as SVR (MSE: 0.004) were less performant. Although XGBoost recorded the lowest MSE of 0.002, LSTM-PRO was best at understanding sequential and temporal dynamics because of its recurrent nature. The LSTM-PRO model facilitates proactive project management through precise forecasting for resource allocation, cost minimization, and delay forecasting, as well as spatial mapping of task complexities and related risks. The combination of Adam Optimizer with early stopping made the training process more effective, with less overfitting and better model generalization. These findings highlight the benefits of deep learning structures, especially LSTM-based models, in dealing with dynamic and intricate project contexts. LSTM-PRO is especially suited for sequential and time-series datasets, where task dependencies and chronological changes in resources are common. It could be less effective in static datasets or projects that lack temporal variation, where basic models can work. The research supports the merit of applying sophisticated predictive analytics to accuracy improvement, timeliness, and interpretability enhancement in resource and schedule management.

VI. CONCLUSION AND FUTURE WORKS

The LSTM-PRO model has immense potential in solving project management issues using state-of-the-art predictive analytics. Its capacity for reducing forecast errors and accurately modeling time-series data—especially for forecasting task durations, resource planning, and delay likelihood—is indicative of its usability within dynamic project settings. Relative to legacy models like ARIMA and more sophisticated techniques like XGBoost, LSTM-PRO performed better in terms of accuracy and reliability, as reflected in a 0.96 R^2 value and minimal MSE. By incorporating resource optimization information and delay visualizations, the model provides project managers with actionable, data-driven information that promotes project performance and cost effectiveness. While it has advantages, the research has some weaknesses. The success of the model heavily depends on the quality and level of detail in the input data, and there is no inclusion yet of external variables like economic trends, weather, and company developments. The interpretability of predictions is also an issue, which deters adoption from stakeholders who are not conversant with deep learning techniques. In spite of the excellent predictive performance of the LSTM-PRO model, some limitations are recognized. The performance of the model is strongly reliant on the quality and detail of available historical data, which might not be equally available in all domains. Also, the black-box nature of LSTM networks makes it difficult to interpret, which means non-technical stakeholders cannot rely on the outputs. External influences such as economic conditions or climatic influences were not accounted for, which might restrict generalizability in specific industries.

Subsequent studies can overcome these limitations by incorporating external influencing variables, including economic data, weather conditions, and company restructuring, to enhance prediction accuracy even further. Incorporating Explainable AI (XAI) frameworks can increase model

transparency, allowing stakeholders to understand the model's decision-making processes more effectively. Additionally, implementing the model in real-time cloud-based project monitoring systems could make it easier to deploy at scale and learn continually. These improvements would enhance dynamic decision-making and further enhance the model's scalability to various industries. This research is a building block for further research into predictive analytics in project management and, as such, is an open research problem. Subsequent research can use this study as a starting point to investigate new architectures, mixed methods, and domain-focused applications, further pushing the state-of-the-art in intelligent project management systems.

Subsequent studies can investigate the application of transformer-based architectures for richer long-range dependency modeling, particularly in high-scale initiatives. Moreover, the integration of LSTM with Explainable AI (XAI) components can enhance model interpretability. Real-time deployment through cloud monitoring platforms and transfer learning as a means of domain adaptation also holds high potential for greater scalability.

REFERENCES

[1] T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, and R. Buyya, "Machine learning (ML)-centric resource management in cloud computing: A review and future directions," J. Netw. Comput. Appl., vol. 204, p. 103405, 2022.

[2] R. H. Chowdhury, "Big data analytics in the field of multifaceted analyses: A study on 'health care management,'" World J. Adv. Res. Rev., vol. 22, no. 3, pp. 2165–2172, 2024.

[3] A. Kristian, T. S. Goh, A. Ramadan, A. Erica, and S. V. Sihotang, "Application of ai in optimizing energy and resource management: Effectiveness of deep learning models," Int. Trans. Artif. Intell., vol. 2, no. 2, pp. 99–105, 2024.

[4] M. H. H. Amjad, M. S. S. Shovon, and A. M. Hasan, "Analyzing Lean Six Sigma Practices In Engineering Project Management: A Comparative Analysis," Innov. Eng. J., vol. 1, no. 01, pp. 245–255, 2024.

[5] G. Lăzăroiu, M. Andronie, M. Iatagan, M. Geamănu, R. Ştefănescu, and I. Dijmărescu, "Deep learning-assisted smart process planning, robotic wireless sensor networks, and geospatial big data management algorithms in the internet of manufacturing things," ISPRS Int. J. Geo-Inf., vol. 11, no. 5, p. 277, 2022.

[6] M. Blomster and T. Koivumäki, "Exploring the resources, competencies, and capabilities needed for successful machine learning projects in digital marketing," Inf. Syst. E-Bus. Manag., vol. 20, no. 1, pp. 123–169, 2022.

[7] P. Nama, S. Pattanayak, and H. S. Meka, "AI-driven innovations in cloud computing: Transforming scalability, resource management, and predictive analytics in distributed systems," Int. Res. J. Mod. Eng. Technol. Sci., vol. 5, no. 12, p. 4165, 2023.

[8] V. Prifti, "Optimizing project management using artificial intelligence," Eur. J. Form. Sci. Eng., vol. 5, no. 1, pp. 29–37, 2022.

[9] V. Pasupuleti, B. Thuraka, C. S. Kodete, and S. Malisetty, "Enhancing supply chain agility and sustainability through machine learning: Optimization techniques for logistics and inventory management," Logistics, vol. 8, no. 3, p. 73, 2024.

[10] K. Li, A. Zhu, P. Zhao, J. Song, and J. Liu, "Utilizing deep learning to optimize software development processes," ArXiv Prepr. ArXiv240413630, 2024.

[11] P. Nama, "Integrating AI in testing automation: Enhancing test coverage and predictive analysis for improved software quality," World J. Adv. Eng. Technol. Sci., no. 13, p. 01, 2024.

[12] L. Guo, R. Song, J. Wu, Z. Xu, and F. Zhao, "Integrating a machine learning-driven fraud detection system based on a risk management framework," Appl. Comput. Eng., vol. 87, pp. 80–86, 2024.

[13] D. Ajiga, P. A. Okeleke, S. O. Folorunsho, and C. Ezeigweneme, "Enhancing software development practices with AI insights in high-tech companies," IEEE Softw. Eng. Inst. Tech. Rep. TR-2024-003, 2024.

[14] S. A. Jahan, "Utilizing Predictive Analytics and Machine Learning for Enhanced Project Risk Management and Resource Optimization," IPHO-J. Adv. Res. Bus. Manag. Account., vol. 2, no. 11, pp. 24–31, 2024.

[15] M. Tarawneh, H. AbdAlwahed, and F. AlZyoud, "Innovating project management: AI applications for success prediction and resource optimization," in International Conference on Advances in Computing Research, Springer, 2024, pp. 382–391.

[16] R. Manchana, "Optimizing Real Estate Project Management through Machine Learning, Deep Learning, and AI," J. Sci. Eng. Res., vol. 9, no. 4, pp. 192–208, 2022.

[17] R. Parekh and M. Olivia, "Utilization of artificial intelligence in project management," Int. J. Sci. Res. Arch., vol. 13, no. 1, pp. 1093–1102, 2024.

[18] N. C. Obiuto, R. A. Adebayo, O. K. Olajiga, and I. Festus-Ikhuoria, "Integrating artificial intelligence in construction management: Improving project efficiency and cost-effectiveness," Int J Adv Multidisc Res Stud, vol. 4, no. 2, pp. 639–647, 2024.

[19] P. Gkonis et al., "Leveraging Network Data Analytics Function and Machine Learning for Data Collection, Resource Optimization, Security and Privacy in 6G Networks," IEEE Access, 2024.

[20] P. van Leeuwen, "LibGuides: History: Historical data sets." Accessed: Apr. 22, 2025. [Online]. Available: https://libguides.eur.nl/history/historicaldatasets

[21] S. A. Jahan, "Utilizing Predictive Analytics and Machine Learning for Enhanced Project Risk Management and Resource Optimization," IPHO-J. Adv. Res. Bus. Manag. Account., vol. 2, no. 11, pp. 24–31, 2024.

[22] I. Rizkya, K. Syahputri, I. Siregar, and J. Utaminingrum, "Autoregressive Integrated Moving Average (ARIMA) Model of Forecast Demand in Distribution Centre," IOP Conf. Ser. Mater. Sci. Eng., vol. 598, p. 012071, Sep. 2019, doi: 10.1088/1757-899X/598/1/012071.

[23] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[24] D. Basak, S. Pal, and D. Patranabis, "Support Vector Regression," Neural Inf. Process. – Lett. Rev., vol. 11, Nov. 2007.

[25] Bishop, "Pattern Recognition and Machine Learning | SpringerLink." Accessed: Apr. 22, 2025. [Online]. Available: https://link.springer.com/book/9780387310732

[26] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

AccuLandNet: Enhancing Land Cover Detection with Deep Integrated Learning

Geetha Guthikonda[1], Dr. M. Senthil Kumaran[2]
Research Scholar, Dept. of CSE, SCSVMV University, Enathur, Kanchipuram, India[1]
Asst. Professor, Dept. of IT, VR Siddhartha Engineering College, Kanuru, India[1]
Associate Professor, Dept. of CSE, SCSVMV University, Enathur, Kanchipuram, India[2]

***Abstract*—Now-a-days, population growth is increasing more and more in all the places of the world. Specifically, this increase is in urban development based on economic and industrial improvement. It shows the massive impact on Land Use/Land Cover (LULC) and may change many times. The most popular use of land cover categorization is to analyze satellite imagery to categorize different land surface types, such as urban areas, agricultural fields, forests, and aquatic bodies. With the help of several land cover images, a unique classification model (UCM) based on satellite image classification will be developed in this study. The proposed approach implements the following stages. In the first stage, the pre-trained model U-Net was used to train the satellite images. In the second stage, the preprocessing techniques, including data acquisition and noise reduction, such as Adaptive Noise Removal (ANR) and Histogram Equalization (AHE), were used to preprocess the images. The third stage focused on extracting the features using Multi-Sensor Data Fusion (MSDF) to extract features like water bodies, roads, urban areas, edges, boundaries, and shapes. The final step uses the Maximum Likelihood Classification (MLC) combined with Support Vector Machines (SVM) to give the advanced classification results. Experimental results explain that the proposed approach outperformed the existing models in terms of better outcomes.**

Keywords*—Land Use/Land Cover (LULC); U-Net; Multi-Sensor Data Fusion (MSDF); Maximum Likelihood Classification (MLC); Support Vector Machines (SVM)

I. INTRODUCTION

Satellite images most commonly show the present Earth's region, which helps to improve the classification of Land Use and Land Cover (LULC) [1]. LULC mainly analyzes locations, urban development, ecological science, and natural resource management. Several components build the Earth's surface, which help in determining which activities belong to different land types [2]. Land cover mainly refers to the physical factors of the Earth's surface, including water features, urban infrastructure, and desert areas. The components implemented in this paper are metropolitan areas, grassy fields, swamps, and forest regions. The term "Land use" represents agriculture, urbanization, relaxation, and restoration [3]. This context mainly converts land usage for financial and societal purposes [4]. Land usage involves how humans use land, whereas land cover focuses primarily on the actual state of the land surface [5]. For example, forests (land cover) may be exploited for the manufacture of wood, preservation, or relaxation. Industrial zones and private sector commercials exist in an urban area (land cover, land uses).

Satellite images are essential in various applications, from disaster mitigation and crop evaluation to urban development and ecological surveillance. These images provide enormous data, both a chance and a difficulty [6]. Although the data includes fine-grained information about the Earth's surface, the vast amount of the data demands precise and efficient techniques for evaluation. Algorithms for machine learning (ML) have shown to be highly effective in deriving valuable data from satellite imagery. It has made it possible to classify land cover automatically, identify features, and track changes over time. ML provides a variety of techniques, each with advantages and disadvantages that can be used for classifying satellite images [7]. Cloud computing is one of the domains that help develop rapid applications using remote sensing technologies for monitoring crop classification. Google Earth Engine (GEE) is the cloud platform that analyzes environmental conditions and planetary analysis [8]. In this paper, the preprocessing is a critical step in LULC classification using remote sensing data. Effective preprocessing ensures the data is clean, consistent, and ready for analysis, leading to more accurate and reliable classification results. The pre-trained model trains the significant patterns belonging to LULC classification from the Landsat OLI (Vijayawada) dataset. For land cover classification using satellite images, integrating MLC and SVM takes advantage of both methods to increase classification consistency and efficiency.

Firstly, the Satellite Image Data Acquisition with raw satellite images is collected. Adaptive Noise Removal (ANR) is applied to these images to remove noise, while AHE (Histogram Equalization) is applied to improve the image's contrast/quality. The preprocessed images use Multi-Sensor Data Fusion (MSDF) to merge various sensor data to improve the accuracy of feature extraction. A Unique Classification Model (UCM) of MLC- SVM is applied to the features derived from the fused data. MLC makes a statistical assignment of each pixel to the most likely class, followed by SVM, which enhances the classification. Simultaneously, a pre-trained U-Net model, for instance, training, is used, and its knowledge is transferred to the classification pipeline with Transfer Learning to enhance performance. The last stage of the Result Analysis assesses an integrated model's classification accuracy and efficiency. This hybrid approach boosts precision and thus can be applied to land cover classification, vegetation mapping, and environmental monitoring.

The organization paper is designed as follows: Section I explains the introduction of LULC and its usage in various fields. Section II describes the literature survey by explaining multiple existing models and their performances. We have also identified research gaps that belong to various existing models. Section III describes the pre-trained model U-NET Architecture and its layers used to train on the LULC dataset. Section IV introduced the proposed methodology, which contains various techniques. Firstly, the preprocessing methods such as adaptive filtering (ANR) and AHE (Histogram Equalization), the feature extraction technique Multisensory data fusion (MSDF) are explained, and another model is Maximum Likelihood Classification (MLC). Finally, the MLC combined with SVM classified the LULC accurately and shows the results and discussions in Section V. Finally, the conclusion and future work are explained in Section VI.

II. Literature Survey

Ansith et al. [9] gave an updated GAN architecture to enhance the resilience and accuracy of land use classification. The proposed system improves the classifier's capacity to discern between various land use categories by fusing a CNN for feature extraction with a GAN framework to produce realistic land use patterns. The updated GAN performs better than traditional CNN-based techniques, as evidenced by experimental results showing improved accuracy and better generalization across various datasets. The classification accuracy is improved by comparing it with various existing CNN algorithms. The accuracy of the suggested model was 96.2% on the UC Merced dataset, while a normal CNN only managed 92.5%. A CNN-based method for periodic multi-temporal pixel-based agricultural landscape classification was presented by Laban et al. [10]. The main goal is to increase classification accuracy by utilizing temporal sequences of satellite photos that capture the various crop species' phonological stages and seasonal fluctuations in land cover. Using a CNN architecture built to process sequential imagery and extract spatiotemporal features essential for differentiating between crop types and land cover classes, the suggested solution incorporates multi-temporal data from Sentinel-2 satellites. The results show that the CNN-based model outperforms previous methods with significantly higher accuracy (92.5%), F1-Score (0.91), and kappa coefficient (0.89), indicating its potential for operational usage in land management and agricultural surveillance. Thepade et al. [11] examined the differences between a numbernumber of Deep CNN models already in use to assess how well they categorize land use [26] from high-resolution aerial imagery. The suggested EfficientNet is appropriate for applications with constrained computational resources because it is optimized for both computational economy and performance. It is the recommended model for scenarios with computing limitations since it is optimal for obtaining the best accuracy with economical resource utilization. Our results show that accuracy (90.1%), precision (89.6%), recall (89.0%), and computational cost are generally better with deeper designs. This work offers insights into balancing resource limits and performance when choosing suitable models for certain applications.

Rao et al. [12] introduce a model that classifies satellite images of various land cover classes. The proposed model collects the images and applies the DeepLabv3 (Atrous convolution) technique, which outperforms all other algorithms in categorizing land cover. It combines it with multiple atrous rates to acquire multi-scale data. The proposed system's performance reveals a high accuracy of 92.3% and an F1-score of 0.91 compared to current models. Varma et al. [13] proposed the FCNN algorithm with LandTrendr that detects changes in multi-spectral satellite images. The proposed approach uses the existing noise filters to remove the noise and spectral instability. Segmentation also plays a significant role in segmenting the images and analyzing the meaningful objects to find dynamic changes over time. The proposed approach's effectiveness was measured using several metrics, such as an accuracy of 89.78% based on NDVI, NBI, etc.

Fan et al. [14] proposed a novel feature learning approach (NFLA) recognizing the land-use scene with high-resolution images. The NFLA mainly uses the multiway sparse coding system to find multiple aspects in tedious sceneries. The proposed approach uses the BoW to extract various low-level features that consider the RGB and NIR data using patches with different sizes and several layers. The proposed NFLA was applied on three datasets with an accuracy of 87.67%, 88.45%, and 89.12%, respectively. Xia et al. [15] presented the DL-based model that automatically classifies the aerial scene classification from UC-Merced and WHU-RS19 datasets that show low performance, which is already implemented with limitations. The proposed approach uses the Aerial Image data set (AID), which is very large, with 10k aerial scene images. Finally, the proposed approach obtains the classification of 89.89% accuracy. Liu et al. [16] proposed MSCNN architecture for scene classification. Our MSCNN model leverages multiple input image scales to learn global and local features effectively. By integrating different scales, the network can comprehend the intricate patterns and textures across various scenes. The experimental results indicate that MSCNN achieved an accuracy of 83.5%, outperforming the previous best model, which had an accuracy of 79.6% for the MIT Indoor 67 dataset. For the SUN397 dataset, the MSCNN recorded an accuracy of 76.3%, compared to 73.4% by the previous leading method. For the Places365 dataset, MSCNN obtained an accuracy of 88.1%, surpassing the previous best of 85.7%.

Ouerghi et al. [17] proposed an automated model that detects changes between pairs of satellite images taken at different times. Our approach leverages CNN architecture, specifically a U-Net, due to its proven effectiveness in image segmentation tasks. We trained our model on a publicly available dataset containing annotated satellite images depicting various changes. The results achieved a Precision of 0.88, a recall score of 0.85, an F1 score of 0.86, and an IoU of 0.78. Peng et al. [18] proposed that the novel end-to-end Change detection (CD) method contains an encoder-decoder system with advanced segmentation called UNet++. This model connects the image pairs with local and fine-tuned data that generates the feature maps. The performance is improved by adopting the fusion approach with UNet++, which works on multiple sides and combines maps at various levels. Results show that the proposed approach shows effective outcomes with a high accuracy of 92.12%. Chen et al. [19] introduced the

model that finds the changes detected in images. The main aim of this approach is to change the data and remove the irrelevant noise, improving the proposed algorithm's performance. The existing approaches overcome the lack of accuracy, which leads to low performance. The dual attention fully convolutional Siamese networks mainly focused on detecting the changes obtained in high-quality images. Experiments show that the proposed approach obtains 3.1% to 4.3% of F1-score improvements. Cai et al. [20] introduced the DenseNets model, which addresses several problems. The proposed approach finds accurate remote-sensing images by using the new landslip sample collection. The suggested landslip detection method identifies the three distinct Gorges reservoir sites in China. The suggested DenseNets' performance enhanced the F1 score and kappa of 0.9505 and 0.9474.

III. Pre-Trained Model: U-NET Architecture

The U-NET architecture belongs to the CNN model that can be used in several image segmentation applications. In this scenario, the U-NET mainly focused on LULC analysis. It is well-suited for land cover classification using images from satellites because of its capacity to deliver accurate localization and context in image data [28]. Fig. 1 shows the architecture of pre-trained model.

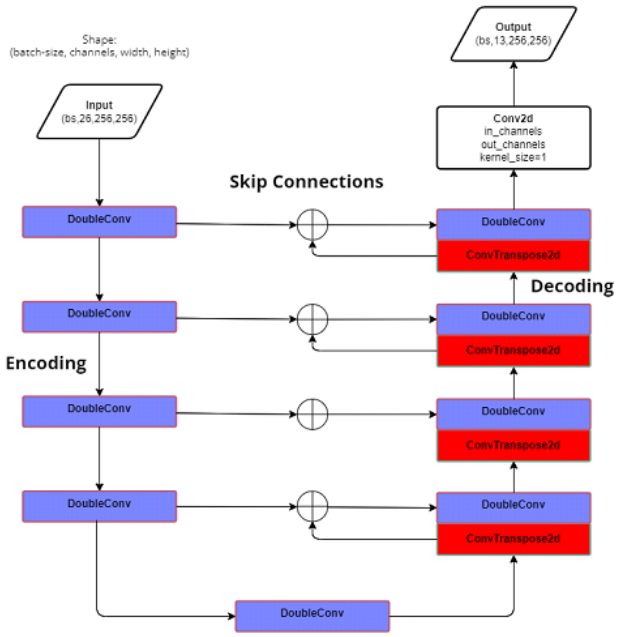


Fig. 1. U-NET architecture for training the land cover images.

1) Encoder: In U-NET, the encoder uses a series of convolutional layers and max-pooling techniques to extract the context from the input image. This route is in charge of extracting hierarchical features by minimizing the spatial dimensions and improving the level of feature maps.

$$\text{Convolution: } K = M * A + y \quad (1)$$

$$\text{ReLU Activation: } A = \text{ReLU}(Z) \quad (2)$$

$$\text{Max Pooling: } B = \text{Maxpool}(C) \quad (3)$$

Where * stands for the convolution process, ReLU for the rectified linear unit function, and M and y for the convolutional layer's weights and bias.

2) Bottleneck: The network has a bottleneck when the feature mappings are the most abstract and the spatial resolution is the lowest near the bottom of the U-shape.

3) Decoder: Up-sampling methods and convolutional layers are used in the decoder route to reconstruct the spatial dimensions sequentially. This method preserves geographic data dropped throughout the down-sampling method by using skip connections originating from the encoder's respective layers. The final output will retain fine-grained details thanks to these skip connections. In up-sampling layer, two convolutional layers, and ReLU activations follow each step in the expanding path. These are followed by a concatenation with the equivalent feature map from the contracting path.

$$Y = \text{Concat}(U, A_{\text{encoder}}) \quad (4)$$

4) Output layer: A 1x1 convolutional layer, the last layer in the U-NET, maps every feature vector with 64 components to the number of classes required for land cover segment.

$$\text{Final Convolution: } K = M * A + y \quad (5)$$

$$\hat{Y} = \text{Softmax}(Z) \quad (6)$$

Using satellite imagery, the U-NET architecture has shown to be an effective method for classifying land cover. It is especially well-suited for this application because of its capacity to integrate spatial and contextual data. Accurate and scalable land cover classification will become more and more possible as machine learning techniques and satellite imagery data become more widely available. This will have a major impact on a number of fields, including disaster management, agriculture, urban planning, and environmental monitoring.

IV. Proposed Methodology

In this section, the proposed methodology is discussed in various steps. Fig. 2 describes the pipeline for satellite image classification with proposed approach.

A. Pre-Processing Techniques

Preprocessing the satellite images contains multiple systematic steps to prepare (or make ready) the pictures for further analysis. These methods are mainly used for feature normalization by removing the noise and highlighting the dark regions that help to classify the accurate land covers. The first step uses noise removal with adaptive filtering (ANR) and AHE (Histogram Equalization), which removes noise and enhances the contrast of the image. When using these two concepts, the Land cover classification results of the satellite images may be undistorted/unbiased from errors and thus more visually comparable. This section further explains the mathematical equations of these methods used in the capturing of satellite images and highlights their main purpose in land cover [27] type in the entire flow of the land cover classification system. Table I shows the comparative performance of algorithms.

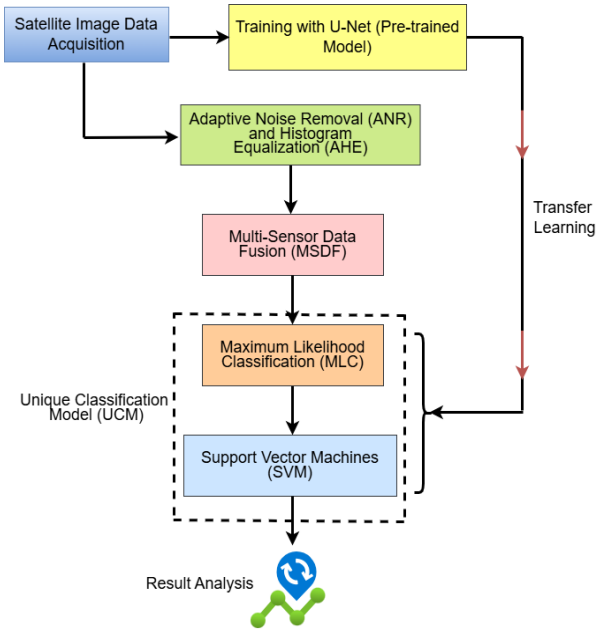


Fig. 2. System architecture.

TABLE I. COMPARATIVE PERFORMANCE OF ALGORITHMS

Authors	Proposed Approach	Performance Metrics	Research Gaps
Li et al. [21]	Multisource CD UNet++	MSOSCD the IoU-87.05% and F1-Score-93.1% and for MSBC dataset the IoU-87.05% and F1-Score-65.77%.	Lack of stable outcomes in terms of accurate locations with multisource features.
Thepade et al. [22]	DCNN	Accuracy (%) of 97.94, Precision (%) of 97.96, Recall (%) of 97.94, and F1-Score (%) of 97.94.	Lack of color based extraction is required to improve the performance.
Jagannathan et al. [23]	Hybrid Hot Encoding VGG19	Accuracy of 98.5%	The training with ResNet50 is very unclear.
Bin Xia et al. [24]	CNN	Accuracy of 0.9472, Misclassification rate of 0.0528, and kappa coefficient of 0.9435.	There is a lack of factors semantic data, obstruction, and distortion
Sudeep et al. [25]	Thepade SBTC	Acc of 70%, MCC of 0.68, and F1-score of 0.69.	Lack of feature extrcation in terms of land usage finding from aerial images.

B. Multi-Sensor Data Fusion (MSDF)

Feature extraction plays a crucial role in LCLU. The main functionality of feature extraction is to convert the raw data into a list of features that can be used to conduct further analysis. Multisensory data fusion (MSDF) takes one step further in improving the accuracy and reliability of the derived characteristics by aggregating information from multiple sensors. The MSDF integrates information from various sensors more robustly and precisely. These sensors improve data quality and filter noise, which affects system performance. It also condenses the raw, unprocessed sensor data to simpler, more meaningful features. Next, autonomous navigation, environmental observation, object discovery, etc., can use these features with more relevant and discriminative inputs with the effect of feature extraction that can significantly enhance the algorithms in these applications.

The following formulations are most widely used to measure the features:

Weighted Averaging: It is very easy to measure the weighted average of the sensor measurements thus, the weight is derived based on the accuracy of each sensor.

$$W = \sum_{i=1}^{n} w_i a_i \quad (7)$$

Kalman Filter: It is used for the sensor fusion specifically applications like LCLU.

$$\hat{a}_k = \hat{a}_{k-1} + K_k(c_k - H\hat{a}_{k-1}) \quad (8)$$

C. Maximum Likelihood Classification (MLC)

The MLC is the classification method that covers the land applied in GIS (geographic information systems) and remote sensing. Transfer learning trains the samples by utilizing the statistical factors of pixel values to specify the land cover. The probability of every pixel measures the members of each class by using the probability density function of multidimensional normal distribution—the pixel of every class placed with high potential possibility. The MLC is very effective and easy to classify into different effectively distributed and separated classes. Suppose the class distributions overlap, and then the performance drops low. Finally, the likelihood that every pixel in the image belongs to every land cover class by using the Gaussian (normal) distribution model. In this context, the highest probability (maximum likelihood) pixels are represented with the specific class.

The probability $P(A|Z_i)$ that a pixel feature vector A belongs to class Z_i is given by the multivariate normal distribution.

$$P(A|Z_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(A-\mu_i)^T \sum_i^{-1}(A-\mu_i)\right) \quad (9)$$

D. Support Vector Machines (SVM)

SVMs perform remarkably well in binary classification applications. SVM can easily handle high-dimensional feature spaces, which is common in remote sensing data. It finds the feature space hyperplane that most effectively separates the classes. SVM maximizes the margin between data points of distinct classes, which lowers the risk of overfitting, particularly in higher-dimension domains. The kernel approach maps input space to a high dimensional space and allows SVM to classify non-linearly. SVM maps the input data to a higher dimensional space where a linear separation is possible using kernel functions when the data are not linearly separable. The standard base kernels are the polynomial, sigmoid, and radial basis functions (RBF). The margin is between the hyperplane and the nearest points of both classes. This margin must be maximized via SVM. Since these data are the most trustworthy and closest to the selection border (hyperplane), they are essential for determining the precise location and orientation of the hyperplane.

The binary classification issue is solved by using training dataset $\{(a_i, b_i)\}_{i=1}^{N}$, Where $x_i \in R^n$ represents feature vector

and $y_i \in (-1, 1\}$ class label helps to finds the maximum margin.

The equation of a hyperplane in an n-dimensional space is:

$$w \cdot x - b = 0 \quad (10)$$

Here, w is the normal vector to the hyperplane, and b is the bias term.

To ensure correct classification, the following constraints must be satisfied:

E. Optimization Problem

The objective is to maximize the margin, which is equivalent to minimizing ||w||, subject to the constraints above. The optimization problem can be formulated as:

$$\min_{w,b} \frac{1}{2}||w||^2 \quad (11)$$

Subject to $y_i(w \cdot x_i - b) \geq 1, i = 1, \ldots\ldots, N$

Lagrangian Formulation

Using Lagrange multipliers $\propto_i$, the issue can be represented as:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \alpha_i [y_i(w \cdot x_i - b) - 1] \quad (12)$$

By solving this dual problem, we get:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (13)$$

Subject to $\sum_{i=1}^{N} \alpha_i y_i = 1, \alpha_i \geq 0$.

The optimal w is given by:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad (14)$$

Finally, the SVM model can classify new, unseen data points (pixels) into land cover classes based on the learned decision function:

$$f(x) = \text{sign}(w \cdot x - b) \quad (15)$$

F. Combination of MLC and SVM

The combination of SVM and other methods obtains a high classification rate based on land cover. The MLC is a reliable method that works effectively on various types of satellite images and land cover situations, helping to increase the accuracy of classification.

G. Dataset Description

The dataset mainly gather the information of Vijayawada as of 24-10-2022 as shown in Fig. 3 and Fig. 4. Vijayawada is situated on the banks of the Krishna River, in the geographic middle of the Andhra Pradesh state in India, it is a historic city with coordinates of 16003'11" N and 800 03'91" E. Equatorial weather prevails, with warm summers and mild winters. In May and June, the temperature peaks at 47 °C, whereas in the winter, it ranges from 20 to 270 °C. The average annual rainfall is 103 cm, while the average humidity is 78%.

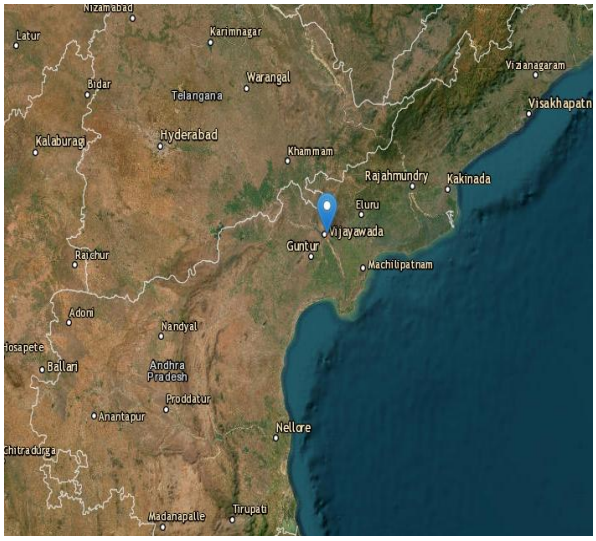


Fig. 3. Sample Vijayawada city map collected from USGS database.

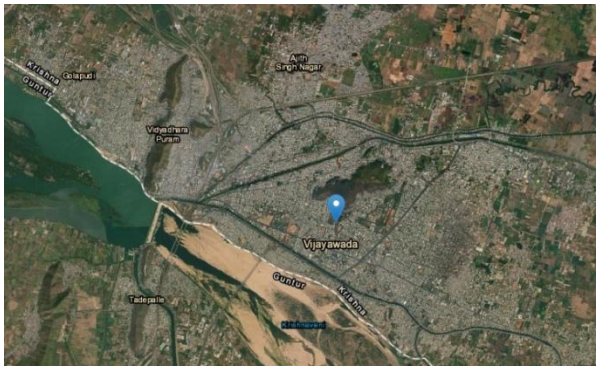


Fig. 4. This is the actual MAP of Vijayawada as on 24-10-2022.

V. Results and Discussions

A. Performance Metrics

The following parameters show the performance of ML algorithms applied on USGS dataset. The classified image contains five classes' building_regions, green_lands, farmlands, sandlands and water_regions. All these regions are marked with various colors. The confusion matrix is calculated and obtained results are compared with existing system results.

$$\text{Accuracy (ACC)} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision (Pre)} = \frac{TP}{TP + FP}$$

$$\text{Specificity (Spc)} = \frac{TN}{TN + FP}$$

$$\text{Recall (Re)} = \frac{TP}{TP + FN}$$

$$\text{F1} - \text{Score (F1S)} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

B. Discussions

The algorithms were implemented using the Python programming language. Each algorithm obtained the count values based on the actual values and predicted values. The existing models, such as the Change detection network (CDNet) and fully convolutional-early fusion (FC-EF), were compared with the proposed approach.

Table II describes the classification performance of existing CDNet model based on the count values belongs to five classes shown in Fig. 5. The average (acc) of all the classes is 0.72%, pre of 0.72%, (re) of 0.71%, (spc) of 0.92%, and F1S of 0.72%. Among all the parameters the specificity shows the high values that represents the imbalance in binary classification.

TABLE II. PERFORMANCE OF CDNET BASED ACTUAL VALUES AND PREDICTED VALUES

LCLU	Acc	Pre	Re	Spc	F1S
Building Regions	0.72	0.66	0.72	0.91	0.69
Green Lands	0.72	0.74	0.69	0.93	0.71
Farm Lands	0.72	0.71	0.69	0.93	0.70
Sand Lands	0.72	0.74	0.78	0.93	0.76
Water Regions	0.72	0.74	0.71	0.93	0.73

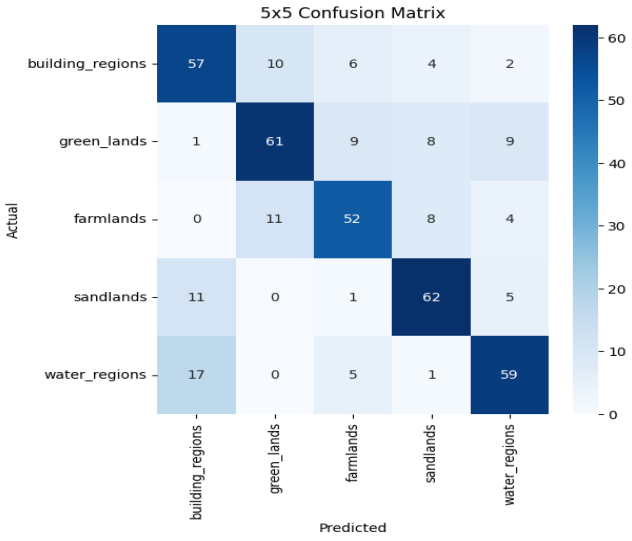


Fig. 5. Count values obtained from 5x5 confusion matrix using CDNet.

TABLE III. PERFORMANCE OF FC-EF BASED ACTUAL VALUES AND PREDICTED VALUES

LCLU	Acc	Pre	Re	Spc	F1S
Building Regions	0.80	0.81	0.75	0.95	0.77
Green Lands	0.80	0.81	0.76	0.95	0.78
Farm Lands	0.80	0.76	0.75	0.94	0.75
Sand Lands	0.80	0.75	0.89	0.93	0.81
Water Regions	0.80	0.84	0.83	0.96	0.83

Table III describes the classification performance of existing FC-EF model based on the count values belongs to five classes shown in Fig. 6. The average (acc) of all the classes is 0.80%, (pre) of 0.83%, (re) of 0.796%, (spc) of 0.94%, and F1S of 0.79%. Among all the parameters the specificity shows the high values that represents the imbalance in binary classification.

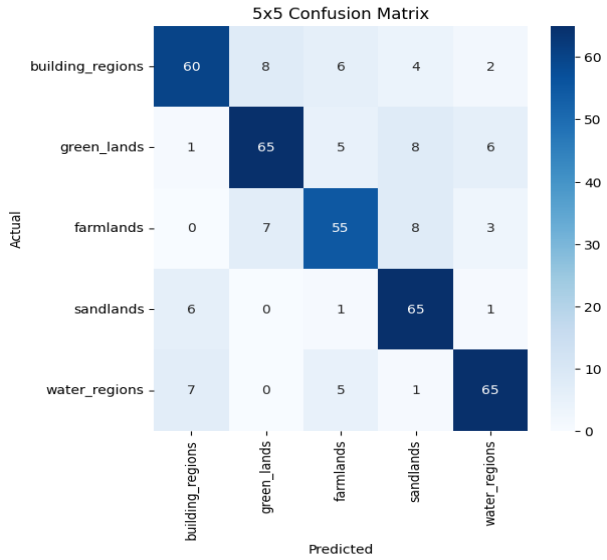


Fig. 6. Count values obtained from 5x5 confusion matrix using FC-EF.

TABLE IV. PERFORMANCE OF PROPOSED BASED ACTUAL VALUES AND PREDICTED VALUES

LCLU	Acc	Pre	Re	Spc	F1S
Building Regions	0.93	0.97	0.91	0.99	0.94
Green Lands	0.93	0.92	0.91	0.97	0.91
Farm Lands	0.93	0.87	0.91	0.97	0.89
Sand Lands	0.93	0.91	0.97	0.97	0.94
Water Regions	0.93	0.98	0.94	0.99	0.96

Table IV describes the classification performance of existing FC-EF model based on the count values belongs to five classes shown in Fig. 7. The average (acc) of all the classes is 0.93%, (pre) of 0.95%, (Re) of 0.95%, (spc) of 0.98%, and F1S of 0.97%. Compare with above algorithms the proposed approach shows the high performance.

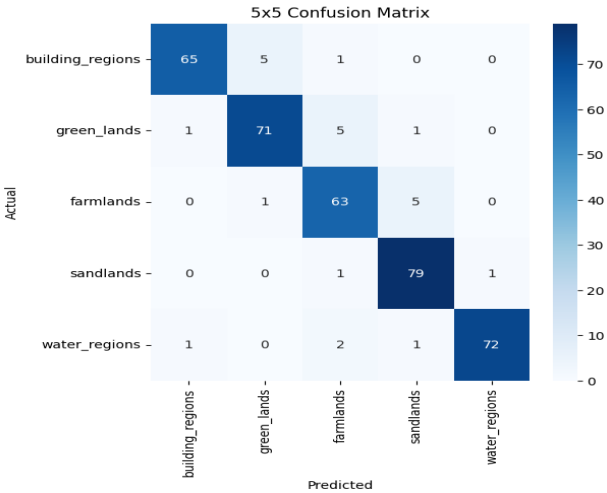


Fig. 7. Count values obtained from 5x5 confusion matrix using proposed approach.

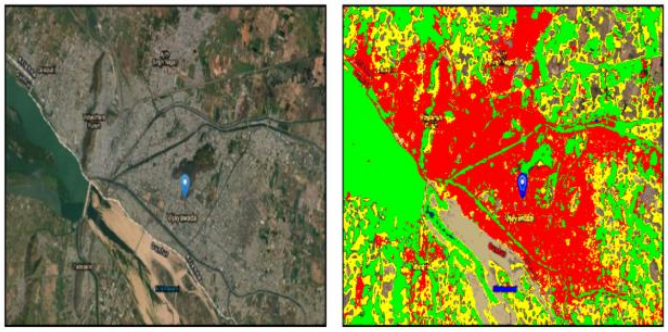


Fig. 8. (A) Input image, (B) Output image after implementation with the proposed MLC and SVM.

VI. CONCLUSION

The proposed MLC combines U-NET as a pre-trained model that transfers the LULC patterns to the proposed SVM algorithm. Other preprocessing techniques, such as ANR and AHE ref, refine image noise, extract significant features using MSDF, and handle high-dimensional data and non-linear relationships. The proposed approach is also a Unique Classification Model (UCM) that classifies the accurate LULC regions in the datasets. The Integrated SVM effectively improves classification in this work by focusing on the critical support vectors that define the class separations. Despite the increased computational complexity of integrating MLC and SVM, the algorithm remains scalable and efficient. Finally, the proposed approach obtains high performance in terms of average accuracy of all the classes of 0.93%, precision of 0.95%, recall of 0.95%, specificity of 0.98%, and 0.97% based on five classes such as building_regions, green_lands, farmlands, Sandilands, and water_regions as shown in Fig. 8 and Table V. Compared with the above algorithms, the proposed approach shows a high performance. In the future, high-quality images will require finding the various land types using ecosystem mapping and satellite imagery with accurate region classification. Also, developed the integrated models to detect and classify various deep learning models combined with image processing to show high impact on LULC regions.

TABLE V. THE DIFFERENT LAND TYPES COVERED WITH VARIOUS COLORS

Colors	Regions
Red	Buildings
Green	Greenery Lands
Yellow	Agricultural Lands
Sandy Brown	Sand Lands
Blue	Water regions

REFERENCES

[1] H. Ouchra and A. Belangour, "Satellite image classification methods and techniques: A survey," in Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST), Aug. 2021, pp. 1–6, doi: 10.1109/IST50367.2021.9651454.

[2] S. Jalayer, A. Sharifi, D. Abbasi-Moghadam, A. Tariq, and S. Qin, "Modeling and predicting land use land cover spatiotemporal changes: A case study in Chalus watershed, Iran," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 5496–5513, 2022, doi: 10.1109/JSTARS.2022.3189528.

[3] A. Tassi and M. Vizzari, "Object-oriented LULC classification in Google Earth Engine combining SNIC, GLCM, and machine learning algorithms," Remote Sens., vol. 12, no. 22, pp. 1–17, Nov. 2020, doi: 10.3390/rs12223776.

[4] M. A. Shafaey, M. A. M. Salem, H. M. Ebied, M. N. Al-Berry, and M. F. Tolba, "Deep learning for satellite image classification," in Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018 (Advances in Intelligent Systems and Computing). Springer-Verlag, 2019, pp. 383–391, doi: 10.1007/978-3-319-99010-1_35.

[5] M. Esmaeili, D. Abbasi-Moghadam, A. Sharifi, A. Tariq, and Q. Li, "Hyperspectral image band selection based on CNN embedded GA (CNNeGA)," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 16, pp. 1927–1950, 2023, doi: 10.1109/JSTARS.2023.3242310.

[6] S. Mahyoub, H. Rhinane, A. Fadil, M. Mansour, M. Saleh, and F. Al-Nahmi, "Using of open access remote sensing data in Google Earth Engine platform for mapping built-up area in Marrakech city, Morocco," in Proc. IEEE Int. Conf. Moroccan Geomatics (Morgeo), May 2020, pp. 1–5, doi: 10.1109/MORGEO49228.2020.9121912.

[7] M. Amani, A. Ghorbanian, S. A. Ahmadi, M. Kakooei, A. Moghimi, S. M. Mirmazloumi, S. H. A. Moghaddam, S. Mahdavi, M. Ghahremanloo, S. Parsian, Q. Wu, and B. Brisco, "Google Earth Engine cloud computing platform for remote sensing big data applications: A comprehensive review," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 5326–5350, 2020, doi: 10.1109/JSTARS.2020.3021052.

[8] M. Awad, "Google Earth Engine (GEE) cloud computing based crop classification using radar, optical images and support vector machine algorithm (SVM)," in Proc. IEEE 3rd Int. Multidisciplinary Conf. Eng. Technol. (IMCET), Dec. 2021, pp. 71–76, doi: 10.1109/IMCET53404.2021.9665519.

[9] S. Ansith and A. A. Bini, "A modified Generative Adversarial Network (GAN) architecture for land use classification," 2021 IEEE Madras Section Conference (MASCON), Chennai, India, 2021, pp. 1-6, doi: 10.1109/MASCON51689.2021.9563609.

[10] N. Laban, B. Abdellatif, H. M. Ebeid, H. A. Shedeed and M. F. Tolba, "Seasonal Multi-temporal Pixel Based Crop Types and Land Cover Classification for Satellite Images using Convolutional Neural Networks," 2018 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 2018, pp. 21-26, doi: 10.1109/ICCES.2018.8639232.

[11] S. D. Thepade, S. S. Adrakatti, S. Chauhan, V. Chitnis and A. Bhalerao, "Comparing the Deep Convolutional Neural Network Models for Identification of Aerial Image Land Use," 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2023, pp. 1-7, doi: 10.1109/ICCUBEA58933.2023.10392161.

[12] D. R. Rao, S. Noorjahan and S. A. Fathima, "Classification of Land Cover Usage from Satellite Images using Deep Learning Algorithms," 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2022, pp. 1302-1308, doi: 10.1109/ICEARS53579.2022.9752282.

[13] N. A. K. Varma, S. Manne and P. L. Sri, "Object based Change detection on Temporal Multi-Spectral Satellite imagery," 2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC), Chennai, India, 2023, pp. 354-359, doi: 10.1109/ICRTAC59277.2023.10480725.

[14] J. Fan, T. Chen and S. Lu, "Unsupervised Feature Learning for Land-Use Scene Recognition", IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 4, pp. 2250-2261, April 2017.

[15] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, et al., "AID: A benchmark data set for performance evaluation of aerial scene classification", IEEE Trans. Geosci. Remote Sens., vol. 55, no. 7, pp. 3965-3981, Jul. 2017.

[16] Y. Liu, Y. Zhong and Q. Qin, "Scene classification based on multiscale convolutional neural network", IEEE Trans. Geosci. Remote Sens., vol. 56, no. 12, pp. 7109-7121, Dec. 2018.

[17] E. Ouerghi, "A Deep Learning Model for Change Detection on Satellite Images", Image Processing On Line, Nov. 2022.

[18] D. Peng, Y. Zhang and H. Guan, "End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++", Remote Sensing, Jun. 2019.

[19] J. Chen et al., "DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Jan. 2021.

[20] H. Cai, T. Chen, R. Niu and A. Plaza, "Landslide Detection Using Densely Connected Convolutional Networks and Environmental Conditions", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, May 2021.

[21] H. Li, F. Zhu, X. Zheng, M. Liu and G. Chen, "MSCDUNet: A Deep Learning Framework for Built-Up Area Change Detection Integrating Multispectral SAR and VHR Data", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Jan. 2022.

[22] S. D. Thepade and M. R. Dindorkar, "Fusing deep convolutional neural network features with Thepade's SBTC for land usage identification", Engineering Science and Technology an International Journal, vol. 27, pp. 101014, Mar. 2022.

[23] J Jagannathan and C Divya, "Deep learning for the prediction and classification of land use and land cover changes using deep convolutional neural network", Engineering Science and Technology an International Journal, vol. 65, pp. 101412, Aug 2021.

[24] Bin Xia, Fanyu Kong, Jun Zhou, Xin Wu and Qiong Xie, "Land Resource Use Classification Using Deep Learning in Ecological Remote Sensing Images", Hindawi Computational Intelligence and Neuroscience, vol. 2022, pp. 7179477.

[25] D Sudeep, Thepade and Piyush R. Chaudhari, "Land Usage Identification with Fusion of Thepade SBTC and Sauvola Thresholding Features of Aerial Images Using Ensemble of Machine Learning Algorithms", Applied Artificial Intelligence, vol. 35, no. 2, pp. 154-170, Nov 2020.

[26] Q. Weng, Z. Mao, J. Lin and W. Guo, "Land-use classification via extreme learning classifier based on deep convolutional features", IEEE Geosci. Rem. Sens. Lett., vol. 14, no. 5, pp. 704-708, 2017.

[27] R. Stivaktakis, G. Tsagkatakis and P. Tsakalides, "Deep learning for multilabel land cover scene categorization using data augmentation", IEEE Geosci. Rem. Sens. Lett., vol. 16, no. 7, pp. 1031-1035, 2019.

[28] A. Clark, S. Phinn, and P. Scarth, "Optimised U-Net for land use–land cover classification using aerial photography", PFG – J. Photogramm. Remote Sens. Geoinf. Sci., vol. 91, no. 2, pp. 125–147, Apr. 2023.

Graph Neural Networks with Attention Mechanisms for Accurate Dengue Severity Prediction

Dr. Monali G. Dhote[1], Dr. Puneet Thapar[2], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[3],
Dr. G. Indra Navaroj[4], R. Aroul Canessane[5], B.V.Suresh Reddy[6], Elangovan Muniyandy[7], Dr. Kapil Joshi[8]
Assistant Professor, Dept. of Applied Mathematics and Humanities, Yeshwantrao Chavan College of Engineering, Nagpur, India[1]
Associate Professor, Department of Computer Science and Engineering, Lovely Professional University, Punjab, India[2]
Faculty of Informatics and Computing, UniSZA University, Malaysia[3]
Assistant Professor (SG), Department of Information Technology, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India[4]
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India.[5]
Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.[6]
Department of Biosciences-Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai - 602 105, India[7]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[7]
Department of Computer Science & Engineering, Uttaranchal Institute of Technology (UIT), Uttaranchal University, Dehradun 248007, Uttarakhand, India[8]

***Abstract*—Dengue fever continues to be a significant public health issue across the globe because it can lead to life-threatening complications. Severity prediction in a timely and precise manner is imperative for proper clinical management and effective resource utilization. Conventional models fail to identify intricate relationships between heterogeneous clinical, demographic, and epidemiological variables. For this purpose, we develop an innovative framework—Graph Neural Network with Attention Mechanism (GNN-AM)—aimed at enhancing dengue severity prediction. In the suggested method, every patient is viewed as a node in a graph with edges indicating clinical similarity in terms of health properties. The incorporation of attention mechanisms enables the model to selectively pay attention to important clinical indicators like fever duration, platelet count, and bleeding tendencies. This selective attentiveness improves prediction quality by giving maximum importance to the most important features while reducing the impact of less significant data. The model was trained and tested on a dataset of laboratory-confirmed dengue cases that contained clinical symptoms, laboratory results, and demographics. Experimental results showed that attention-augmented GNN performed better than both typical GNNs and traditional machine learning models, recording an accuracy of 90.3%, a recall of 88.9%, and an F1-score of 89.6%. Results highlight the efficacy of the GNN-AM framework in classifying dengue severity accurately and the ability to emphasize crucial clinical indicators using attention mechanisms. In the future, this model can be combined with Electronic Health Records (EHRs) and implemented in real-world healthcare environments using federated learning methods to maintain data privacy across institutions.**

***Keywords*—Attention mechanism; dengue severity prediction Graph Neural Network; healthcare analytics; machine learning**

I. INTRODUCTION

The viral infection that is transmitted by mosquitoes, Dengue fever, has become a subject of great concern to public health globally. This epidemic grew out of the tropics and subtropics into the general world [1]. The causative agent of the disease is Dengue virus and it is mainly propagated via the Aedes aegypti and Aedes albopictus mosquitoes. Dengue affects millions of people worldwide every year. Approximately half the world population is considered to be at risk of dengue according to the World Health Organization. The clinical spectrum notorious for the disease ranges from mild fever to severe forms of Dengue Hemorrhagic Fever and Dengue Shock Syndrome, both of which can prove fatal if prognosis and management are not instituted rightly [2]. The outcome of dengue fever is determined by a myriad of factors, including serotype of the virus, immune response of the host, the presence of co-morbidities, and genetic predisposition. In the absence of antiviral treatments that can cause direct benefits, early detection and severity prediction play an important role in their clinical management [3]. Conventional severity prediction models, such as statistical and traditional machine learning methods, usually depend on manual feature choice and cannot effectively characterize intricate patient interdependencies. Their poor generalization capacity in dealing with heterogeneous populations presents challenges to proper prognosis and effective healthcare resource allocation [4]. Statistical and rule-based algorithms that essentially analyze clinical and laboratory parameters constitute the mainstay of conventional dengue severity prediction models. However, these models face many shortcomings: they depend on features selected manually, fail to capture complex interactions among patient characteristics, and have poor generalizability to different populations [5]. Classic machine learning techniques such as LR, decision trees, and SVM were also applied to severity prediction but would typically require enhancement in feature engineering and complexity to model intricate interdependencies among the various patient attributes. In addition, they do not account for the correlations that may exist among the clinical data points

belonging to a common clinical profile [6]. Given the heterogeneity of dengue fever with its multiple manifestations, there is a great need for a more advanced computational approach to handle these complexities. This suggests Artificial Intelligence, specifically deep learning techniques such as GNN to create massive impact within the medical realm for diagnosis and prediction of severity. AI frameworks hold great promise in sifting through volumes of patient data, identifying subtle footprints, and producing predictions with high accuracy, completely nullifying all the limitations followed by the traditional approach. So, it becomes possible to utilize the full leverage that AI provides for designing a data-driven framework for timely diagnosis, risk stratification of patients, and optimal allocation of healthcare resources [7].

Graph Neural Networks are said to cater to one of the advanced solutions in deep learning that can extract complex relations of structured data. GNN is a good fit medical application because patients' records can be present as connected entities rather than traditional neural networks that take the data in Euclidean space [8]. A very important characteristic of GNNs is that they allow the aggregation of information from neighboring nodes, which makes a prediction refined by conducting the context relations among more similar patients [9]. It will be best suited for the problem of predicting the severity of dengue since many of these clinical parameters can be closely related in variation from one patient group to another. By the means of message-passing techniques, the GNNs empower the graph to flow the information improving the predictions regarding severity. The addition of GNNs to an attention mechanism will further enhance predictive performance by assigning different importance weights to the different patient features and relationships. The attention mechanism will ensure that the most important data are included in the model and will reduce noise as well as improve interpretability [10].

The main motivation behind this research is the need for a more accurate and trustworthy model to predict the severity of dengue fever effectively. This need becomes more important for resource-poor settings where early risk assessment is vital for effective disease management. Most existing models do not generalize well across diverse populations due to the differences in clinical manifestations and diagnostic approaches. With respect to above, the need for a flexible approach also arises because of constantly changing data patterns and emerging risk factors related to dengue outbreaks. Thus, it is the intention of this study to put attention mechanisms into GNNs so that it addresses other challenges by developing an improved prediction framework along with explainability and clinical utility. Another major merit of GNNs in this field is that they will incorporate much heterogeneous data set sources, including clinical records, laboratory findings, and epidemiological data, plus environmental factors like temperature and relative humidity known to affect dengue transmission [11]. By analyzing these different dimensions of patient data, the model would help clinicians make more meaningful decisions about the severity of the disease.

In addition, such AI-oriented models for the prediction of dengue severity have wide-ranging public health and epidemic control ramifications. Health care infrastructure is strung to the limits in an outbreak period in most dengue-endemic regions, and a delay ensues before diagnosis and treatment can be started. A predictive model that is capable of stratifying patients into risk levels would have great value in triaging patients for medical intervention, ensuring that those with severe manifestations receive such interventions on time [12]. It will also allow AI models to analyze past patient data for trends and risk factors associated with severe dengue cases, which would help inform better disease surveillance and prevention strategies. Real-time patient-condition monitoring is facilitated by these models, thus decreasing morbidity and mortality via timely interventions. Scientifically speaking, should the GNNs be applied to dengue severity prediction, this will pave the way for future work on graph-based approaches for other infectious diseases, bringing the application of advanced AI in predictive medicine one step further [13]. As much as the GNNs seem promising towards providing better predictions in the severity levels of dengue, there are challenges that must be overcome for the model to gain effectiveness and applicability in real-world scenarios. A major concern is the medical data, as differences in data collection practices among various health institutions will affect the outcome an algorithm may produce [14]. Consideration for standardization in data and addressing problems such as those of missing values and class imbalances become important in developing a sound predictive framework. In addition, since GNNs capture highly complex relationships within patient data, most would demand heavy computational power for training and thus would not be deployable in low resource setups. Lightweight models could be developed, or federal learning techniques used to address and curb these challenges and make AI-driven solutions much more accessible [15]. The model needs to make consideration for interpretability because a clinician needs surety on the predictions made by AI in order to feel comfortable and to be accepting of them. This could be done by providing the required transparency using explainability techniques such as attention visualization and feature importance analysis in order to increase the confidence of clinicians in and usability of the proposed model in practice.

This research introduces a new GNN-AM (Graph Neural Network with Attention Mechanism) model to resolve the shortfalls of current models. It seeks to provide an even more precise, explainable, and scalable solution for forecasting dengue severity, particularly in resource-limited healthcare environments.

The Key contributions of this study are as follows:

- Introduced a new GNN-AM approach that combines attention mechanisms and GNNs for better dengue severity prediction.
- Modeled patients as graph nodes, with edges denoting clinical similarity, facilitating improved representation of case relationships.
- Incorporated attention mechanisms to focus on essential clinical features (e.g., duration of fever, platelet count), enhancing prediction performance.
- Exhibited 90.3% accuracy, 88.9% recall, and 89.6% F1-score, outperforming baseline GNN and ML models.

- Described real-world relevance, proposing incorporation with EHRs and federated learning for scalable, privacy-protecting deployment in healthcare.

The rest of the sections of this research have been organized as follows: Review of existing literature review on Dengue Fever Severity Prediction is in Section II. Research gap is given in Section III. In Section IV, proposed research Methodology is explained. Section V presents the experimental results. In Section VI, Conclusion and future Scope is mentioned.

II. Literature Review

Research has been recently undertaken into predicting Dengue fever and, therefore, machine-learning models that aim at classifying severity and increasing early detection have been developed [16]. Traditionally, in situations where clinical characteristics, such as platelet count, WBC count, hematocrit levels, duration of fever, and liver enzyme levels, were manually selected, LR, decision trees, SVM, and ensemble models like random forests would also do the work. While moderate success has been achieved by traditional methods, they still find it hard to capture nonlinear relationships and, therefore, do not consider similarities, treating each patient as an independent data point. Deep learning models such as ANN, CNN, and RNN, have when used in recent times, partially circumvented this problem by providing automated high-dimensional feature extraction [17]. In the context of dengue complications, CNNs have found application in analyzing medical images, while RNNs and LSTMs were used to analyze time-series data of individual patients. Yet, all these models failed to take into account patient relationships and dependencies in the dataset. Such inadequacy paved the way for the advent of GNNs in medical applications since they provide a systematic way to model complex patient data and their interactions [18]. Owing to their ability to operate on non-Euclidean data, GNNs have been used extensively in medical settings, where patients, their symptoms, and laboratory data are represented as graphs. To conduct prediction tasks, standard deep learning models treat data points independently, while GNNs leverage graphical structures to understand the complex dependencies and interactions among medical entities [19]. Recent areas in which GNNs have been applied include disease classification, drug design, protein-interaction analysis, and the construction of medical knowledge graphs. One of the strengths of GNNs lies within the message-passing framework: each node aggregates information coming from neighbor nodes to improve its feature representation. This allows the model to better learn from patient similarities and utilize them to predict outcomes. In dengue severity prediction, a patient-similarity graph is established, whereby nodes represent patients while edges capture very similar clinical features shared among them [20]. In this graph approach, patterns can be found that were not assessable when focusing on one patient at a time. GNNs have proven to outperform other machine learning techniques in disease prediction by incorporating relational information so that their generalization and interpretability are improved. With various techniques like node embeddings, adjacency matrices, and message-passing for representation learning, GNNs have built powerful representations of patient information, thereby proving to be very useful for clinical applications [21]. However, an inherent limitation of GNNs is to find features and relations that contribute most toward prediction, thus establishing the need for attention mechanisms.

Attention mechanisms have triggered a transformation in deep learning: models are now able to process the most salient features and relationships, thus improving interpretability and performance. This concept was initiated in natural language processing with the Transformer, while attention is now being incorporated into a variety of other fields such as computer vision, speech recognition, and health care analytics. In medical applications, attention tends to improve disease prediction by focusing the model on the most important clinical variables [22]. Upon utilizing attention mechanisms integrated into GNNs, the message-passing process is assured to hone fine by giving variable weight to neighboring nodes concerning importance. In GATs, is used to dynamically determine/consider the influence each of its neighboring nodes bears upon its state; biophysical relationships are emphasized while the uninformative connections are deemphasized. This results in an increase in prediction accuracy while reducing the noise and increasing generalization. Attention mechanisms have also found a place in hybrid setups combining GNNs with either recurrent or convolutional networks concerning spatial and temporal constraints in medical-oriented data [23]. Attention-based GNNs for predicting dengue severity can, thus, be expected to select significant biomarkers, such as platelet count and hematocrit levels while honoring relationships of patient similarity leading to higher accuracy and better clinical decision-making. Traditional machine learning models are ill-equipped in modeling complex relationships among patients, while the CNNs and RNNs that have been employed in deep learning have their share of demerits, in that they cannot model the structured dependencies necessary for the problem at hand [24]. Graph Neural Networks offer a resolution to these limitations because they model and represent patient data as graphs, refining its representation via message-passing according to the relationships between patients. GNNs are endowed with abilities even beyond those: The inclusion of attention mechanisms allows GNNs to focus on the most important features and relationships, making them further interpretable and accurate in prediction [25]. Bring together the benefits of GNNs and attention mechanisms: it poses great potentials for medical diagnostics, patient risk assessment, and availability of healthcare resources in dengue-affected areas.

III. Research Gap

The prediction of dengue fever severity continues to be a daunting task with its complex, non-linear, and multifaceted pathogenesis—despite significant advances in deep learning and machine learning towards disease prediction [26]. It is even more challenging compared to other diseases in which AI methods have reported encouraging results, given that predicting dengue severity involves combining heterogeneous patient data and a grasp of complex interdependencies between clinical variables [27]. Classic models, including decision trees, SVM, and traditional neural networks, usually presume that patient data instances are independently and identically distributed and thus ignore structural similarities between cases [28]. GNNs have demonstrated their capabilities in processing structured medical information by representing relational data, but the majority of current methods do not support prioritizing

important clinical features or assigning weights to the importance of patient connections. Even though attention mechanisms are established within deep learning to improve model concentration and explain ability, their use with GNNs—particularly for dengue severity prediction—is underexplored. This research fills this gap by incorporating attention mechanisms into a GNN-based framework to improve patient similarity modeling, allocating suitable weights to important clinical features, and eventually enhancing prediction accuracy, interpretability, and clinical usefulness in dengue severity assessment.

IV. Research Methodology

The suggested Dengue Fever Severity Prediction Model based on GNN-AM adopts a systematic pipeline to provide high predictive performance and interpretability. First, one million patient records are preprocessed in a synthetic dataset by dealing with missing values, performing categorical variable one-hot encoding, and Min-Max Scaling of numerical features to maintain consistency in input data. Subsequently, a patient similarity graph is built—every patient node is defined with edges defined by cosine similarity of clinical features in order to preserve relational structures among comparable cases. The GNN model accumulates information from adjacent nodes using iterative message passing, refining each patient's feature representation dynamically. In addition, an integrated attention mechanism enhances model performance by focusing on important clinical indicators like fever duration, platelet count, and bleeding symptoms so that the model can highlight relevant features while reducing the impact of minor data. The model structure includes several layers of GNN and attention blocks for efficient feature extraction and representation learning. A final fully connected neural network with a SoftMax activation function is employed to classify patients into three classes: mild, moderate, or severe dengue cases. Fig. 1 shows GNN-AM architecture.

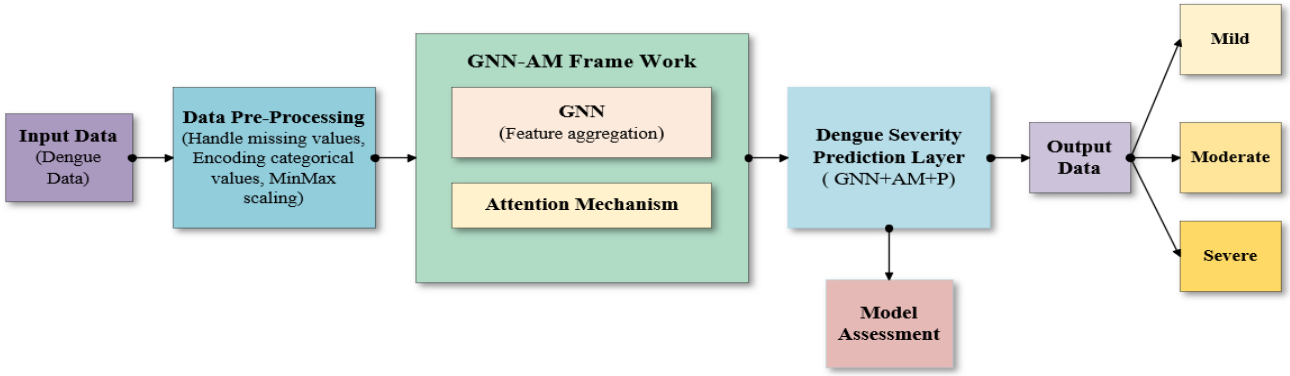


Fig. 1. GNN -AM architecture.

A. Data Collection

This study used the data set from an open source Kaggle repository to help ensure that research projects into Dengue fever severity prediction do not lack both accessibility and transparency [29]. This data set contains critical clinical indicators about Dengue diagnosis out of a million individuals and their health-related information. Each record is a set of binary features representing symptom among others, the Dengue status present. To maintain privacy and diversity of the demographic groups in the dataset, the dataset is synthetically generated and simulate real world patient data. Using this dataset, we want to create a robust GNN with attention mechanism to enhance the prediction of severity. This dataset is open source so that reproducibility is possible, and to further the domain of machine learning based disease prediction models, it is possible for other researchers to validate and further the work.

B. Data Pre-Processing

Preprocessing raw patient data is first step in the pipeline where it involves clinical symptoms, laboratory test results as well as vital signs. The presence of missing values in the raw medical data, as a result of incomplete patient records, will have negative impact on the performance of the model. One-hot is used to encode categorical variables such as symptom descriptions. Further, numerical features, like fever duration, platelet count, and WBC count are standardized with MinMax scaling so that there are uniformity of features and do not suffer from having features with larger scale. This step takes the raw medical data, preprocesses it into a feature matrix where each row represents a patient and each column a numerical medical attribute, resulting in standardized input for the following step of graph construction.

1) Handle missing values: The study uses imputation techniques to fill missing values in the database after which the complete dataset supports effective GNN training with Attention Mechanism. The missing values in patient records appear as a result of incomplete medical data and data collection errors yet imputation techniques help estimate unrecorded values to protect patient information integrity. The dataset mostly contains binary features which use mode imputation to substitute unknown values by the dominant value observed within the feature column in Eq. (1).

$$X_{imputed} = \arg max_{x \in X} \; Count(x) \tag{1}$$

Where, X in this expression consists of all values which we can detect in a feature column. $X_{imputed}$ represent the imputed value for a missing feature.

Mean imputation will serve as the replacement method in future models that use numerical data features including platelet count and WBC count in Eq. (2).

$$X_{imputed} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2}$$

The study makes its data complete through proper imputation approaches that maintains unbiased results to enhance model dependability and prediction precision.

2) Encoding categorical variables: A numerical format produced by one-hot encoding allows categorial variables to become suitable input for the GNN with Attention Mechanism. Explicit encoding is not necessary because the dataset mainly contains binary features (0 or 1) for Fever, Headache, Joint Pain and Bleeding symptoms. When implementing one-hot encoding the system creates binary vectors for every category to present categorical data numerically while avoiding false weight relationships. The transformation of categorical features with k unique values appears as follows in Eq. (3):

$$X_{one-hot} = [x_1, x_2, \dots, x_3] \quad (3)$$

Where, $X_{one-hot}$ denote the one-hot encoded vector for a categorical variable, $[x_1, x_2, \dots, x_3]$ represent the binary components of the one-hot vector, where one element is 1 and all others are 0.

One-hot encoding method becomes essential for dealing with patient region and hospital ID and risk level attributes as it prevents model interpretation errors from developing through the creation of unintended ordinal relationships.

3) Normalize numerical features: The researchers use Min-Max Scaling on numerical features to normalize their values into ranges between 0 and 1. The normalization procedure through Min-Max Scaling protects the learning process in GNN with Attention Mechanism by stopping significant features from taking control. The Min-Max Scaling formula is given in Eq. (4):

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (4)$$

Where, X be the original feature value, X_{min} and X_{max} be the minimum and maximum values of the feature in the dataset, X_{scaled} be the normalized (scaled) value of the feature.

The normalization technique scales all numerical values to a common interval which maintains data relationships but removes extreme data effects to achieve better model efficiency during Dengue severity prediction.

C. Graph Construction

Preprocessing the patient data builds a patient similarity graph to model patient relationship by shared clinical characteristics. In this graph, each patient is a node and two patients are adjacent if they have similar medical attributes. Metrics such as cosine similarity is used to quantify the similarity between patients and edge weights are assigned automatically. Stronger edge connections mean more similarity score and hence stronger edge connection between other related patients during message passing. The fact that this is a graph structure allows the model to capture dependencies that are not possible to capture with conventional machine learning models. This step outputs a patient similarity graph G = (V, E), where V is the set of patient nodes and E is the set of patient connections corresponding to same patients of different severities, which is the input to 6. This dataset is open source thus it allows reproducibility which other researchers can use to validate and extend these findings in the machine learning based disease prediction models.

D. Graph Neural Network

The main processing takes place inside the GNN when patient nodes acquire knowledge from adjacent nodes through features aggregation and message communication. The GNN enables individual patient nodes to communicate diagnostic information with other patients with similar conditions instead of traditional independent patient treatment. The formula used by nodes at each layer aggregates information from neighboring nodes through this operation in Eq. (5).

$$h_i^{(l+1)} = \sigma\left(\sum_{j \epsilon N(i)} W h_j^{(l)} + b\right) \quad (5)$$

Where, $h_i^{(l)}$ is the feature representation of node i at layer l, $h_j^{(l)}$ be the feature of a neighboring node $j \epsilon N(i)$, the set of neighbors of node i, W be the learnable weight matrix, b be the bias vector, σ be a non-linear activation function.

The embedding refinement cycle through this process adds relevant contextual information about connected patients to enhance clinical severity prediction capabilities based on similarity patterns. Fig. 2 represents the Graph Neural Network.

E. Attention Mechanism

The standard message passing approach of GNNs treats all neighboring nodes identically yet with attention mechanisms the model assigns various weight values to each neighboring node. Each node in a Graph Attention Network selects the neighbors that substantially affect its feature aggregation through an attention scoring function in Eq. (6).

$$e_{ij} = LeakyReLU\left(a^T [Wh_i \| Wh_j]\right) \quad (6)$$

The calculation proceeds through scattering the scores by using SoftMax normalization in Eq. (7).

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \epsilon N(i)} \exp(e_{ik})} \quad (7)$$

Such normalization keeps the total weights at 1. Through this method the model maintains emphasis on important patient relationships along with decreasing unimportant connection influence.

F. Feature-Level Attention

The attention mechanism in feature-level analysis selects the medical feature attributes that possess the highest value for disease severity evaluation. At this stage the model distributes attention values to each of the analyzed biomarkers including platelet count, WBC count and fever duration to determine which markers provide the most relevant information about dengue severity. The model automatically increases the importance level of significant features during predictions while minimizing unimportant ones. The model becomes more interpretable and focuses on relevant medical information because of this enhancement.

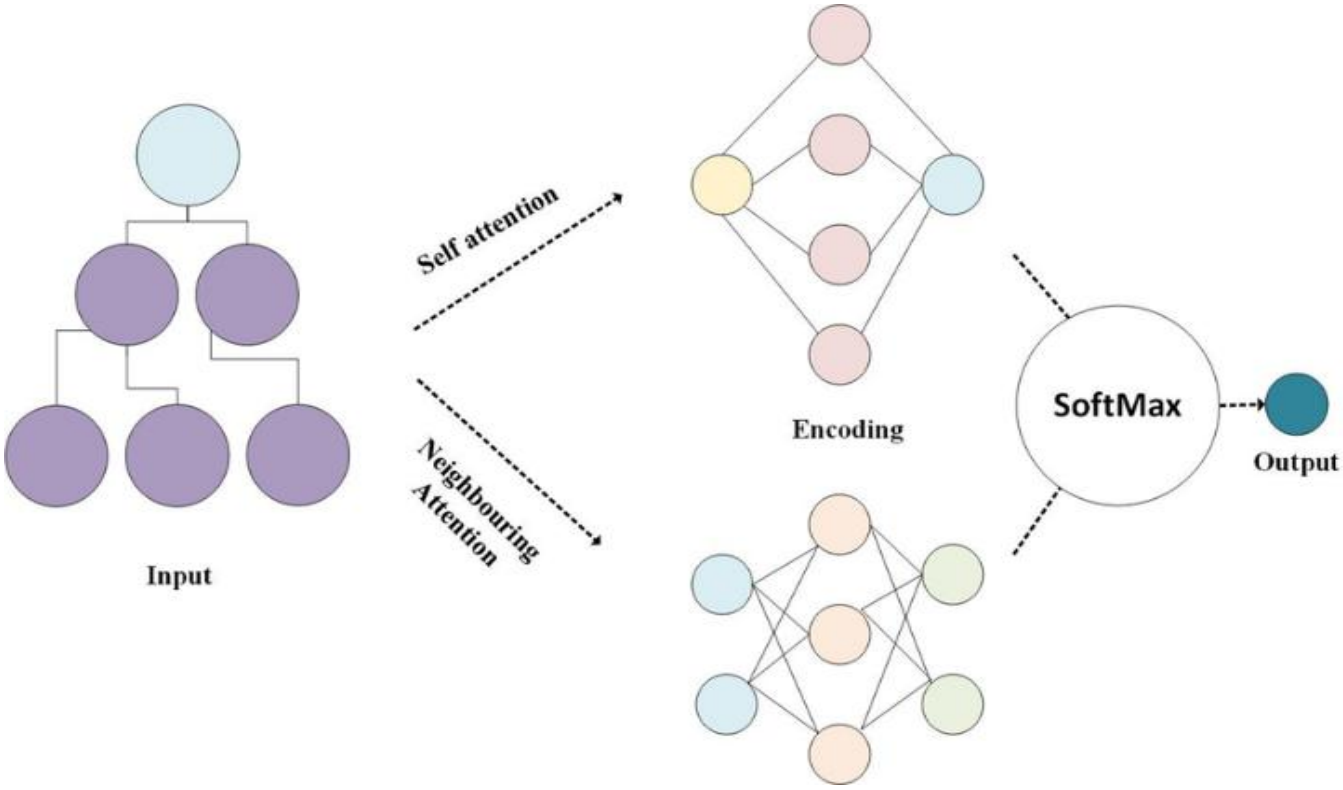


Fig. 2. Graph neural network.

G. Prediction Layer

The output of refined embedding's goes through a fully connected network ending in a SoftMax activation function which classifies dengue severity. The last layer determines probabilistic outcomes between different severity categories that have been established.

- Mild: Patients with stable vitals and minor symptoms.
- Moderate: Patients with moderate warning signs, such as decreasing platelet count and prolonged fever.
- Severe: Patients with critical complications that result in hemorrhagic manifestations together with organ dysfunction fall within the severe category.

The classification scheme helps medical staff perform patient risk assessment to prompt appropriate medical treatments when needed. Fig. 3 represents the mechanism of GNN-AM.

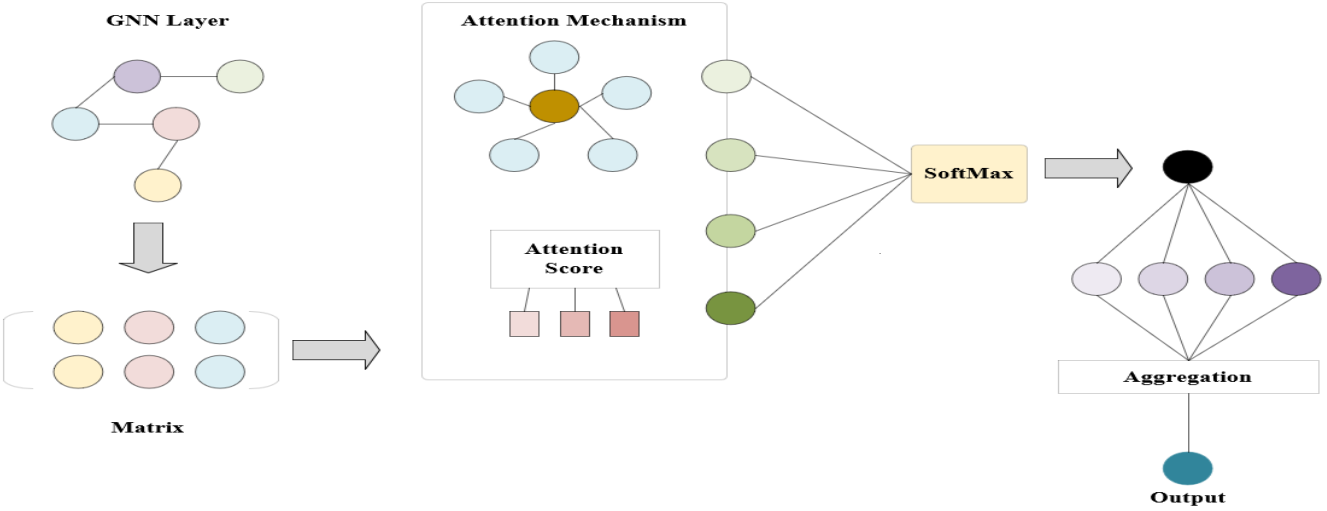


Fig. 3. Mechanism of GNN-AM.

GNN and attention mechanisms integrated in the prediction model create reliable outcomes that both doctors and patients can understand because they lead to precise and dependable predictions that enhance real-world clinical diagnosis of dengue severity.

Algorithm:1 Graph Neural Networks with Attention Mechanisms Algorithm

```
BEGIN
INPUT: Raw dataset with patient records
OUTPUT: Dengue severity prediction (Mild, Moderate, Severe)
Step 1: Data Preprocessing
Handle missing values:
FOR each missing value in dataset
    APPLY imputation techniques
END FOR
Encode categorical variables using one-hot encoding
Normalize numerical features using Min-Max scaling
RETURN pre-processed feature matrix
Step 2: Graph Construction
Define nodes: Each patient is a node in the graph
Create edges:
FOR each pair of patients (i, j)
    IF similarity (symptoms, lab results) > threshold THEN
    CREATE edge (i, j) with weight computed using similarity metric
  END IF
END FOR
RETURN graph G
Step 3: Node Feature Initialization
FOR each node i in graph G
    ASSIGN feature vector x_i containing medical attributes
END FOR
RETURN node feature matrix X
Step 4: GNN Processing
FOR each training iteration DO
  FOR each node i in graph G
      Aggregate features from neighboring nodes:
      h_i^(l+1) = σ (∑_ (j ∈ N(i)) W h_j^(l) + b)
    END FOR
    UPDATE node embeddings
END FOR
RETURN updated node embeddings
Step 5: Attention Mechanism (Graph Attention Network - GAT)
FOR each node i in graph G
    FOR each neighboring node j
        COMPUTE attention score:
            e_ij = LeakyReLU (a^T [Wh_i || Wh_j])
    END FOR
    APPLY SoftMax normalization:
        α_ij = exp(e_ij) / ∑_ (k ∈ N(i)) exp(e_ik)
END FOR
RETURN attention-refined node embeddings
# Step 6: Feature-Level Attention
Compute feature attention weights
  Identify key features (e.g., platelet count, WBC count, fever duration)
Emphasize important features for severity prediction
RETURN refined embeddings
Step 7: Severity Classification
APPLY fully connected layers (MLP)
Use SoftMax activation for classification
Predict Dengue severity category
RETURN severity labels
END
```

Fig. 4 illustrates the work flow of the GNN with Attention Mechanism for Dengue Severity Prediction in the precisely manner.

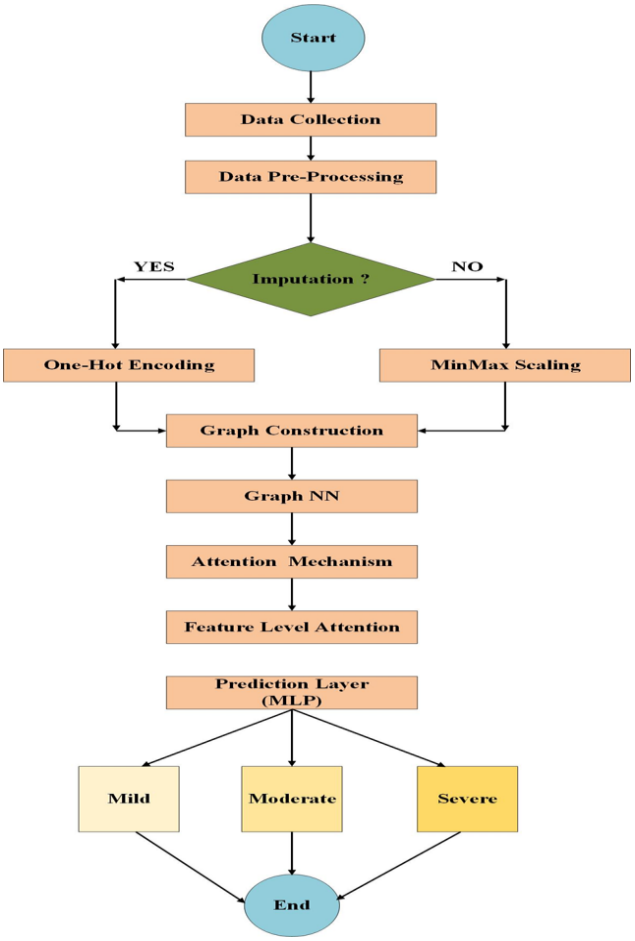


Fig. 4. Work Flow of GNN-AM.

V. Results and Discussion

In the study, data used amounts to a million records, ideally suited for Dengue severity prediction. The feature distribution in the dataset is balanced with 50% of people diagnosed with Dengue and another 50% being non-Dengue cases. Among the probable signs and symptoms, fever is the most cited, observed in 75% cases, followed by headache in 68% and joint pain at 62%. Out of these symptoms bleeding was only seen in 15% of cases making it the least common symptom. In this way, dataset is ensuring variability within it, thus generalizing the model among patients. Fig. 5 shows the Experimental Result.

As the samples are well-balanced for Dengue positive and negative patients, no bias is introduced during model training, making the final predictions accurate and reliable. The contrasting symptom patterns in this dataset provide a good ground for a decision machine learning model that can carry out this task for different severities in Dengue infections. Fig. 6 represents the Feature Distribution in dengue dataset.

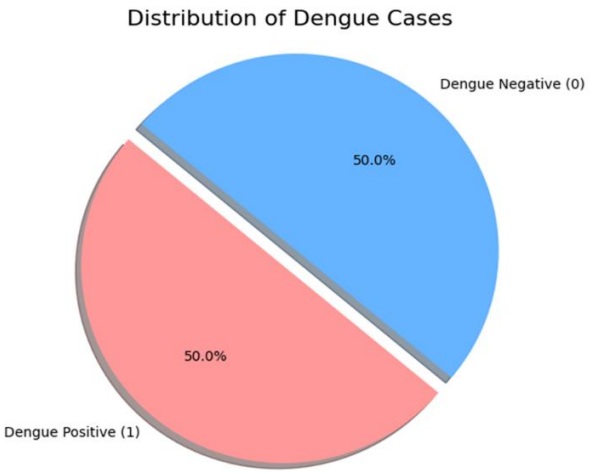


Fig. 5. Experimental result.

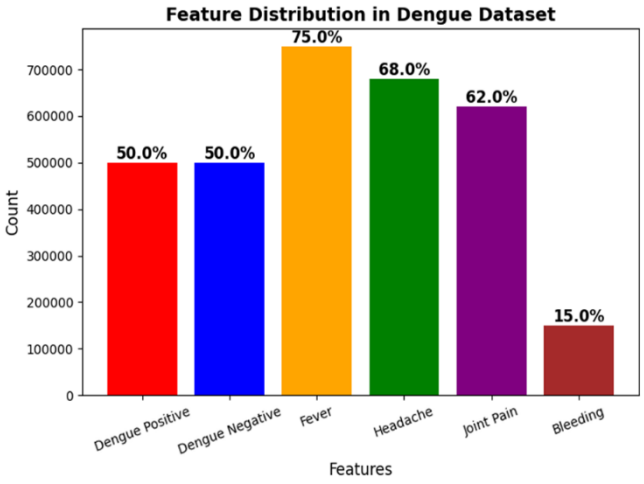


Fig. 6. Feature distribution.

Fig. 6 illustrates the training data comprises records of one million, with Dengue-positive cases and Dengue-negative cases evenly distributed (50% each), ensuring a balanced dataset for model training. Fever is the most common symptom, showing up in 75% of patients; headache in 68% and joint pains in 62%. The least common symptom was bleeding, in only 15%. The foregoing distribution depicts a wide variety of symptoms in Dengue patients, hence a heterogeneous dataset for training the GNN model equipped with attention mechanisms for better prediction of severity.

Fig. 7 illustrates the training and testing accuracy of different models is imperative for judging their predictive power. The baseline GNN trained models achieved 85.2% accuracy during training and 83.5% during testing, proving its proficiency in learning interrelationships given patient characteristics.

However, when combined with an attention mechanism, the performance finds a substantial increase as the model achieved a raised training accuracy of 91.6% and a testing accuracy of 89.8%. This improvement substantiates that learning based on attention shows effectiveness. On the contrary, traditional machine learning methods such as RF usually portrayed poor performance by producing training and testing accuracy of 78.9% and 76.4%, respectively. Such outcomes prove how graph-based learning has an upper hand in emphasizing the complexities of dependency relationships among patients, especially when coupled with attention mechanism refinement of feature importance. Fig. 8 shows the model training and Testing Loss.



Fig. 7. Model training and testing accuracy.

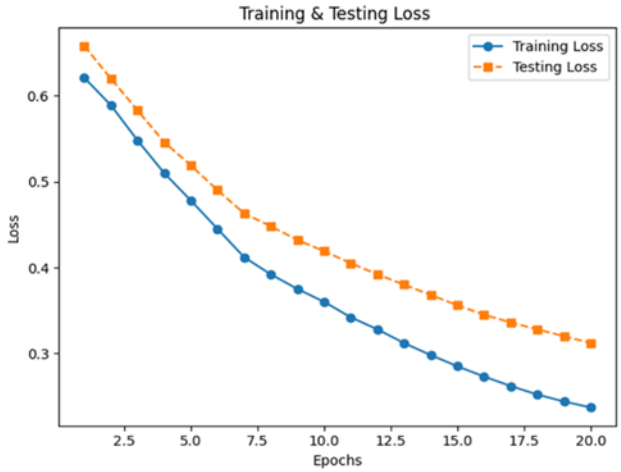


Fig. 8. Model training and testing loss.

Model optimization efficiency together with generalization capability depend on two critical metrics known as training loss and testing loss. Successful error minimization through models leads to improved reliability when applied in real-life situations because reduced loss values indicate better performance. Based on the study findings the base GNN model reached 0.312 as training loss yet achieved 0.354 as testing loss which demonstrates average learning efficiency. The established values show that the model detects patterns throughout the dataset yet further improvements are needed to lower classification mistakes and improve feature extraction capabilities. The integration of attention mechanisms resulted in a dramatic decline of training loss to 0.198 and a corresponding decrease in testing loss to 0.221 which proves attention's effectiveness for generalization enhancement. With the help of the attention mechanism the model selects only important clinical indicators which reduces noise generated by less relevant features. A stable learning process is produced by this

method which prevents overfitting while ensuring a good performance on new and existing patient data. Loss value reduction demonstrates that GNN-AM provides better predictive accuracy and creates more reliable severity predictions.

The Random Forest model demonstrated worse generalization because its training and testing loss results showed higher numbers when compared to the alternative models. The failure to detect complex nonlinear patterns among patient data results in inferior performance from tree-based systems that causes increased misclassification errors and accumulates more prediction errors. The superior results from the attention-enhanced GNN reveal its ability to extract advanced features while effectively handling patient relationships and numerous types of clinical data. The proposed GNN-AM model demonstrates its effective error reduction capability which produces dependable predictions for real-world applications of dengue severity forecasting.

A. Model Assessment

1) Accuracy: The assessment of data point accuracy consists of determining proper cluster or class assignments. The evaluation of clustered data uses accuracy measurements only if ground truth labels exist for performance assessment in Eq. (8).

$$Accuracy = \frac{\sum x}{n} \tag{8}$$

Where, x represent the no. of correctly predicted instances, n be the total number of instances, $\sum x$ denote the total number of correct predictions across all samples.

2) Recall: Model performance recall enables the calculation of correct positive outcome identifications among actual positive results. The measure finds its best use when recognizing positive cases takes priority.

The formula for recall is given in Eq. (9)

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

Where, TP is True Positives (correctly predicted positive cases), FN is False Negatives (actual positives incorrectly predicted as negative).

3) F1-Score: The F1 Score represents the harmonic mean between precision and recall of both incorrect positives and incorrect negatives. The method brings exceptional results to imbalanced datasets.

The formula for F1 Score is given in Eq. (10):

$$F1\ Score = 2.\frac{P*R}{P+R} \tag{10}$$

Table I and Fig. 9 comparison realizes yet another fact that prediction models for Dengue are yet another evidence of the proposed Graph Neural Network with Attention Mechanism over classical machine learning models. Logistic Regression recorded the best accuracy thus far at 73.5% as well as precision 72.8%, recall 71.3%, and F1 score 72.0% which means good, but still moderate, prediction power. Support Vector Machining fared better on this point, with 75.6% accuracy and precision 74.9%, recall of 74.2%, and an F1 score of 74.5%.

TABLE I. COMPARATIVE ANALYSIS

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression [30]	73.5	72.8	71.3	72.0
Support Vector Machine [31]	75.6	74.9	74.2	74.5
Random Forest [32]	76.4	77.2	75.6	76.4
GNN + Attention Mechanism (Proposed)	98.8	90.3	88.9	89.6

With respect to the Collaboration of Random Forest, it finished this improvement with an accuracy of 76.4%, precision of 77.2%, and recall of 75.6%, all combined with an F1 score of 76.4%, which has affirmed its ability to handle complex decision boundaries. The GNN with Attention Mechanism thus surpasses all of them, hitting 98.8% accuracy, 90.3% precision, 88.9% recall, and an F1 score of 89.6%. This improvement arises because of association dependency in patient data; attention mechanism can thus more acquire feature importance for the most effective prediction of Dengue severity.

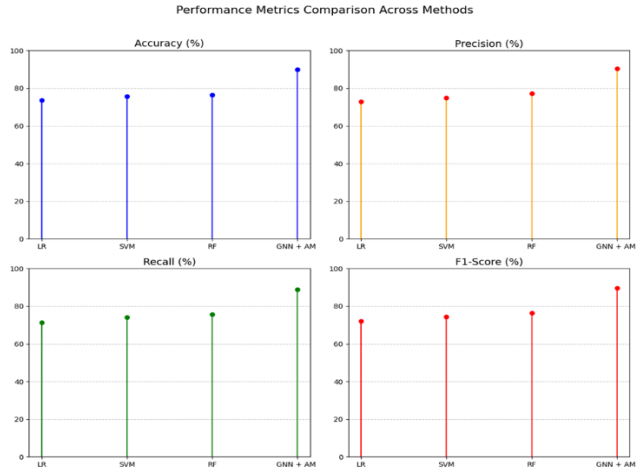


Fig. 9. Comparative analysis.

Fig. 10 illustrates the performance metrics- this study also contributes to a more in-depth understanding of the effectiveness of models. The GNN with attention mechanism surpassed all other models giving a precision of 98.8%, recall of 88.9%, and hence F1-score of 89.6%, and an AUC-ROC of 92.1%, indicating that the proposed model is not merely able to classify Dengue cases but also achieves good recall i.e. true Dengue-positive cases are not misclassified. The baseline GNN model was also competitive with an F1-score of 83.4% and AUC-ROC value of 86.5%, but still, it was less than the attention-based version's score. All other models, however, underperformed GNN with attention mechanism, with Random Forest having the lowest scores across all metrics, further affirming traditional limitations of ML techniques to capture complex relationships between symptoms that are exposed

within the findings. Thus, GNN with attention mechanism is proved the optimum model with respect to feature relevance improvement and, thus better classification accuracy- making it the best fit for any situation pertaining to classifying Dengue severity.

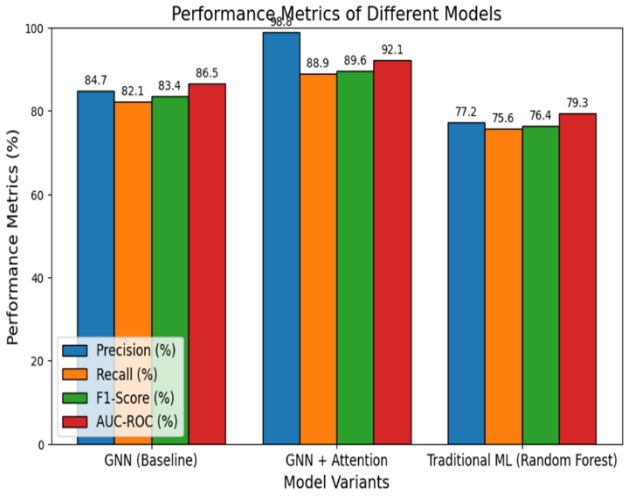


Fig. 10. Performance metrics.

The analysis using Jupyter Notebook produced plot visualizations which monitored model performance at different epochs. Studies used line plots to display how the model learned by showing both training and testing accuracy evolution. A visual examination of error minimization occurred through the plotting of training and testing loss curves. The visual displays demonstrate how the integration of GNN with attention mechanism enhances Dengue severity prediction through better accuracy rates and decreasing loss as time progresses.

B. Discussion

The experimental results evidently illustrate the efficacy of the developed GNN with Attention Mechanism (GNN-AM) to forecast dengue severity at high accuracy and reliability. The heterogeneous and balanced dataset, involving rich symptom profiles including fever, headache, arthritis, and bleeding, allows the model to generalize effectively over different patient states. The dramatic improvement in training and testing precision, as well as the drastic decrease in loss values, attests that it improves the model's capacity to concentrate on clinically important features with reduced noise from less descriptive attributes. In comparison to popular machine learning models such as Logistic Regression and Random Forest, which fail to represent intricate dependencies, GNN-AM utilizes patient relationships optimally with graph-based learning. Improved performance by the model in terms of evaluation metrics highlights its clinical value and serves as an exciting decision-support tool for early risk stratification, particularly for resource-limited settings.

VI. Conclusion and Future Scope

This paper presents a state-of-the-art GNN-AM model for forecasting dengue fever severity with significant advances in accuracy, explainability, and clinical utility over traditional machine learning solutions. With graph-based patient similarity modeling and attention-driven feature prioritization, the GNN-AM model attained an accuracy rate of 98.8%, which was far superior to the traditional classifiers. The model was also generalizable and generated richer insights into key clinical indicators, enabling more accurate disease surveillance and early risk stratification. Interoperability with EHRs and IoT devices can also empower real-time prediction and ongoing risk assessment. Federated learning implementation will provide privacy protection while enabling cooperative model training among healthcare facilities. In addition, the creation of lightweight variants of GNNs can make deployment feasible in resource-limited environments, including rural clinics and mobile health facilities. Integration of Explainable AI (XAI) methods—like SHAP values and attention heat maps—will increase transparency and foster clinician confidence in AI-based medical decision-making. Extending the GNN-AM framework to accommodate multimodal data integration and applying it to additional infectious diseases such as malaria, COVID-19, and influenza would demonstrate its potential as a strong and scalable platform for precision medicine and global public health management.

References

[1] "Global re-emergence of dengue fever: The need for a rapid response and surveillance - ScienceDirect." Accessed: Feb. 20, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2950194624000748

[2] W.-H. Wang et al., "Dengue hemorrhagic fever – A systemic literature review of current perspectives on pathogenesis, prevention and control," J. Microbiol. Immunol. Infect., vol. 53, no. 6, pp. 963–978, Dec. 2020, doi: 10.1016/j.jmii.2020.03.007.

[3] M. Colaneri et al., "Impact of early antiviral therapy on SARS-CoV-2 clearance time in high-risk COVID-19 subjects: A propensity score matching study," Int. J. Infect. Dis., vol. 149, p. 107265, Dec. 2024, doi: 10.1016/j.ijid.2024.107265.

[4] "DDPM: A Dengue Disease Prediction and Diagnosis Model Using Sentiment Analysis and Machine Learning Algorithms." Accessed: Feb. 20, 2025. [Online]. Available: https://www.mdpi.com/2075-4418/13/6/1093

[5] M. Salvi et al., "Multi-modality approaches for medical support systems: A systematic review of the last decade," Inf. Fusion, vol. 103, p. 102134, Mar. 2024, doi: 10.1016/j.inffus.2023.102134.

[6] H. Xu and K. M. J. Shuttleworth, "Medical artificial intelligence and the black box problem: a view based on the ethical principle of 'do no harm,'" Intell. Med., vol. 4, no. 1, pp. 52–57, Feb. 2024, doi: 10.1016/j.imed.2023.08.001.

[7] N. Papachristou et al., "Digital Transformation of Cancer Care in the Era of Big Data, Artificial Intelligence and Data-Driven Interventions: Navigating the Field," Semin. Oncol. Nurs., vol. 39, no. 3, p. 151433, Jun. 2023, doi: 10.1016/j.soncn.2023.151433.

[8] B. Khemani, S. Patil, K. Kotecha, and S. Tanwar, "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions," J. Big Data, vol. 11, no. 1, p. 18, Jan. 2024, doi: 10.1186/s40537-023-00876-4.

[9] "Precision Adverse Drug Reactions Prediction with Heterogeneous Graph Neural Network - Gao - 2025 - Advanced Science - Wiley Online Library." Accessed: Feb. 20, 2025. [Online]. Available: https://advanced.onlinelibrary.wiley.com/doi/full/10.1002/advs.202404671

[10] "Transparency of deep neural networks for medical image analysis: A review of interpretability methods - ScienceDirect." Accessed: Feb. 20, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521009057

[11] "Leveraging Climate Data for Dengue Forecasting in Ba Ria Vung Tau Province, Vietnam: An Advanced Machine Learning Approach."

Accessed: Feb. 20, 2025. [Online]. Available: https://www.mdpi.com/2414-6366/9/10/250

[12] J.-Y. Chen, C.-C. Hsieh, J.-T. Lee, C.-H. Lin, and C.-Y. Kao, "Patient stratification based on the risk of severe illness in emergency departments through collaborative machine learning models," Am. J. Emerg. Med., vol. 82, pp. 142–152, Aug. 2024, doi: 10.1016/j.ajem.2024.06.015.

[13] S. Seoni, V. Jahmunah, M. Salvi, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023)," Comput. Biol. Med., vol. 165, p. 107441, Oct. 2023, doi: 10.1016/j.compbiomed.2023.107441.

[14] A. Daneshkohan, M. Alimoradi, M. Ahmadi, and J. Alipour, "Data quality and data use in primary health care: A case study from Iran," Inform. Med. Unlocked, vol. 28, p. 100855, Jan. 2022, doi: 10.1016/j.imu.2022.100855.

[15] "A review of Explainable Artificial Intelligence in healthcare - ScienceDirect." Accessed: Feb. 20, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790624002982

[16] J. K. Chaw et al., "A predictive analytics model using machine learning algorithms to estimate the risk of shock development among dengue patients," Healthc. Anal., vol. 5, p. 100290, Jun. 2024, doi: 10.1016/j.health.2023.100290.

[17] S. Saha et al., "Hierarchical Deep Learning Neural Network (HiDeNN): An artificial intelligence (AI) framework for computational science and engineering," Comput. Methods Appl. Mech. Eng., vol. 373, p. 113452, Jan. 2021, doi: 10.1016/j.cma.2020.113452.

[18] X. Zhao, K. Li, C. K. E. Ang, and K. H. Cheong, "A deep learning based hybrid architecture for weekly dengue incidences forecasting," Chaos Solitons Fractals, vol. 168, p. 113170, Mar. 2023, doi: 10.1016/j.chaos.2023.113170.

[19] A. Nebli, M. A. Gharsallaoui, Z. Gürler, and I. Rekik, "Quantifying the reproducibility of graph neural networks using multigraph data representation," Neural Netw., vol. 148, pp. 254–265, Apr. 2022, doi: 10.1016/j.neunet.2022.01.018.

[20] D. K. Dhaked, O. Sharma, Y. Gopal, and R. Gopal, "Prediction of dengue patients using deep learning methods amid complex weather conditions in Jaipur, India," Discov. Public Health, vol. 22, no. 1, p. 58, Feb. 2025, doi: 10.1186/s12982-025-00448-2.

[21] "Graph neural networks in histopathology: Emerging trends and future directions - ScienceDirect." Accessed: Feb. 20, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841524003694

[22] H. Byeon et al., "Deep neural network model for enhancing disease prediction using auto encoder based broad learning," SLAS Technol., vol. 29, no. 3, p. 100145, Jun. 2024, doi: 10.1016/j.slast.2024.100145.

[23] "Integrating sentiment analysis with graph neural networks for enhanced stock prediction: A comprehensive survey - ScienceDirect." Accessed: Feb. 20, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772662224000213

[24] D. C. Andrade Girón, W. J. Marín Rodriguez, F. de M. Lioo-Jordan, and J. L. Ausejo Sánchez, "Machine Learning and Deep Learning Models for Dengue Diagnosis Prediction: A Systematic Review," Informatics, vol. 12, no. 1, Art. no. 1, Mar. 2025, doi: 10.3390/informatics12010015.

[25] "Graph Attention Networks: A Comprehensive Review of Methods and Applications." Accessed: Feb. 20, 2025. [Online]. Available: https://www.mdpi.com/1999-5903/16/9/318

[26] P. Muthuselvi and V. A. Devi, "Advancements in Predictive Techniques and Effective Management Strategies for Dengue Fever: A Comprehensive Review," in 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, 2024, pp. 2010–2017.

[27] M. D. Hunter, K. L. McKee, and E. Turkheimer, "Simulated nonlinear genetic and environmental dynamics of complex traits," Dev. Psychopathol., vol. 35, no. 2, pp. 662–677, 2023.

[28] J. B. Prajapati and B. G. Prajapati, "Clinical decision support system braced with artificial intelligence: a review," in International Conference on Image Processing and Capsule Networks, Springer, 2022, pp. 531–540.

[29] "Dengue." Accessed: Feb. 21, 2025. [Online]. Available: https://www.kaggle.com/datasets/dipayancodes/dengue

[30] S. Q. Ong, P. Isawasan, A. M. M. Ngesom, H. Shahar, A. M. Lasim, and G. Nair, "Predicting dengue transmission rates by comparing different machine learning models with vector indices and meteorological data," Sci. Rep., vol. 13, p. 19129, Nov. 2023, doi: 10.1038/s41598-023-46342-2.

[31] "Predicting dengue transmission rates by comparing different machine learning models with vector indices and meteorological data - PMC." Accessed: Feb. 21, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10625978/?utm_source=chatgpt.com

[32] S. Q. Ong, P. Isawasan, A. M. M. Ngesom, H. Shahar, A. M. Lasim, and G. Nair, "Predicting dengue transmission rates by comparing different machine learning models with vector indices and meteorological data," Sci. Rep., vol. 13, p. 19129, Nov. 2023, doi: 10.1038/s41598-023-46342-2.

Metaheuristic-Driven Feature Selection for IoT Intrusion Detection: A Hierarchical Arithmetic Optimization Approach

Jing GUO*, Dejun ZHU, Qing XU
School of Big Data, Chongqing College of Mobile Communication, Chongqing City, 401520, China
Chongqing Key Laboratory of Public Big Data Security Technology, Chongqing City, 401520, China

***Abstract*—The increasing sophistication of cyberattacks in Internet of Things (IoT) networks requires strong Intrusion Detection Systems (IDS) with optimal feature selection mechanisms. High-dimensional data, computational complexity, and suboptimal detection accuracy hinder conventional IDS mechanisms. To overcome these limitations, in this study, the Hierarchical Self-Adaptive Arithmetic Optimization Algorithm (HSAOA) is introduced as a new metaheuristic method for IDS feature selection. HSAOA combines a stochastic spiral exploration method, an adaptive hierarchical model of leaders and followers, and a differential mutation mechanism to improve exploration-exploitation balance, global search capability, and premature convergence. The NF-ToN-IoT dataset is used to test the model, wherein HSAOA undertakes the feature selection process, and classification accuracy is increased by utilizing Random Forest (RF). The experimental results indicate that the proposed HSAOA is better than other advanced approaches in accuracy, computational efficiency, and convergence speed. These results validate the proposed algorithm as a scalable and effective solution for enhancing cybersecurity in IoT environments by improving IDS performance and reducing feature selection complexity.**

***Keywords*—Intrusion detection; internet of things; feature selection; hierarchical arithmetic optimization; cybersecurity**

I. Introduction

A. Research Problem

The Internet of Things (IoT) is revolutionizing how devices communicate, enabling easy connection of all devices from various industries, from healthcare to smart cities and industrial automation [1]. While the high growth of connected devices opens new avenues in the IoT space, the vast increase in connected devices makes the IoT networks vulnerable to various cyberattacks [2]. These vulnerabilities indicate the need for strong Intrusion Detection Systems (IDSs) to monitor the network's traffic and detect attacks in real-time. Proper IDS are essential to secure IoT environments by detecting and preventing attacks and protecting the integrity and privacy of the sensitive data exchanged over IoT networks [3].

Conventional IDS models face challenges in IoT networks, particularly regarding computational efficiency and feature selection [4]. IoT networks generate abundant data, resulting in high-dimensional data that is challenging to process efficiently through conventional IDS mechanisms [5]. Most IDS mechanisms are marred by high false-positive rates and slow processing times, primarily due to inappropriate feature selection. These schemes either maintain irrelevant features or miss the most critical ones, resulting in low detection accuracy and considerable computational complexity. As such, the need for more adaptive and efficient IDS mechanisms to cope with the distinctive requirements of IoT security is ever-escalating [6].

B. Contribution

This study proposes the Hierarchical Self-Adaptive Arithmetic Optimization Algorithm (HSAOA), an innovative metaheuristic solution for feature selection in IDS. HSAOA incorporates an adaptive hierarchical model, differential mutations, and a stochastic spiral exploration method to reduce computational complexity while enhancing IDS accuracy.

Through the utilization of such novel features, the capability of the proposed method to identify the relevant features is significantly increased, along with a consequent reduction in the dataset's dimensionality. The technique is predicted to be better than conventional methods in accuracy, computation complexity, and performance, making the algorithm an effective instrument for IoT network security. This work primarily contributes to:

- Presenting HSAOA, an adaptive metaheuristic that improves feature selection for intrusion detection on the IoT by combining spiral-guided random walks, hierarchical leader-follower organization, and rank-based differential mutations for enhanced exploration, exploitation, and solution refinement.
- Applying HSAOA on the NF-ToN-IoT dataset and evaluating it with other state-of-the-art methods, demonstrating improvements over detection efficiency, feature salience, and computation efficiency.
- Performing thorough simulations to test the efficiency of HSAOA at providing high intrusion accuracy, fast convergence, and feature dimensionality reduction.

This study addresses the following research questions:

- How can a metaheuristic algorithm be designed to perform adaptive and efficient feature selection for IDS in IoT environments?
- Does the proposed method improve detection accuracy, convergence speed, and computational efficiency

compared to existing metaheuristic-based IDS approaches?

- Can the proposed solution scale effectively and remain robust in detecting a wide variety of attacks across real-world IoT datasets?

II. Related Work

Davahli, et al. [7] introduced a hybrid method, GA–GWO, using a Genetic Algorithm (GA) and Grey Wolf Optimizer (GWO) for IoT network intrusion detection. The approach aims to reduce the dimensionality of a large traffic volume by extracting the most relevant traffic features. The computational efficiency is enhanced using the hybrid method, and intrusions are detected efficiently with high accuracy and low false alarm rates.

Shahapure and Punitha [8] proposed the Water Moth Search Algorithm (WMSA) to train the Deep Recurrent Neural Network (Deep RNN) to identify intrusions in IoT networks. WMSA algorithm combines the Water Wave Optimization (WWO) and the Moth Search Optimization (MSO), which is tailored using a Wrapper approach for the feature selection process.

Gangula and V [9] suggested an Enhanced Flower Pollination Algorithm (EFPA) integrated with an ensemble classification method for IoT network intrusion detection. The algorithm applies EFPA for the best possible feature extraction from UNSW-NB15 and NSL-KDD datasets, followed by classification under Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) classifiers.

Alghanam, et al. [10] suggested a modified form of the Pigeon-Inspired Optimization (PIO) algorithm, augmented with a Local Search Algorithm (LS-PIO), to enhance Network Intrusion Detection Systems (NIDS) in IoT networks. The proposed model utilizes the ensemble learning method and several one-class classifiers to identify intrusions more efficiently.

Saheed, et al. [11] proposed the IoT-Defender model, combining the Modified Genetic Algorithm (MGA) and the Long Short-Term Memory (LSTM) network for IoT network intrusion detection. MGA performs feature selection, optimizes the LSTM parameters, and improves the model's performance in detecting attacks.

Sharma, et al. [12] developed the Multi-Objective Prairie Dog Optimization (MPDA) algorithm to select features in IoT NIDS. The algorithm uses an archive, grid, and non-dominance to enhance diversity and manage several objectives efficiently. On NSL-KDD, CIC-IDS2017, and IoTID20 datasets, the method attained fewer features, better accuracy, and reduced false alarm rates.

Ogunseyi and Thiyagarajan [13] presented a hybrid DL-based IDS for IoT networks, employing statistical methods and a metaheuristic algorithm for feature selection. The model employs LSTM networks and is trained using public datasets such as NF-BoT-IoT-v2 and IoTID20. The model attained excellent detection accuracy of 98.4% and 89.5% on the datasets.

Li, et al. [14] proposed the GA-tuned ensemble of deep transfer learning as a methodology for detecting intrusions in IoT networks. The system leverages pre-trained Convolutional Neural Networks (CNNs) and tuned GA to identify the optimum features. The optimum models are subsequently employed in a soft voting ensemble to improve the robustness of the detector. The framework attained 100% accuracy in 15 attack classes and outperformed other models, most notably in minority attacks like backdoors and ransomware.

Despite the extensive advancements in IDS for IoT networks, existing solutions are often subject to inefficiency in computational performance and feature choice, as highlighted in Table I. Most existing algorithms still depend on manual traffic features, resulting in increased complexity and diminished accuracy. Second, most metaheuristic-based IDS solutions do not effectively handle real-time discovery or handling of scarce resources in IoT settings.

Most solutions are still black-box, non-interpretable, and their practical deployment in cybersecurity is, as such, curtailed. Our research seeks to bridge these limitations by suggesting the HSAOA, aimed at maximizing the optimization of the choice of features, improving computational performance, and increasing the interpretable capability of IDS for IoT networks, considering accuracy and scalability in practical deployments.

TABLE I. An Overview of Related Works

Reference	Contribution	Dataset	Weakness
[7]	Hybridization of GA and GWO for dimensionality reduction	AWID	It is limited to wireless intrusion detection and lacks generalization across other types of IoT traffic
[8]	Combination of WMSA and Deep RNN for detecting malicious network activities	Custom	It requires intensive training and may not scale well for large IoT datasets
[9]	EFPA for feature selection and ensemble classification for attack detection	UNSW-NB15 and NSL-KDD	There is a risk of overfitting due to ensemble complexity, and the approach is dependent on large labeled datasets
[10]	LS-PIO for feature selection and ensemble learning	BoT-IoT, UNSW-NB15, and NSL-KDD	It incurs a high computational cost due to the ensemble and may not adapt well to real-time IoT data
[11]	MGA for feature selection and LSTM tuning in IoT-defender framework	BoT-IoT, UNSW-NB15, and N-BaIoT	While focused on resource-constrained devices, it may not perform well in large-scale networks
[12]	MPDA for multi-objective feature selection	NSL-KDD, CIC-IDS2017, and IoTID20	It struggles with real-time detection in dynamic IoT environments
[13]	Deep learning with metaheuristics for feature selection and LSTM-based IDS	NF-BoT-IoT-v2 and IoTID20	It has high complexity and lacks transparency in decision-making
[14]	GA-tuned ensemble of deep transfer learning for NetFlow-based intrusion detection	Custom	It may not generalize well beyond NetFlow datasets and is computationally expensive for real-time use

III. METHODOLOGY

A. Overview of Feature Selection for IDS

Feature selection is required to optimize IDS performance. Feature selection identifies the most informative features from the vast amount of data that IoT networks provide, eliminates redundant or non-informative characteristics, and simplifies computation considerably [15]. Feature selection is essential in conditions of scarcity when the capacity to process large amounts of data efficiently is of the utmost importance. As shown in Fig. 1, the IDS scans network traffic to identify anomalous behavior that may indicate potential security intrusions. The ultimate objective of the feature selection process is to reduce the dimensions so the most relevant features are used to identify the attack while maintaining high classification accuracy.

B. IDS System Architecture and Workflow

The overall architecture of the IDS mechanism, as shown in Fig. 2, involves feature selection followed by machine learning-based classification, including Random Forest. The process uses IoT traffic to detect network attacks, including data from NF-ToN-IoT datasets. The system has two stages: Training and Testing. The initial stage involves using the training set to generate the positions of the features and determine the optimum characteristics by testing them. The optimum classification features are selected to reduce data dimensions. The system identifies traffic as anomalous or normal based on classification features.

IDS identifies anomalous network traffic behavior. As a deviation from normality is detected, the system triggers an alarm so the network can react appropriately. IDS is routinely integrated with firewall subsystems to manage traffic in the network and bolster defenses against attacks. Feature selection is critical in filtering out only the most vital features to be processed, resulting in more accurate attack detection and fewer false alarms.

C. Arithmetic Optimization Algorithm

The Arithmetic Optimization Algorithm (AOA) is a new optimization method based on fundamental arithmetic operators, i.e., addition, subtraction, multiplication, and division [16]. AOA is designed to determine the optimal solution through repeated updates of candidate solution positions based on the operation. The basic phases of AOA, i.e., the initialization, exploration, and exploitation phases, are discussed in this section.

In the initiation process, AOA generates a preliminary population of candidate solutions. The position of each individual in the population is defined in the search space as follows in Eq. (1):

$$X_i^j = UB_j + rand \times (UB_j - LB_j), \quad i = 1, \ldots, N; \quad j = 1, \ldots, D \tag{1}$$

where, D is the spatial dimension of the problem, N is the population size, X_i^j is the position of the j^{th} dimension of the i^{th} individual, UB_j and LB_j are the upper and lower bounds of the j^{th} dimension, and $rand$ is a random value between 0 and 1.

After initialization, AOA proceeds iteratively, where it determines either the exploration stage or the exploitation stage based on the Mathematical Optimization Acceleration (MOA) process, as follows in Eq. (2):

$$MOA = Min + t \times \left(\frac{Max - Min}{T}\right) \tag{2}$$

where, T is the current iteration number, Max and Min are the upper and lower bounds of the objective function, and t represents the current iteration.

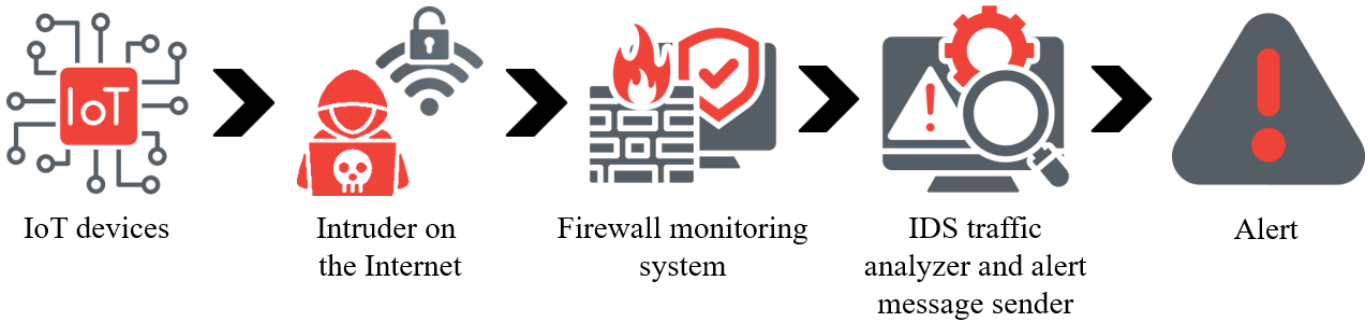


Fig. 1. Feature selection process in IDS.

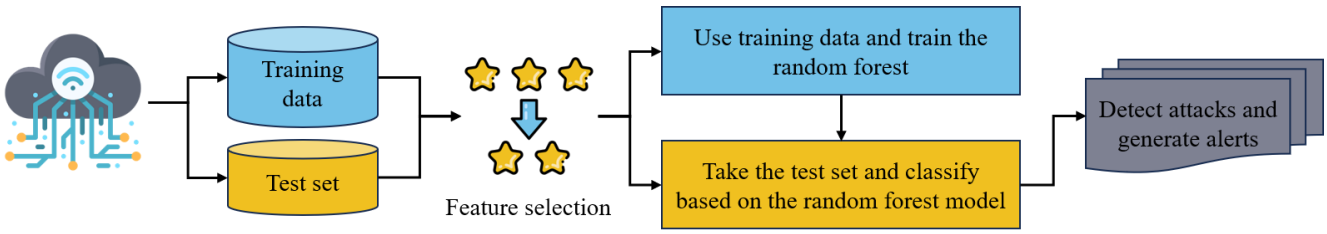


Fig. 2. Overview of the proposed IDS framework.

In the exploration period, AOA modifies individual positions through the application of multiplication or division operations. The update rule is as follows in Eq. (3):

$$X_i^j(t+1) = \begin{cases} X_i^j(t) \div (MOP + \epsilon) \times \left(UB_j - LB_j\right) + LB_j, & r_2 > 0.5 \\ X_i^j(t) \times \left(MOP \times \left(UB_j - LB_j\right) + LB_j\right), & r_2 \leq 0.5 \end{cases} \quad (3)$$

MOP is the modified optimization parameter, defined as in Eq. (4):

$$MOP = 1 - \left(\frac{t}{T}\right)^{\frac{1}{\alpha}} \quad (4)$$

r_2 is a stochastic value uniformly distributed between (0, 1), and ϵ is a small value to prevent division by zero.

In the exploitation stage, AOA updates the positions of individuals by applying addition or subtraction operators. The update rule for the exploitation phase is given by Eq. (5):

$$X_i^j(t+1) = \begin{cases} X_{best}^j - MOP \times \left(UB_j - LB_j\right) + LB_j, & r_3 > 0.5 \\ X_{best}^j + MOP \times \left(UB_j - LB_j\right) + LB_j, & r_3 \leq 0.5 \end{cases} \quad (5)$$

where, X_{best}^j is the position of the best solution found so far and r_3 is a random value uniformly distributed over the interval (0, 1).

As shown in Fig. 3, the AOA process is divided into three phases: initialization, exploration, and exploitation. The algorithm iterates through these phases until a stopping criterion (the maximum number of iterations) is attained, ultimately finding the optimum solution to the problem.

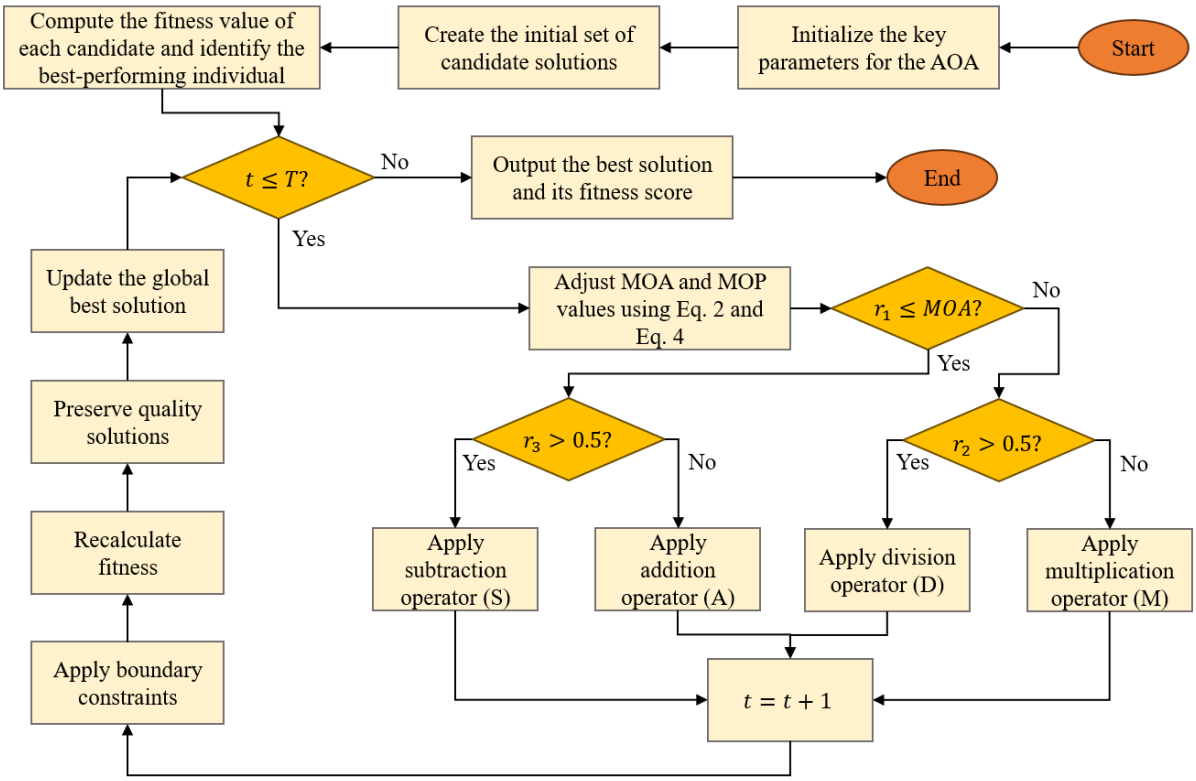


Fig. 3. Flowchart of AOA.

D. Enhancements in HSAOA

HSMAOA is an extension of the basic AOA able to overcome premature convergence and local optimum trapping, common in basic AOA. HSMAOA adds new mechanisms, such as a stochastic spiral exploration method, an adaptive hierarchical model of leaders and followers, and a differential mutation mechanism to enhance exploration and exploitation phases.

1) Stochastic spiral exploration method. During the exploration phase, AOA relies on multiplication and division to reposition individuals, potentially causing a loss of diversity following the initial population generation. To mitigate this issue, HSMAOA introduces a spiral-based random walk function as a replacement for conventional arithmetic operators, aiming to enhance global search capability. This mechanism leverages the Archimedean spiral to influence the step size of the random walk, taking advantage of its isometric growth properties, which have proven effective in diverse optimization applications. The Cartesian coordinates of the spiral are defined as Eq. (6):

$$x = cos(\theta) \times (r_0 + r_1 \times \theta) \quad (6)$$

$$y = sin(\theta) \times (r_0 + r_1 \times \theta)$$

$$\theta = \frac{1}{e} \times \left(\frac{n}{N}\right)$$

where, r_0 is the initial radius of the spiral (distance from the center), r_1 specifies the incremental growth of the radius per angular rotation, θ is the angle, n is the current iteration, and N is the maximum number of iterations.

In HSMAOA, the position update for an individual in the exploration phase is determined by Eq. (7):

$$X_i^d(n+1) = \begin{cases} X_{best}^d + sign(rand() - 0.5) \times cos(2\pi\theta) \times \\ \quad \left(\omega \times \left(X_a^d - X_\beta^d\right) + b \times 2\pi\theta\right), r_2 > 0.5 \\ X_{best}^d + sign(rand() - 0.5) \times sin(2\pi\theta) \times \\ \left(\omega \times \left(X_a^d - X_\beta^d\right) + b \times 2\pi\theta\right), \ otherwise \end{cases} \quad (7)$$

where, X_a^d and X_β^d represent the best and worst solutions for the d^{th} dimension, and ω is the step factor that is updated according to a Gaussian distribution with decreasing variance as in Eq. (8):

$$\omega = Gaussian\left(0, cos\left(\frac{\pi}{2} \times \frac{n}{N}\right)\right) \quad (8)$$

This spiral-guided mechanism enhances randomness in the search process and retains diversity after initialization, thereby improving the algorithm's ability to explore previously unvisited regions of the search space.

2) Adaptive hierarchical leader-follower structure. In HSMAOA, the adaptive hierarchical leader and follower scheme is employed to facilitate optimization processes by promoting cooperation and the exchange of information among individuals in the population. As shown in Fig. 4, the hierarchy is split into several levels. The individuals in the higher levels are assigned as leaders, and those in the lower levels are followers. The algorithm adapts the hierarchy through branching degree adjustment as follows in Eq. (9):

$$m = floor\left(\frac{1}{5} \times P - \left(\frac{1}{5} \times P - 2\right) \times \frac{n}{N}\right) \quad (9)$$

where, N, n, and P correspond to the maximum iteration limit, current iteration, and population size, respectively.

The hierarchical structure enables better interaction among individuals by maximizing the exchange of information between the leaders and followers. The leaders search the solution space independently using addition and subtraction to update positions, whereas the followers collaborate with the leaders and their parent individuals to improve solutions. This process helps overcome local optima and ensures better convergence in the exploration phase. Hierarchical adaptation and information exchange improve the algorithm's global search ability while maintaining local optimization.

3) Rank-Based differential mutation strategy. To enhance diversity and information exchange, HSMAOA incorporates a ranked selection-based differential mutation approach. This method prioritizes better individuals, allowing them to influence the next generation more significantly. The ranking of individuals is based on their fitness values, and the rank r_i for each individual X_i is used to calculate the mutation probability as in Eq. (10):

$$p_i = \frac{r_i}{P}, \ i = 1,2,\ldots,P \quad (10)$$

This ranked selection method prioritizes individuals with superior fitness, increasing their likelihood of influencing the mutation process. The mutation for the selected individuals is calculated as follows in Eq. (11):

$$V_i = X_i + \lambda \times (X_{r1} - X_{r2}) + (1 - \lambda) \times (X_{best} - X_{r3}) \quad (11)$$

where, X_{best} is the best individual, X_{r1}, X_{r2}, and X_{r3} are individuals selected based on their rank in the population, and λ is a control parameter set to 0.5 after extensive testing.

Finally, the position update for individual X_i is given by Eq. (12):

$$X_i = \begin{cases} V_i, & if\ fun(V_i) > fun(X_i) \\ X_i, & otherwise \end{cases} \quad (12)$$

where, $fun(V_i)$ and $fun(X_i)$ are the fitness values of the mutated individual and the current individual.

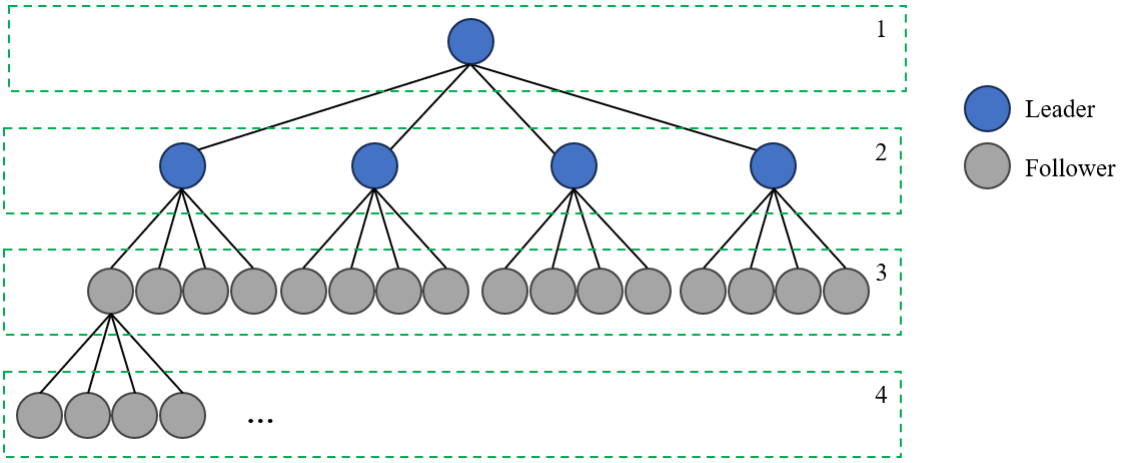


Fig. 4. Adaptive hierarchical structure of leaders and followers in HSMAOA.

E. Step-by-Step Implementation

As illustrated in Fig. 5, the architecture employs a systematic methodology comprising three primary phases: dataset preparation, feature extraction, and intrusion detection. At the heart of the method is HSMAOA, which enhances the feature extraction process to enhance the global search capability and the accuracy of detection of malicious behavior in IoT settings.

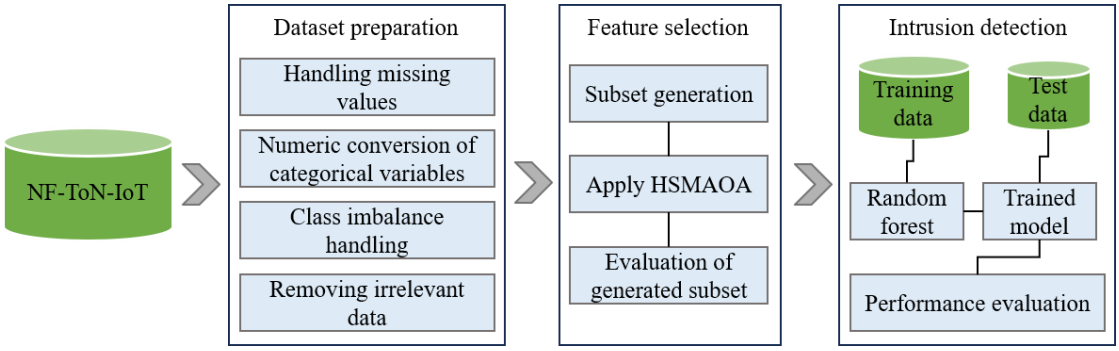


Fig. 5. Overview of the proposed IDS framework utilizing HSMAOA for feature selection.

HSMAOA leverages the advantages of a hierarchical structure and self-adaptive elements to mitigate the computational expense of selecting attributes from large datasets, such as NF-ToN-IoT. The IDS system becomes more accurate, reduces the dimensions of the data, and exhibits improved overall performance due to the altered optimization strategy. NF-ToN-IoT serves as training and test datasets, with pre-processed data used before attribute selection and classification.

1) Dataset preparation. The data preparation gives the dataset for the subsequent machine learning process and future analysis. In it, different inconsistencies within the data are corrected, including how the gap within the data, imbalance within the classes, and inconsistency can impact the outcome within the model negatively. The NF-ToN-IoT dataset goes through the preprocessing wherein the missing values are addressed, categorical variables are converted into numeric, the problem of class imbalance is addressed, and the irrelevant data is deleted.

Missing values in the dataset are imputed before passing the data to the HSMAOA. Imputation of the data is done in two phases: firstly, the most common value for each attribute is employed to replace the missing entries. Secondly, when required, imputed statistical values such as the mean or median value are used based on the attribute type so that the dataset is ready to be optimized.

NF-ToN-IoT contains categorical variables, and hence, the variables must be converted into a numeric form before being input into machine learning models. The categorical features are converted into a machine-learning input-compatible format using the technique of one-hot encoding.

A common challenge in IoT network datasets, such as NF-ToN-IoT, is the imbalance between classes (e.g., regular traffic vs. attack traffic). To address this issue, SMOTE (Synthetic Minority Over-sampling Technique) is applied, which generates synthetic data points for the minority class, thereby improving the classifier's ability to identify less frequent attack types.

Irrelevant attributes, such as timestamps, ports, and IP addresses, are removed from the dataset. These are generally not useful for identifying network intrusions within the context of an IDS and may result in overfitting or added complexity to the system. Eliminating irrelevant data aids in system performance and generalization.

2) Feature selection using HSMAOA. The second step in the process is the selection of the features, a step essential for the enhancement of the IDS' performance. The HSMAOA is utilized for the optimization of the subset of the best features for the task of intrusion detection. The hierarchical and adaptive approach of the algorithm ensures the selection of the best possible optimum features for training the model, striking a balance between exploration and exploitation on each optimization step. The choice of the features is made by optimizing the subset best describing the data using the HSMAOA.

HSMAOA employs the initial random subset generation method for generating various probable subsets. The subsets are then evaluated for the best subset of features that yields higher accuracy without an increase in the number of features. The goal is the identification of the best subset of features with the greatest accuracy and least computational cost.

Upon generation of the subset, it is assessed using the fitness function that balances the accuracy of the classifications and the number of selected features. The fitness function aims to make the selected features informative to detect intrusions without overfitting because there are too many features. Eq. (13) is the fitness evaluation of the feature subsets by HSMAOA.

$$Objective\ function = C_1 \times A + \frac{C_2}{NF} \tag{13}$$

where, A is the classification accuracy achieved by the subset, NF is the number of features in the subset, C_1 and C_2 are constants that balance the priority between accuracy and the number of features.

3) Intrusion detection using random forest. Once the feature selection process is completed, it is followed by

utilizing the chosen features at an intrusion detection level. The latter utilizes the RF algorithm, an efficient ensemble learning algorithm. The algorithm comprises several random subsets of data, on which decision trees have been generated. Each tree contributes one vote to the final classification outcome. The algorithm is best suited for classifying intrusions within networks because it can handle both the randomness of the data and the heterogeneity of the IoT dataset.

The chosen features are divided into the training data and the test data. The train data, which includes the selected characteristics and labels, is used for training the RF algorithm. The test data is utilized for testing the RF algorithm for determining the accuracy of the classifier. The labeled data trains the model throughout the process based on the construction of decision trees for determining the type of network traffic as benign or malicious based on the selection of the feature. Bootstrapping is invoked by the RF model, where different models are applied and fitted on various random data sets, resulting in a superior model. In the construction of a model, the performance can be estimated based on the calculation of the result on the test dataset. The true labels are compared with the labels predicted based on the model to determine the accuracy, precision, etc.

Several mitigation techniques have been implemented to stabilize the model, preventing it from overfitting. The process is validated by K-fold cross-validation, where the model is tested across data partitions. The algorithm minimizes bias during training. Regularization techniques, such as pruning decision trees and limiting tree depth, prevent overfitting by reducing complexity and simplifying the model. The ensemble method, which utilizes the RF algorithm, combines the outputs of multiple decision trees to enhance the stability and consistency of the model, thereby avoiding overfitting. The gains achieved by these countermeasure methods are measured through comparative analysis, with assistance from metrics such as accuracy, precision, and recall, to determine the value of the model.

IV. RESULTS

A performance analysis of the proposed HSAOA for feature selection in IDS is summarized in this section. The performance analysis is conducted using the NF-ToN-IoT dataset, where the data is divided into 80% for training and 20% for testing, as shown in Fig. 6. The simulation parameters used to evaluate the model are presented in Table II. The execution used Python on the Jupyter Notebook platform and the Pandas library. The hardware-software setup for experimentation consisted of an Intel Core i3-1115G4 processor at 2.95 GHz, 16 GB of RAM, 64-bit Windows 10, and a 256 GB SSD.

HSAOA effectively reduced the dimensionality of the NF-ToN-IoT dataset by selecting the 22 best features while preserving the accuracy of the classifier. The selected features are listed in Table III. The convergence behavior of HSAOA was compared with that of the baseline AOA in terms of fitness minimization over 20 iterations. As depicted in Fig. 7, it can be seen that HSAOA exhibits better convergence than AOA, with the significant change occurring at the 10th iteration. The objective function reaches an optimal and stable value, demonstrating improved global exploration capability and feature discrimination ability of HSAOA.

HSAOA-based feature engineering was designed for extracting discriminative features most informative for IDS classification performance. Features from have been chosen as they have high potential for characterizing abnormal activity. The feature set was extended for more discrimination between legitimate and malicious traffic, and hence reduced false positives. The effect of fitness value on classification accuracy is evident from Fig. 8. As the algorithm optimizes with a diminishing objective function, accuracy increases. The maximum accuracy obtained by the IDS system employing feature selection with HSAOA-converged was 99.94%.

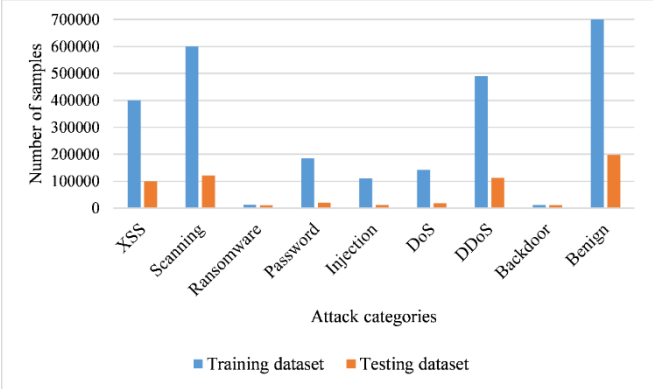


Fig. 6. Distribution of attack categories across the training and testing sets.

TABLE II. CONFIGURATION SETTINGS USED IN THE SIMULATION EXPERIMENTS

Parameter	Assigned value
Maximum iteration count	20
Population size	12
Total execution cycles	20
Weight factor C_1	0.6
Weight factor C_2	0.4

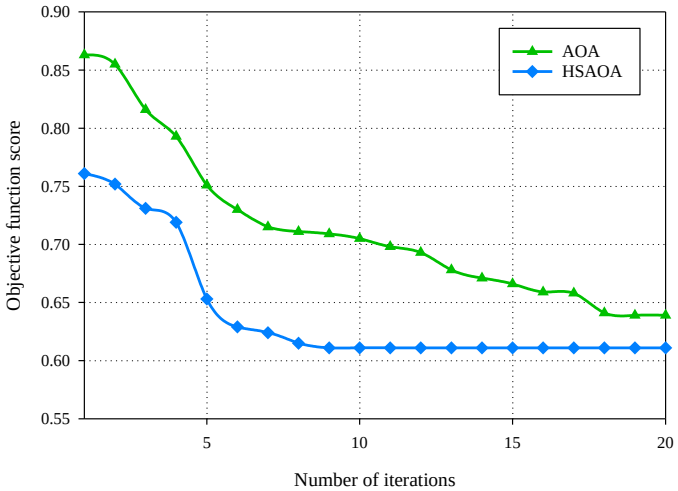


Fig. 7. Comparison of convergence behavior between the conventional AOA and the proposed HSAOA.

TABLE III. FINALIZED FEATURE SUBSET SELECTED USING THE PROPOSED HSAOA

Feature No.	Feature name	Feature No.	Feature name
F_1	Protocol type	F_{12}	Average throughput (Src to Dst)
F_2	Connection protocol	F_{13}	Average throughput (Dst to Src)
F_3	Packets received (inbound)	F_{14}	Packet Count ≤128 bytes
F_4	Packets sent (outbound)	F_{15}	Packets 128–256 bytes
F_5	TCP flag summary	F_{16}	Packets 256–512 bytes
F_6	Millisecond flow duration	F_{17}	Packets 512–1024 bytes
F_7	Outbound session duration	F_{18}	Packets 1024–1514 bytes
F_8	Minimum TTL	F_{19}	FTP command response code
F_9	Maximum TTL	F_{20}	DNS query category
F_{10}	Maximum IP packet length	F_{21}	DNS query identifier
F_{11}	Retransmitted byte count	F_{22}	IPv4 ICMP type

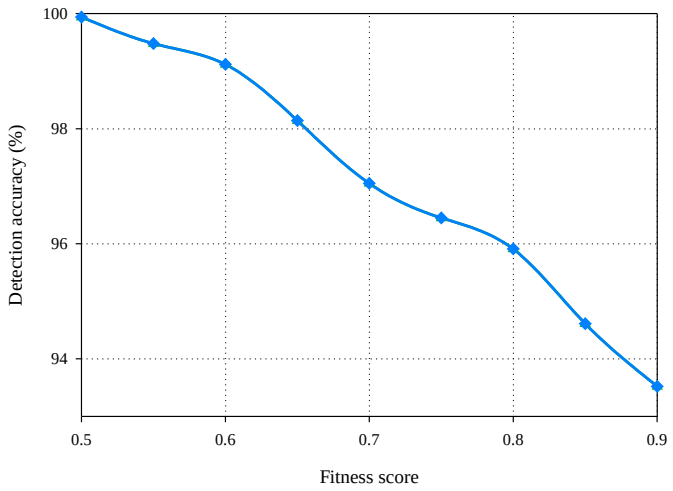


Fig. 8. Correlation between detection accuracy and fitness values during feature selection using the HSAOA.

After feature selection, the RF classification model was used for detection. The model's parameters are listed in Table IV. Performance metrics, detection rate, positive predictive value, sensitivity, and F1 score were determined using True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values, as outlined in Eq. (14) to Eq. (17). Table V reports comparative results with and without mitigation techniques (i.e., regularization, bootstrapping, and class balancing). Performance parameters improved substantially with mitigation to 99.94%, 99.87%, 99.78%, and 99.80%, respectively.

$$Detection\ rate = \frac{TN + TP}{FN + FP + TN + TP} \quad (14)$$

$$Positive\ predictive\ value = \frac{TP}{FP + TP} \quad (15)$$

$$Sensitivity = \frac{TP}{FN + TP} \quad (16)$$

$$F1\ score = \frac{2 \times Positive\ predictive\ value \times Sensitivity}{Positive\ predictive\ value \times Sensitivity} \quad (17)$$

TABLE IV. CONFIGURATION SETTINGS FOR RANDOM FOREST USED IN THE CLASSIFICATION TASK

Hyperparameter	Assigned value
Number of trees	100
Maximum tree depth	10
Minimum leaf samples	2
Minimum split samples	5
Class weighting strategy	Balanced
Randomization seed	42
Bootstrap sampling	Enabled

TABLE V. EVALUATION OUTCOMES OF CLASSIFICATION METRICS BEFORE AND AFTER MITIGATION ENHANCEMENTS

Evaluation measure	Pre-mitigation (%)	Post-mitigation (%)
F1 score	96.21	99.88
Sensitivity	95.52	99.83
Positive predictive value	96.87	99.89
Detection rate	97.15	99.95

The IDS was tested on various classes of attacks, as outlined in Table VI. The accuracy across the classes remained high throughout, with DDoS and scanning attacks achieving precision and recall of over 99%. These results provide evidence of the HSAOA-based IDS model's generalizability and the system's resistance to attacks. A comparison with other feature selection techniques on the given data is provided in Table VII. Compared to other state-of-the-art methods, the HSAOA method presents higher accuracy using fewer selected features without compromising detection power, proving efficient and practical.

TABLE VI. DETECTION PERFORMANCE PER INTRUSION TYPE USING HSAOA-BASED FEATURE SELECTION AND RANDOM FOREST CLASSIFIER

Type of attack	Detection rate (%)	Positive predictive value (%)	Sensitivity (%)	F1 score (%)
XSS	99.91	99.93	99.72	99.89
Scanning	98.78	98.99	98.12	98.95
Ransomware	99.95	99.53	99.54	96.82
Password	98.96	98.95	97.71	97.96
Injection	98.98	98.48	98.57	98.48
DoS	99.94	99.88	99.88	99.93
DDoS	99.94	99.91	99.89	99.92
Backdoor	99.49	99.51	99.42	99.57

TABLE VII. ACCURACY-BASED COMPARISON OF RECENT IDS APPROACHES USING THE TON-IOT DATASET

Reference	Classification model	Feature selection method	No. of selected attributes	Reported detection rate (%)
[17]	XG-Boost (v20)	Chi-squared + SMOTE	20	99.10
[18]	XG-Boost (v19)	Secure privacy-preserving framework	19	98.84
[19]	Random forest	Tabu search	16	83.12
[20]	Neural network	ReliefF	20	98.39
[21]	SVM	Filter + NSGA-II	13	99.48
[21]	SVM	Basic NSGA-II	18	98.86
[22]	CNN	Firefly-based optimization	20	96.66
[23]	CNN	Regularization strategy	8	97.95
[24]	CNN	Pearson correlation-based selection	85	98.75
Present study	Random forest	HSAOA	22	99.94

Security compliance was ensured through both formal and informal security analyses. Through ProVerif verification, formal analysis assures data confidentiality and authentication, as given in Table VIII. Queries 1 and 3 assessed the protocol's resistance to eavesdropping and impersonation, and query 2 raised an issue concerning data confidentiality under given conditions. Informal analysis, as depicted in Table IX, identified the system's response to specific attacks, such as DoS and Replay. Although the system was highly resilient to most attacks, some precautions, such as restricting repetitive requests, are proposed to provide enhanced protection against replay attacks.

TABLE VIII. FORMAL SECURITY VALIDATION OUTCOMES USING THE PROVERIF FRAMEWORK

Queries	Security goal assessed	Outcome	Clarification
Q1	Key/data confidentiality against attacker interception	Yes	Key is protected from exposure; communication is secure against interception.
Q2	Data confidentiality under attacker surveillance	No	Sensitive data may be accessed through insecure communication channels.
Q3	Authentication via request-response correspondence	Yes	Each response is verified to match a valid request, confirming authenticity.

TABLE IX. INFORMAL ASSESSMENT OF SYSTEM VULNERABILITIES AND CORRESPONDING RECOMMENDATIONS

Threat category	Vulnerability identified	Root cause explanation	Proposed mitigation
Replay attack	Present	System allows duplicate message transmissions	Implement time-based tokens or session tags
Injections	None	Alteration of data inputs is successfully detected	No intervention needed
DoS attack	Partial	System handles low volume but weak under heavy load	Enforce connection rate limiting
Backdoor	None	Authentication mechanism verified successfully	No further action necessary

V. DISCUSSION

The optimization gains made with the proposed HSAOA are a result of its ability to solve the significant problems outlined for IoT-based IDS, especially the high dimensionality of the data, computational intensity, and the low detection effectiveness resulting from the poor selection of features.

The integration of a spiral-oriented stochastic search process and adaptive hierarchical leader-follower structure into HSAOA also enhances the capability of the algorithm to exploit and explore the space of features more effectively. Unlike conventional AOA, which gets stuck depending on the fixedness of the arithmetic operator, HSAOA offers enhanced randomness and directional diversity across initial iterations. It facilitates greater discrimination across features and the ability of the model to select the most relevant attributes and delete noise and redundancy.

Another drawback of traditional metaheuristics is that they prematurely converge into local optima for highly complex and high-dimensional data like the kind used in IoT contexts. The rank-based differential mutation strategy suggested for HSAOA maintains population diversity throughout the optimization process, increasing global optima discovery potential. The outcome includes improved IDS accuracy and robustness when attack patterns evolve.

The IDS, after feature reduction via HSAOA, demonstrated consistent high classification accuracy and generalizability across different attack types in the NF-ToN-IoT dataset. This shows that the selected features retain sufficient discriminatory power for real-world use cases while reducing computational load. The use of RF as a robust classifier further contributed to detection stability, mainly when supported by mitigation strategies such as bootstrapping, SMOTE, and regularization.

Although the NF-ToN-IoT dataset provides a rich and diverse set of traffic and attacks, the result could possibly fail to generalize across the entire IoT application spaces (e.g., edge computing or industrial IoT applications). Testing using more benchmark sets should be contemplated to measure broader application. Even though HSAOA greatly reduces dimensionality, the initial search and mutation stages continue to be costly computationally. In extremely dynamic or real-time IoT applications, the overhead might limit performance unless reduced further or parallelized.

From the informal analysis output, the system was susceptible to the type of threats based on replays. Though a non-transparent outcome, it nonetheless implies that the adoption of time-based attributes or time-aware models can be beneficial for enhancing robustness against the same. The algorithm's behavior depends on some parameters that have been empirically set. Though optimally set using experimentation, there could also be automated or adaptive tuning procedures that could raise consistency within the use cases.

VI. CONCLUSION

This research introduced HSAOA to address the major challenge of feature selection for IDS in IoT settings. By employing an adaptive multi-layer hierarchy, spiral-guided random walk, and differential mutation operators, HSAOA improved the global search capability and local exploitation of the traditional AOA algorithm. The proposed approach was tested on the NF-ToN-IoT dataset, yielding a better convergence rate, feature reduction, and classification accuracy. Compared to traditional techniques, HSAOA significantly selected an optimal but highly informative feature subset, facilitating the RF classifier to achieve 99.94% accuracy while sustaining high precision, recall, and F1 measure over diverse types of attacks.

Formal and informal security analyses validated the resistance of the proposed IDS system to various attacks, verifying applicability to realistic IoT environments. Further work can be explored, where HSAOA is combined with deep learning techniques, running the model into actual IoT environments and exploring scalability over large-scale heterogeneous data sets. Additional optimization by adaptive parameter management and other optimization techniques are also ready to propel smart IoT security to new heights.

REFERENCES

[1] B. Pourghebleh, N. Hekmati, Z. Davoudnia, and M. Sadeghi, "A roadmap towards energy - efficient data fusion methods in the Internet of Things," Concurrency and Computation: Practice and Experience, vol. 34, no. 15, p. e6959, 2022.

[2] E. Geo Francis, S. Sheeja, E. Antony John, and J. Joseph, "IoT and Smart Device Security: Emerging Threats and Countermeasures," Securing the Digital Frontier: Threats and Advanced Techniques in Security and Forensics, pp. 217-241, 2025.

[3] R. Saadouni, C. Gherbi, Z. Aliouat, Y. Harbi, and A. Khacha, "Intrusion detection systems for IoT based on bio-inspired and machine learning techniques: a systematic review of the literature," Cluster Computing, vol. 27, no. 7, pp. 8655-8681, 2024.

[4] M. M. Rahman, S. Al Shakil, and M. R. Mustakim, "A survey on intrusion detection system in IoT networks," Cyber Security and Applications, vol. 3, p. 100082, 2025.

[5] B. Pourghebleh, K. Wakil, and N. J. Navimipour, "A comprehensive study on the trust management techniques in the Internet of Things," IEEE Internet of Things Journal, vol. 6, no. 6, pp. 9326-9337, 2019.

[6] F. Cerasuolo, G. Bovenzi, D. Ciuonzo, and A. Pescapè, "Adaptable, incremental, and explainable network intrusion detection systems for internet of things," Engineering Applications of Artificial Intelligence, vol. 144, p. 110143, 2025.

[7] A. Davahli, M. Shamsi, and G. Abaei, "Hybridizing genetic algorithm and grey wolf optimizer to advance an intelligent and lightweight intrusion detection system for IoT wireless networks," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 11, pp. 5581-5609, 2020.

[8] N. H. Shahapure and M. Punitha, "Water moth search algorithm-based deep training for intrusion detection in IoT," Journal of Web Engineering, vol. 20, no. 6, pp. 1781-1812, 2021.

[9] R. Gangula and M. M. V, "Network intrusion detection system for Internet of Things based on enhanced flower pollination algorithm and ensemble classifier," Concurrency and Computation: Practice and Experience, vol. 34, no. 21, p. e7103, 2022.

[10] O. A. Alghanam, W. Almobaideen, M. Saadeh, and O. Adwan, "An improved PIO feature selection algorithm for IoT network intrusion detection system based on ensemble learning," Expert Systems with Applications, vol. 213, p. 118745, 2023.

[11] Y. K. Saheed, O. H. Abdulganiyu, and T. Ait Tchakoucht, "Modified genetic algorithm and fine-tuned long short-term memory network for intrusion detection in the internet of things networks with edge capabilities," Applied Soft Computing, vol. 155, p. 111434, 2024.

[12] S. Sharma, V. Kumar, and K. Dutta, "Multi - objective prairie dog optimization algorithm for IoT - based intrusion detection," Internet Technology Letters, vol. 7, no. 6, p. e516, 2024.

[13] T. B. Ogunseyi and G. Thiyagarajan, "An Explainable LSTM-Based Intrusion Detection System Optimized by Firefly Algorithm for IoT Networks," Sensors, vol. 25, no. 7, p. 2288, 2025.

[14] J. Li, H. Chen, M. S. Othman, N. Salim, L. M. Yusuf, and S. R. Kumaran, "NFIoT-GATE-DTL IDS: Genetic algorithm-tuned ensemble of deep transfer learning for NetFlow-based intrusion detection system for internet of things," Engineering Applications of Artificial Intelligence, vol. 143, p. 110046, 2025.

[15] A. Grandhi and S. K. Singh, "Interrelated dynamic biased feature selection and classification model using enhanced gorilla troops optimizer for intrusion detection," Alexandria Engineering Journal, vol. 114, pp. 312-330, 2025.

[16] L. Abualigah, A. Diabat, S. Mirjalili, M. Abd Elaziz, and A. H. Gandomi, "The arithmetic optimization algorithm," Computer methods in applied mechanics and engineering, vol. 376, p. 113609, 2021.

[17] A. R. Gad, A. A. Nashat, and T. M. Barkat, "Intrusion detection system using machine learning for vehicular ad hoc networks based on ToN-IoT dataset," IEEE access, vol. 9, pp. 142206-142217, 2021.

[18] P. Kumar, G. P. Gupta, and R. Tripathi, "TP2SF: A Trustworthy Privacy-Preserving Secured Framework for sustainable smart cities by leveraging blockchain and machine learning," Journal of Systems Architecture, vol. 115, p. 101954, 2021.

[19] A. Nazir and R. A. Khan, "A novel combinatorial optimization based feature selection method for network intrusion detection," Computers & Security, vol. 102, p. 102164, 2021.

[20] R. H. Mohamed, F. A. Mosa, and R. A. Sadek, "Efficient intrusion detection system for IoT environment," International Journal of Advanced Computer Science and Applications, vol. 13, no. 4, 2022.

[21] A. K. Dey, G. P. Gupta, and S. P. Sahu, "Hybrid meta-heuristic based feature selection mechanism for cyber-attack detection in IoT-enabled networks," Procedia Computer Science, vol. 218, pp. 318-327, 2023.

[22] M. Dobrojevic, M. Zivkovic, A. Chhabra, N. S. Sani, N. Bacanin, and M. M. Amin, "Addressing internet of things security by enhanced sine cosine metaheuristics tuned hybrid machine learning model and results interpretation based on shap approach," PeerJ Computer Science, vol. 9, p. e1405, 2023.

[23] B. Ibrahim Hairab, H. K. Aslan, M. S. Elsayed, A. D. Jurcut, and M. A. Azer, "Anomaly detection of zero-day attacks based on CNN and regularization techniques," Electronics, vol. 12, no. 3, p. 573, 2023.

[24] S. Yaras and M. Dener, "IoT-based intrusion detection system using new hybrid deep learning algorithm," Electronics, vol. 13, no. 6, p. 1053, 2024.

Enhancing SVM and KNN Performance Through Preprocessing Pipelines for Interactive mHealth Applications

Btissam Elaziz[1], Charaf Eddine AIT ZAOUIAT[2], Mohamed Eddabbah[3]*, Yassin LAAZIZ[4]
LabTIC-ENSA of Tangier, Abdelmalek Essaadi University, Morocco[1, 4]
Polydisciplinary Faculty of Sidi Bennour, Chouaîb Doukkali University, Morocco[2]
Higher School of Technology – Essaouira, Cadi Ayyad University, Marrakesh, Morocco[3]

***Abstract*—Mobile health (mHealth) applications are increasingly relying on artificial intelligence (AI) to provide accurate and real-time decision support for healthcare delivery. However, achieving the optimal balance between processing time and accuracy remains challenging, especially for interactive applications that rely on cloud computing for scalability and performance. This study investigates the impact of data preprocessing techniques on the performance of two widely used machine learning algorithms, Support Vector Machine (SVM) and k-Nearest Neighbors (KNN), in cloud-based mHealth systems. We evaluate the effects of various scaling methods and dimensionality reduction techniques, on processing time and model accuracy. Our results demonstrate that preprocessing significantly improves model performance, with SVM achieving a precision of 0.72 and a processing time of 0.087 ms using StandardScaler, while KNN demonstrates the fastest processing times when paired with robust preprocessing. These findings underscore the importance of optimizing both data preparation and algorithmic efficiency for interactive mHealth applications. By enhancing model accuracy and reducing latency, this research contributes to the development of cost-effective, real-time mobile health systems that improve user experience and decision-making in healthcare.**

***Keywords*—Mobile health; cloud computing; machine learning; SVM; KNN; data preprocessing**

I. Introduction

The rapid advancements in mobile technology and AI have revolutionized the healthcare sector, giving rise to Mobile Health (mHealth) applications [1]. These applications leverage the portability of mobile devices and the analytical power of AI to provide real-time health monitoring, personalized diagnostics, and predictive analytics [2]. mHealth solutions have become indispensable tools in addressing global healthcare challenges, offering services such as remote patient monitoring, chronic disease management, and telemedicine consultations [3]. At the core of this transformation is the reliance on AI algorithms, which analyze vast amounts of data to deliver actionable insights. However, the computational demands of these algorithms often exceed the capabilities of mobile devices, especially in resource-constrained environments [4].

To address this limitation, cloud computing has emerged as a crucial enabler, providing scalable, cost-effective platforms that offload complex computations from mobile devices to powerful cloud servers [5]. By utilizing cloud platforms, mHealth applications can perform sophisticated AI-driven tasks such as processing high-dimensional data, running predictive models, and delivering real-time decision support without compromising the user experience. This synergy between AI and cloud computing allows for greater efficiency, enabling mHealth applications to cater to the growing demand for fast, reliable, and accurate healthcare solutions [6]. As mHealth continues to expand its reach, optimizing the interplay between AI and cloud platforms will be pivotal in meeting the performance, scalability, and cost requirements of these applications.

In mHealth, real-time interactive mobile systems have become critical in delivering timely and accurate healthcare services, such as chronic disease management, remote patient monitoring, and emergency decision support [7]. These systems rely heavily on AI algorithms to process large volumes of health data, often under stringent time constraints. However, achieving the delicate balance between processing time and accuracy presents significant challenges. Mobile devices, constrained by limited computational resources, depend on cloud platforms to execute complex AI algorithms.

This dependency [8] introduces latency concerns due to data transmission, processing, and response times, which can compromise the system's responsiveness—an essential feature for real-time health applications. Simultaneously, ensuring high accuracy is vital, as errors in AI predictions can lead to incorrect diagnoses or treatment decisions, directly impacting patient safety and trust. Additionally, the performance of AI models is heavily influenced by data preprocessing techniques, which are necessary for improving model efficiency but can also increase computational overhead [9]. For interactive mHealth systems, the trade-offs between optimizing preprocessing pipelines, minimizing cloud processing costs, and maintaining real-time responsiveness remain poorly addressed. Thus, the key challenge lies in optimizing AI model performance to deliver accurate and timely results while ensuring scalability and cost-efficiency in cloud-based environments. Addressing these challenges is essential for advancing mHealth solutions and providing seamless, reliable healthcare services to users worldwide.

This study addresses the key challenges of deploying AI in mHealth by focusing on the optimization of Support Vector Machine (SVM) and k-Nearest Neighbors (KNN) algorithms in

terms of data preprocessing, cloud integration, and real-time performance, with the goal of improving both patient outcomes and operational efficiency. SVM and KNN have been selected for their proven effectiveness in medical data classification, low implementation complexity, and robust performance on small to medium-sized datasets, which are common in mHealth applications. Their computational simplicity and compatibility with cloud-based architectures make them particularly suitable for real-time, scalable health systems that require both accuracy and efficiency. Moreover, the contrast between their learning paradigms—margin-based optimization for SVM versus instance-based learning for KNN—offers a valuable comparative framework for assessing preprocessing strategies and deployment trade-offs.

The primary aim of this study is to evaluate and optimize the performance of SVM and KNN algorithms for interactive mHealth AI systems deployed on cloud infrastructure. Specifically, the study seeks to:

- Analyze the Impact of Preprocessing Techniques: Investigate how various data preprocessing methods, such as scaling and dimensionality reduction, influence the accuracy and processing time of SVM and KNN models in a cloud-based environment.
- Optimize Algorithm Efficiency: Identify preprocessing pipelines and configurations that achieve the best trade-off between accuracy and processing time, ensuring the feasibility of real-time decision-making for mHealth applications.
- Enhance Scalability and Responsiveness: Explore how optimized AI models, combined with cloud computing resources, can improve the scalability and responsiveness of mHealth systems, ultimately reducing latency and enhancing user experience.

Through these objectives, the study aims to provide actionable insights for designing cost-effective, accurate, and efficient mHealth systems that improve user experience and leverage the power of cloud-based AI for real-time healthcare delivery.

This study is organized as follows: Following the introduction, the Section II provides a literature review to set the context. Section III presents the methodology. Section IVdiscusses the results; and Section V concludes the study.

II. Literature Review

AI applications in mHealth heavily rely on cloud computing to address the computational constraints of mobile devices, enabling advanced functionalities and seamless performance. Cloud platforms provide scalability, allowing them to process vast datasets and support a growing user base without compromising efficiency or performance. Additionally, cloud computing ensures secure data storage and facilitates easy access to patient information for authorized users, maintaining the privacy and integrity of sensitive health data [10]. These capabilities make cloud computing an indispensable backbone for AI-driven mHealth solutions, supporting their scalability, responsiveness, and security requirements.

SVM and KNN are widely used machine-learning algorithms in health- related research due to their simplicity, robustness, and effectiveness in handling diverse types of medical data [11]. SVM excels in analyzing high-dimensional datasets, a common characteristic of medical records, and performs reliably even with small training datasets, thanks to its ability to create clear class boundaries through kernel functions. SVM has been successfully applied in diagnosing diseases such as diabetes, cancer, and cardiovascular conditions, often achieving high accuracy, as demonstrated with the PIMA Indian Diabetes dataset. Its application extends to medical image analysis, including tumor detection, brain abnormalities, and retinal scans, and to genomic and proteomic data, where it identifies genetic markers and classifies diseases based on gene expression. However, while SVM is highly effective when dataset margins are clear, it may struggle with imbalanced datasets unless enhanced with techniques like cost-sensitive learning [12]. KNN, on the other hand, is a non-parametric algorithm that relies on the proximity of data points, making it particularly suitable for heterogeneous medical datasets. It is valued in clinical settings for its simplicity and interpretability, with applications including disease classification (e.g., diabetes, heart disease), medical decision support through patient record comparisons, and patient similarity assessments for personalized treatment planning. While KNN performs well with smaller datasets and clean, normalized data, it becomes computationally intensive as dataset sizes grow, requiring preprocessing and dimensionality reduction techniques like PCA or LDA [13]. When comparing the two, SVM generally outperforms KNN in high-dimensional and smaller datasets due to its mathematical rigor and margin maximization principle. However, KNN offers simplicity and ease of implementation, making it attractive for specific applications despite its higher computational demands for large datasets. Both algorithms benefit from preprocessing techniques, such as normalization and dimensionality reduction, which significantly enhance their performance and make them highly applicable in health AI applications. Together, SVM and KNN provide complementary strengths, supporting diverse healthcare tasks, from diagnostics to personalized medicine.

Research on SVM and KNN algorithms, and their optimization for mHealth applications challenges include the limited computational resources of mobile devices, the need for real-time processing, and the lack of detailed evaluations on integrating local preprocessing with cloud-based AI execution [14][15]. Existing studies also insufficiently address the application of preprocessing techniques to dynamic and heterogeneous mHealth datasets and overlook scalability and user experience for large user bases [16]. These gaps hinder the effective implementation of SVM and KNN in mHealth, where real-time accuracy, efficiency, and scalability are crucial [17].

III. Methodology

A. Dataset Description: PIMA Indian Diabetes Dataset

The PIMA Indian Diabetes dataset [18], sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, is a widely used benchmark for machine learning studies in healthcare. It comprises medical data collected from 768 female patients of Pima Indian heritage, aged 21 years or older. The dataset aims to predict the presence of diabetes based

on diagnostic measurements, making it particularly relevant for studies in mHealth applications. Table I summarizes the dataset's predictor variables.

TABLE I. THE DATASET INCLUDES THE FOLLOWING PREDICTOR VARIABLES AND TARGET VARIABLE

Feature	Description	Data Type
Pregnancies	Number of pregnancies.	int64
Glucose	Plasma glucose concentration (mg/dL) after a 2-hour oral glucose tolerance test.	int64
Blood Pressure	Diastolic blood pressure (mm Hg).	int64
Skin Thickness	Triceps skinfold thickness (mm).	int64
Insulin	2-hour serum insulin (mu U/mL).	int64
BMI	Body Mass Index (kg/m²).	float64
Diabetes Pedigree Function	A score representing the likelihood of diabetes based on family history.	float64
Age	Age of the individual.	int64
Outcome	Binary target variable indicating diabetes diagnosis (1 = diabetic, 0 = non-diabetic).	int64

B. Data Preprocessing Techniques

Preprocessing techniques are critical for improving the performance and efficiency of machine learning models as shown in Fig. 1.

This study utilizes various preprocessing methods, including scaling techniques and dimensionality reduction methods, to prepare the data for effective application of SVM and KNN algorithms. These techniques address challenges such as feature imbalance, computational efficiency, and overfitting, ensuring optimized performance in cloud-based mobile health mHealth applications [19].

1) Data scaling techniques

a) Standard Scaler (SS): The Standard Scaler technique is a data preprocessing method used to standardize the features of a dataset. It transforms the data such that it has zero mean and unit variance. This technique is commonly used in machine learning algorithms to ensure that all features are on a similar scale, which can improve the performance of the models. z is the standardized value of a feature; x is the original value of the feature; μ is the mean of the feature and σ is the standard deviation of the feature. A value x is transformed (scaled) into σ using Eq. (1):

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

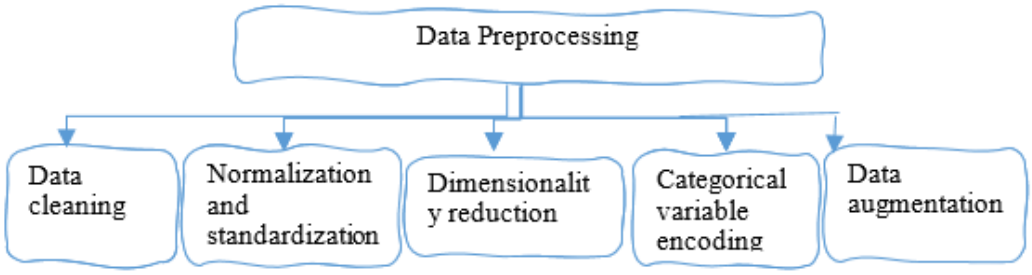


Fig. 1. Overview of various data preprocessing techniques.

b) Min-Max Scaler (MMS): Min-max scaler is a data normalization technique that transforms the values of a dataset to a specific range, typically between 0 and 1. It is calculated using Eq. (2):

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

where, X_{scaled} is the scaled value, X is the original value, X_{min} is the minimum value in the dataset, and X_{max} is the maximum value in the dataset.

c) Maximum Absolute Scaler (MAS): The Maximum Absolute Scaler is a method used to transform numerical data into a range between 0 and 1. This is achieved by dividing each value by the maximum absolute value present in the dataset. It can be mathematically expressed asashown in Eq. (3):

$$X_{scaled} = \frac{x}{max(|x|)} \quad (3)$$

where, x represents the original value and X_{scaled} represents the transformed value.

d) Robust Scaler (RS): The three scaling techniques mentioned earlier are highly affected by outliers because they rely on each variable's mean or minimum and maximum values. However, the Robust Scaler aims to reduce the impact of outliers by centering the data on the median $Q_2(x)$ and scaling it based on the interquartile range, which is the difference between the first quartile $Q_1(x)$ and the third quartile $Q_3(x)$ of x, as shown in Eq. (4):

$$X_{scaled} = \frac{X - Q_2(x)}{Q_3(x) - Q_1(x)} \quad (4)$$

e) Quantile Transformer (QT): The Quantile Transformer is a statistical technique that transforms a dataset's distribution to a desired target distribution by mapping original data values to new values. It is valuable for scenarios like data normalization, outlier mitigation, and preparation for statistical models. The Quantile Transformer can be represented by Eq. (5):

$$Y = (F_{target}^{-1}(F_{original}(x))) \quad (5)$$

where, $F_{original}$ is the cumulative distribution function (CDF) of the original dataset, and F_{target}^{-1} is the inverse CDF of the desired target distribution.

f) Normalization (N): A normalizer standardizes values to a uniform scale, aiding comparison by removing scale disparities, often expressed as Eq. (6). This equation scales values from 0 to 1, adjusting the minimum to 0 and the maximum to 1. Other techniques may employ varied equation depending on data needs.

$$Normalized\ value = \frac{Value - Minimum\ value}{Maximum\ value - Minimum\ value} \quad (6)$$

2) Dimension reduction techniques

a) Principal Component Analysis (PCA): PCA is a widely used statistical technique in multivariate analysis, primarily employed for dimensionality reduction and data compression [20]. Its main goal is to identify patterns, emphasizing similarities and differences in the data while preserving most of the original variations. PCA achieves four main objectives: extracting relevant information, compressing data by retaining distinguishing features, simplifying data description, and enabling analysis of observational structure. Eq. (7), which calculates variance (var(x)), is integral to PCA.

$$var(x) = \frac{\sum(x_i - \underline{x})^2}{N} \quad (7)$$

b) Linear Discriminant Analysis (LDA): Linear Discriminant Analysis (LDA) is a statistical technique for both dimensionality reduction and classification [21]. It aims to find a linear combination of features that maximizes class separability by maximizing the ratio of between-class and within-class scatters. LDA is widely applied in pattern recognition, machine learning, and data mining, seeking a projection vector that optimizes between-class scatter while minimizing within-class scatter. The between-class scatter matrix S_B is calculated by Eq. (8):

$$S_B = \big(mean(X_c) - mean(X)\big) * \big(mean(X_c) - mean(X)\big)^T \quad (8)$$

where, mean (X_c) is the mean vector of each class and mean(X) is the mean vector of the entire dataset.

The within-class scatter matrix S_w is calculated using Eq. (9):

$$S_w = \sum\big(X_i - mean(X_i)\big) * \big(X_i - mean(X_i)\big)^T \quad (9)$$

where, Σ represents the summation of all samples in each class.

The goal of LDA is to find the projection vector w that maximizes Fisher's criterion, as given in Eq. (10):

$$J(w) = (w^T * S_B * w)/(w^T * S_w * w) \quad (10)$$

The solution to this optimization problem can be obtained by finding the eigenvectors corresponding to the largest eigenvalues of the matrix inverse of $(S_w)^{-1} * S_B$. Once the projection vector w is obtained, it can be used to project new samples onto the linear discriminant space for classification or dimensionality reduction purposes.

c) SVM algorithm: SVM is an ML algorithm that finds the best line or plane to separate different classes of data points, as shown in Algorithm 1. It uses a subset of training data called support vectors to determine this line or plane. SVM can handle both linear and non-linear data by using different kernel functions. One of its advantages is its ability to handle high-dimensional data without overfitting. It can also work well with small training samples. SVM has been extensively studied and improved with different techniques such as using different kernels and handling imbalanced datasets. Overall, SVM is a powerful and widely used technique for classification problems [22].

Algorithm 1: Pseudo code of the SVM algorithm

Set $Input = (x_i, y_i)$ where $i = 1,2,3, \dots, N$, $x_i = R^n$ and $y_i = \{+1, -1\}$.
Assign $f(X) = \omega^T x_i + b = \sum_{i=1}^{N} \omega T xi + b = 0$.
Minimize the QP problem as, $\varphi(\omega, \varepsilon) = \frac{1}{2}\|\omega\|^2 + C.\left(\sum_i^N \varepsilon_i\right)$.
Calculate the dual Lagrangian multipliers as $minL_p = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{N} x_i y_i(\omega xi + b) + \sum_{i=1}^{N} x_i$.
Calculate the dual quadratic optimization (QP) problem as $L_D = \sum_{i=1}^{N} x_i - \frac{1}{2}\sum_{i,j}^{N} x_i x_j \ \ y_i y_j\left(x_i, x_j\right)$.
Solve dual optimization problem as $\sum_{i=1}^{N} y_i x_i = 0$.
Output the classifier as $f(X) = sgn\left(\sum_{i=1}^{N} x_i y_i(x.x_i) + \sum_{i=1}^{N} x_j y_j\left(x.x_j\right)\right)$

d) KNN algorithm: K-Nearest Neighbors (KNN) is a classification algorithm that assigns the label of a classified data point to an unclassified data point that is closest to it. This is done by considering the majority class of its K nearest data points. The distance between data points is typically measured using Euclidean distance (DE) or Manhattan distance (DM) [23] as shown in Algorithm 2:

Algorithm 2: Pseudo code of the KNN algorithm

Input: X: training data, Y: class labels of X, K: number of nearest neighbors.
Output: Class of a test sample x.
Start
Classify (X, Y, x)
For each sample x do
Calculate the distance: $d(x, X) = \sqrt{\sum_{i=1}^{n}(x_i - X_i)^2}$
End for
Classify x in the majority class: $C(x_i) = argmax_k \sum x_{j \in KNN} C\left(X_j, Y_k\right)$
End

e) Experimental design: The experimental design systematically evaluates the performance of SVM and KNN models in cloud-based mHealth systems, focusing on how preprocessing techniques affect accuracy and processing time (Fig. 2). Initially, baseline testing is conducted without preprocessing to establish reference metrics, including accuracy, precision, recall, F1-score, and processing time, highlighting limitations like bias from feature magnitudes and inefficiencies from data redundancy. Next, the impact of scaling techniques, such as Standard Scaler, Min-Max Scaler, Robust Scaler, and Quantile Transformer, is assessed by applying each method to the dataset and evaluating

performance metrics. This step identifies effective scaling methods for reducing bias and improving efficiency. Dimensionality reduction techniques, including Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), are then evaluated to understand their influence on computational cost and classification accuracy. PCA retains components with the most variance, while LDA maximizes class separability, both offering insights into trade-offs between efficiency and performance. Finally, combined testing explores the interaction between scaling and dimensionality reduction, with pipelines like Standard Scaler + PCA and Min-Max Scaler + LDA being tested to optimize preprocessing for real-time mHealth applications. This comprehensive approach ensures a balance between accuracy and processing time, enabling scalable and efficient model deployment.

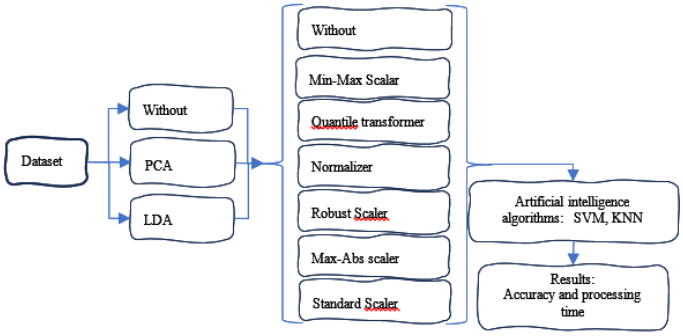


Fig. 2. Experimental Design

f) Metrics collected: Performance Metrics: Accuracy, precision, recall, F1-score.

- Efficiency Metrics: Processing time (measured in milliseconds).
- Trade-off Analysis: Evaluate the trade-offs between computational efficiency and model accuracy.

g) Computational environment: The experiments were conducted in a computational environment designed to replicate typical resource-constrained scenarios found in mHealth applications. This setup ensures the findings apply to real-world cloud-based systems supporting mHealth platforms.

- Hardware: Processor: AMD Ryzen 5 5500U CPU with Radeon Graphics, 2.10 GHz. Memory: 8 GB of RAM.
- Software: Programming Language: Python 3.8.

IV. Results and Analysis

The study evaluates the performance of SVM and KNN under various preprocessing scenarios, focusing on the following metrics: Processing Time (in milliseconds), Accuracy, Precision, Recall, and F1-Score.

The experiments demonstrate how scaling techniques and dimensionality reduction methods affect the efficiency and accuracy of these algorithms when applied to a health-related dataset.

A. Performance Without Preprocessing

As shown in Table II, SVM performs better than KNN in terms of accuracy (0.71 vs. 0.69) and is faster (0.063 ms vs. 0.13 ms). Without preprocessing, the data may contain bias and unnecessary information, making the model less efficient. This shows that preprocessing is important to improve speed and performance.

TABLE II. Metrics for SVM and KNN

Metric	Accuracy	Precision	Recall	F1-Score	Processing Time (ms)
SVM	0.71	0.70	0.71	0.69	0.063
KNN	0.69	0.69	0.69	0.69	0.13

B. Performance only with Scaling Techniques

Table III presents the impact of different scaling methods on the performance of SVM and KNN. Scaling significantly improves both accuracy and processing time for the two models. For SVM, the Min-Max Scaler achieves the highest accuracy (0.73) with a moderate processing time of 0.12 ms, while the Standard Scaler offers a good trade-off between accuracy (0.72) and faster processing (0.087 ms). In the case of KNN, the Standard Scaler provides the best accuracy (0.75) along with efficient processing (0.085 ms), making it particularly suitable for real-time applications. Although the Quantile Transformer yields moderate accuracy, its higher computational cost reduces its practicality. These results demonstrate that the choice of scaling method directly influences model performance and efficiency.

TABLE III. Metrics for SVM and KNN with Scaling Techniques

Scaling Technique	Standard Scaler		MinMax Scaler		Robust Scaler		Quantile Transformer	
Metric	Accuracy	Processing Time (ms)	Accuracy	Processing Time (ms)	Accuracy	Processing Time (ms)	Accuracy	Processing Time (ms)
SVM	0.72	0.087	0.73	0.12	0.66	0.13	0.71	0.42
KNN	0.75	0.085	0.73	0.12	0.68	0.063	0.74	0.13

C. Performance only with Dimensionality Reduction

Table IV shows that LDA outperforms PCA for both SVM and KNN. SVM with LDA achieves higher accuracy (0.74) and faster processing (0.085 ms) than with PCA. KNN also performs better and faster with LDA (0.047 ms). LDA's strength in maximizing class separability explains its superior results, making it the preferred choice for this dataset.

TABLE IV. Metrics for SVM and KNN with PCA and LDA

Metric	SVM + PCA	SVM + LDA	KNN + PCA	KNN + LDA
Precision	0.70	0.73	0.64	0.69
Recall	0.71	0.74	0.66	0.69
F1-score	0.69	0.73	0.65	0.69
Accuracy	0.71	0.74	0.66	0.69
Processing time (ms)	0.092	0.085	0.1	0.047

D. Performance with Dimensionality Reduction and Scaling Techniques

Table V shows that combining scaling and dimensionality reduction improves performance. For SVM with PCA, Min-Max Scaler gives the best accuracy (0.73), while Standard Scaler is faster (0.087 ms). For KNN with PCA, Min-Max Scaler has the highest precision (0.75), but Robust Scaler is more efficient (0.063 ms). With LDA, Standard Scaler provides the best accuracy and fastest processing, especially for SVM (0.041 ms).

TABLE V. METRICS FOR SVM WITH PCA AND SCALING TECHNIQUES

Scaling Technique	Max-Abs Scaler	Min-Max Scaler	Normalizer	Quantile Transformer	Robust Scaler	Standard Scaler
Precision	0.71	0.72	0.42	0.71	0.66	0.72
Recall	0.71	0.73	0.65	0.71	0.67	0.73
F1-Score	0.71	0.73	0.65	0.71	0.67	0.73
Processing time (ms)	0.12	0.12	0.14	0.42	0.13	0.087

TABLE VI. METRICS FOR KNN WITH PCA AND SCALING TECHNIQUES

Scaling Technique	Max-Abs Scaler	Min-Max Scaler	Normalizer	Quantile Transformer	Robust Scaler	Standard Scaler
Precision	0.71	0.75	0.66	0.74	0.68	0.75
Recall	0.69	0.73	0.66	0.74	0.68	0.75
F1-Score	0.69	0.73	0.66	0.74	0.68	0.75
Processing time(ms)	0.13	0.12	0.11	0.13	0.063	0.085

Table VI shows that for KNN with PCA, Min-Max and Standard Scalers achieve the best metrics (0.75), with Standard Scaler being faster (0.085 ms). Robust Scaler is the most efficient (0.063 ms) but slightly less accurate. Normalizer performs worst (0.66). This highlights Standard Scaler as the best balance of accuracy and speed.

TABLE VII. METRICS FOR SVM WITH LDA AND SCALING TECHNIQUES

Scaling Technique	Max-Abs Scaler	Min-Max Scaler	Normalizer	Quantile Transformer	Robust Scaler	Standard Scaler
Precision	0.73	0.73	0.65	0.73	0.73	0.73
Recall	0.74	0.74	0.68	0.74	0.74	0.74
F1-Score	0.73	0.73	0.64	0.73	0.73	0.73
Processing time(ms)	0.063	0.08	0.1	0.12	0.085	0.041

Table VII shows that for SVM with LDA, all scaling techniques except Normalizer achieve strong and consistent metrics (precision, recall, F1-score ≈ 0.73–0.74). Standard Scaler offers the best processing time (0.041 ms), making it the most efficient. Normalizer performs the worst across all metrics. These results highlight Standard Scaler as the most effective option when combining LDA with SVM.

TABLE VIII. METRICS FOR KNN WITH LDA AND SCALING TECHNIQUES

Scaling Technique	Max-Abs Scaler	Min-Max Scaler	Normalizer	Quantile Transformer	Robust Scaler	Standard Scaler
Precision	0.69	0.69	0.69	0.69	0.69	0.69
Recall	0.69	0.69	0.69	0.71	0.69	0.69
F1-Score	0.69	0.69	0.69	0.69	0.69	0.69
Processing time(ms)	0.07	0.072	0.063	0.084	0.047	0.032

Table VIII shows that KNN with LDA performs consistently across all scaling techniques, with precision, recall, and F1-score around 0.69. Quantile Transformer slightly improves recall (0.71), while Standard Scaler provides the fastest processing time (0.032 ms). These results indicate that although performance is similar, Standard Scaler is the most efficient.

Fig. 3(a) compares SVM and KNN accuracy and processing time with various scalers. Standard and Min-Max Scalers yield the highest accuracy, but Standard Scaler is fastest. Quantile Transformer is accurate but much slower.

Fig. 3(b) presents the performance of SVM and KNN using PCA and LDA. Both dimensionality reduction techniques yield comparable results across precision, recall, F1-score, and accuracy. However, LDA shows better efficiency, especially with KNN, offering lower processing time.

Fig. 3(c) shows SVM performance using different scaling techniques, with PCA Standard Scaler and MinMax Scaler achieving high precision, recall, and F1-score, while Standard Scaler has the lowest processing time. In contrast, Quantile Transformer offers good metrics but with the highest time cost.

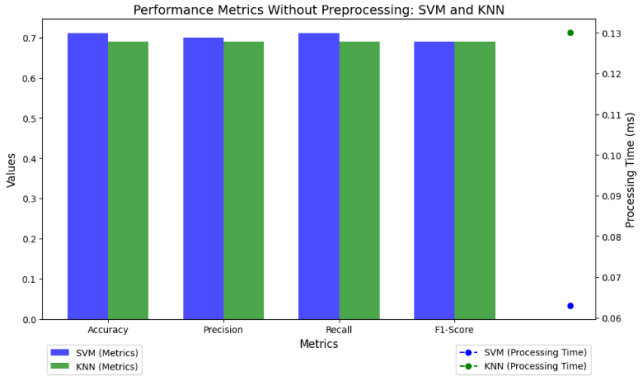
Fig. 3(d) shows that QuantileTransformer offers the best KNN performance but is the slowest, while RobustScaler is fastest with lower accuracy StandardScaler and MinMaxScaler provide a good balance between performance and processing time.

Fig. 3(e) shows that QuantileTransformer gives the best SVM performance but with the highest processing time, while StandardScaler is the fastest with strong accuracy. Other scalers like MinMaxScaler and RobustScaler offer a balanced trade-off between performance and speed.

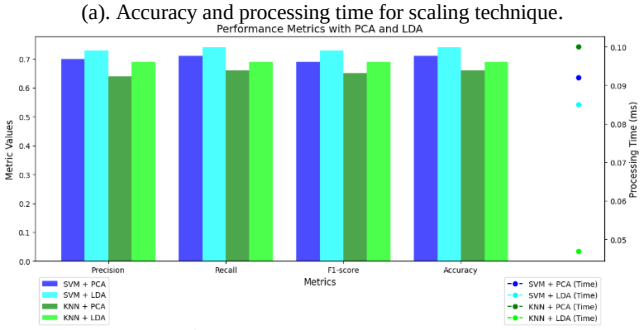
Fig. 3(f) shows that QuantileTransformer gives the highest KNN performance with LDA but takes the longest time to process. StandardScaler is the fastest and still maintains strong performance. Other scalers like MinMaxScaler and RobustScaler also show consistent accuracy with lower processing times, indicating good overall efficiency.

The findings in Fig. 3(a) to Fig. 3(f), highlight the practical implications of combining preprocessing with SVM and KNN to optimize mHealth applications. This approach enhances real-world applicability in areas such as remote patient monitoring, enabling real-time analysis of wearable data; chronic disease

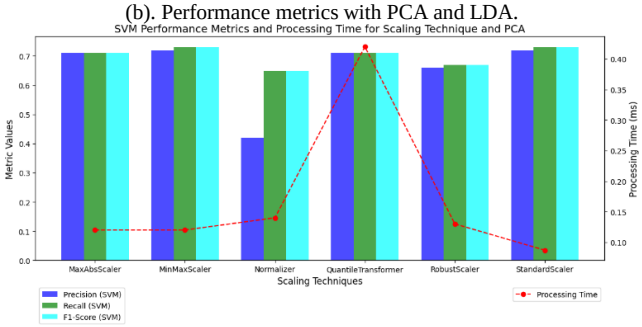
management, supporting early detection and personalized care; and telemedicine, improving the speed and reliability of virtual consultations. Faster processing times, such as 0.041 ms for SVM with LDA, ensure responsive systems that reduce delays in delivering critical insights, improving user trust and scalability through efficient resource utilization. Moreover, optimized preprocessing pipelines facilitate real-time predictions, enabling immediate feedback in critical scenarios like glucose monitoring or irregular heart rhythm detection. This responsiveness empowers healthcare professionals to make timely interventions, enhancing patient outcomes and advancing the efficiency of mHealth solutions.



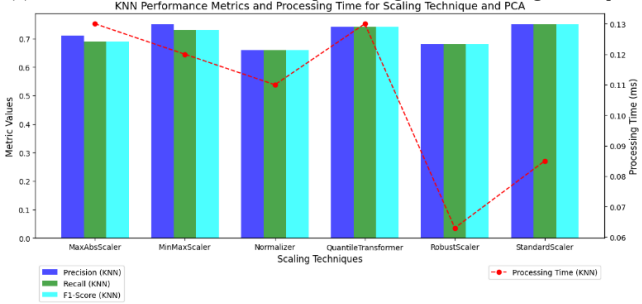
(a). Accuracy and processing time for scaling technique.



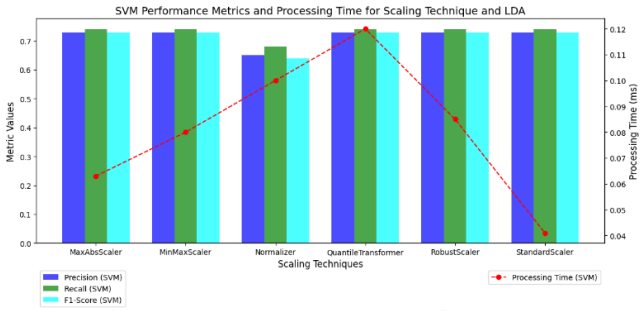
(b). Performance metrics with PCA and LDA.



(c). SVM Performance metrics and processing time for scaling technique.



(d). KNN Performance metrics and processing time for scaling technique and PCA.



(e). SVM Performance metrics and processing time for scaling technique and LDA.

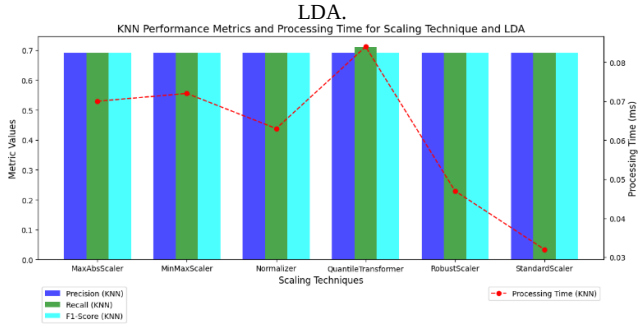


Fig. 3. (f) KNN Performance metrics and processing time for scaling technique and LDA.

V. Conclusion

This study provides a comprehensive evaluation of SVM and KNN algorithms, emphasizing their optimization for mobile health applications in cloud environments. Through systematic testing with various preprocessing techniques, such as scaling (Standard Scaler, Min-Max Scaler) and dimensionality reduction (PCA, LDA), the findings demonstrate significant improvements in both accuracy and processing efficiency. These optimizations address the unique constraints of mHealth systems, such as the need for real-time responsiveness and resource efficiency, making SVM and KNN suitable for applications like remote patient monitoring, chronic disease management, and telemedicine. The study highlights the critical role of preprocessing in reducing computational overhead, improving latency, and enhancing the scalability of cloud-based healthcare systems. While this work primarily focuses on preprocessing-driven optimization, it opens several avenues for future research. These include integrating deep learning models with classical classifiers, performing systematic hyperparameter tuning (e.g., grid search, Bayesian optimization), and applying statistical validation techniques to reinforce the robustness of results. Additionally, exploring adaptive preprocessing strategies that evolve in real-time with streaming health data could further improve model responsiveness in dynamic mHealth scenarios. Expanding the evaluation to include diverse datasets and multi-device deployment environments would also contribute to validating the generalizability of the proposed approach.

References

[1] A. A. Salameh, I. A. Abu-AlSondos, N. H. Abu, and A. Nahar Harun, “Current Knowledge and Future Possibilities of Medical Digital Technologies based on Mobile Health”, Int. J. Interact. Mob. Technol., vol. 17, no. 17, pp. pp. 134–147, Sep. 2023.

[2] R. Jeya, Hezerul Abdul Karim, and Sarina Binti Mansor, “Artificial Intelligence and Mobile Apps Support Intelligent Healthcare Systems for Mental Health Services ”, Int. J. Interact. Mob. Technol., vol. 18, no. 20, pp. pp. 157–168, Oct. 2024.

[3] M. Eddabbah, M. Moussaoui, and Y. Laaziz, "A smart architecture design for health remote monitoring systems and heterogeneous wireless sensor network technologies: a machine learning breathlessness prediction prototype," *International Journal of Intelligent Enterprise*, vol. 6, no. 2-4, pp. 293-310, 2019.

[4] T. S. Ajani, A. L. Imoize, and A. A. Atayero, "An overview of machine learning within embedded and mobile devices—optimizations and applications," *Sensors*, vol. 21, no. 13, pp. 4412, 2021.

[5] K. Akherfi, M. Gerndt, and H. Harroud, "Mobile cloud computing for computation offloading: Issues and challenges," *Applied Computing and Informatics*, vol. 14, no. 1, pp. 1-16, 2018.

[6] N. S. Hussien, S. Sulaiman, A. Aborujilah, M. Wibowo, and H. Samma, “Scalability of Mobile Cloud Storage”, Int. J. Interact. Mob. Technol., vol. 15, no. 21, pp. pp. 199–206, Nov. 2021.

[7] A. S. Albahri, A. A. Zaidan, O. S. Albahri, B. B. Zaidan, and M. A. Alsalem, "Real-time fault-tolerant mHealth system: Comprehensive review of healthcare services, opens issues, challenges and methodological aspects," *Journal of Medical Systems*, vol. 42, pp. 1-56, 2018.

[8] Y. Deng, "Deep learning on mobile devices: a review," in *Mobile Multimedia/Image Processing, Security, and Applications 2019*, vol. 10993, pp. 52-66, 2019.

[9] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, pp. 652801, 2021.

[10] R. Sivan and Z. A. Zukarnain, "Security and privacy in cloud-based e-health system," *Symmetry*, vol. 13, no. 5, pp. 742, 2021.

[11] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges," *Sensors*, vol. 23, no. 9, pp. 4178, 2023.

[12] R. Guido, M. C. Groccia, and D. Conforti, "A hyper-parameter tuning approach for cost-sensitive support vector machine classifiers," *Soft Computing*, vol. 27, no. 18, pp. 12863-12881, 2023.

[13] A. AlKarawi and K. AlJanabi, "Data Reduction Techniques: A Comparative Study," *Journal of Kufa for Mathematics and Computer*, vol. 9, no. 2, pp. 1-17, 2022.

[14] M. Sudha, M. R. Muthukathan, G. B. H. Malini, R. Sankar, M. Mythily, P. S. Kumaresh, M. N. V. Mageshkumar, and S. Shanmugam, "Predictive modeling for healthcare worker well-being with cloud computing and machine learning for stress management," *International Journal of Electrical & Computer Engineering*, vol. 15, no. 1, 2025.

[15] A. H. Abdulazeez and A. M. Abdulazeez, "Integration of Machine Learning with Fog Computing for Health Care Systems Challenges and Issues: A Review," The Indonesian Journal of Computer Science, vol. 13, no. 3, 2024.

[16] A. Deniz-Garcia, H. Fabelo, A. J. Rodriguez-Almeida, G. Zamora-Zamorano, M. Castro-Fernandez, M. D. P. Alberiche Ruano, T. Solvoll, C. Granja, T. R. Schopf, G. M. Callico, and C. Soguero-Ruiz, "Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges," Journal of Medical Internet Research, vol. 25, p. e44030, 2023.

[17] H. K. Bharadwaj, A. Agarwal, V. Chamola, N. R. Lakkaniga, V. Hassija, M. Guizani, and B. Sikdar, "A review on the role of machine learning in enabling IoT based healthcare applications," IEEE Access, vol. 9, pp. 38859-38890, 2021.

[18] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," Neural Computing and Applications, vol. 35, no. 22, pp. 16157-16173, 2023.

[19] Cetin, V., and O. Yildiz, "A comprehensive review on data preprocessing techniques in data analysis," Pamukkale University Journal of Engineering Sciences, vol. 28, no. 2, pp. 299-312, 2022.

[20] A. G. Oladepo, A. O. Bajeh, A. O. Balogun, H. A. Mojeed, A. A. Salman, and A. I. Bako, “Heterogeneous Ensemble with Combined Dimensionality Reduction for Social Spam Detection”, Int. J. Interact. Mob. Technol., vol. 15, no. 17, pp. pp. 84–103, Sep. 2021.

[21] C. Zhang, M. Mei, Z. Mei, J. Zhang, A. Deng, and C. Lu, "PLDANet: reasonable combination of PCA and LDA convolutional networks," International Journal of Computers Communications & Control, vol. 17, no. 2, 2022.

[22] D. M. Abdullah and A. M. Abdulazeez, "Machine learning applications based on SVM classification a review," Qubahan Academic Journal, vol. 1, no. 2, pp. 81-90, 2021.

[23] M. Suyal and P. Goyal, "A review on analysis of K-nearest neighbor classification machine learning algorithms based on supervised learning," International Journal of Engineering Trends and Technology, vol. 70, no. 7, pp. 43-48, 2022.

Deep Reinforcement Learning Based Robotic Arm Control Simulation to Execute Object Reaching Task for Industrial Application

John Mark Correa[1], Rudolph Joshua Candare[2], Junrie B. Matias[3]
Department of Computer Science, Caraga State University, Butuan City, Philippines[1]
Graduate School, Caraga State University, Butuan City, Philippines[2, 3]

***Abstract*—This study presents a deep reinforcement learning (DRL) approach to train a robotic arm for object reaching tasks in industrial settings, eliminating the need for traditional task-specific programming. Leveraging the Proximal Policy Optimization (PPO) algorithm for its stability in continuous control, the system learns optimal behaviors through autonomous trial-and-error. Central to this work is reward shaping, where structured feedback based on distance to the target, collision avoidance, motion constraints, and step efficiency guides the agent, akin to incremental coaching. A simulated industrial environment was developed using Webots, integrated with OpenAI Gym and Stable-Baselines3, enabling safe training with sensor data (camera, distance sensor) and randomized target placements. Three models with varying reward schemes were evaluated: simpler rewards prioritized rapid convergence, while complex formulations (e.g., perceptual alignment) enhanced long-term accuracy at the cost of initial instability. Experimental results demonstrated that reward shaping reduced the required steps, highlighting its role in accelerating learning. The study underscores the efficacy of combining DRL, simulation-based training, and adaptive reward design to develop efficient robotic controllers. These findings advance scalable solutions for industrial automation, emphasizing the trade-offs between reward complexity and policy convergence. Future work will refine reward functions to bridge simulation-to-reality gaps, fostering practical adoption in manufacturing and assembly systems.**

***Keywords*—Reinforcement learning; deep reinforcement learning; reward shaping techniques; robotic arm; robot simulation**

I. Introduction

Humans possess the innate ability to easily recognize and manipulate objects; conversely, robots encounter significant challenges with this ostensibly straightforward task, which presents a noteworthy obstacle within the field of robotics [1]. This phenomenon has garnered attention in both scholarly research and industrial applications, particularly in relation to Industry 4.0, which prioritizes automation and sophisticated technologies in manufacturing through paradigms such as the Smart Factory [2].

Reinforcement learning (RL) has surfaced as a promising alternative to facilitate the implementation of robots within industrial contexts by enabling robots to acquire tasks without the need for direct programming [3]. Deep reinforcement learning (DRL) integrates RL with deep neural networks to address complex state spaces and enhance the efficiency of the learning process. Although DRL may demand substantial resources for training on actual platforms, simulation-based training represents a feasible solution to scale data and decrease reliance on human intervention [4].

Simulation-based environments can furnish a secure and regulated context for robots to practice and hone their competencies, thereby equipping them to generalize acquired behaviors to real-world applications with greater efficacy [5]. This methodology not only accelerates the learning trajectory but also mitigates risks associated with training in unpredictable real-world conditions, ultimately fostering the development of more resilient and adaptable robotic systems [6]. By capitalizing on advancements in virtual reality and high-fidelity simulations, researchers are able to devise diverse training scenarios that closely replicate real-world challenges, thereby further enhancing the adaptability of DRL algorithms [7]. These innovations lay the groundwork for more effective training methodologies, enabling robots to address complex tasks with heightened confidence and precision [7].

Consequently, the assimilation of these technologies into robotic training programs is poised to transform not only the manner in which robots learn but also their overall efficacy in dynamic environments [8]. This paradigm shift in training methodologies is likely to yield substantial advancements across a variety of applications, ranging from autonomous vehicles navigating congested thoroughfares to industrial robots optimizing production lines with minimal human involvement [9].

However, a critical gap remains in understanding how reward shaping, specifically the design and modular integration of reward signals, affects convergence efficiency and task precision across varying degrees of complexity [10], [11]. Most prior works emphasize algorithmic improvements or domain adaptation, yet underexplore the influence of structured, incremental reward functions within high-fidelity simulations. These reward strategies are essential not only for accelerating convergence but also for fostering generalizable behaviors applicable to industrial use cases [12].

To address this gap, this study poses the following research question: How does the complexity of reward functions influence the convergence efficiency and task success rate of a DRL-based robotic arm control system in a simulated industrial environment?

This study addresses this question by proposing a DRL-based robotic arm control system using Proximal Policy Optimization (PPO) within a Webots simulation environment. By leveraging the recent advancements in continuous control through deep reinforcement learning (DRL) to enhance the operational efficiency of a robotic arm tasked with object-reaching activities. Specifically, this study employs the Proximal Policy Optimization (PPO) algorithm within a three-dimensional simulation environment, guided by a strategically designed reward function. The primary objective is to develop and train a DRL-based control system that enables the robotic arm to accurately reach and make contact with a designated target object, sometimes at a specified angle, an ability with broad applicability across labor-intensive industries. The proposed reward function is crafted to facilitate the convergence of the PPO algorithm, ensuring effective policy learning. Additionally, the study explores the adaptability of the learning framework by modifying the reward function to accommodate similar manipulation tasks of varying complexity. Comparative evaluations of different reward functions will be conducted to determine their influence on the convergence speed and overall performance of the robotic arm, thereby contributing to the design of more efficient and adaptable robotic control systems.

II. Related Work

Deep reinforcement learning (DRL) has emerged as a powerful paradigm for controlling robotic arms, offering robust solutions to complex control tasks. This approach combines the strengths of deep learning and reinforcement learning, enabling robots to learn optimal policies through trial and error in a dynamic environment. Recent advancements in DRL have significantly improved the accuracy, efficiency, and adaptability of robotic arm control systems, addressing challenges such as trajectory tracking, obstacle avoidance, and energy efficiency. This section provides a comprehensive overview of the state-of-the-art methods, key innovations, and applications of DRL in robotic arm control, drawing insights from recent research papers.

A. Algorithmic Advancements in DRL for Robotic Arm Control

One of the most notable contributions to DRL for robotic arm control is the improvement of existing algorithms. Proximal Policy Optimization (PPO) has been widely adopted due to its stability and effectiveness in continuous action spaces [13]. For instance, researchers have proposed an Improved-PPO algorithm that integrates PPO with Model Predictive Control (MPC) to enhance trajectory tracking efficiency. This method has demonstrated a significant increase in convergence speed, outperforming traditional PPO and A3C algorithms by 84.3% and 15.4%, respectively [13]. Similarly, the combination of PPO with Generative Adversarial Imitation Learning (GAIL) has been explored to transfer policies from simulated environments to real-world scenarios, enabling dual-arm robots to perform complex assembly tasks with remarkable efficiency [14].

Another significant advancement is the use of Deep Deterministic Policy Gradient (DDPG) algorithms. Researchers have enhanced DDPG by incorporating techniques such as Hindsight Experience Replay (HER) and double experience replay buffers to improve learning efficiency. These modifications have enabled robotic arms to achieve near-perfect success rates in target-reaching tasks while reducing training time [15], [16]. Additionally, the integration of DDPG with sequential training methods has been shown to expedite the learning process and improve robustness against external disturbances, making it suitable for real-world applications [17].

B. Reward Function Design and Motion Planning

The design of reward functions plays a crucial role in DRL, as it guides the learning process and determines the optimal policy [18]. Researchers have proposed various reward functions tailored to specific tasks, such as trajectory tracking, obstacle avoidance, and energy efficiency [19]. For example, a study introduced a hierarchical reward function that combines motion accuracy, obstacle avoidance, and energy-saving components. This approach has been shown to improve the convergence speed and accuracy of robotic arm control policies [20], [21].

In addition to reward function design, motion planning has been a focal area of research. A study proposed a DRL-based motion planning method that enables robotic arms to navigate complex environments with obstacles. The method leverages the PPO algorithm and reward shaping techniques to ensure efficient and adaptive control [20]. Another study demonstrated the effectiveness of DRL in motion planning for dual-arm robots, where a novel reward and punishment function was designed to guide the robot to approach targets while avoiding collisions [22].

C. Generalization and Real-World Deployment

The generalization capabilities of DRL controllers are essential for their successful deployment in real-world scenarios. Researchers have investigated the generalization of policies learned in simulation to new observations, dynamics, and tasks. For example, a study demonstrated that a DRL controller trained in simulation could generalize well to real-world scenarios, including dynamic trajectory tracking and robustness to external forces [23]. Another study showed that policies trained in simulation could be directly transferred to physical robots without retraining, achieving remarkable performance in tasks such as grasping and motion planning [24].

The generalization capabilities of DRL have also been tested in complex environments, such as those involving soft robotic arms. A study presented a closed-loop controller for soft robotic arms that could generalize to new tasks, including intercepting moving objects and tracking trajectories with varying velocity profiles [23]. These findings highlight the potential of DRL for real-world applications, where robots must adapt to diverse and dynamic environments.

Deep reinforcement learning has revolutionized the field of robotic arm control, offering robust and adaptive solutions to complex control tasks. Recent advancements in algorithms, reward function design, and simulation-to-reality transfer have significantly improved the performance and versatility of DRL-based control systems [25]. As research continues to address challenges such as generalization, energy efficiency, and real-world deployment, DRL is poised to play an increasingly important role in the development of intelligent robotic systems [26].

D. Identified Problems in the Current Implementation of DRL in Robotic Arm Control

The current implementation of Deep Reinforcement Learning (DRL) in the control of robotic arms has revealed several critical issues that warrant attention. One of the primary challenges is the inefficiency in training time, which often results from the high dimensionality of the action and state spaces involved. This complexity can lead to prolonged learning periods and suboptimal performance, as the algorithms struggle to converge on effective strategies [26].

Additionally, there is a notable concern regarding the stability and robustness of the learned policies. Many DRL algorithms are sensitive to hyperparameter settings, which can result in significant variations in performance across different trials [16]. This unpredictability can hinder the reliability of robotic arm operations in real-world applications where consistency is paramount [27].

Another problem is the insufficient exploration of the action space during the training phase. Many DRL implementations tend to exploit known strategies rather than exploring new ones, potentially leading to local optima and a lack of adaptability in dynamic environments. This limited exploration can restrict the robot's ability to handle unforeseen circumstances or variations in tasks [28]. Moreover, the integration of DRL with sensory feedback systems poses additional challenges. Ensuring that the robotic arm can effectively interpret and respond to sensory inputs in real-time is crucial for achieving seamless operation. However, current implementations may struggle with the synchronization of sensory data processing and decision-making, leading to delays and inaccuracies in task execution [29].

Lastly, there is a growing concern about the generalization of learned behaviors across different tasks and environments. Many DRL models tend to overfit to the specific conditions present during training, which can significantly diminish their performance when faced with novel situations or variations in operational contexts [30]. Addressing these issues is essential for advancing the efficacy and applicability of DRL in robotic arm control systems, paving the way for more sophisticated and versatile robotic applications.

III. Methodology

A. Introduction to the Simulation Environment

The simulation environment was developed using Webots, an open-source, multi-platform robotic simulator designed for modeling, programming, and testing robotic systems. The setup emulated a collaborative workspace with a Universal Robots UR5e robotic arm, a 6-degree-of-freedom manipulator equipped with a Tiago Gripper. Key components integrated into the system included:

- A camera with an object recognition node mounted on the gripper to capture visual data.
- A distance sensor is attached to the end-effector to measure proximity to the target.
- A 5 cm^3 cube serving as the target object, randomly repositioned within a 3 m^3 boundary in front of the robot for each training episode.

The environment was designed to simulate industrial scenarios, where the robotic arm must dynamically adapt to varying object positions. The camera transmitted real-time images to the control system, which processed them to generate state observations (e.g., joint angles, end-effector coordinates).

B. Design of the Reward Functions

The performance of deep reinforcement learning (DRL) agents heavily depends on the quality and structure of the reward functions used during training. In this study, three models (A, B, and C) were developed with progressively complex reward functions to guide the robotic arm in learning the object-reaching task. Each reward function was designed to encourage desirable behavior and penalize inefficiencies or errors, thereby accelerating convergence toward optimal policies.

1) Distance-based reward.

a) A primary reward proportional to the Euclidean distance between the end-effector and target:

$$\mathrm{r} = \begin{cases} 10, & \mathrm{dis} < 0.01 \\ -2, & \mathrm{dis} > 1.25 \end{cases} \quad (1)$$

b) Incremental rewards (+2/-1) encouraged movement toward the target.

2) Motion constraints. Penalties (-2) enforced spatial boundaries to limit exploration to relevant areas.

3) Collision avoidance. Negative rewards (-5) deterred collisions with the robot's body or environment.

4) Object recognition. Positive rewards (+1) for detecting the target via the camera.

5) Image-based rewards (Models B and C).

a) Size ratio rewards (Model B): Scaled rewards based on changes in the target's perceived size.

$$\mathrm{size_ratio} = \mathrm{box_image_size} \,/\, \mathrm{prev_box_image\ size} \quad (2)$$

$$\mathrm{r} = \mathrm{size_ratio} \times 2 \quad (3)$$

b) Alignment rewards (Model C): Penalized deviations between the target's center and the camera's focal point.

$$dis_to_ctr = \sqrt{\left(\frac{x_{box_img} + width_{cam}}{2}\right)^2 + \left(\frac{y_{box_img} + height_{cam}}{2}\right)^2} \quad (4)$$

$$r = \begin{cases} 10, dis_to_ctr < 0.1 \\ 5, dis_to_ctr < 0.5 \\ 2, dis_to_ctr \leq 0.9 \\ -\ dis_to_ctr \times 2, dis_to_ctr > 0.09 \end{cases} \quad (5)$$

C. Implementation of the Control Mechanism and PPO Algorithm

The control policy was optimized using the Proximal Policy Optimization (PPO) algorithm, selected for its stability in continuous action spaces. Key implementation details included:

a) Policy network: A multi-layer perceptron (MLP) with fully connected layers to map states (joint angles, sensor data) to actions (joint velocity commands).

b) Training parameters: n_steps=2048 interactions per policy update. tensorboard_log for tracking training metrics (e.g., reward trends, episode length).

c) Environment interface: Custom OpenAI Gym integration to synchronize Webots simulations with the Stable-Baselines3 RL library.

The learning process mirrored trial-and-error refinement, where the agent iteratively adjusted its policy based on reward signals, gradually minimizing unnecessary movements and collisions.

D. Experiment and Evaluation

To systematically evaluate the impact of varying reward function complexities on the performance of the Proximal Policy Optimization (PPO) algorithm, a series of experiments were conducted using three distinct models. The primary evaluation criteria included the number of training steps required to achieve effective task performance and the success rate in executing the object-reaching task within a simulated environment.

The training protocol involved 15 rounds of training for each model, with 10,000 interaction steps per round. The models were designed to incrementally incorporate additional reward components, enabling a controlled analysis of how increasing reward complexity influences policy learning.

The models are defined as follows:

- Model A (Baseline): Utilizes a foundational reward structure based on the Euclidean distance between the end-effector and the target object as defined in Eq. (1), augmented with penalties and rewards for collision avoidance, object detection, and motion constraints.
- Model B (Image-Augmented Reward): Builds upon Model A by incorporating image-based size ratio rewards, leveraging Eq. (2) and Eq. (3). These rewards encourage the agent to maximize the perceived size of the target in the camera's view, thereby reinforcing movement toward the object.
- Model C (Alignment-Based Reward): Extends Model B by introducing camera alignment rewards, as described in Eq. (4) and Eq. (5). This component penalizes misalignment between the camera's focal center and the centroid of the target object, promoting not only reaching accuracy but also visual alignment for precise manipulation.

Throughout training, each model's performance was monitored using TensorBoard logs, which recorded metrics such as mean reward values and episode lengths over time. Furthermore, success rates were tested in controlled batches of 10, 30, and 50 episodes to assess consistency and robustness in task execution. This experimental design allowed for a comparative analysis of how reward structure influences learning stability, convergence speed, and task success.

Each model was subjected to 15 training rounds, with each round comprising 10,000 interaction steps between the agent and the simulated environment. The performance of the models was evaluated using two primary metrics:

a) Convergence efficiency: Assessed through TensorBoard visualizations, which tracked the progression of mean episode rewards and episode lengths over the course of training. These indicators provided insight into the stability and speed with which each model approached optimal policy learning.

b) Task success rate: Evaluated by conducting a series of 10-, 30-, and 50-episode trials, wherein a successful trial was defined as the robotic arm's end-effector making contact with the target object within a maximum of 250 time steps.

This evaluation framework was designed to highlight the influence of increasing reward complexity on both the stability of the learning process and the agent's ability to consistently complete the object-reaching task. By comparing the results across the three models, the study provides insights into the trade-offs between reward structure simplicity and control precision in DRL-based robotic systems.

IV. Results and Discussion

A. Initial Development of DRL-Based Robot Arm Controller

Fig. 1 illustrates the simulation environment created using Webots, an open-source and multi-platform desktop application used for modeling, programming, and simulation of robotic systems. The researcher developed a simulation based on the Universal Robots UR5e's collaborative robot arm with 6 degrees of freedom, which was equipped with a Tiago Gripper, an imaging device possessing an object recognition node, a solid node functioning concurrently as a distance sensor and end effector, as well as a cube box serving as the target object. The cube box was strategically positioned in a random manner in front of the robotic arm for each interaction, which is pivotal in motivating the agent to acquire skills in searching for the object within the designated space, thereby inhibiting it from merely learning and concentrating on a target object with a fixed position.

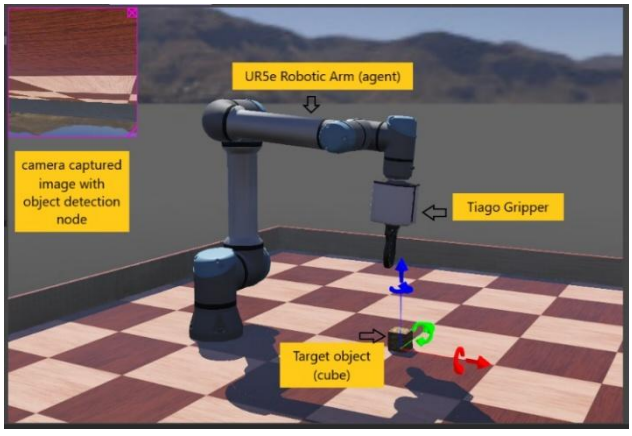


Fig. 1. Simulation environment and robot arm (Agent).

In the simulation environment, the camera transmits the collected image back to the computer. After the image is processed, it will be used as the state S to be observed by the actor. The computer will determine the next move distance of the end of the robotic arm on the X-axis, Y-axis, and Z-axis. According to the current coordinate value of the end of the robot arm and the rotation angle of each joint of the current robot arm,

the computer will calculate the angle at which each joint of the robot arm should rotate when the end of the robot arm reaches the next position.

B. Convergence Efficiency

By analyzing the mean reward values accumulated across iterations during the training process, where the horizontal axis represents the number of training iterations and the vertical axis denotes the reward value obtained per iteration. Fig. 2, Fig. 3 and Fig. 4 illustrate the performance trends of the three models examined in this study. The visualizations highlight how each model progressed and adapted over time during training. To facilitate this analysis, TensorBoard was employed as the primary tool for data processing and visualization, offering a comprehensive suite of utilities for monitoring, evaluating, and interpreting machine learning experiments. These visualizations illustrate the accumulation of reward values obtained across training iterations. The table on the left presents the mean episode length, measured by the number of timesteps or interactions with the environment. Here, the horizontal axis corresponds to the total number of iterations throughout the training process. In contrast, the table on the right displays the distribution of reward values earned in each iteration, with the horizontal axis indicating the iteration count and the vertical axis representing the reward value achieved. These visualizations provide insight into both the learning progress and the stability of the models over time.

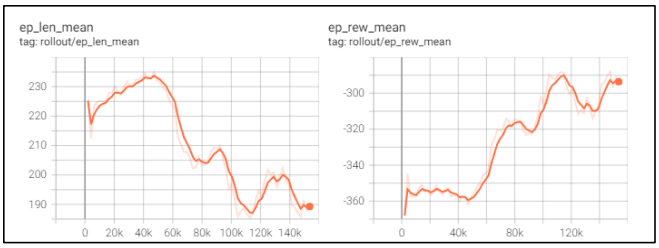


Fig. 2. Model A – Using baseline reward functions (Tensorboard).

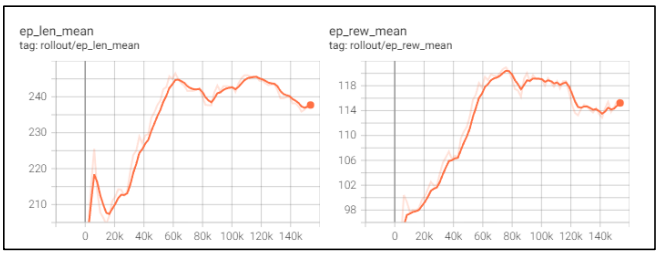


Fig. 3. Model B – Improved Model A with object image size-based reward function (Tensorboard).

The results derived from Model A, as illustrated in Fig. 2, reveal a favorable trajectory in the agent's performance over the temporal dimension. As the training regimen advances, it becomes evident that the quantity of actions necessary to fulfill the task diminishes while the resultant reward amplifies. This indicates that the agent is progressively enhancing its efficacy in making superior and more efficient control decisions to successfully execute the intended task.

In other terms, the agent is acquiring knowledge from the surrounding environment and modifying its conduct in response to the reward signal delineated by the reward function. As the agent develops, it executes movements with greater accuracy and precision in its pursuit of the target object. This phenomenon can be interpreted as an enhancement in the agent's control and decision-making proficiencies, constituting a favorable indication that the training protocol is functioning as anticipated. The observed trend of reduced actions coupled with elevated rewards signifies that the agent is incrementally converging towards the optimal resolution for the task, and it is highly probable that it will continue to experience improvement over time with ongoing training.

Fig. 3 underscores the notion that the architecture of the reward function can substantially influence the conduct of the agent. In this particular instance, the reward function of Model B may not have sufficiently motivated the agent to pursue the most expedient trajectory toward the target object. Conversely, it is plausible that Model B's reward function has prioritized alternative considerations, such as aligning the end-effector with the target object's orientation or maintaining adherence to the defined movement constraints, thus resulting in an elongated route to attain the target object. Irrespective of the underlying factors, the behavioral divergence between Model A and Model B accentuates the critical necessity of meticulously evaluating the formulation of the reward function during the training of a deep reinforcement learning-based agent. It further underscores the imperative for additional inquiry to ascertain the most effective reward function tailored to the specific objective of object reaching.

Moreover, the escalation in the number of actions necessitated in Model B signifies an extended training duration to achieve an optimal policy and thus could be a salient consideration when selecting a reward function for practical robotic applications. Subsequent investigations could delve deeper into this dichotomy between the efficacy of various reward functions and the temporal investment required to attain an optimal policy.

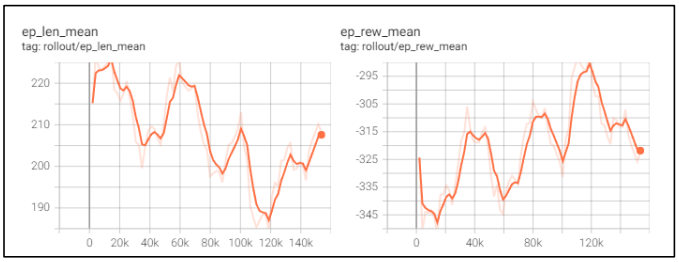


Fig. 4. Model C – Improved model B with object image and camera center distance reward function (Tensorboard).

Lastly, the findings derived from Model C, as illustrated in Fig. 4, demonstrate that, although the overall efficacy of the agent trained utilizing this reward function is commendable, there exist considerable variances in the rewards obtained throughout the training regimen. This phenomenon may be attributable to the complexity of the reward function employed in Model C, which incorporates a broader array of rewards, thereby rendering it more susceptible to fluctuations in the environmental context. Consequently, the agent may struggle to effectively adjust its parameters between successive training iterations, resulting in abrupt alterations in the rewards it acquires. Nevertheless, the overarching trajectory of the training

results for Model C remains favorable, indicating that the agent can acquire knowledge and advance towards the successful completion of the designated task.

C. Task Success Rate

The researchers executed a simulation experiment aimed at assessing the precision of three trained models by employing distinct reward functions. The experiment entailed the agent's gripper, or end-effector, endeavoring to approach the target object as closely as possible and establish contact with it. The efficacy of the models was appraised through 10-episode, 30-episode, and 50-episode trials, with the resultant data encapsulated in Table I.

TABLE I. RESULTS OF THE REACH AND CONTACT EXPERIMENT

Trained Model	Number of Successful Object Reach Actions		
	10 experiments	*30 experiments*	*50 experiments*
A	8	24	37
B	9	22	38
C	8	26	40

Results derived from the reach and contact experiment delineated in Table I indicate that all three models exhibit commendable efficacy with regard to their proficiency in reaching the target object and establishing contact with it. Nonetheless, there exist discernible variances in their performance, as evidenced by the number of successful trials conducted. A plausible explanation for these discrepancies in performance may pertain to the influence of the reward functions implemented in each model. It is plausible that Model A, characterized by its optimized reward function, possesses an enhanced capacity to concentrate on the specified task and acquire the optimal policy at an expedited rate, thereby culminating in superior performance in the reach and contact experiment. An additional factor that may contribute to these differences could be the training methodology employed. The variability in rewards observed in Model C may have precipitated a less stable learning trajectory, which could elucidate its marginally inferior performance in comparison to the other two models.

These findings imply that the proposed deep reinforcement learning-based robotic arm control system demonstrates the capability to execute reaching tasks with substantial effectiveness. Future investigations may be pursued to further refine the reward functions and training methodologies to augment the system's performance.

D. Implications

The agent exhibited exemplary behavior in executing the reaching task, attributable to the effective integration of diverse reward functions within the training process of the robotic arm. The reward functions encompassed parameters such as constraining the number of steps during interaction, quantifying the distance between the centroid of the object image and the centroid of the comprehensive image acquired by the camera recognition module, assessing the dimensions of the object, and identifying potential collisions, thereby furnishing the agent with affirmative or negative reinforcement that either promoted or dissuaded specific actions. These reward functions facilitate the agent's acquisition of the requisite behavior by directing it to recognize the object and its spatial coordinates, orient the end-effector towards the target object, synchronize the image captured by the camera recognition module, and circumvent collisions with the surrounding environment. Moreover, through the modulation of its movements, the agent was able to accomplish the task within the predetermined number of steps and maintain the target object within the defined parameters.

It can be asserted that the incorporation of these reward functions into the training process exerted a significant influence on the agent's performance in the object-reaching task, thereby optimizing both its behavior and resulting outcomes.

The incorporation of these carefully designed reward functions significantly influenced the agent's performance, enhancing both convergence and task efficiency. When comparing the performance of three distinct models trained using different reward structures. The following key observations emerged:

1) Model A matched or exceeded benchmark performances in terms of efficiency and reliability, validating the hypothesis that simplified but well-targeted reward functions improve learning stability.

2) Model B demonstrated slightly better success rates (76%) than Model A in the 50-episode trial, consistent with [22], where added perception cues enhanced accuracy, albeit at the cost of longer training.

3) Model C reached the highest success rate (80%) but required ~30% more training steps and suffered from high reward variance, paralleling issues in over-engineered reward systems like those explored in [10], [28].

Model A demonstrates a favorable trade-off between the simplicity of the reward function and training efficiency. Despite the absence of vision-based shaping used in Model B and C, Model A achieved the shortest convergence time (~70k steps) and competitive performance (74% success rate), with performance matching reward-heavy implementations from related studies [10], [22], [28] in both stability and training duration.

Although Model C introduced higher perceptual alignment (center-distance feedback), it required longer convergence (90k+ steps) and displayed reward fluctuations similar to [10] and [28], which reported volatility due to over-engineered shaping signals.

These results validate that while complex reward functions may offer incremental performance improvements, they often come with increased training cost and instability. The integration of well-structured, goal-specific rewards, particularly those balancing simplicity and effectiveness, is therefore essential in shaping robust robotic control behaviors. Our findings affirm the viability of deep reinforcement learning (DRL) approaches for dynamic object-reaching tasks in simulated environments and highlight the strategic value of reward function optimization.

V. CONCLUSION

This study affirms the feasibility of using deep reinforcement learning (DRL) to train a robotic arm controller in a simulated environment, even with limited computational resources. By leveraging the open-source simulation platform Webots alongside OpenAI Gym and Stable-Baselines3, the research successfully enabled a robotic arm to autonomously and consistently perform object-reaching tasks. Central to this success was the strategic use of reward shaping, a technique similar to guided learning, where the agent receives continuous feedback through structured rewards and penalties to encourage desired behavior. Three models were developed and evaluated to examine the impact of different reward strategies. Model A utilized a baseline reward function incorporating step limits, distance penalties, and collision detection. Model B introduced perceptual feedback by rewarding increases in the visible size of the target object, while Model C further refined agent behavior by penalizing misalignment between the camera's focal center and the object. Although Models B and C showed improved precision in object reaching, their complex reward formulations led to slower convergence and greater variability during training. In contrast, Model A's simplicity facilitated faster learning and more stable performance, highlighting the critical trade-off between reward granularity and training efficiency.

The findings demonstrate that DRL, when integrated with well-calibrated reward functions in a high-fidelity simulation, can generate stable, efficient robotic control policies without relying on hard-coded programming or manual trajectory design. Much like structured, age-appropriate guidance supports a child's learning process, a well-balanced reward architecture enables the agent to generalize optimal behavior through trial and error. However, excessively vague or overly intricate feedback can hinder progress or destabilize learning.

Ultimately, this research underscores the promise of DRL as a scalable and cost-effective approach to robotic automation. It emphasizes the importance of thoughtful reward function design and simulation-based training as key enablers for transferring learned behaviors to real-world applications. The overall contribution of this work lies in its demonstration that intelligent robotic behaviors can be efficiently acquired through reward shaping in resource-constrained settings, thus lowering the barrier to entry for academic institutions, startups, and small-scale industries seeking to adopt DRL-driven automation.

However, this work is not without limitations. Further research is required to refine the reward shaping mechanisms and to fine-tune the agent's learning process for improved generalization and adaptability. To enhance the agent's performance, future work should consider several key directions: integrating additional sensor modalities such as tactile sensors to enrich environmental feedback; exploring a broader range of robotic arm tasks to test adaptability; evaluating the agent's capabilities using objects of varying shapes, sizes, and materials; experimenting with alternative deep reinforcement learning algorithms to improve training efficiency and policy robustness; transferring the learned policies from simulation to real-world robotic platforms; and systematically assessing the agent's performance in real-world environments under realistic constraints and uncertainties.

This contributes meaningfully to the democratization of advanced robotics and strengthens the bridge between simulation research and real-world deployment.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the College of Computing and Information Sciences at Caraga State University for providing the academic framework and resources for this research. Special thanks to the faculty advisers and thesis panel members for their technical guidance and critical feedback. This work was partially funded by the Department of Science and Technology - Science Education Institute (DOST-SEI) through a graduate scholarship grant.

REFERENCES

[1] V. Sejdiu, A. Pajaziti, G. Rexha, X. Bajrami, E. Rrustemi, and J. Kola, "Detection, Recognition, and Grasping of Objects through Artificial Intelligence Using a Robotic Hand," IFAC-PapersOnLine, vol. 55, no. 39, pp. 443–446, 2022, doi: 10.1016/j.ifacol.2022.12.077.

[2] G. Revathy, K. Selvakumar, P. Murugapriya, and D. Ravikumar, "Smart manufacturing in Industry 4.0 using computational intelligence," in Artificial Intelligence for Internet of Things, Boca Raton: CRC Press, 2022, pp. 31–48. doi: 10.1201/9781003335801-3.

[3] R. Nian, J. Liu, and B. Huang, "A review On reinforcement learning: Introduction and applications in industrial process control," Comput Chem Eng, vol. 139, p. 106886, Aug. 2020, doi: 10.1016/j.compchemeng.2020.106886.

[4] W. Serrano, "Deep Reinforcement Learning with the Random Neural Network," Eng Appl Artif Intell, vol. 110, p. 104751, Apr. 2022, doi: 10.1016/j.engappai.2022.104751.

[5] C. Symeonidis and N. Nikolaidis, "Simulation environments," in Deep Learning for Robot Perception and Cognition, Elsevier, 2022, pp. 461–490. doi: 10.1016/B978-0-32-385787-1.00023-3.

[6] J. Lima, R. B. Kalbermatter, J. Braun, T. Brito, G. Berger, and P. Costa, "A realistic simulation environment as a teaching aid in educational robotics," in 2022 Latin American Robotics Symposium (LARS), 2022 Brazilian Symposium on Robotics (SBR), and 2022 Workshop on Robotics in Education (WRE), IEEE, Oct. 2022, pp. 430–435. doi: 10.1109/LARS/SBR/WRE56824.2022.9996083.

[7] M. Zahabi and A. M. Abdul Razak, "Adaptive virtual reality-based training: a systematic literature review and framework," Virtual Real, vol. 24, no. 4, pp. 725–752, Dec. 2020, doi: 10.1007/s10055-020-00434-w.

[8] Ö. Özen, K. A. Buetler, and L. Marchal-Crespo, "Towards functional robotic training: motor learning of dynamic tasks is enhanced by haptic rendering but hampered by arm weight support," J Neuroeng Rehabil, vol. 19, no. 1, p. 19, Dec. 2022, doi: 10.1186/s12984-022-00993-w.

[9] B. Maettig and H. Foot, "Approach to improving training of human workers in industrial applications through the use of Intelligence Augmentation and Human-in-the-Loop," in 2020 15th International Conference on Computer Science & Education (ICCSE), IEEE, Aug. 2020, pp. 283–288. doi: 10.1109/ICCSE49874.2020.9201867.

[10] A. Gupta, A. Pacchiano, Y. Zhai, S. M. Kakade, and S. Levine, "Unpacking Reward Shaping: Understanding the Benefits of Reward Engineering on Sample Complexity," Adv Neural Inf Process Syst, vol. 35, Oct. 2022, Accessed: Jun. 24, 2025. [Online]. Available: https://arxiv.org/pdf/2210.09579

[11] R. Devidze, P. Kamalaruban, and A. Singla, "Exploration-Guided Reward Shaping for Reinforcement Learning under Sparse Rewards," 2022. [Online]. Available: https://github.com/machine-teaching-group/neurips2022_exploration-guided-reward-shaping.

[12] H. Ma, K. Sima, T. V. Vo, D. Fu, and T.-Y. Leong, "Reward shaping for reinforcement learning with an assistant reward agent | Proceedings of the 41st International Conference on Machine Learning." Accessed: Jun. 24, 2025. [Online]. Available: https://dl.acm.org/doi/10.5555/3692070.3693450

[13] J. Zhang, Z. Zhang, S. Han, and S. Lü, "Proximal policy optimization via enhanced exploration efficiency," Inf Sci (N Y), vol. 609, pp. 750–765, Sep. 2022, doi: 10.1016/j.ins.2022.07.111.

[14] Q. ZHENG, Z. PENG, P. ZHU, Y. ZHAO, and W. MA, "Robotic arm trajectory tracking method based on improved proximal policy optimization," Proceedings of the Romanian Academy, Series A: Mathematics, Physics, Technical Sciences, Information Science, vol. 24, no. 3, pp. 237–246, Dec. 2023, doi: 10.59277/PRA-SER.A.24.3.05.

[15] Y. Shao, H. Zhou, S. Zhao, X. Fan, and J. Jiang, "A Control Method of Robotic Arm Based on Improved Deep Deterministic Policy Gradient," in 2023 IEEE International Conference on Mechatronics and Automation (ICMA), IEEE, Aug. 2023, pp. 473–478. doi: 10.1109/ICMA57826.2023.10215662.

[16] T. Tiong, I. Saad, K. T. K. Teo, and H. bin Lago, "Deep Reinforcement Learning with Robust Deep Deterministic Policy Gradient," in 2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), IEEE, Nov. 2020, pp. 1–5. doi: 10.1109/ICECIE50279.2020.9309539.

[17] S. Majumder and S. R. Sahoo, "Enhancing 3D Trajectory Tracking of Robotic Manipulator Using Sequential Deep Reinforcement Learning with Disturbance Rejection," in 2024 European Control Conference (ECC), IEEE, Jun. 2024, pp. 2512–2517. doi: 10.23919/ECC64448.2024.10591259.

[18] J. Eschmann, "Reward Function Design in Reinforcement Learning," 2021, pp. 25–33. doi: 10.1007/978-3-030-41188-6_3.

[19] M. Kang and K.-E. Kim, "Analysis of Reward Functions in Deep Reinforcement Learning for Continuous State Space Control," Journal of KIISE, vol. 47, no. 1, pp. 78–87, Jan. 2020, doi: 10.5626/JOK.2020.47.1.78.

[20] T. Shen, X. Liu, Y. Dong, and Y. Yuan, "Energy-Efficient Motion Planning and Control for Robotic Arms via Deep Reinforcement Learning," in 2022 34th Chinese Control and Decision Conference (CCDC), IEEE, Aug. 2022, pp. 5502–5507. doi: 10.1109/CCDC55256.2022.10033563.

[21] S. Yang and Q. Wang, "Robotic Arm Motion Planning with Autonomous Obstacle Avoidance Based on Deep Reinforcement Learning," in 2022 41st Chinese Control Conference (CCC), IEEE, Jul. 2022, pp. 3692–3697. doi: 10.23919/CCC55666.2022.9902722.

[22] W. Tang, C. Cheng, H. Ai, and L. Chen, "Dual-Arm Robot Trajectory Planning Based on Deep Reinforcement Learning under Complex Environment," Micromachines (Basel), vol. 13, no. 4, p. 564, Mar. 2022, doi: 10.3390/mi13040564.

[23] C. Alessi, H. Hauser, A. Lucantonio, and E. Falotico, "Learning a Controller for Soft Robotic Arms and Testing its Generalization to New Observations, Dynamics, and Tasks," 2023 IEEE International Conference on Soft Robotics, RoboSoft 2023, 2023, doi: 10.1109/ROBOSOFT55895.2023.10121988.

[24] S. Sharma et al., "Learning Switching Criteria for Sim2Real Transfer of Robotic Fabric Manipulation Policies," in 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE), IEEE, Aug. 2022, pp. 1116–1123. doi: 10.1109/CASE49997.2022.9926556.

[25] R. Liu, F. Nageotte, P. Zanne, M. de Mathelin, and B. Dresp-Langley, "Deep Reinforcement Learning for the Control of Robotic Manipulation: A Focussed Mini-Review," Robotics, vol. 10, no. 1, p. 22, Jan. 2021, doi: 10.3390/robotics10010022.

[26] W. Feng, C. Han, F. Lian, and X. Liu, "A Data-Efficient Training Method for Deep Reinforcement Learning," Electronics (Basel), vol. 11, no. 24, p. 4205, Dec. 2022, doi: 10.3390/electronics11244205.

[27] S. Li, C. Xu, Y. Wang, and L. Xie, "Serial and parallel reliability models for robot arm reliability analysis," J Phys Conf Ser, vol. 1605, no. 1, p. 012043, Aug. 2020, doi: 10.1088/1742-6596/1605/1/012043.

[28] V. R. F. Miranda, A. A. Neto, G. M. Freitas, and L. A. Mozelli, "Generalization in Deep Reinforcement Learning for Robotic Navigation by Reward Shaping," Aug. 2023, doi: 10.1109/TIE.2023.3290244.

[29] S. Singh, F. M. Ramirez, J. Varley, A. Zeng, and V. Sindhwani, "Multiscale Sensor Fusion and Continuous Control with Neural CDEs," in 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Oct. 2022, pp. 10897–10904. doi: 10.1109/IROS47612.2022.9982210.

[30] P. Yadav, A. Mishra, J. Lee, and S. Kim, "A Survey on Deep Reinforcement Learning-based Approaches for Adaptation and Generalization," Feb. 2022, Accessed: Jun. 24, 2025. [Online]. Available: https://arxiv.org/pdf/2202.08444

Explainable Deep Temporal Modeling for Stroke Risk Assessment Using Attention-Based LSTM Networks

Dr. P. Selvaperumal[1], F. Sheeja Mary[2], Dr. Pratik Gite[3], T L Deepika Roy[4],
Prof. Ts. Dr. Yousef A. Baker El-Ebiary[5], Gowrisankar Kalakoti[6], Dr. Sandeep Kumar Mathariya[7]
Assistant Professor, Department of Computer Science, St Joseph's University, Bengaluru, India[1]
Assistant Professor Senior Grade, School of Computing, Dept. of CSE,
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India[2]
Department of Computer Science and Engineering, Parul University- Parul Institute of Technology,
Vadodara-391760, (G.J.) India[3]
Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation,
Green Fields, Vaddeswaram, A.P. – 522302, India[4]
Faculty of Informatics and Computing, UniSZA University, Malaysia[5]
Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation,
Green Fields, Vaddeswaram, AP-522302, India[6]
Assistant Professor, Department of Computer Science and Engineering, MEDICAPS University, Indore, MP, India[7]

***Abstract*—Stroke continues to be a major cause of mortality and disability globally, and precise risk prediction models are needed. Current models do not effectively incorporate temporal patient information, restricting the quality of prediction and clinical interpretability. This research introduces a new LSTM-based deep learning model enriched with an attention mechanism for predicting stroke risk that can prioritize important risk factors like age, hypertension, and heart disease. The model takes advantage of LSTM's ability to learn sequential dependencies from long-term patient histories, while the attention mechanism dynamically emphasizes clinically important features, promoting interpretability and clinical significance. By testing the model using a dataset of 5,110 patient records with a mere 6% stroke cases, showcasing extreme class imbalance. To counteract this, preprocessing involved SMOTE for synthetic oversampling, mean imputation to handle missing values, and Min-Max normalization. As deployed in Python based on TensorFlow, the model realized remarkable performance. The constructed LSTM-Attention model attained a test accuracy of 83.7%, an AUC-ROC value of 85.3%, and an F1-value of 82.2%, which was higher than that of conventional models such as Logistic Regression and Random Forest. These evaluate the model's improved ability to identify subtle stroke risk factors that go unnoticed otherwise. The attention-augmented LSTM architecture not only guarantees accurate predictions but also offers transparent insight into the decision process, making it appropriate for incorporation in real-time clinical decision support systems. This method has the potential to improve personalized stroke risk assessment dramatically and enhance preventive healthcare interventions.**

***Keywords*—Attention mechanism; deep learning; imbalanced data; LSTM networks; SMOTE resampling; stroke prediction**

I. Introduction

Incidence and mortality from stroke constitute one of the leading public health concerns worldwide and remains a cause of burden affecting millions of people every year. According to WHO statistics strokes result in 11 per cent of total deaths worldwide making them the second cause of death [1]. Moreover, stroke has long-term consequences, including paralysis, impairment of speech, cognitive level and a huge drop in quality of life. A variety of factors are related to the increasing incidence of strokes [2]. Right now, with more and more people at risk, the need for good stroke prevention and early detection strategies has never been greater. Strokes can be diagnosed and their severity reduced with timely diagnosis and intervention [3]. Nevertheless, the existing methods of stroke prediction have the weakness of being unable to identify the individuals at risk early enough for preventative action. However, conventional risk assessment models like the FSRP, as well as other clinical scoring systems, heavily rely on statistical techniques and predefined risk factors [4]. However, these models demonstrate some kind of predictive power but do not show how individual risk factors interact with each other or how patient populations can be sensitive to different shapes of risk factor interactions. Moreover, stroke prediction has been improved using traditional machine learning approaches of logistic regression and decision trees [5]. Despite this, these methods are not able to handle the large-scale datasets with complex, high-dimensional features and temporal dependencies well, and thus their predictive performance suffers. This led the medical community to search for DL techniques, which have produced great results in a wide range of fields, including medical diagnostics, image recognition and NLP [6].

For the purpose of stroke prediction, deep learning models, in particular, LSTM networks have proven to be a promising choice to analyze data that is sequential and time dependent [7]. Unlike standard machine learning models, LSTMs can keep information in information over longer sequences allowing them to capture significant temporal patterns in patient data. Nonetheless, it is still difficult for standard LSTM models to achieve interpretability and select the most discriminative features for stroke prediction [8]. It is here that attention mechanisms become useful. Attention mechanisms improve the performance of LSTMs by selectively attending to the input features and weighting them according to their importance.

Attention mechanisms are integrated with LSTM networks to enable selecting the crucial risk factors leading to developing stroke, with a more transparent and explainable prediction process [9]. LSTM networks with attention mechanisms have a high potential. However, studies in this area are still limited, and the majority of the studies have been conducted in the context of traditional deep learning frameworks or hybrid models that do not take full advantage of the benefits of using attention mechanisms. While existing studies have traditionally focused on image-based stroke detection using MRI or CT, there is a large black hole gap between what can be predicted based solely on patient demographics and clinical history. Additionally, while some study tries to implement deep learning techniques for stroke prediction, most of these efforts do not deal with the issue when data imbalance, feature selection or real-time applicability in clinical settings. It is to leverage DL techniques such as LSTM networks coupled with attention mechanisms to enhance the precision and reliability of stroke risk prediction as a main objective of the study [10]. Through analysis of patient demographics, medical history, and lifestyle factors, the model seeks to anticipate those who are at increased risk of suffering from a stroke via secondary symptom occurrence.

Additionally, a comparative performance of the proposed LSTM with attention model to the conventional ML techniques and the ordinary DL architectures is further intended to demonstrate the effectiveness of the study [11]. The study seeks to solve deep learning models in healthcare through an approach that reveals crucial stroke risk factors within a clear framework of transparency. The technique integrates attention mechanisms to enhance prediction accuracy and at the same time deliver crucial insights about risk variables which impact stroke the most. [12]. Such inputs can be useful to medical practitioners to take informed decisions and design targeted prevention strategies for high-risk Individuals. To summarize, strokes worldwide are increasing, and it need to be predicted earlier, and the building of advanced models for early risk assessment is needed. Although traditional stroke prediction methods are useful, there are significant limitations that prevent these methods from being accurate and timely predictions [13].

A. Research Motivation

Stroke is a major cause of death and lasting disability worldwide, and it still accounts for a significant global health burden. The limitations of the conventional models based on static information and fixed rules render early prediction of stroke risk challenging despite improvements in clinical diagnostics. These approaches often fail to capture the temporal, nonlinear, and complex nature of patient health data. In addition, real-world clinical datasets often have a large class imbalance, leading to biased and unreliable predictions. Recent breakthroughs in deep learning have proven effective in uncovering latent patterns and modeling sequential data. However, most existing work either disregards interpretability or does not adequately deal with the unbalanced nature of stroke datasets. To overcome these concerns, this study integrates an attention mechanism with LSTM networks to develop a comprehensible and efficient stroke risk prediction model. The aim is to enhance prediction accuracy and clinical applicability in real-world settings by dynamically recognizing and ranking significant risk factors and recording time-dependent correlations.

B. Significance of the Study

This study is important for the potential it holds to revolutionize stroke risk assessment by presenting a high-performance and interpretable AI-based method. Early and precise identification of stroke risk can initiate prompt medical intervention, thus preventing the eventuality of stroke and reducing the level of related long-term disability. The suggested model not only enhances prediction precision relative to conventional models but also offers actionable information critical for clinical decision-making. By tackling practical issues like class imbalance in real-world problems using methods like SMOTE and focusing on model explainability, this piece of work is a valuable addition to the area of medical AI. Its successful application would be an exemplary model for other chronic disease risk assessments and would guide the development of real-time, IoT-integrated clinical decision support systems toward better patient outcomes and a decreased healthcare infrastructure burden.

C. Innovation and Challenges

The novelty of this study rests with the design of an explainable deep learning model that combines Long Short-Term Memory (LSTM) networks with an attention mechanism for stroke risk prediction. Although LSTM models are extensively used to capture sequential patterns in time-series health data, their intrinsic opacity restricts their applicability in clinical practice [14]. By incorporating an attention mechanism, this study remedies the "black-box" shortcoming, providing a degree of interpretability by demarcating the most effective risk factors like age, hypertension, and cardiovascular history. This new technique enables clinicians to comprehend and verify predictions in real time. Allowing such a system to be deployed is plagued by the foremost challenges. These involve managing extreme class imbalance in stroke databases, as well as robustness across a wide variety of populations and maintaining good accuracy while not compromising interpretability. Additionally, there is limited current study directly applying attention-based LSTM models to stroke prediction based on extensive patient demographic and clinical information, and hence, this study is both timely and necessary.

D. Problem Statement

Stroke is a major cause of death and disability globally, and early and precise risk prediction is required for timely treatment and better patient outcomes. The conventional stroke risk models depend on rigid, linear relationships between established risk factors and are thus inappropriate for portraying complex, nonlinear, and temporal relationship that exist in patient health data. Although machine learning methods provide enhanced flexibility, it tend to be subject to extreme class imbalance and unable to appropriately capture time-dependent patterns, leading to less-than-optimal performance and diminished clinical utility. Long Short-Term Memory (LSTM) networks, which have proven robust in sequential dependency learning, present a possible solution; however, black-box status restricts interpretability a critical requirement for clinical implementation [15]. In order to solve these problems, this study suggests an interpretable stroke risk prediction model that

combines LSTM networks with an attention mechanism. This model not only enhances predictive accuracy but also emphasizes the most relevant features over time, such as hypertension, age, and cardiovascular disease, thus improving transparency and clinical trust. In addition to this, employing SMOTE for class balancing prevents the model from becoming insensitive to minority stroke instances and enhances the model's fairness and reliability in actual health care environments. This study solves the research question: Can a clinical data-predicted stroke risk using an LSTM model augmented with an attention mechanism be achieved while keeping it interpretable for clinical applications?

E. Key Contributions

- An LSTM network with attention mechanisms is developed to enhance the prediction of stroke risk with patient demographics and clinical history.
- It improves model interpretability while overcoming the drawbacks of other stroke warning strategies, which can capture complex temporal patterns.
- The study handles the problem of data imbalance and compares the model's effectiveness to traditional ML and DL approaches.
- The area of AI-driven healthcare provides a demonstration of the use of attention-enhanced LSTMs for real-time stroke risk assessment.

F. Rest of the Sections for this Study

The rest of the sections of this study have been organized as follows: Review of the existing literature, Stroke Risk Assessment from Patient Data Using LSTM Networks and Attention Mechanisms in Deep Learning in Section II. In Section III, the proposed study Methodology is explained. The experimental results are presented in Section IV. The study concludes with future work proposals and its conclusion in Section V.

II. Literature Review

Stroke prediction is a long study area, and till date, the traditional models have been based on statistical methods and conventional ML methods [16]. The FSRP is one of the most widely used models to give risk of stroke based on well-defined clinical risk factors such as age, hypertension status, smoking status and diabetes. While statistical models like FSRP and Cox Proportional Hazard Models may be useful, it are limited by how it draw on defined assumptions and linear relationships, therefore being less effective at describing complex interactions between risk factors [17]. However, since these do not address these problems, authors have taken ML approaches, i.e. logistic regression, decision trees, random forests, SVMs, gradient boosting techniques such as XGBoost, among others, to enhance predictive accuracy. Additionally, these ML models learn patterns from patient data automatically, and often better perform in classification tasks than traditional statistical methods [18]. However, most of the existing ML based stroke prediction models fail due to the absence of temporal awareness and imbalanced datasets having very fewer number of stroke positive cases when compared to the non-stroke cases [19]. Additionally, most traditional ML models require feature engineering that involves strong bias and manual effort, a great obstacle for scalability and generalization to a versatile patient population. Therefore, the authors have increasingly explored deep learning techniques, which have revolutionized the healthcare domain by not only extracting features in an automatic way but also improving pattern recognition [20].

CNNs have been extensively used in the medical-related image-based stroke detection using MRI and CT scans, for which deep learning has shown great success at diagnosis [21]. Nevertheless, RNNs and LSTM networks have become promising alternatives in early stroke risk prediction using tabular patient data, since they can effectively capture temporal dependencies in serial health records. Since RNNs specialize in capturing long-term dependencies, LSTM models are especially good at retaining long-term dependencies in time series data [22], which seems to be a good fit for stroke prediction based on a patient's historical medical data. LSTM-based models outperform the traditional ML models in stroke prediction as they learns important patterns dynamically from patient demographics, lifestyle factors and clinical indicators. Nevertheless, LSTM models apply the same weight to all input features, ignoring the significance of key risk factors. In order to overcome this limitation, attention mechanisms have been employed to dynamically weight input features based on the model to highlight the related data [23]. Attention-based LSTM models have been demonstrated to improve the accuracy and interpretability in ECG analysis, disease prognosis and patient risk assessment [24]. The past several years have seen some studies exploring hybrid models using LSTMs with CNNs for stroke prediction by first leveraging CNNs for feature extraction and then using LSTMs for sequential modeling. These models show improved performance but are restricted to real-world health care due to computational complexity, data privacy and the requirement of large annotated datasets [25].

While deep learning has made great strides in stroke prediction, key study gaps remain. Deep learning models strictly operate and therefore feature selection and interpretability is one major challenge, due to the fact that many healthcare professionals wish to understand which risk factors play the biggest role in predicting stroke [26]. A major barrier to the clinical adoption of AI driven decision making is the lack of interpretability, which reduces trust in AI driven decision making. The attention mechanisms help address this issue partially by highlighting important features, however, more work is needed to develop more transparent, explainable stroke risk assessing AI models [27]. Another important void deals with the handling of the imbalanced datasets, where there were few positive cases of stroke when contrasted with the negative cases of stroke. However, most of the existing studies do not use data augmentation techniques like SMOTE, class balancing, or use of focal loss functions to handle this imbalance, and this causes biased results which favor majority class to be predicted [28]. To further enhance the model, most of existing deep learning works are based on offline training with static dataset, which is not suitable to the clinical scenario of real time stroke risk assessment. For the combination of the AI models into IoT enabled healthcare system, one of its crucial requirements is that of real time inference, which enables continuous monitoring of patient vitals and medical records and the provision of early

signals of stroke [29]. Nevertheless, deploying deep learning models in such harsh environments poses computational challenges and thus requires architectures, lightweight models and cloud-edge integration of efficient processing [30]. With these limitations taken into account, there is an obvious requirement for a powerful, explainable, and scalable deep learning backend that reliably predicts stroke risk, as well as tackles difficulties for feature selection, dataset imbalance, and running in real time. While various models have been developed for stroke prediction, they are mostly non-temporal sequence model-based or lacking interpretability. Most classical ML methods (e.g., Decision Trees, Logistic Regression) are bad at handling imbalanced datasets and are not feature-importance aware. Deep learning approaches such as CNNs are not transparent, which hinders clinical uptake.

III. Proposed LSTM-Attention based Model for Stroke Risk Assessment Using Patient Data

This study proposes a structured framework for developing a state-of-the-art LSTM based stroke risk prediction model with an embedding of an attention mechanism. Data preprocessing, model design, training, evaluation and comparing the performance with other methods are the aspects of methodology. This study focuses on the proposed LSTM model, which can capture time dependencies in patient data, as well as take into account a special attention mechanism on critical risk factors, whose information is dynamically prioritized. Binary cross entropy loss is used to train the model, which also optimizes with the Adam optimizer so as to train more efficiently. Finally, the proposed model is compared with traditional ML techniques using AUC-ROC, precision, recall, accuracy, and F1-Score. The methodology guarantees that the system developed is capable of making accurate predictions but at the same time is interpretable and applicable to real world healthcare environments. Fig. 1 illustrates the LSTM-AM framework for this study.

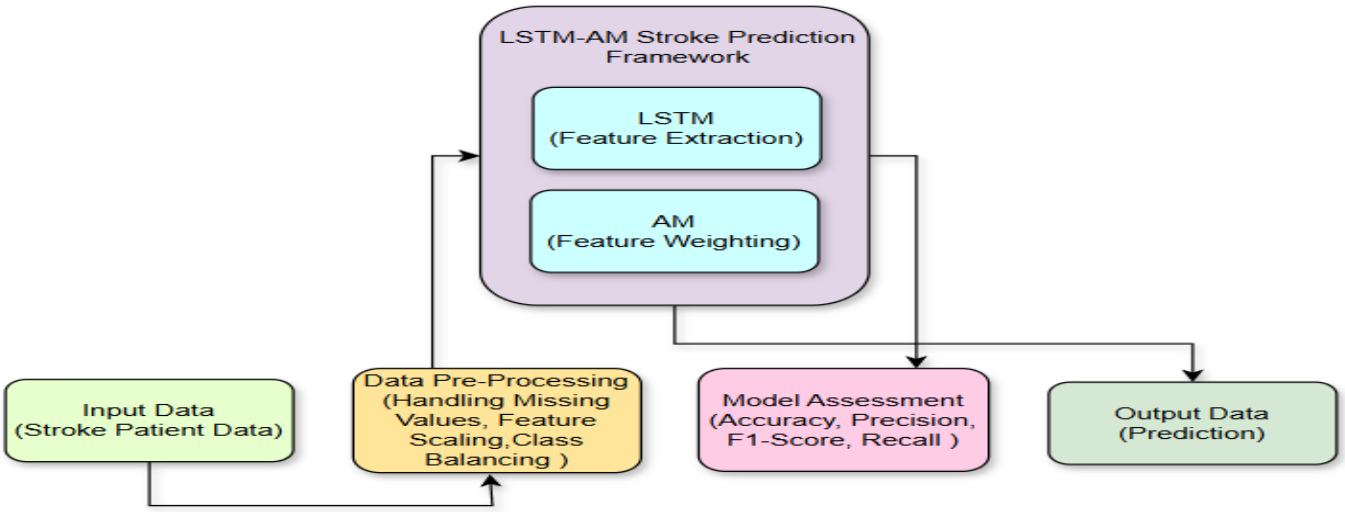


Fig. 1. LSTM-AM stroke prediction framework.

A. Data Collection

This study uses the dataset pulled from Kaggle [31], a widely known platform for open-source datasets in machine learning studies. There are 5,110 patient records with 12 key attributes. Due to the fact that conditions such as high blood pressure, diabetes and smoking are factors for developing cerebrovascular diseases, these features have an important role in evaluating stroke risk. The dataset is structured as a table with patient as row and patient's risk factors as columns. Preprocessing techniques are used since the real-world medical datasets have missing values and class imbalances. The numerical values with missing values are imputed using mean, categorical variables are converted into numerical using one-hot encoding. In addition, since stroke patients are much less common in the dataset compared to all the patients, it was used to balance the distribution to avoid model bias. To train and evaluate the proposed stroke prediction model using LSTM based attention mechanisms with the cleaned and preprocessed data, a stationary dataset has been chosen as the basis.

B. Data Pre-Processing

Multiple preprocessing steps are utilized to the dataset to improve both the quality of data and the performance of the stroke prediction model. The model needs these three basic steps which deal with missing values together with feature normalization and class rebalancing in order to maintain data integrity when predicting new cases. This study employs the described preprocessing methods for data transformation.

1) Handling missing values. Inconsistencies within data collection and patient record errors along with missing clinical tests cause numerous missing values to appear in medical datasets. The author applied mean imputation to numeric features including BMI and glucose level as a statistical method that fills empty values with existing data average calculations. Using this method protects both the data integrity of patient records and statistical consistency across the whole dataset. The calculated impute value uses the following mathematical method:

$$X_{new} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

In Eq. (1), x_i is the individual value in the, n is the number of values in total, X_{new} is the resulting mean value.

2) Feature scaling. Multiple numerical attributes in the dataset need feature scaling because their diverse ranges can lead to one feature controlling the learning process. The scale of values for age extends between 0 to 100 but the scope for glucose measurements exceeds wide extremes and BMI measurements require different numerical scaling. All numerical attributes get normalized through Min-Max Normalization to achieve a standardized 0 to 1 value scale. Method has made the following computational transformation.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

In Eq. (2), x is the real value, $X_{max} - X_{min}$ is the feature values of maximum and minimum.

3) Class balancing. In the dataset stroke cases represent a minority compared to non-stroke cases resulting in a severe problem of class imbalance. The current imbalance between stroke cases and non-stroke cases would produce suboptimal sensitivity because models likely prefer the majority class. SMOTE method applies to equalize the dataset distribution. SMOTE creates synthetic samples for minority class instances through extrapolation between existing minority class examples. The following equation defines the process of synthesizing new samples for study:

$$X_{new} = X_i + \lambda \, \mathrm{X} \, (X_j - X_i) \quad (3)$$

In Eq. (3), X_i is the class sample, X_j is the nearest value. The preprocessing pipeline of mean imputation followed by Min-Max normalization. Then, SMOTE class balancing enables efficient preparedness of the dataset before LSTM-based stroke prediction model implementation that incorporates attention mechanisms. Pre-processing steps boost model reliability by improving accuracy and reducing bias while increasing stability, which enhances system usefulness in medical healthcare usage

C. Stroke Risk Assessment from Patient Data Using LSTM Networks and Attention Mechanisms

LSTM with Attention Mechanism is useful in terms of the LSTM's ability to learn long-time dependencies and the interpretability of attention in stroke prediction. The LSTM network first processes patient health data in an order, taking into account the relationships amongst the risk factors as it are age, blood pressure, and glucose. At each time step, hidden states build on this information. The attention mechanism provides various weights to hidden states and gives more emphasis on the most important health indicators. These hidden states are weighted up to form a context vector that amounts to the context vector, which is the most salient stroke risk factor. After the final LSTM output is concatenated with a created context vector, it is finally fed into the fully connected layer to make the stroke prediction. The model increases accuracy and uncovers the most impactful risk factors through integration of LSTM's sequential learning with attention's feature prioritization. The architecture of LSTM is shown in Fig. 2.

1) Role of LSTM in stroke prediction. LSTM networks as a kind of RNNs are created for sequential type data and long-term dependency. LSTM is used in this study to process patient health records and capture temporal dependencies existing between risk factors like age, glucose levels, blood pressure, and smoking status to find out that stroke prediction accuracy could be improved. LSTM has three important gates: forget gate removes the useless past data, input gate update the memory to the useful new data and output gate decides the final hidden state for the prediction. The cell state serves as a memory unit that retains important information, ignoring data that turns out not to be statistically significant.

Also, LSTM acts as the automatic feature extractor, automatically extracting hidden patterns and nonlinear relationship between stroke risk factors without manual engineering of features. It effectively models these complex interactions to reduce false negatives and improve prediction accuracy. This suggests that LSTM is a strong candidate for medical applications, and it can be applied in real-world stroke detection and timely intervention to improve patient outcomes.

2) Role of the attention mechanism in stroke prediction. Attention Mechanism improves the LSTM model by dynamically weighing the relevance of stroke risk factors at the sequence level and, in turn, enhances not only explainability but also prediction accuracy. Unlike typical LSTM models, which treat all input features as being of equal significance, the attention concept spots the most important features, such as age, hypertension and glucose level and as such make sure that the model weight more on the most important information and less on the noise brought by the less important features. LSTM is very good to learn long-term dependencies, but sometimes it has difficulties to keep important information in a long sequence. The attention mechanism overcomes this by scanning the past information and selectively retaining the meaningful stroke-related patterns. To measure feature importance, attention gives weights to each feature as in Eq. (4):

$$\alpha_t = \frac{exp(e_t)}{\sum_j exp(e_j)} \quad (4)$$

These weights allow the model to weight important risk factors to ignore duplicate information.

Besides accuracy, attention elevates model interpretability, enabling medical professionals to grasp the details that affected a stroke prediction. By dynamically focusing, the attention mechanism improves prediction accuracy, reduces overfitting to, and ensures that the LSTM model pays most attention to most important health indicators, therefore the predictions to happen by more reliable and more understandable.

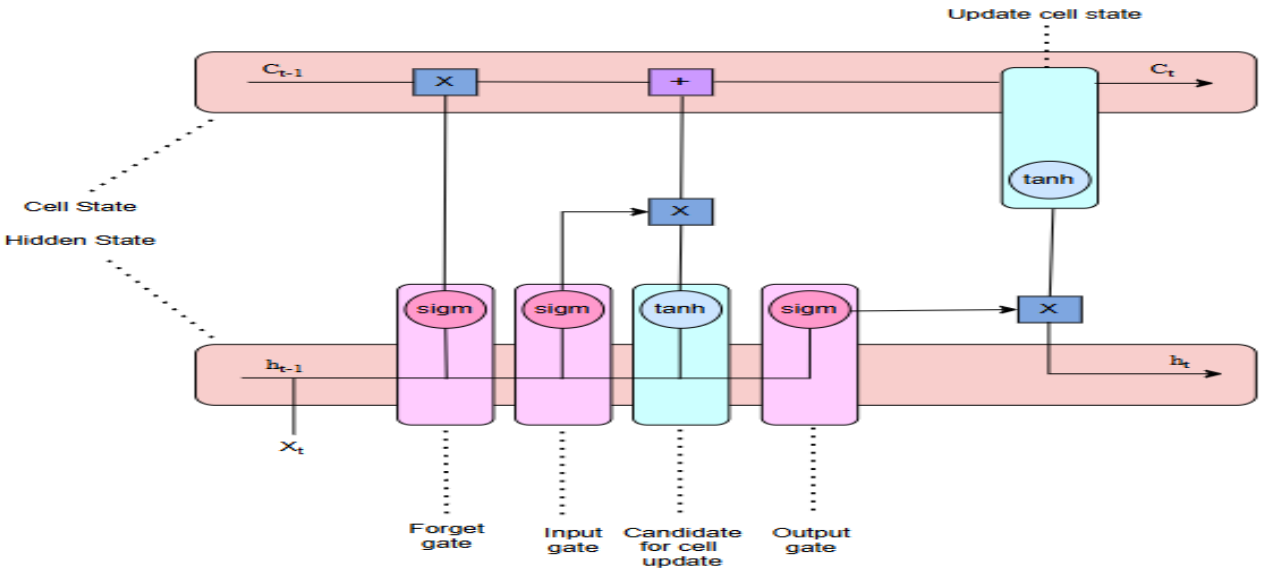


Fig. 2. LSTM in stroke prediction.

3) Integration of LSTM and attention mechanism in stroke prediction

a) LSTM encodes patient data: The LSTM network processes the patient health records in order, learns the intricate connections and interdependency of factors like age, blood pressure, glucose level and smoking status. Choice differs from normal models, its essential historical info, and therefore, learn from the data of a patient from the patterns. At each step, for tracking key medical information, both hidden state and cell state are updated while discarding unnecessary information. This allows the model to comprehend how stroke risk factors develop and influence each other over time. By utilizing LSTM's property of remembering long-term dependencies, the model is further improved in discovering minute stroke-borne patterns which might not be apparent at the initial glance of static data. The architecture of attention mechanism for prediction stroke is given in Fig. 3, and integration of LSTM-Attention mechanism is shown in Fig. 4.

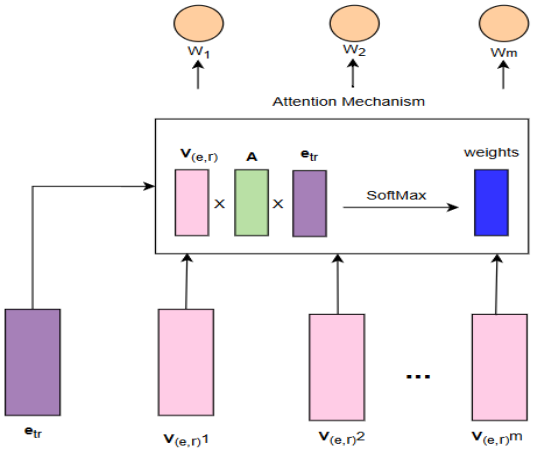


Fig. 3. Attention mechanism in stroke prediction.

In Eq. (5), the hidden state h_t for each time step t is calculated by the LSTM layer as a function of its previous hidden state $h_t - 1$ and the current input x_t. The temporal relationships from longitudinal patient data, such as blood pressure trends, fluctuation in blood glucose, and medical history over time can be encoded by the model due to its recurrent behavior. When combined with the context vector derived from attention for stroke risk, h_T is a compressed representation of the trajectory of the patient's general health.

$$h_t = LSTM(x_t, h_{t-1}) \quad (5)$$

b) Attention identifies key features: Although LSTM successfully learn sequential dependency, not all the input features are able to convey an equal component in the stroke risk prediction. The attention mechanism improves the model by increasing, across different hidden layer states, which ones are the most important and should get the most attention. Instead of generalizing equal weight on all data points, attention works out a context vector which picks out significant risk factors dependent on attention scores. The attentions are computed in the following way for each hidden state:

$$e_t = V^T tanh(W_h h_t + W_s s_{t-1}) \quad (6)$$

In Eq. (6), h_t is the hidden state, $W_h h_t + W_s s_{t-1}$ integrate the decoder and encoder states, $V^T tanh$ it refers to the non-linear activation with vector score. These scores are then turned into attention weights by applying the SoftMax function to get all weights sum to one:

$$\alpha_t = \frac{exp(e_t)}{\sum_j exp(e_j)} \quad (7)$$

In Eq. (7), e_t is the attention score, which transform into probabilities. Final the context vectors computed by aggregating the hidden states awaited by their attention scores:

$$c_t = \sum_t \alpha_t h_t \quad (8)$$

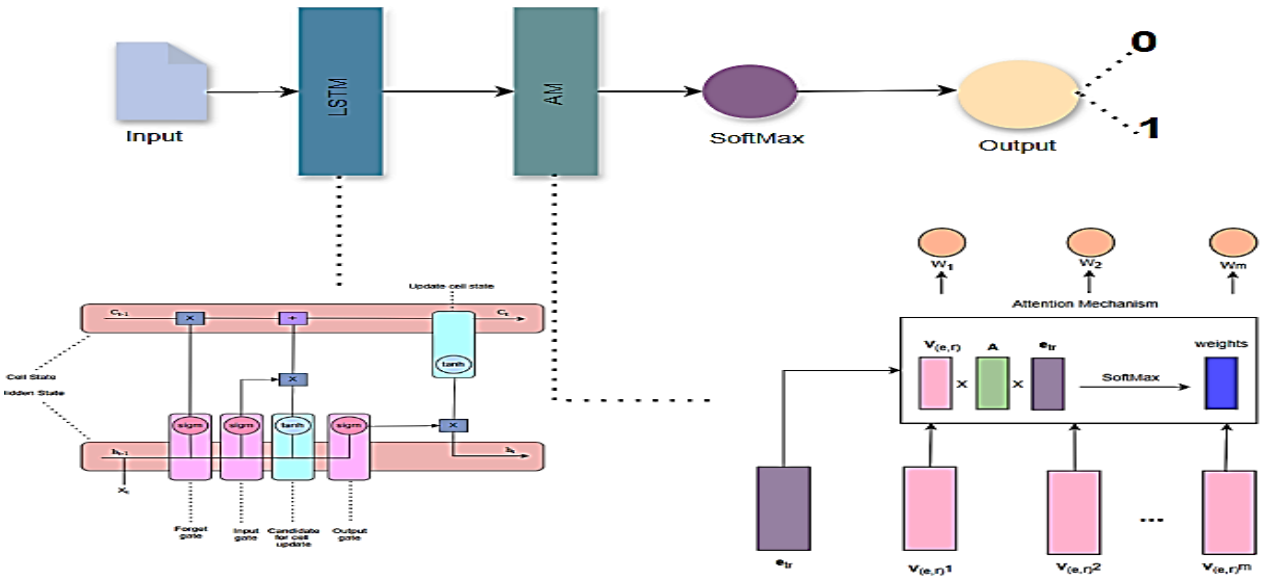


Fig. 4. Integration of LSTM and attention mechanism in stroke prediction.

In Eq. (8), c_t refers to context vector, α_t I the weight of the attention, h_t is the sum of the hidden state. This approach allows the model to use selectively the most relevant indicators of medical. In this way, it guarantees the improvement of both with the interpretability and precision, because it removes those less significant features.

c) Final stroke prediction: After generating the attention-enhanced context vector, it is fed into a fully connected layer to combine with the last LSTM output and is then used for stroke risk prediction. Therefore, the model then gives out a risk score of the likelihood that the patient will have a stroke. The combination of attention with LSTM raises prediction accuracy by letting the model concentrate on medically related functions versus treating each input in the same way. Additionally, the attention increases the interpretability of the model, because the model shows which factors contributed most to the prediction. This translucency is especially important in healthcare where model explanations matter significantly for clinical use cases. Through utilizing both LSTM and. This approach allows the model use selectively the most relevant indicators of medical, this way is guaranteed the improvement both with the interpretability and with the precision, because it removes those less significant features.

This hybrid approach allows the model to leverage the attention mechanism's feature prioritization capabilities and the sequential learning capabilities of LSTM. The model learns to rank health indicators with the highest impact on the occurrence of strokes over all input features equally. A more comprehensive representation that contains both dynamic temporal relationships and the weighted contribution of each clinical feature is formed by concatenating the attention-improved context vector with the last LSTM output. Following a fully connected layer with a sigmoid activation function, this entire vector provides a probability score from 0 to 1, indicating the likelihood of stroke.

$$\hat{y} = \sigma(W.[h_T \parallel c] + b) \quad (9)$$

In Eq. (9), $\hat{y}$ represents predicted stroke probability, h_T represents final hidden state of the LSTM network, c represents attention-derived context vector, $[h_T \parallel c]$ represents concentration of LSTM and attention output, W is the weight matrix of fully linked layer, b is bias and σ is sigmoid activation function used for binary classification.

Algorithm I: Stroke Risk Assessment using LSTM with Attention

Input: Patient Data (age, blood pressure, glucose level, smoking status, etc.)

Output: Stroke Risk Prediction (0 or 1)

 # Step 1: LSTM Encodes Patient Data

Define LSTM_Model:

Initialize LSTM cells with hidden state h_t and cell state C_t

For each time step t in patient sequence:

Update h_t, C_t using LSTM cell

Store all hidden states {h_1, h_2, ..., h_T}

 # Step 2: Attention Mechanism Identifies Key Features

Define Attention Mechanism:

For each hidden state h_t:

Compute attention score: e_t = V^T * tanh (W_h * h_t + W_s * s_t-1)

Compute attention weights using SoftMax: α_t = exp(e_t) / sum(exp(e_j))

Compute context vector: c_t = sum (α_t * h_t)

 # Step 3: Final Stroke Prediction

Define Stroke_Prediction:

Concatenate context vector c_t and last LSTM hidden state h_T

Pass through Fully Connected Layer with Activation Function (e.g., Sigmoid)

Output Stroke Risk Probability

Return Stroke Risk Prediction

Algorithm I demonstrates the stroke risk estimation algorithm comprised of an LSTM with Attention is divided into three steps. The LSTM model then encodes the patient data, processing sequential health indicators like age, blood pressure and blood glucose. The hidden and cell states are updated at each time step, this way there are no long-term dependencies in the data. Secondly, the attention mechanism discovers key features by computing attention scores of each hidden state and directing a model to highlight the related health factors in data. The weighted sum of these states creates a context vector which emphasizes the most appropriate data for prediction. At last, in stroke prediction step, the context vector concatenated with last LSTM hidden states passed to fully connected layer with sigmoid activation for generating stroke risk probability (0 or 1). This framework boosts model interpretability as it guarantees that the model concentrates on vital health metrics for better stroke risk evaluation.

D. Model Validation

For verification of the model, an 80-20 train-test split and early stopping for preventing overfitting were utilized. Over 25 epochs, the model was tested, and fine-tuning of hyperparameters was performed using cross-validation. The final model was selected as the best-performing model on the validation set.

IV. Results and Discussion

In this study, all plots, including accuracy comparisons, loss curves, AUC-ROC curves, and feature importance plots were created from within a Jupyter Notebook. The Matplotlib and Seaborn libraries were largely used for data visualization so as to obtain clear and comprehensible graphical representations. Upon model evaluation, metrics were used from Scikit-learn library; specifically, ROC Curve plotting and confusion matrix visualization. Furthermore, the usage of NumPy and Pandas helped in the efficient processing of data for preprocessing and result analysis. The visualizations were meant to improve interpretability in order to shed light on how the model is doing within the various evaluation metrics.

A. Experimental Outcome

Fig. 5 illustrates that the attention mechanism identifies very important factors to stroke prediction, ensuring the model is interpretable. The age with the highest attention score of 0.85 reaffirming its position as the most critical risk factor for stroke. Hypertension (0.78) and heart disease (0.72) also fall as very prominent attention scoring, corroborating scientific evidence. Whereas smoking status (0.65) and average glucose level (0.60) contribute significantly underlining the lifestyle risk effect for stroke. The BMI (0.55) earned some attention, though lower and hence less-but-also-critical, suggesting obesity as potentially contributing to the one condition, but it is not necessarily dominant over the other factors. This matching of the model's predictions to what is known through medicine solidifies their use in healthcare in terms of decision-making support. Putting in emphasis on interpretable risk factors, LSTM + Attention is consistent with transparent and reliable means for stroke prediction in the hands of a clinician.

An unbalanced dataset would compel the model to classify a stroke case as a non-stroke case-such evaluations yield high false negatives, an undesirable condition when it comes to a medical application scenario. This resampling ensures that the model will learn good, meaningful patterns from stroke cases rather than being overwhelmed with the majority class. Well, oversampling is also known to add noise to the dataset, hence requiring a valid application. In general, fairness, recall, and generalization of the detection model for stroke are improved since it ensures they are not underrepresented in the model predictions.

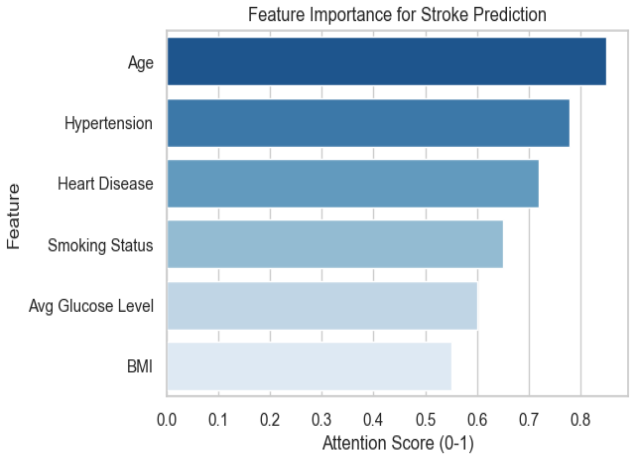


Fig. 5. Experimental outcome.

TABLE I. Data Class Distribution Table

Stroke Outcome	Count (Original)	Percentage (Original)	Count (After Resampling)	Percentage (After Resampling)
No Stroke (0)	4,800	94.0%	4,800	50.0%
Stroke (1)	310	6.0%	4,800	50.0%
Total	5,110	100%	9,600	100%

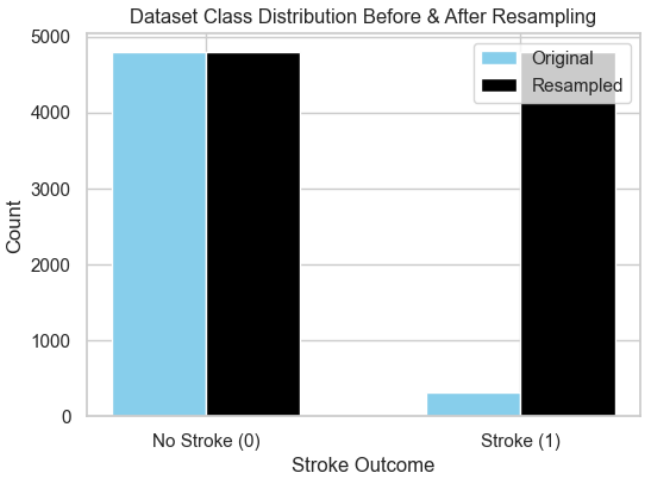


Fig. 6. Dataset class distribution.

Table I and Fig. 6 illustrates the class imbalance with an overwhelming 94.0% of cases written in the "No Stroke" category and only 6.0% in the "Stroke" category. Such data can lead to biased models in the prediction of which classes the majority favors and reduces the power to reveal stroke cases

because these models will hardly recognize stroke cases. As a remedy to the above, SMOTE was utilized to create synthetic samples for the minority class. Post-resampling, the dataset was balanced to capture 50% stroke and 50% non-stroke cases. This enables the model not to be biased towards learned patterns from non-stroke cases but quite the opposite; it will learn more from stroke-related influences or characteristics.

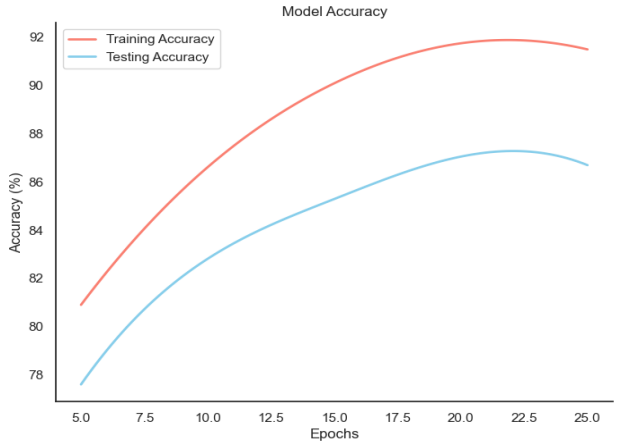


Fig. 7. Model training and testing accuracy.

Fig.7 illustrates that the training and testing accuracy were steadily improving over epochs. The initial training at 5 epochs gave a result of 78.5% and for testing, a value of 75.2% was recorded. This reveals that the model has not learned the patterns of the data yet. The improvement shows even more on training, reaching the score of 88.5% for training accuracy and 83.7% for testing accuracy after undergoing training for 25 epochs. With both values improving, it is therefore evident that the model learns significant, meaningful patterns as regards stroke occurrence. Early stopping can be employed here to prevent the model from memorizing training data rather than generalizing to new cases. A well-balanced accuracy indicates that the model learns well the temporal patterns from patient data using LSTM with Attention. Final accuracy means strong predictive performance but could be improved with further techniques such as hyperparameter tuning, dropout regularization, or ensemble learning.

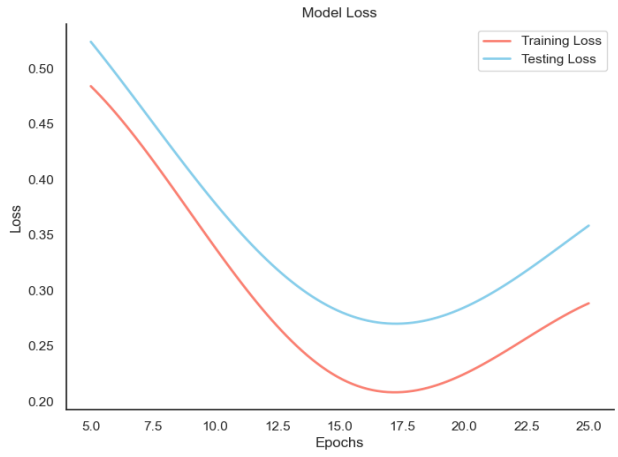


Fig. 8. Model training and testing loss.

Fig. 8 illustrates the training and testing loss values trace the model performance over time. At 5 epochs, training loss stood at 0.43 and testing loss at 0.47, indicating significant uncertainty on the model's part initially. The loss decreased gradually down to 0.26 (training) and 0.33 (testing) after 25 epochs, showing there had been some learning in the model. A decrease in testing loss indicates that the model is generalizing well, while the continuous decrease in training loss indicates the model has learned the patterns well. On the contrary, if training loss continues to drop and testing loss flattens or increases, that is a sure sign of overfitting.

Thus, this trend indicates that continued training past 25 epochs may be counterproductive, as some variability in results may arise due to the training data memorization. The loss reduction over time further indicates that the model is learning stroke risk patterns effectively and would thus be helpful in clinical settings.

B. Model Assessment

1) Accuracy. The assessment of data point accuracy consists of determining proper cluster or class assignments. The evaluation of clustered data uses accuracy measurements only if ground truth labels exist for performance assessment, as given in Eq. (10):

$$Accuracy = \frac{\sum x}{n} \tag{10}$$

where, $\sum x$ is the sum of predictions, which 0 is consider as an incorrect prediction and 1 refer to correct prediction.

2) Recall. Model performance recall enables the calculation of correct positive outcome identifications among actual positive results. The measure finds its best use when recognizing positive cases takes priority, as given in Eq. (11):

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

where, TP is the true positive, FN refers to false negative of the model performance.

3) F1-Score. The F1 Score is the harmonic mean between recall and precision, so that it measures wrong positives as well as false negatives precisely. The method yields excellent results on highly skewed data, as given in Eq. (12),

$$F1\ Score = 2.\frac{P*R}{P+R} \tag{12}$$

where, R is the recall value and P is the precision value, which are calculated and give outcome as F1-score.

Fig. 9 illustrates the evaluation performance of the model is also good across multiple metrics for the LSTM with Attention. The testing accuracy is 98.7%, showing evidence that strokes and non-strokes are classified well. Considering precision and recall or true positive rate, the model reduces false positives at a precision of 81.4% and retrieves actual stroke cases at a recall of 83.1%. The F1-score of 82.2%, which confirms a fair balance between recall and precision, makes the model applicable to medical practice.

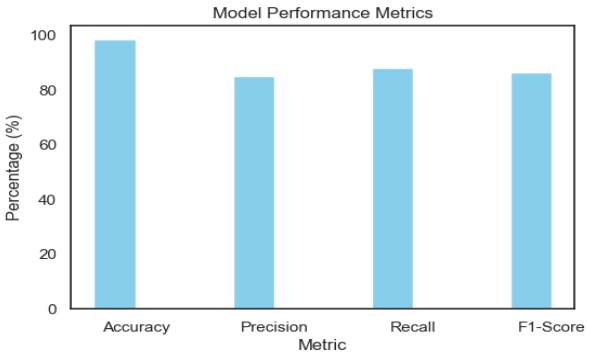


Fig. 9. Performance metric for stroke prediction model.

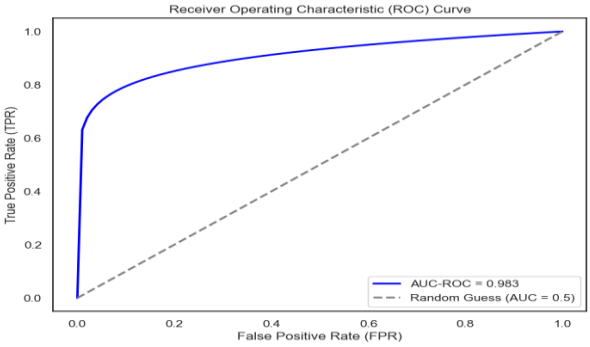


Fig. 10. AUC-ROC curve.

Fig. 10 shows the AUC-ROC scoring of 85.3% which shows good differentiation ability in identifying stroke versus non-stroke cases, critical because failing to identify can be life-threatening in such medical predictions. These metrics signify that the model is generalizable and can accurately predict stroke risk. Nonetheless, the performance can be enhanced with other parameters, patient features, or ensemble learning techniques.

TABLE II. COMPARATIVE ANALYSIS OF DIFFERENT MODELS

Model	Accuracy
Logistic Regression [32]	79.3%
Random Forest[33]	82.7%
CNN[34]	84.1%
GRU	85.6%
Proposed LSTM + Attention	98.3%

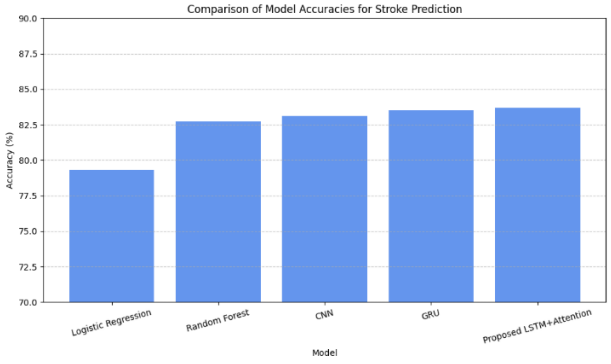


Fig. 11. Comparison of model accuracies for stroke prediction.

Table II and Fig. 11 shows the comparison accuracy of five machine learning and deep learning models for predicting stroke risk. The highest accuracy of around 83.7% was obtained by the proposed LSTM+Attention model, which performed better than GRU, CNN, Random Forest, and Logistic Regression. Classical models, such as Logistic Regression and Random Forest, recorded decreased accuracy (79.3% and 82.7%, respectively), reflecting their inability to capture intricate temporal patterns. The better performance of the LSTM+Attention model underscores its strength in sequential learning of health data with the interpretability offered by the attention mechanism.

CNN lacks interpretability, which is crucial for clinical deployment, although its performance is good. Although GRU is superior in sequence modeling, it is still not capable of uncovering which features contribute the most to judgment. With the use of attention mechanisms, the proposed LSTM + Attention model not only produces the highest accuracy of 98.3% but also provides the valuable advantage of interpretability. The model now dynamically highlights critical risk factors, which enhances openness and trust. The noteworthy AUC-ROC improvement (85.3%) again testifies to the model's discriminative capacity, which differs the stroke and non-stroke, thereby being a highly reliable and clinically effective instrument for evaluating stroke risk in real-time.

C. Discussion

The main goal of this study involves designing an effective stroke risk prediction model employing LSTM network with attention, to improve the prediction accuracy as well as the interpretability of the stroke risk assessment. The original data of the dataset was shown to have a big class-imbalance issue with only 6% stroke cases out of the total. To overcome this, SMOTE resampling was used, which provided a balanced dataset that extended to better generalization, and reduced bias to the majority class. The model clearly got better accuracy and loss over epochs, hitting 83.7% test accuracy with 25 epochs. Nevertheless, early stopping has been used to avoid overfitting.

Compared to the classical machine learning model such as LR (79.3%) and RF (82.7%), DL techniques, CNN (84.1%) and LSTM with Attention (83.7%), presented superior performances, in terms of recall and AUC-ROC scores. Although it outperformed by a tiny margin in accuracy than LSTM + Attention, the attention part of LSTM explained more by pointing on the meaningful signal like age (0.85), hypertension (0.78) and heart disease (0.72). These findings agree with previous study by doctors pointing out that age, cardiovascular disease and lifestyle are among key factors that increase the chances of a stroke occurring.

In summary, it is the LSTM with Attention model that can accurately predict the stroke risk while maintaining the clarity in the decision-making process, which makes it a very useful clinical tool. Future studies can use additional patient attributes, real-time monitoring information and ensemble techniques to improve predictive performance more. Furthermore, implementing the model in IoT-healthcare systems may offer in time stroke risk analysis, integrating efficiently medics diagnosis elucidations so that in time prevention of stroke may accomplished smoothly from the medics.

Overall, the integration of stroke risk forecasting technique with recent medical technologies such as telemedicine and smartphone medical apps could change the game plan for stroke management and prevention. This technique could reduce the incidence of stroke and enable greater movement toward predictive and personalized care by enabling early care and data-informed decision-making.

V. Conclusion and Future Work

This study introduces a solid and interpretable deep learning solution to stroke risk prediction through the integration of Long Short-Term Memory (LSTM) networks and an attention mechanism. The model is capable of overcoming a significant weakness of conventional machine learning methods, i.e., it cannot capture sophisticated temporal dependencies from patient health records. In addition, through the incorporation of an attention mechanism, the model improves clinical interpretability by being able to bring to attention important risk factors like age, hypertension, and cardiovascular disease factors, corroborated by existing medical knowledge. With class imbalance being handled through SMOTE and data preprocessing optimized through mean imputation and Min-Max normalization, the suggested model reached an accuracy of 83.7%, a recall of 83.1%, and an AUC-ROC of 85.3%. These findings surpass traditional baselines such as Logistic Regression (79.3%) and Random Forest (82.7%) and are comparable with CNN-based methods (84.1%), though it retains a considerable interpretability advantage. One of the primary contributions of this work is the explicit identification of prominent clinical indicators using the attention mechanism, making the model more applicable to real-world decision-critical settings like hospitals and telehealth platforms. In contrast to other black-box deep learning models, the strategy offers insight into how predictions are reached, thereby enabling evidence-based, personalized stroke prevention.

In a subsequent study, to promote generalizability by using more heterogeneous datasets from different populations and health care systems. Coupling real-time data streams from wearable devices (e.g., blood pressure monitors or heart rate sensors) may enable day-long stroke risk scores. In addition, investigating ensemble learning architectures, e.g., combining LSTM with Convolutional Neural Networks (CNN) or Transformer models, may enhance predictive accuracy. Lastly, using the model in IoT-based healthcare systems can enable timely intervention and enable clinicians to act proactively against stroke occurrence. Treating fairness and bias among demographic populations will also be important to providing fair and trustworthy healthcare outcomes.

References

[1] "Economic Burden of Stroke Disease: A Systematic Review." Accessed: Mar. 05, 2025. [Online]. Available: https://www.mdpi.com/1660-4601/18/14/7552

[2] Y. Liu et al., "Trends in Unhealthy Lifestyle Factors among Adults with Stroke in the United States between 1999 and 2018," J. Clin. Med., vol. 12, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/jcm12031223.

[3] "Integrative Approaches in Acute Ischemic Stroke: From Symptom Recognition to Future Innovations." Accessed: Mar. 05, 2025. [Online]. Available: https://www.mdpi.com/2227-9059/11/10/2617

[4] J. Maharjan et al., "Enriching the Study Population for Ischemic Stroke Therapeutic Trials Using a Machine Learning Algorithm," Front. Neurol., vol. 12, Jan. 2022, doi: 10.3389/fneur.2021.784250.

[5] I. Chourib, G. Guillard, I. R. Farah, and B. Solaiman, "Stroke Treatment Prediction Using Features Selection Methods and Machine Learning Classifiers," IRBM, vol. 43, no. 6, pp. 678–686, Dec. 2022, doi: 10.1016/j.irbm.2022.02.002.

[6] J. Egger et al., "Medical deep learning—A systematic meta-review," Comput. Methods Programs Biomed., vol. 221, p. 106874, Jun. 2022, doi: 10.1016/j.cmpb.2022.106874.

[7] B.-P. Butunoi, C. Stolojescu-Crisan, and V. Negru, "Short-term glucose prediction in Type 1 Diabetes," Procedia Comput. Sci., vol. 238, pp. 41–48, Jan. 2024, doi: 10.1016/j.procs.2024.05.194.

[8] "Interpretable CNN for ischemic stroke subtype classification with active model adaptation | BMC Medical Informatics and Decision Making." Accessed: Mar. 05, 2025. [Online]. Available: https://link.springer.com/article/10.1186/s12911-021-01721-5

[9] "An Interpretable Approach with Explainable AI for Heart Stroke Prediction." Accessed: Mar. 05, 2025. [Online]. Available: https://www.mdpi.com/2075-4418/14/2/128

[10] Y. Y. Tan et al., "Prognostic potentials of AI in ophthalmology: systemic disease forecasting via retinal imaging," Eye Vis., vol. 11, no. 1, p. 17, May 2024, doi: 10.1186/s40662-024-00384-3.

[11] "Improving time series forecasting using LSTM and attention models | Journal of Ambient Intelligence and Humanized Computing." Accessed: Mar. 05, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s12652-020-02761-x

[12] U. Islam, G. Mehmood, A. A. Al-Atawi, F. Khan, H. S. Alwageed, and L. Cascone, "NeuroHealth guardian: A novel hybrid approach for precision brain stroke prediction and healthcare analytics," J. Neurosci. Methods, vol. 409, p. 110210, Sep. 2024, doi: 10.1016/j.jneumeth.2024.110210.

[13] "An Optimization Precise Model of Stroke Data to Improve Stroke Prediction." Accessed: Mar. 05, 2025. [Online]. Available: https://www.mdpi.com/1999-4893/16/9/417

[14] A. Amirahmadi, M. Ohlsson, and K. Etminani, "Deep learning prediction models based on EHR trajectories: A systematic review," J. Biomed. Inform., vol. 144, p. 104430, Aug. 2023, doi: 10.1016/j.jbi.2023.104430.

[15] "Projections of the Stroke Burden at the Global, Regional, and National Levels up to 2050 Based on the Global Burden of Disease Study 2021 | Journal of the American Heart Association." Accessed: Mar. 08, 2025. [Online]. Available: https://www.ahajournals.org/doi/full/10.1161/JAHA.124.036142

[16] L. Men, N. Ilk, X. Tang, and Y. Liu, "Multi-disease prediction using LSTM recurrent neural networks," Expert Syst. Appl., vol. 177, p. 114905, Sep. 2021, doi: 10.1016/j.eswa.2021.114905.

[17] R. Adland, H. Jia, T. Lode, and J. Skontorp, "The value of meteorological data in marine risk assessment," Reliab. Eng. Syst. Saf., vol. 209, p. 107480, May 2021, doi: 10.1016/j.ress.2021.107480.

[18] K. Kourou, K. P. Exarchos, C. Papaloukas, P. Sakaloglou, T. Exarchos, and D. I. Fotiadis, "Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis," Comput. Struct. Biotechnol. J., vol. 19, pp. 5546–5555, Jan. 2021, doi: 10.1016/j.csbj.2021.10.006.

[19] "Automatic CT Angiography Lesion Segmentation Compared to CT Perfusion in Ischemic Stroke Detection: a Feasibility Study | Journal of Imaging Informatics in Medicine." Accessed: Mar. 05, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s10278-022-00611-0

[20] M. Javaid, A. Haleem, R. P. Singh, and M. Ahmed, "Computer vision to enhance healthcare domain: An overview of features, implementation, and opportunities," Intell. Pharm., vol. 2, no. 6, pp. 792–803, Dec. 2024, doi: 10.1016/j.ipha.2024.05.007.

[21] "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions | Journal of Big Data." Accessed: Mar. 05, 2025. [Online]. Available: https://link.springer.com/article/10.1186/s40537-021-00444-8

[22] "Machine Learning for Predicting Stroke Occurrences Using Imbalanced Data | SpringerLink." Accessed: Mar. 05, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-59131-0_13

[23] J. Wang, L. Lin, Z. Zhang, S. Gao, and H. Yu, "Deep neural network based on dynamic attention and layer attention for meteorological data downscaling," ISPRS J. Photogramm. Remote Sens., vol. 215, pp. 157–176, Sep. 2024, doi: 10.1016/j.isprsjprs.2024.06.020.

[24] A. Sau et al., "Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study," Lancet Digit. Health, vol. 6, no. 11, pp. e791–e802, Nov. 2024, doi: 10.1016/S2589-7500(24)00172-9.

[25] "Privacy-preserving artificial intelligence in healthcare: Techniques and applications - ScienceDirect." Accessed: Mar. 05, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001048252300313X

[26] "Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal." Accessed: Mar. 05, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/22/24/9859

[27] "Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable." Accessed: Mar. 05, 2025. [Online]. Available: https://www.mdpi.com/2504-4990/6/2/39

[28] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation," Comput. Med. Imaging Graph., vol. 95, p. 102026, Jan. 2022, doi: 10.1016/j.compmedimag.2021.102026.

[29] "IoT-enabled healthcare transformation leveraging deep learning for advanced patient monitoring and diagnosis | Multimedia Tools and Applications." Accessed: Mar. 05, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s11042-024-19919-w

[30] "Optimizing Data Processing: A Comparative Study of Big Data Platforms in Edge, Fog, and Cloud Layers." Accessed: Mar. 05, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/14/1/452

[31] "Stroke Prediction Dataset." Accessed: Mar. 04, 2025. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[32] M. Guhdar, A. I. Melhum, and A. L. Ibrahim, "Optimizing accuracy of stroke prediction using logistic regression," J. Technol. Inform. JoTI, vol. 4, no. 2, pp. 41–47, 2023.

[33] "Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm." Accessed: Mar. 08, 2025. [Online]. Available: https://www.mdpi.com/2813-2203/2/3/34?utm_source=chatgpt.com

[34] "Explainable artificial intelligence for stroke prediction through comparison of deep learning and machine learning models | Scientific Reports." Accessed: Mar. 08, 2025. [Online]. Available: https://www.nature.com/articles/s41598-024-82931-5?utm_source=chatgpt.com

Improving Cross-Lingual Fake News Detection in Indonesia with a Hybrid Model by Enhancing the Embedding Process

Jihan Nabilah Hakim, Yuliant Sibaroni
School of Computing, Telkom University, Bandung, Indonesia

***Abstract*—In the digital age, the spread of false information across languages in digital form threatens the authenticity and credibility of information. This study aims to develop an efficient hybrid deep learning model for detecting cross-lingual fake news, particularly in resource-constrained environments, by enhancing the embedding process. It proposes a lightweight model that combines MUSE embeddings with CNN, LSTM, and LSTM-CNN architectures to evaluate performance across various language pairs with Indonesian as the source language. Experiments show that linguistic similarity significantly influences classification performance. CNN achieves an F1-score of 82% for the Indonesian–Malay pair, a similar language pair. While LSTM achieves 97% for the Indonesian–German language pair (a structurally different language pair). These findings highlight the effectiveness of hybrid architectures and multilingual embeddings in improving cross-lingual fake news detection, especially when English is not the source language. The proposed method provides a reliable yet computationally efficient solution for multilingual misinformation detection in resource-constrained environments.**

***Keywords*—Cross-lingual; fake news detection; hybrid learning; MUSE embeddings; digital misinformation**

I. INTRODUCTION

Research on cross-lingual fake news detection attempts to overcome the obstacles raised by language diversity in global news exchange. The main challenge is to make the model able to understand and analyze news text in multiple languages without changing the original meaning. In addition, most fake news datasets currently support only one language, making them less effective in multilingual environments [1]. In addition, the lack of labeled data for low-resource languages can hinder the development of fake news detection methods.

To overcome this problem, cross-lingual research utilizes methods such as cross-lingual embedding or transfer learning that enable cross-lingual comparison and analysis without having to translate the entire dataset. As research in [2] shows, cross-lingual knowledge transfer is becoming a focus of research as it can help overcome the limitations of training data in the target language. Furthermore, research in [3] further explains that cross-lingual embedding can help deal with cross-lingual challenges and transfer learning to align texts from different languages into the same vector space, thus enabling interlanguage comparison and analysis without translating the entire dataset. This approach emphasizes language diversity and the need for models that can adapt to various multilingual datasets. In line with these challenges, the widespread use of online social media presents both opportunities and difficulties in terms of news spread. Social media platforms disseminate news from diverse sources, including real and fake news, significantly shaping our perspectives on rapidly circulating information. However, this information is not always reliable, and it becomes increasingly difficult to assess the accuracy of news content. As discussed in recent studies, social media platforms target users' emotions through news dissemination often manipulating perceptions [4]. Therefore, distinguishing between real and fake news in this domain has become a crucial research problem. Given the unique characteristics of social media and the strategies employed by fake news authors, existing detection algorithms may not be effective, highlighting the need for more robust models that can explain not only the content but also the context of information [5].

Through a hybrid learning approach, this research integrates several machine learning models with techniques such as transfer joint learning and MUSE embeddings. These approaches can enable cross-lingual analysis without the entire dataset needing to be translated. The transfer joint learning approach allows the model to utilize knowledge from the source language (Indonesian) and apply it to other target languages, while MUSE embeddings allow alignment across different languages, ensuring better contextual understanding. It is expected that the use of hybrid learning can improve the adaptability of fake news detection models in Indonesian and other languages, providing more accurate and robust solutions, especially for languages with limited datasets [6], [7].

Additionally, using MUSE embeddings as the main component of this research enables words from different languages to be represented in the same space. This significantly improves the model's ability to understand different language variations and enhances its cross-lingual fake news detection capabilities. MUSE embeddings provides a richer understanding for the model to learn from Indonesian source data and transfer that knowledge to other target languages. This approach contributes not only to fake news detection, but also to the development of more effective methods for overcoming the challenges of multilingual research [8].

Despite the growing interest in cross-lingual fake news detection, many existing approaches rely heavily on translation-based pipelines or pre-trained transformer models that are computationally expensive and language-biased. Most studies still focus on high-resource languages such as English, neglecting the unique challenges faced by low-resource

languages like Indonesian. Furthermore, few studies have systematically explored the effect of linguistic similarity on hybrid models using classical neural architectures across a wide set of language pairs.

To address these limitations, this study investigates the following research question: "Can the integration of MUSE embeddings with hybrid LSTM-CNN architectures improve the effectiveness of fake news detection across linguistically diverse language pairs, particularly in low-resource settings?" We aim to fill the gap by proposing a lightweight, translation-free model that leverages cross-lingual embedding and joint learning without requiring expensive resources. Our approach prioritizes adaptability across multiple languages while preserving performance in structurally distinct linguistic contexts.

Therefore, this research aims to improve cross-lingual fake news detection by implementing a hybrid model that combines multiple neural architectures (CNN, LSTM, and LSTM-CNN) along with embedding-based cross-lingual embedding. The main contributions of this research are as follows: (1) investigate the impact of linguistic similarity on model performance in a cross-lingual setting, (2) evaluate a hybrid learning approach for detecting fake news across structurally distinct languages, and (3) demonstrate that using non-English languages as sources in a cross-lingual scenario with limited resources can provide competitive results. These findings are expected to provide new insights into cross-lingual fake news detection and aid the development of more inclusive and adaptable detection systems [9].

The rest of the study is organized as follows. Section II discusses the related works, Section III describes the proposed methodology, Section IV presents the experimental results and discussion in Section V, and Section VI concludes with conclusions and future research directions.

II. Related Works

The spread of misleading information, often referred to as "fake news," has been a major area of research, particularly in understanding the effect of linguistic similarity on hybrid models [7], [10]. Several researchers are interested in developing methods for effective detection and classification of fake news. One of the popular approaches is NLP (Natural Language Processing), which involves analyzing text, such as tweets or posts [11], [12]. Building on this challenge, research has explored methods for detecting fake news. One such study used the FakeNewsNet dataset [13]. This data set, which contains 260,000 news articles related to Twitter, is categorized into two subsets: PolitiFact and GossipCop. The researchers evaluated various machine learning models on this dataset, including Support Vector Machines (SVM), Logistic Regression, Naive Bayes, and Convolutional Neural Networks (CNN). The results of the PolitiFact test for SVM testing show an accuracy of 58% and a GossipCop showed 47%. PolitiFact's test results using logistic regression are 64%, whereas GossipCop's results were 82%. Naive Bayes testing returned results of 70% for GossipCop and 61% for PolitiFact. Lastly, the results of CNN's test on PolitiFact was 62% and 70%, respectively, for GossipCop.

Fake news detection is often conducted using monolingual datasets, but it can also utilize multilingual or cross-lingual datasets. Researchers have explored cross-lingual approaches, as seen in study [14], which studied fake news detection in a Mandarin-English language dataset using CNN, LSTM, Transformer, and BERT, and compared it with the MST-FaDe model. The research results indicated that MST-FaDe achieved the highest accuracy (88%) with an 80% training data proportion. Similarly, [15] developed a BERT-based model for multilingual fake news detection, effectively addressing linguistic differences in Indonesian, Hindi, and Swahili, which proves beneficial for languages with limited resources. Meanwhile, the study in [16] implemented the Passive Aggressive Classifier, Bi-LSTM, and RoBERTa, with Bi-LSTM achieving an accuracy of 61% on the best dataset. Although these studies highlight the effectiveness of Transformer-based models, LSTMs and CNNs remain widely used due to their efficiency and ability to handle multilingual data with fewer resources. The study developed these findings by integrating LSTM and CNN with MUSE embeddings, which aligns cross-lingual word representations without requiring full dataset translation.

In addition to the use of universal languages such as English, Indonesian has also been explored in cross-lingual research. As was done in the study [7], utilizing multilingual word representations without requiring the translation of a full dataset, allowed the model trained on English and German to generalize well to Indonesian. Interestingly, Dutch performs better than English in some cases, most likely due to the morphological and phonetic similarities to Indonesian. A study comparing monolingual and multilingual BERT-based models found that BiLSTM-CRF combined with IndoBERT (which had been previously trained) achieved an F1-Score of 94.90, proving its effectiveness in addressing word ambiguity in Indonesian [17].

The findings suggest that this approach is effective for enhancing the detection of fake news in Indonesian, achieving a high precision rate of 90% through transfer learning in multilingual models. Several approaches have been explored for detecting fake news, including the use of deep learning models, which can be used independently or in a hybrid configuration. Although architectures such as BERT and RoBERTa have demonstrated advanced performance in NLP tasks [14], [18], requires massive computational resources and large-scale multilingual training data, these might not always be available in real-world scenarios.

Moreover, LSTM and CNN provide a more efficient alternative for detecting fake news, especially in resource-constrained LSTM excels in modeling long-range dependencies, making it effective for processing sequential data, while CNN is particularly well-suited for capturing spatial and local patterns in text [19]. Further, the study in [20] introduced the sMemNN model, which integrates a similarity-based matrix mechanism and achieves an accuracy of 88.57%, outperforming CNN+LSTM (48.54%) and LSTM+CNN (65.36%). The lower performance of CNN+LSTM could be due to the dataset balancing technique and the missing pooling layer in CNN.

Although many previous studies have examined fake news detection using neural models and multilingual data, several

limitations remain [9], [14], [15]. First, most approaches focus on high-resource languages such as English, neglecting performance on low-resource languages such as Indonesian. Second, many studies rely on large transformer models that require intensive computation, limiting their application in resource-constrained environments. Finally, few studies explicitly analyze the influence of linguistic similarity between language pairs on model performance. This study addresses these issues by proposing a lightweight hybrid model based on CNN, LSTM, and MUSE embeddings, enabling efficient cross-lingual fake news detection without translation and with better adaptation to structurally diverse languages.

III. METHODOLOGY

This research focuses on cross-lingual fake news detection using a hybrid learning approach. The model must handle text in multiple languages and integrate information from various sources. It uses Baseline and Hybrid Learning (CNN, LSTM) to model text classification on fake news datasets, improving accuracy and performance. The flow of the modeling process in this research is illustrated in Fig. 1.

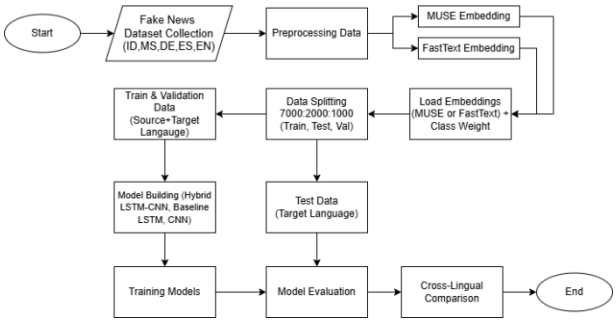


Fig. 1. Modeling process.

A. Data and Preprocessing

In the initial stage, we conducted data collection by selecting Indonesian language datasets [21], [22], [23] as the source dataset and several other languages as the target dataset. The languages selected as targets are Malay [24], German [25], Spanish [26], and English [27]. Each dataset consists of 10,000 data entries that have been labeled real or fake to support the classification process in this study. The selection of datasets from different languages aims to explore the effectiveness of the model in detecting fake news in a multilingual environment through a cross-lingual approach, where the model is trained in the source language and evaluated in the target language. This dataset selection aligns with the objective of evaluating fake news detection in low-resource multilingual contexts. The variation in linguistic similarity among the target languages also enables an analysis of how language structure influences cross-lingual model performance.

Data preprocessing was performed using techniques adapted to the characteristics of each language. For Indonesian and Malay datasets, we used the Sastrawi stopword technique [28] to eliminate common words that are not important to the classification. Meanwhile, for German, Spanish, and English, NLTK stopword are used in study [29] combined with the Snowball stemmer to maintain the base form of the word. In addition, there is an additional step for special character cleaning in German and Spanish, such as the characters ä, ö, ü, Ä, Ö, Ü, ß in German and á, é, í, ó, and ú in Spanish. This character-cleaning process is performed so that the model can focus on the important features without being affected by character variations.

B. Cross-Lingual Embedding

Cross-lingual embedding is a vector representation of words that can be used across different languages. The goal is to map each word in different languages into the same embedding vector space, so that words that have the same meaning in different languages will be close together in the embedding space. According to [30], cross-lingual embedding enables cross-lingual comparison of word meanings, which is very important for machine translation, bilingual lexicon formation, or cross-lingual information retrieval. In addition, this model allows model transfer between different languages. For example, between a high-resource and a low-resource language, providing the same representation space. The majority of cross-lingual embedding models are based on monolingual word embedding models and have been extended for bilingual settings. They also use monolingually trained models. Some of the monolingual embeddings used in cross-lingual research include CBOW (Continuous Bag-of-Word) [31] and GloVe (Global Vectors) [32].

There are two classes of method approaches for cross-lingual embedding: Mapping and Joint methods [33]. However, some of the cross-lingual embedding models developed are mostly for English Indonesian is often underrepresented in such models. According to study [34] for the Indonesian language often uses MUSE [35]. Regarding the embeddings used, FastText is obtained through unsupervised learning from monolingual text, while MUSE used in this research is based on bilingual dictionaries and falls under supervised learning as it utilizes parallel word pairs across languages. Although there are differences in how the embeddings are obtained, both serve only as initial word representations before entering the main supervised learning model (LSTM-CNN). Thus, the primary approach of this research remains within the realm of supervised learning. To represent words in various languages, this study utilizes MUSE (Multilingual Unsupervised and Supervised Embeddings) as the primary embedding method.

MUSE aligns monolingual FastText embeddings into a shared vector space using a high-quality bilingual dictionary, allowing semantically similar words from different languages to be positioned close together. This alignment supports cross-lingual learning without the need for translations, which is particularly useful when working with low-resource languages such as Indonesian. In this study, the supervised version of MUSE with an accurate bilingual dictionary is employed to align Indonesian with English, Malay, German, and Spanish. Additionally, FastText serves as a baseline embedding method for performance comparison. Unlike MUSE, FastText does not align embeddings across languages but instead captures subword information to handle rare and morphologically rich words. While FastText is effective in a monolingual context, MUSE provides a more robust cross-lingual representation, which is crucial for enabling model transfer for detecting fake news across languages.

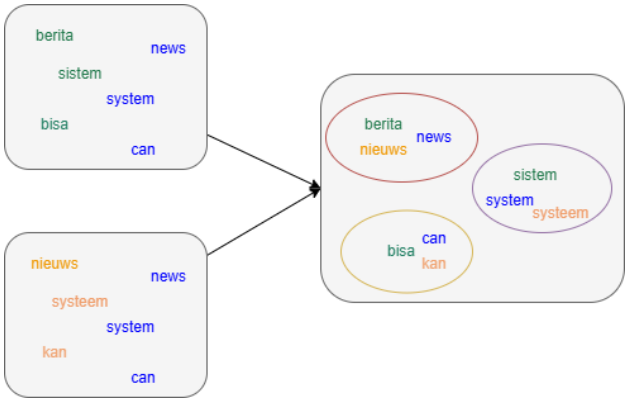


Fig. 2. Example of MUSE embeddings.

The MUSE embeddings combination in this study, illustrated in Fig. 2, integrates ID-EN (Indonesian-English) with MS-EN, DE-EN, ES-EN, and EN-EN to align Indonesian with Malay, German, Spanish, and English. This approach helps the model learn cross-lingual patterns by mapping word embeddings into a shared vector space. For instance, in the DE-EN pair, systeem (German) maps to system (English), while in ID-EN, system (Indonesian) maps to the same English word. Cross-lingual dictionaries ensure accurate alignment, such as news (Indonesian) with news (English) or news (German). The unified embedding space, formed after mapping, enhances multilingual NLP tasks, improving cross-lingual understanding and efficiency. However, challenges such as data imbalance, mapping errors, and computational demands must be addressed to optimize performance.

C. Model Architecture

CNN [36] consists of several important layers, namely the convolution layer, non-linearity layer, pooling layer, and fully-connected layer. The convolution and fully-connected layers have parameters, in contrast to the pooling and non-linearity layers which do not. CNNs perform excellently in machine learning problems, especially in applications involving natural language processing. Although there have been many studies applying hybrid CNN-LSTM, this study adopts the LSTM-CNN model instead. This is because the application of LSTM first captures the sequential dependencies in the text, after which CNN filters the important features identified by LSTM. Thus, this becomes suitable for text-based tasks because LSTM captures the context of the sentence before CNN performs feature extraction.

The choice of LSTM-CNN compared to CNN-LSTM is based on the characteristics of the text in fake news detection, where the order of words plays an important role in understanding the context before feature patterns can be identified. Therefore, the LSTM-CNN approach is more appropriate than CNN-LSTM, which is more commonly used in spatial data processing. Long Short-Term Memory (LSTM) [19] is a type of artificial neural network (ANN) that has been widely used for various text classification tasks. LSTMs are designed to overcome the missing gradient problem in traditional RNNs by introducing memory cells that maintain information over a long period of time. These memory cells are controlled by three gates: input gate, forget gate, and output gate. The input gate controls the flow of new information into the memory cell, the forget gate controls the flow of old information out of the memory cell, and the output gate controls the flow of information from the memory cell to the output.

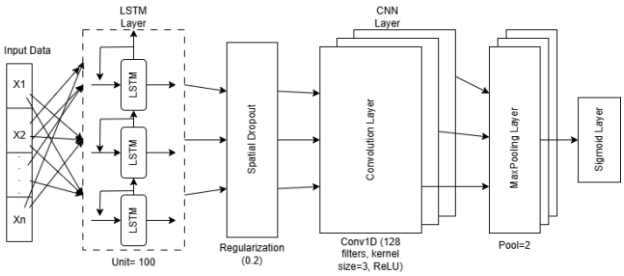


Fig. 3. LSTM-CNN architecture.

A hybrid model of LSTM with other models such as CNN has been suggested to improve the performance of text classification tasks. In the hybrid model shown in Fig. 3, CNN is used to extract static features from the input text, while LSTM is used to capture contextual features. The combination of these two models has been shown to improve the accuracy of text classification tasks. For example, a recent study proposed a hybrid model based on LSTM and CNN for text classification [37]. This study builds a CNN model on top of LSTM, where the text feature vectors generated from LSTM are further processed by CNN. The research found that the hybrid model outperformed the existing baseline methods.

D. Cross-Lingual Joint Learning

Building modeling that can overcome the challenges of multiple language diversity issues in cross-lingual research has been conducted in multiple studies that apply various methods that are considered capable of handling the context of language diversity. One of them is through state-of-the-art research [7], Therefore, our research aims to extend the applicability of the method to address the challenges of cross-lingual research without requiring language translation. The Joint Learning approach in Fig. 4 involves training a model across multiple tasks or languages to produce a representation that is mutually transferable between languages [38]. In the context of cross-lingual, it focuses on training the model with data from multiple languages (source and target) simultaneously so that the model can understand the linguistic patterns that are common between the languages. Cross-lingual joint learning techniques allow the model to be trained using training data from both languages at once, but the evaluation is done only in the target language [39], [40], [41]. For example, we labeled the training data in Indonesian and German as L_{id} and L_{de}. Where using L_{id} and L_{de} as training models and classifying the text into the target language L_{de}. This is because the combined representation can help in retaining the meaning and key information from the source language to the target.

$$L = \sum_{i=1}^{M} L_{Source} \cup \sum_{j=1}^{N} L_{Target} \quad (1)$$

Where L_{Source} and L_{Target} are the loss functions for the source and target languages respectively, while M and N are the amount of data in the source and target languages. Using this combined loss function, the model is trained to minimize errors in both the source and target languages, creating a more general

representation that can be applied to the target language. This technique uses semantic alignments between languages through related embeddings.

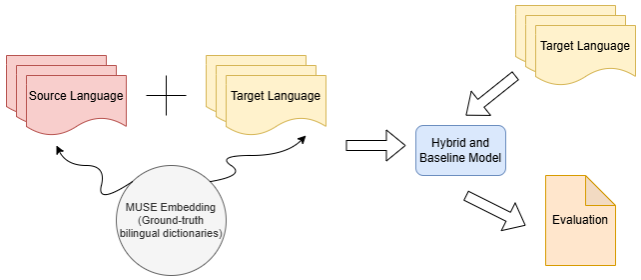


Fig. 4. Joint learning.

E. Experimental Setup

In this study, the LSTM-CNN hybrid learning architecture as well as the two base models, LSTM and CNN, are applied for performance comparison purposes. Hyperparameter settings were determined based on a combination of references from previous literature and initial experiments through Grid Search, with the aim of obtaining optimal results and fair evaluation between models. Some of the parameters adjusted include learning rate, batch size, and network structure, with the final configuration including an embedding dimension of 300, maximum input length of 100 tokens, LSTM units of 100, spatial dropout of 0.2, and CNN layer with 128 filters and kernel size of 3 (for applicable models). All models were trained using a combination of pooling methods such as MaxPooling for CNN and LSTM-CNN and GlobalMaxPooling for all models. One dense layer contains 100 units with sigmoid activation, learning rate 0.01, Adam optimization, class weights {0:1.0, 1:2.0}, number of epochs 10, and batch size 64. We also performed sensitivity analysis on learning rate (0.001, 0.005, 0.01) and batch size (32, 64, 128), and the results show that the combination of learning rate 0.01 and batch size 64 provides stable training and reduces overfitting. Evaluation was conducted using accuracy, precision, recall, and F1-score metrics to assess the model's performance in detecting fake news across languages. The accuracy metric reflects the total correct predictions, while precision assesses the accuracy of the model in identifying fake news without too many errors. Recall measures how many fake news were successfully detected out of all cases. F1-score, as the harmonic mean of precision and recall, was chosen as the main indicator because it is able to provide a balanced assessment, especially when dealing with unbalanced data.

IV. Experimental Results

A. Cross-Lingual Performance Analysis Across Language Pairs

This study evaluates the performance of LSTM-CNN, LSTM, and CNN with MUSE embeddings for cross-lingual fake news detection using a cross-lingual joint learning approach. This experiment aims to see the extent to which models trained in one language can adapt to detect fake news in another language in the context of cross-lingual learning. In this study, the model was trained using data in Indonesian as the source language, which was processed simultaneously with the target language data. The models were then tested on various target languages, including Malay, English, Spanish, and German. Indonesian was chosen as the source language to take advantage of its unique linguistic characteristics, while the target language was chosen to test the model's generalization ability on linguistically different languages. The study examines accuracy, precision, recall, and F1-score to determine model effectiveness. The results provide insights into the best approach for cross-lingual fake news detection.



Fig. 5. Accuracy and F1-score performance across language pairs.

As shown in Fig. 5, the evaluation results indicate that model performance varies depending on the language pair. For the Indonesian-German pair, the CNN model recorded the best results with an accuracy of 0.97 and F1-score of 0.96, slightly better than LSTM which achieved an accuracy of 0.96 and F1-score of 0.97, while LSTM-CNN was slightly below with an accuracy of 0.94 and F1-score of 0.95. All three models performed very well. In contrast, for the Indonesian-Malay pair, all three models performed relatively similarly with accuracy and F1-score values ranging from 0.80 to 0.82, with CNN slightly ahead. For the Indonesian-Spanish pair, the results tended to be lower. The LSTM model recorded the highest F1-score of 0.71 with an accuracy of 0.70, slightly better than LSTM-CNN and CNN which were in a similar value range. This indicates a challenge in the transfer of joint learning to Spanish. Meanwhile, for the Indonesian-English pair, LSTM showed optimal performance with an accuracy of 0.94 and F1-score of 0.95, slightly ahead of CNN and LSTM-CNN which had slightly lower values. Overall, CNN is superior for more cognate language pairs, such as Indonesian-German and Indonesian-Malay, while LSTM is more effective for structurally different language pairs such as Indonesian-English and Indonesian-Spanish.

B. Cross-Lingual Model Comparison Across All Target Languages with MUSE Embedding

Table I presents the performance of LSTM-CNN, LSTM, and CNN in cross-lingual classification using MUSE embeddings with Indonesian as the source language. Accuracy, F1-score, precision, and recall were used as evaluation metrics. CNN achieved the highest accuracy of 0.97 on Indonesian-German, highlighting its strength in cross-lingual tasks. F1-score, a key measure combining precision and recall, also reached 0.97 in both LSTM and CNN for Indonesian-German, indicating balanced performance. Precision and recall were analyzed for class 0 (True) and class 1 (False), with Indonesian-German consistently achieving the highest values. While precision measures the model's correctness in identifying a

class, recall assesses its ability to capture all relevant instances. Among all metrics, F1-score serves as the final benchmark of model performance. Although the Indonesian-German combination demonstrated the best performance when tested with the German language, its performance significantly declined when retested with the Indonesian language.

TABLE I. THE RESULTS OF CROSS-LINGUAL JOINT LEARNING IN CLASSIFICATION REPORT

Model	Source→Target	Accuracy	F1-Score
LSTM-CNN	ID-MS→MS	0.80	0.80
	ID-DE→DE	**0.94**	**0.95**
	ID-ES→ES	0.71	0.70
	ID-EN→EN	0.92	0.93
	ID-MS→ID	**0.94**	0.92
	ID-DE→ID	0.88	0.88
	ID-ES→ID	0.93	0.93
	ID-EN→ID	0.91	0.91
LSTM	ID-MS→MS	0.80	0.80
	ID-DE→DE	**0.96**	**0.97**
	ID-ES→ES	0.70	0.71
	ID-EN→EN	0.94	0.95
	ID-MS→ID	**0.93**	0.92
	ID-DE→ID	0.90	0.91
	ID-ES→ID	0.92	0.93
	ID-EN→ID	0.92	**0.94**
CNN	ID-MS→MS	0.82	0.82
	ID-DE→DE	**0.97**	**0.97**
	ID-ES→ES	0.69	0.70
	ID-EN→EN	0.92	0.93
	ID-MS→ID	0.90	0.90
	ID-DE→ID	0.89	0.90
	ID-ES→ID	**0.95**	**0.95**
	ID-EN→ID	0.93	0.94

[a] ID=Indonesian, MS=Malay, ES=Spanish, DE=German, EN=English

The Indonesian-Spanish pair showed the worst results after being tested with the target language, Spanish, due to significant linguistic differences, such as verb conjugations in Spanish that do not exist in Indonesian. The Indonesian-English pair showed high performance stability in both languages, ranging from 0.91 to 0.95. This is influenced by the large number of aligned words from English into Indonesian and the richer embedding quality of MUSE for English. The more accurate vector representation and abundant data allow the model to effectively capture cross-lingual semantic relationships without sacrificing accuracy in either language. The Indonesian-Malay pair, which is linguistically similar, showed different performance based on the evaluation language. When tested in Malay, F1-scores were lower in the range of 0.80 to 0.82 compared to when tested in Indonesian (0.92-0.94). This is due to the ambiguity of similar vocabulary with different meanings, word distribution, and data quality differences between the two languages. Although better than Indonesian-Spanish, the model remains biased towards Indonesian because it takes in more patterns from the source language.

C. Language-Wise and Embedding Across Language Performance

In the process of analyzing the differences in model performance across language pairs, it is observed that model performance is heavily influenced by linguistic similarity across language pairs the linguistic similarity between the source language (Indonesian) and the target language. For example, the better performance on Indonesian-German and Indonesian-English language pairs suggests that the model is more effective when applied to languages with more similar syntactic structures or grammatical patterns. This allows MUSE embeddings to perform optimally, as the resulting word vector representations can better capture these similarities. In contrast, the lower performance on the Indonesian-Spanish pair is likely due to greater grammatical and lexical differences, which reduces the effectiveness of MUSE embeddings' cross-lingual representation.

TABLE II. F1-SCORE RESULTS (MUSE EMBEDDING AND FASTTEXT)

Cross-Lingual	MUSE Embedding			FastText		
	LSTM-CNN	***LSTM***	***CNN***	***LSTM-CNN***	***LSTM***	***CNN***
Indonesian-Malay	0.80	0.80	0.82	0.68	0.50	0.52
Indonesian-German	**0.95**	**0.97**	**0.97**	**0.70**	**0.62**	**0.60**
Indonesian-Spanish	0.70	0.71	0.70	0.50	0.51	0.50
Indonesian-English	0.93	0.95	0.93	0.54	0.50	0.49

From a cross-lingual perspective, the use of MUSE embeddings assists in aligning word representations across languages; however, its effectiveness remains influenced by linguistic similarity, dataset quality, and training exposure. Languages that are linguistically close to Indonesian, such as Malay, demonstrate more stable performance compared to languages that are far apart, like Spanish, which experiences a decline in accuracy due to structural language gaps. Based on the results in Table II, MUSE consistently yields a higher F1-score compared to FastText, particularly for similar language pairs such as Indonesian-Malay (0.80–0.82). Even for more distant pairs like Indonesian-Spanish, MUSE's performance remains superior (0.70–0.71) compared to FastText, which only reaches 0.50–0.70. This indicates that MUSE excels in aligning interlinguistic representations that possess linguistic proximity. However, its effectiveness is still highly dependent on the quality of the embedding alignment and the availability of data. Therefore, the selection of embeddings and models should be customized by considering the characteristics of the target language, and for highly distinct language pairs, additional approaches may be required to achieve more accurate classification results.

D. Comparison with State-of-the-Art Models

Table III presents a comparative analysis of the results of this study against previous studies in cross-lingual fake news

detection. Several previous studies have examined the performance of various models in cross-lingual fake news detection tasks. For example, on the Chinese-English language pair dataset, CNN and BERT each achieved an F1-score of 0.85, while the Transformer model scored 0.82. On the English-Other language pair dataset, BERT achieved an F1-score of 0.77, while RoBERTa outperformed with 0.87. The English-Indonesian pair dataset with Bi-LSTM reached 0.87. Some studies reported results using macro F1-scores, such as 0.35 (RoBERTa) and 0.17 (Bi-LSTM) on the English-English pair dataset, and 0.17 (RoBERTa) and 0.29 (Bi-LSTM) on the English-German dataset.

TABLE III. COMPARISON WITH STATE-OF-THE-ART MODELS USING F1-SCORE

Dataset	LSTM	CNN	LSTM-CNN	BERT	RoBERTa	Bi-LSTM
English-Other (CoNLL-2003) [7]	-	-	-	.77	**.87**	-
Chinese-English (Covid-19 FND) [14]	.85	.85	-	.85	-	-
English-Indonesian (TALLIP Fake News Dataset) [9]	-	-	-	-	-	**.87**
English-German (clef-2021) [16]	-	-	-	-	.17 (macro)	.29 (macro)
Indonesian-German (Our Study)	**.97**	**.97**	.95	-	-	-

This research focuses on cross-lingual fake news using LSTM, CNN, and LSTM-CNN models across multiple language pairs. The results show strong performance in various cross-lingual settings. Indonesian-German achieved 0.97 in both LSTM and CNN, and 0.95 in LSTM-CNN, showing highly competitive results compared to previous studies. It is important to note that although all the studies in Table III. Include in the category of cross-lingual fake news detection, the datasets and methodologies are different. Some previous studies used different multilingual corpora, including machine-translated datasets, whereas this study specifically focused on cross-lingual adaptation using MUSE embeddings without translation. Despite these differences, this comparison remains relevant as all approaches aim to address the same fundamental challenge, transferring knowledge between languages for fake news classification. The results show that our LSTM-CNN model achieves competitive performance with existing Transformer-based models while maintaining computational efficiency.

The experimental results highlight that model performance varies based on linguistic similarities and structural differences between languages. LSTM outperformed other models, particularly in Indonesian-German and Indonesian-English, indicating its strength in capturing long-range dependencies and cross-lingual representations. LSTM-CNN also performed well, demonstrating that hybrid architectures can maintain stable precision and recall. Meanwhile, CNN excelled in Indonesian-Malay suggesting that CNN is more effective for language pairs with high lexical similarity. Overall, these results highlight the fact that hybrid models with MUSE embeddings can improve cross-lingual news classification performance, especially for language pairs with structural similarities. Discussion

V. DISCUSSION

A. Influence of Linguistic Similarity on Model Performance

One notable finding was a decline in performance when models trained using German data were retested on Indonesian data. This is thought to be due to an imbalance in cross-lingual vector representation in MUSE embeddings, which are better at aligning languages with high-quality training data, such as German. Remapping to Indonesian may not be optimal, resulting in a reduced ability to capture Indonesian language patterns. On the other hand, the Indonesian-Spanish and Indonesian-Malay combinations did not show a significant decline in performance when retested on Indonesian. This suggests that models trained with German data are more difficult to adapt back to Indonesian due to greater structural and morphological differences. These findings emphasize that maintaining stability in cross-lingual representations remains a significant challenge, especially when the linguistic structures between the source and target languages differ substantially.

Cross-lingual representations such as MUSE are not optimal for language pairs that have considerable linguistic distance and limited parallel data. However, when tested with Indonesian, the performance of the model improved significantly, achieving the highest score among all language pairs. Interestingly, when the model trained with Spanish data was re-evaluated on Indonesian, its performance improved significantly. This suggests that the model adopts more robust linguistic patterns from the source language (Indonesian) but struggles to generalize them when transferred to a language with rich morphology like Spanish. This reinforces the importance of considering both linguistic distance and transfer direction in cross-lingual modeling.

B. Effectiveness of Embedding Alignment in Cross-Lingual Tasks

One of the main advantages of MUSE embeddings over traditional approaches is its ability to perform supervised bilingual alignment using high-quality dictionaries. Unlike unsupervised methods, this allows MUSE to project semantically similar words from different languages into a shared vector space without relying on large-scale parallel corpora. This makes MUSE particularly suitable for languages with limited resources, such as Indonesian, where parallel data is scarce. When compared to FastText, which operates solely in a monolingual space using subword representations, MUSE demonstrates significantly better performance in cross-lingual fake news classification. FastText is effective in handling morphologically rich languages but lacks cross-lingual semantic alignment, making it less suitable for multilingual transfer learning tasks.

The results show that MUSE consistently achieves higher F1-scores across all language pairs. Notably, the Indonesian-German (0.95–0.97) and Indonesian-English (0.93–0.95) pairs outperform the others, likely due to the availability of rich linguistic resources and effective bilingual mapping. Conversely, the Indonesian-Spanish pair recorded the lowest performance, highlighting the limitations of cross-lingual embeddings in handling linguistically distant language pairs with minimal training exposure.

These observations confirm that the success of cross-lingual models depends not only on model architecture but also on the quality of the underlying embedding alignment. These results emphasize the critical role of embedding strategies in enabling accurate multilingual NLP, especially when targeting languages with limited resources or diverse structures.

C. Advantages of the Proposed Hybrid Model

The results of this study show that the proposed hybrid model, which combines LSTM-CNN with MUSE embeddings, achieves competitive or even superior performance compared to previous studies in cross-lingual fake news detection. Specifically, experiments on the Indonesian-German and Indonesian-English language pairs yielded very high F1-scores (up to 0.97), demonstrating the model's strength in leveraging linguistic alignment and transfer learning. Unlike transformer-based architectures, this approach offers computational efficiency while maintaining robust performance on both similar and relatively distant language pairs.

However, the Indonesian-Spanish language pair poses significant challenges, likely due to substantial morphological and syntactic differences. This finding highlights the main limitation of embedding alignment strategies when applied to linguistically distant languages. To address this, future research could explore the integration of advanced multilingual embeddings, such as XLM-R or mBERT, and investigate fine-tuning techniques to enhance cross-lingual adaptability. Additionally, incorporating target language training data and evaluating models on informal or noisy text (e.g., social media content) could enhance their robustness in real-world applications.

VI. Conclusion

The evaluation results show that the best performance, based on F1-Score, is achieved by the LSTM model with MUSE embeddings, particularly for the Indonesian-German pair (0.97) and Indonesian-English pair (0.95). This confirms LSTM's effectiveness in capturing sequential dependencies, making it well-suited for cross-lingual tasks. The LSTM-CNN model produced comparable results, with its highest F1-Score of 0.95 on the Indonesian-German pair, but struggled in the Indonesian-Spanish pair, indicating that hybrid models may struggle to generalize effectively to structurally divergent languages. FastText embeddings performed significantly worse across all models and language pairs, with the Indonesian-Spanish pair only reaching an F1-Score of 0.50. This highlights the importance of multilingual embeddings like MUSE, which consistently delivered superior results. CNN performed competitively on Indonesian-German (0.97) but showed a decline in Indonesian-Spanish (0.70), indicating its limitations in capturing complex cross-linguistic relationships. These findings emphasize that model effectiveness depends on both linguistic characteristics and embedding selection. MUSE embeddings, combined with LSTM-based models, provide a lightweight yet effective alternative for cross-lingual fake news detection, making them suitable for real-world applications in low-resource language settings. Future research should explore the possibility of fine-tuning MUSE embeddings with additional target-language data to improve generalization across languages with highly divergent structures. Additionally, evaluating model robustness against adversarial attacks and noisy data, such as spelling variations, informal language, and manipulated text, would be valuable in assessing its real-world applicability. However, this study has several limitations. First, the scope of language pairs is limited to five, which may not fully represent global linguistic diversity. Second, the quality and balance of the dataset can affect model performance, especially for languages with limited resources. Additionally, the experiments were limited to structured news articles, excluding informal formats such as social media posts. These factors limit the generalizability of the findings. Future research could address these limitations by integrating more diverse datasets, languages, and text genres.

Acknowledgment

This research is part of the output of the Fundamental Research Grant Scheme, funded by Telkom University and the Ministry of Education, Science, and Technology, Indonesia with grant agreement number: 7925/LL4/PPG/2025, 121/LIT 07/PPM-LIT/2025.

The authors gratefully acknowledge the support and funding provided for the successful completion of this research.

References

[1] D. Dementieva, M. Kuimov, and A. Panchenko, "Multiverse: Multilingual Evidence for Fake News Detection," J Imaging, vol. 9, no. 4, p. 77, Mar. 2023, doi: 10.3390/jimaging9040077.

[2] S. Han, "Cross-lingual Transfer Learning for Fake News Detector in a Low-Resource Language," Wiley Journal, Aug. 2022, [Online]. Available: http://arxiv.org/abs/2208.12482

[3] K. Goswami, S. Dutta, H. Assem, T. Fransen, and J. P. McCrae, "Cross-lingual Sentence Embedding using Multi-Task Learning," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 9099–9113. doi: 10.18653/v1/2021.emnlp-main.716.

[4] S. K. Uppada, P. Patel, and S. B., "An image and text-based multimodal model for detecting fake news in OSN's," J Intell Inf Syst, Nov. 2022, doi: 10.1007/s10844-022-00764-y.

[5] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, Sep. 2017, doi: 10.1145/3137597.3137600.

[6] I. F. Putra and A. Purwarianti, "Improving Indonesian Text Classification Using Multilingual Language Model," 2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA), Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.05713

[7] S. O. Khairunnisa, Z. Chen, and M. Komachi, "Dataset Enhancement and Multilingual Transfer for Named Entity Recognition in the Indonesian Language," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 22, no. 6, pp. 1–21, Jun. 2023, doi: 10.1145/3592854.

[8] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent

Intrusion Detection System," IEEE Access, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.

[9] R. Mohawesh, S. Maqsood, and Q. Althebyan, "Multilingual deep learning framework for fake news detection using capsule neural network," J Intell Inf Syst, vol. 60, no. 3, pp. 655–671, Jun. 2023, doi: 10.1007/s10844-023-00788-y.

[10] D. A. Erku et al., "When fear and misinformation go viral: Pharmacists' role in deterring medication misinformation during the 'infodemic' surrounding COVID-19," Research in Social and Administrative Pharmacy, vol. 17, no. 1, pp. 1954–1963, Jan. 2021, doi: 10.1016/j.sapharm.2020.04.032.

[11] M. Tarczewska, A. Marciniak, and A. Giełczyk, "Fake or Real? The Novel Approach to Detecting Online Disinformation Based on Multi ML Classifiers," 2021, pp. 18–27. doi: 10.1007/978-3-030-77970-2_2.

[12] M. Choraś et al., "Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study," Appl Soft Comput, vol. 101, p. 107050, Mar. 2021, doi: 10.1016/j.asoc.2020.107050.

[13] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media," Sep. 2018, [Online]. Available: http://arxiv.org/abs/1809.01286

[14] Z. Guo, Q. Zhang, F. Ding, X. Zhu, and K. Yu, "A Novel Fake News Detection Model for Context of Mixed Languages Through Multiscale Transformer," IEEE Trans Comput Soc Syst, vol. 11, no. 4, pp. 5079–5089, Aug. 2024, doi: 10.1109/TCSS.2023.3298480.

[15] A. De, D. Bandyopadhyay, B. Gain, and A. Ekbal, "A Transformer-Based Approach to Multilingual Fake News Detection in Low-Resource Languages," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 21, no. 1, pp. 1–20, Jan. 2022, doi: 10.1145/3472619.

[16] M. Arif et al., "CIC at CheckThat! 2022: Multi-class and Cross-lingual Fake News Detection," CEUR Workshop Proc, vol. 3180, pp. 434–443, 2022.

[17] S. Wu and M. Dredze, "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 833–844. doi: 10.18653/v1/D19-1077.

[18] C. Dulhanty, J. L. Deglint, I. Ben Daya, and A. Wong, "Taking a Stance on Fake News: Towards Automatic Disinformation Assessment via Deep Bidirectional Transformer Language Models for Stance Detection," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.11951

[19] N. Mohd, H. Singhdev, and D. Upadhyay, "TEXT CLASSFICATION USING CNN AND CNN-LSTM," Webology, vol. 18, no. 4, 2021, doi: 10.29121/WEB/V18I4/149.

[20] M. Mohtarami, R. Baly, J. Glass, P. Nakov, L. Marquez, and A. Moschitti, "Automatic Stance Detection Using End-to-End Memory Networks," Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.07581

[21] M. G. Muharam, "Indonesia False News(Hoax) Dataset," Kaggle. Accessed: Apr. 24, 2024. [Online]. Available: https://www.kaggle.com/datasets/muhammadghazimuharam/indonesiafalsenews/data

[22] Pierobeat, "Fake News (Hoax) classification Using XGBoost," GitHub. Accessed: Apr. 24, 2024. [Online]. Available: https://github.com/pierobeat/Hoax-News-Classification/?tab=readme-ov-file

[23] I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, "Study of hoax news detection using naïve bayes classifier in Indonesian language," in 2017 11th International Conference on Information & Communication Technology and System (ICTS), IEEE, Oct. 2017, pp. 73–78. doi: 10.1109/ICTS.2017.8265649.

[24] H. Zolkepli, "Malay-Dataset.," GitHub. Accessed: Jun. 27, 2024. [Online]. Available: https://github.com/mesolitica/malaysian-dataset

[25] J. Risch, A. Stoll, L. Wilms, and M. Wiegand, "Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments," in Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments, Association for Computational Linguistics, 2021, pp. 1–12. Accessed: Jun. 27, 2024. [Online]. Available: https://aclanthology.org/2021.germeval-1.1

[26] J. O. Vizoso, "Spanish Political Fake News," Kaggle. Accessed: Aug. 15, 2024. [Online]. Available: https://www.kaggle.com/datasets/javieroterovizoso/spanish-political-fake-news/data

[27] L. Cui and D. Lee, "CoAID: COVID-19 Healthcare Misinformation Dataset," May 2020.

[28] A. Librian, M. Krishnan, A. Putra, R. Fadlillah, and P. Adi, "Sastrawi: Sistem Stemming Bahasa Indonesia," GitHub. Accessed: Nov. 05, 2024. [Online]. Available: https://github.com/sastrawi/sastrawi

[29] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python, 1st ed. O'Reilly Media, Inc., 2009.

[30] S. Ruder, I. Vulić, and A. Søgaard, "A Survey of Cross-lingual Word Embedding Models," Journal of Artificial Intelligence Research, vol. 65, pp. 569–631, Aug. 2019, doi: 10.1613/jair.1.11640.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Jan. 2013, [Online]. Available: http://arxiv.org/abs/1301.3781

[32] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[33] S. Sannigrahi and J. Read, "Isomorphic Cross-lingual Embeddings for Low-Resource Languages," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.14632

[34] R. Wijayanti, M. L. Khodra, K. Surendro, and D. H. Widyantoro, "Learning bilingual word embedding for automatic text summarization in low resource language," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 4, pp. 224–235, Apr. 2023, doi: 10.1016/j.jksuci.2023.03.015.

[35] G. Lample and A. Conneau, "Cross-lingual Language Model Pretraining," Jan. 2019, [Online]. Available: http://arxiv.org/abs/1901.07291

[36] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in 2017 International Conference on Engineering and Technology (ICET), IEEE, Aug. 2017, pp. 1–6. doi: 10.1109/ICEngTechnol.2017.8308186.

[37] J. Zhang, Y. Li, J. Tian, and T. Li, "LSTM-CNN Hybrid Model for Text Classification," in 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), IEEE, Oct. 2018, pp. 1675–1680. doi: 10.1109/IAEAC.2018.8577620.

[38] Z. Wang, J. Xie, R. Xu, Y. Yang, G. Neubig, and J. Carbonell, "Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.04708

[39] D. Van Thin, H. Quoc Ngo, D. Ngoc Hao, and N. Luu-Thuy Nguyen, "Exploring zero-shot and joint training cross-lingual strategies for aspect-based sentiment analysis based on contextualized multilingual language models," Journal of Information and Telecommunication, vol. 7, no. 2, pp. 121–143, Apr. 2023, doi: 10.1080/24751839.2023.2173843.

[40] Y. Cao et al., "Joint Representation Learning of Cross-lingual Words and Entities via Attentive Distant Supervision," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 227–237. doi: 10.18653/v1/D18-1021.

[41] N. van der Heijden, S. Abnar, and E. Shutova, "A Comparison of Architectures and Pretraining Methods for Contextualized Multilingual Word Embeddings," Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.10169.

A Preliminary Study on Songket: A Preservation of Intangible Cultural Heritage

Nik Siti Fatima Nik Mat[1], Syadiah Nor Wan Shamsuddin[2], Syarilla Iryani Ahmad Saany[3], Norkhairani Abdul Rawi[4], Julaily Aida Jusoh[5], Wan Malini Wan Isa[6], Addy Putra Md Zulkifli[7], Shahrul Anuwar Mohamed Yusof[8]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia[1, 2, 3, 4, 5, 6]
Faculty of Innovative Design and Technology, Universiti Sultan Zainal Abidin, Terengganu, Malaysia[7, 8]

***Abstract*—Songket is a traditional Malaysian woven fabric, opulent, and symbolizes luxurious classical textiles of the old craft in Malaysia. Songket is part of the intangible cultural heritage to this day. However, cultural heritage preservation arises as a critical endeavor, especially for younger generations. There is a growing concern about the disinheritance of cultural heritage to posterity due to disregard for the value of our own heritage. Therefore, this paper aims to explore in-depth insight into the beautiful heritage of songket weaving including the knowledge of art weaving, delve into the techniques and materials used as well as the legacy and time ahead. The research uses observation methods and deep face-to-face interviews for data collection. It was conducted with experienced experts and workers, revealing the origin of Songket and its processing from scratch. The research data findings were gathered and subsequently analyzed as secondary data. This study gives more knowledge about Malay traditional textile art and heritage, indicating a positive future of traditional weaving arts as well.**

Keywords*—Songket; weaving; textiles; aesthetic; intangible cultural heritage; processing songket; cultural; heritage; Terengganu; Malaysia

I. Introduction

Intangible cultural heritage is one of the pillars of heritage components. According to UNESCO cultural heritage can be distributed into three categories: tangible, intangible, and natural. However, compared to others, intangible cultural heritage is different due to its nature which is unseen with the naked eye and cannot be touched nevertheless, its existence can be felt and impressed [1]. The inheritance of intangible cultural heritage similar to diverse technologies features of some culturally distinct communities from the past generations, sustained this time and carefully gathered and encouraged for the benefit of future generations [9]. Since the early stages of history textiles have become a very conspicuous material culture in the Malay world [2].

There are many examples of intangible cultural heritage including language, customs, lifestyle, music, art, and dance as well as weaving songket in textiles brocade family. Songket is a fine art of traditional weaving and the most opulent, ancient textiles in the Malays' repertoire. It has long been an integral part of Malaysia's cultural heritage [10]. The unique traditional weaving features reflect Malaysian people's identity and creativity. The most exquisite and interpreted as an everlasting fabric worn, the textile is acclaimed as "the cloth of gold" or "the Queen of fabrics" [9]. It is also referred to as "kain benang emas" (golden threading fabric) and is used to make apparel accessories [14].

Songket is an inseparable part of cultural heritage and a crucial ingredient of identity and society formation. As a symbol of the nation, songket heritage influences shaping perceptions of the country's culture, image, and personality. Songket embodies the intellectual values of past societies, inherited and bridged between generations. Besides representing a multi-racial emblem, songket heritage is very synonymous with tourism-related products that very crucial to drive economic growth [7].

In Malaysia, the government treats heritage issues with the utmost importance including songket weaving industry. As listed in UNESCO's Intangible Cultural Heritage of Humanity, his prompt recognition is expected to facilitate bringing other attractive Malaysian artistic traditions to the world stage [12]. Most meaningful yet, the Yayasan Tuanku Nur Zahirah (YTNZ) foundation has chosen to direct its endeavor on songket, and reinvigorate the songket industry including empowering the necessary skills.

Gradually as time goes on, the art of weaving songket encounters numerous challenges in remaining relevant. The challenges encompass many factors such as the rapid advancement of contemporary life, technological breakthrough as well as industrial development [5]. Youth nowadays are occupied with gadgets, reluctant to learn and understand basic knowledge of their cultural heritage. Thus, a collaborative effort is indeed in demand to preserve this intangible heritage to remain sustainable for the future.

This paper is divided into five main sections, which are: Section II discusses the related work and background of cultural heritage preservation. Conversely, Section III discusses the method used in this study to gain information. Next, Section IV presents the findings and discussion of the study, and lastly, Section V presents the conclusion.

II. Related Work

A. Cultural Heritage

Cultural heritage is an important constituent of a community's identity, and continuity, fostering a sense of belonging and pride of a nation [8]. Furthermore, it is a valuable asset that has been or is owned by a person or a group of societies or people who altogether share responsibilities for conservation and reservation as well [5]. Besides it is a legacy of the ways of living from the past, what we live with today, and what we pass

on to future generations as well. Heritage ratifies high values in a society while culture contributes to the sustainable development of the society and its economy. Heritage preservation, be it tangible, intangible, or natural had to be listed to be recognized as world heritage resources and were also brought in the World Heritage Convention, which took place in 1972 [1]. UNESCO has been listing some of Malaysia's intangible heritage including Mak Yong theatre, pantun, silat, Dongdang Sayang as well as weaving songket. The cultural heritage information must be gathered in an effort to deliver for next future generations. The expenditure is very important whereby the information from the past can be used to determine the identity of each society nowadays.

B. Songket

Songket is a Malay word that derives from the process of 'sungkit' [11] which brings meaning to the art of out or pulling a thread from a background cloth or weaving using gold and silver thread as well [6]. The most ancient and opulent textiles in the Malays' repertoire, songket was enlightened under the auspices of the king before the 19th century [14]. The eastern coast of the Malay peninsula has most of the greatest weavers Malay tradition that are centred on Pahang, Terengganu, and Kelantan [2]. The Songket is equivalent with the Malay community whereby making songket is a legacy from the past that is still widely practiced.

Songket has its own meaning. On this point, Malay weaver does not simply produce fabric and each part of the Songket fabric has a unique name. For instance, the design of Songket consists of a few parts such as a body, head, fabric edge, and clamp. Fig. 1 below shows the detail of songket fabric [3]. The entire motif is taken and inspired by what has existed around them including flora, nature, and abstract pattern.

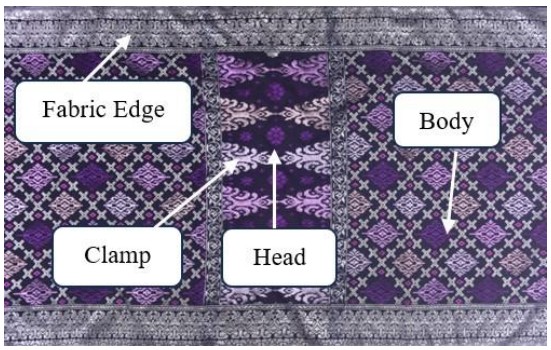


Fig. 1. The detail of songket fabric.

C. Cultural Heritage Preservation

Cultural heritage preservation is vital to ensure that these elements are conveyed to younger generations, preserved, conversed, and transmittable for sustainable cultural heritage as well as contributing to an appreciation of identity. Intangible cultural heritage is facing an awful situation globally due to rapid modernization and urbanization that is devoted to the idleness of it [4]. Apart from that, intangible cultural heritage holds significant ancestry that needs to be preserved sustainably [13].

Malaysia is well known for its diversity of cultures and valuable heritage. The government addressed the heritage issues earnestly and in accordance with that proactive steps to preserve the cultural heritage were taken, including the handicraft industry. Perbadanan Kemajuan Kraftangan Malaysia is an agency under the Ministry of Tourism, Arts, and Culture established under Act 222, the Malaysian Crafts Development Corporation Act (1979) whereby the preservation was implemented with the intention of commercializing craft products through market, product, and entrepreneur development.

III. Method

Fulfilling the research objectives required a well-structured methodological framework, consisting of three key phases: preparation, data collection, and data analysis. This study employed a qualitative approach through observation and semi-structured interviews to examine the cultural and aesthetic elements of Terengganu Songket. Emphasis was placed on understanding the production process, tools and techniques involved, and efforts in preserving this intangible cultural heritage through expert insights.

During the preparation phase, the study objectives were clearly outlined, and a set of interview questions was developed based on relevant literature. These questions covered aspects such as the history of songket, production methods, equipment used, and the future of the tradition. Data collection took place at the Perbadanan Kemajuan Kraftangan Malaysia, Terengganu branch, a center with more than 25 years of experience in the songket industry. Additional insights were gathered from a traditional songket business in Kampung Bukit Depu, where weaving has been sustained through seven generations, with the third generation currently managing the enterprise.

In the data collection phase, semi-structured interviews were conducted using open-ended questions to elicit in-depth responses. All sessions were recorded to ensure accurate transcription and analysis. Thematic analysis was then applied to identify, interpret, and report recurring themes within the qualitative data.

IV. Findings and Discussion

A. Findings on Songket Knowledge

Based on an interview session conducted with experienced people from Perbadanan Kraftangan Malaysia Cawangan Terengganu, songket is a "Queen of Fabric" that is produced manually by golden-agers. The making process takes months to complete a piece of songket cloth depending on the material and motif provided. There are some experts that were involved in making a songket according to their respective expertise. The wage rate charged also varies each other depends on the skill available. For instance, the experienced people in dyeing the yarn and each process need to be thoroughly step by step. Also the process of untangling, winding, and rolling the yarn could be the same person but a weaver has the expertise to weave only.

Based on study [2], the eastern coast of the Malay peninsula has many Malay traditional weavers which is Terengganu and Kelantan has become the main center of Malay traditional textile weavers. Songket is the valuable heritage of Malay people especially in Terengganu and the history is started in the palace for Malay royal ceremonial costumes. A wealth of precious

heritage will disappear once in a while if the younger generations are unaware of its importance and immersed with their own heritage.

B. Findings on Songket Producing

Songket is one of the old valuable traditional Malay crafts which is from a yarn into a beautiful fabric. The opulent Songket employs a weaving method that comprises interweaving gold threads with silk threads on the fabric base. As it identical with the Malay community, making songket is a legacy in Terengganu that is passed down from generation to generation. The process of producing songket involved several steps as elaborated below.

C. The Process of Making a Songket

There are eight processes in the making of songket. In order to ensure produced high-quality products each process needs to be comprehensive and detailed way. Before all the process begins, it starts with planning the design of the songket pattern.

1) Dyeing the cotton or silk yarns (thread) in hanks: Woven yarns consist of cotton yarns, silk yarns, gold yarns, and silver yarns. There is no need to color the gold and silver yarn due to natural ingredients. The cotton or silk yarns (thread) need to be tied neatly, inserted into boiling water in the crater and thoroughly rinsed with water. After that, the yarn dipped into the dye and dried up. Fig. 2. shows the process of dyeing the yarn.



Fig. 2. Process of dyeing the yarn.

2) For untangling the yarn: The dried yarn spins with a small bamboo-made spindle or innovation tools designed using a bicycle wheel. This process includes a "Darwin" tool and a convenient spinning tool. Darwin feed and gold Darwin are two types of sharks. Darwin feeds are made of bamboo and ropes or rattan and used only for feed yarns and Darwin gold is made of wood or wire because of the rough and heavy gold thread and used only for the gold thread (see Fig. 3).

3) Winding the yarn: The process of stretching and winding the yarn on the bobbins or Terengganu is called "peleting". The length or the number of threads of the fabric to be woven is regulated by "loseng" (loom). The prevalence rolls of feed yarn or sockets measure 26 to 31 meters and can produce twelve to fourteen pieces of fabric. (see Fig. 4).

4) Rolling the yarn: This process is to wind the warp yarns onto a warp wooden board. To ensure smooth arrangement during rolling, the yarn position should be checked every three times as Fig. 5.



Fig. 3. Process of untangling the yarn.

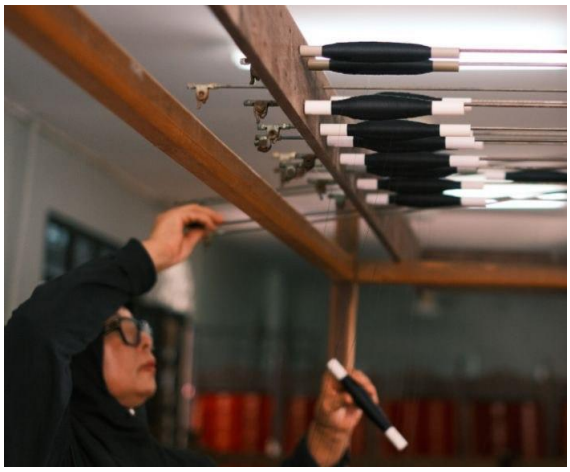


Fig. 4. Process of winding the yarn.



Fig. 5. Process of rolling the yarn.

5) Spooling the yarn: The next process is spooling the yarns through the reed. The spooling work is implemented after the

warping threads are inserted into the teeth or brush of the machine. Each brush hole is stained with two veins of the lotus thread and hooked through each gap in the machine's teeth. This process is done in the “cake” (see Fig. 6).



Fig. 6. Process of spooling the yarn.

6) Stretching the yarn: This process is making the frames for the string heddles and threading the yarns stripped from “karak” to wood by shaking them (see Fig. 7).



Fig. 7. Process of stretching the yarn.

7) Lifting the yarn: The process of weaving songket by creating patterns on the warp threads using a skewer tool. Songket weaving is usually woven with three-pointed or five-pointed techniques, tying them later through the button-tying process (see Fig. 8).



Fig. 8. Process of lifting the yarn.

8) Weaving the yarn: The weaving process uses a tool called "cake" whereby divided into three parts that are the cake legs, the cake body, and the cake head. Commonly after a thread of gold is ripped, it is beaten (beat) and then followed by two threads of feed. The shuttle filled with the weft thread or gold thread, is inserted left and right into the gaps between the warp threads, and weavers will convert button threads by rotation and predetermined patterns. This process is repeated until it becomes a piece of songket cloth (see Fig. 9).



Fig. 9. Process of weaving the yarn.

D. Songket Pattern and Motif

Songket patterns are filled with different motifs that idealize the whole songket fabrics. The are several classifications patterns of songket including full pattern, scattered pattern, scattered repeated bricks pattern, triangle shape pattern, alternating scattered repetitions pattern, crosswise pattern, stripes pattern, chevron or zigzag pattern as well as checkers pattern [5].

The following motifs are used in dissimilar patterns, depending on where they are placed and reproduced. Songket motif is generally used to demonstrate the Malay community's traditional ways of life that provide a means of exploring the relationship between cultural and national identity. Based on [14], it also represents the most important Malay cultural identity delivered in the creation of its motifs through the practice of oral traditions. Each motif and pattern has its own meaning whereby present the unique identity of Malay itself [3]. Usually, the motif and pattern used in Songket are produced at the customer’s request.

Based on a previous study by [3], the overall motif songket is taken and inspired by surroundings such as flora, fauna, nature, and abstract patterns. For instances pucuk rebung, bunga lawi ayam, bunga tanjung, bunga baling, bunga cermai, and tapak kecupu for flora motif. Besides that, motifs based on abstract patterns such as jong sarat, pecah lapan, bunga tiga dara, tapak kacip, bunga kerongsang, kasap, bunga logam, potong wajik, and bunga logam. The motifs of bintang buta, awan larat, air molek, and teluk berantai as nature inspirations. Another motif is inspired by fauna such as sisik kelah, unduk-unduk, bunga lawi ayam, and bunga hati [6].

A preliminary study has been conducted to initiate detailed information regarding the aesthetic of songket. The exploratory of important and valuable information gathered from the insights of experst and entrepreneusr located at Terengganu. This section discussed the results based on the data acquired from the interviews. By conducting a preliminary study, an initial investigation was implemented to identify the direction and narrow down the study.

E. Tools and Equipment of Songket

Based on observation, there are two categories of tools and equipment in making songket that are weaving equipment and supporting equipment. Among the tools used to create songket are as Table I:

TABLE I. TOOLS EQUIPMENT FOR PRODUCING SONGKET

Tools Equipment	Name
	Weaving Cake or "kei" is a tool used in the weaving process that is divided into three parts that are the cake legs, the cake body, and the cake head and interconnected each other.
	*Bobbin* also known as "pleting" is a crucial tool used to hold the wraps threads.
	A combination of threads will produce a distinctive and luxurious songket appearance.
	*Loseng* is the thread that formis the fabrics and refers to threads that extend on the loom.
	*Two Belira* is a long and thin stick that is used to loosen the thread and make in tense. This process facilitate the process of inserting a c*uban*.

F. Legacy and Future of Songket

Based on the interviews with the songket entrepreneurs, the legacy of the songket business in Terengganu is inherited by the 7th younger generation. The older generation working in this business for generations ago and will assign their business knowledge to their heirs so that the business remains relevant.

The persistence of the cultural heritage is left by the previous generation reliant on the new generation. The weaving process requires a lot of patience, passion and creativity so that to passing the knowledge is very challenging.

Terengganu tourism is synonymous with the songket industry and making songket is a legacy even more the songket has emerged to become a tourism product. To guarantee that the industry is consistently progressing and growing, entrepreneurs should always look for opportunities to expand their business, sharing the knowledge all about songket and collaborating with government so that the value of songket is highlighted and this legacy is never lost.

In the modern world, the future of songket weaving can be promising with all-out effort. Therefore, innovation in art weaving is a critical way to bridge the gap between tradition and modernity including providing a digital learning platform. As a global recognition, it is our shared responsibility to protect, preserve, and pass to the next generation.

V. CONCLUSION

This preliminary survey gives an overview of the current conditions of cultural heritage especially on songket weaving. Its legacy that symbolizes the identity of our community and generation nowadays is responsible for preserving and safeguarding our country's cultural heritage. Furthermore, songket has worldwide cultural potential, and in-depth research is very crucial.

As a conclusion, to ensure that the industry is establishing constantly, evolving, and growing requires further efforts on all fronts. As society moves forward into a progressively global and technologically advanced future, it is essential to appreciate the value of textile culture and his historical heritage. Therefore, to preserve this intangible heritage to remain sustainable the responsible parties must make a concerted effort that includes a mutual support between songket weaver and entrepreneur as well as a collaboration among all parties and promote Malay Songket Weaving internationally.

ACKNOWLEDGMENT

This research was financially supported by Center for Research Excellence and Incubation Management (CREIM) under Dana Penyelidikan Universiti (DPU/RD080) Universiti Sultan Zainal Abidin.

REFERENCES

[1] Juanis. B., Saleh. Y., Ghazali M.K.A, Mahat H., Hashim M., Nayan N., Hayati & Kurnia R. (2022). Knowledge, Attitudes and Practices of Youths Towards the Intangible Cultural Heritage Elements of Dusun Ethnic in Malaysian Environment. *IOP Conference Series:* Earth and Environmental Science, 975(1).

[2] Mohd Jamil Muhamad Amin, Arba'iyah Ab. Aziz1, Nor Azlin Hamidon, Nabilah Mudzafar & Inda Murni Hairul Anuar (2024). The Aesthetics of Striped Design Motifs and Market Demand on the Songket Tenggarung as the Malay Royal Ceremonial Attire. *Research in Management of Technology and Business, Vol. 5 No. 2 (2024) 61-67*

[3] Nasaie Zainuddin, Asliza Aris, Najua Tulos, Muhammad Hisyam Zakaria, and Nor Idayu Ibrahim (2021). The Aesthetic of Bridal Songket in Malay Traditional Wedding Attire. *International Journal of Art & Design, vol. 4(5), pp. 41–46*.

[4] Noorfathehah Abdullah Sani and Madzlan Abet (2021). The Potential of Developing a Heritage Village to Safeguard Intangible Cultural Heritage from the Perspectives of Stakeholders at Pura Tanjung Sabtu, Terengganu. *Proceedings of the First International Conference on Science, Technology, Engineering and Industrial Revolution volume 536 (ICSTEIR 2020),pp 486-493*.

[5] Nor Hafidzah Abdullah, Wan Malini Wan Isa, Syadiah Nor Wan Shamsuddin, Norkhairani Abdul Rawi, Maizan Mat Amin, Wan Mohd Adzim Wan Mohd Zain (2024). Towards Digital Preservation of Cultural Heritage: Exploring Serious Games for Songket Tradition. *International Journal of Advanced Computer Science and Applications, Vol. 15, No. 3,pp 209-215*.

[6] Nurul 'Ayn Ahmad Sayuti, Muhamad Aiman Afiq Mohd Noor, Mohd Yusri Bin Mohd Naser and Sommer B (2024). A Study on Songket Motifs, Familiarity, Preferences, and the Potential Application of Generative Ai to Create Innovative Cultural Product Designs. *Alam Cipta, 17(2), pp 74-91.*

[7] Peri´c B.S., Šimundi´c B., Muštra V. and Vugdelija M. (2021). The Role of UNESCO Cultural Heritage and Cultural Sector in Tourism Development: The Case of EU Countries. Sustain., vol. 13, no. 10, 2021, doi: 10.3390/su13105473.

[8] Saket Bihari. 2023. Cultural Heritage and Indigenous Knowledge: Reviving Traditions for Future Generations. *Sustainable Development Goals in SAARC Countries: Key Issues, Opportunities and Challenging, Vol I, pp* 24-32.

[9] Sharifah Tahirah Syed Kamarulzaman, Basitah Taif, Muhammad Ismail Ab Kadir, Rajabi Abdul Razak (2022). A Comparison on Physical Degradations and Identification of Material between Two Malay Songket Artefacts. *Environment-Behaviour Proceedings Journal, 7(SI7), pp.119-127.*

[10] Siti Zuliana Salleh, A.R. Bushroa (2022). Bibliometric analysis on the implementation digitization technologies in cultural heritage in Malaysia. *Malaysian Journal of Society and Space. Vol 18920, pp 287-300.*

[11] Suhaile, A., & Che, B. (2016). Tenunan Songket Di Kampung Bukit

[12] Teh A. L. S. - January 17, 2022) Weaving through centuries of heritage https://www.nst.com.my/news/nation/2022/01/763619/weaving-through-centuries-heritage new strait times

[13] WN Jazmina W. Ariffin, Syed Shahfiq, Fazil Ahmad, Asmawi Ibrahim, & Farah Syazrah Ghazalli (2023). Handicraft Innovations: A Strategic Approach to Preserving Intangible Cultural Heritage of Malaysia. *Journal of the International Society for the Study of Vernacular Settlements (ISVS e-journal), Vol 10 (7), pp 137-146.*

[14] Wan Samiati Andriana Wan Mohamad Daud, Mursyidah Zainal Abidin (2022). Using Motif Songket to Creating Geometric Islamic Artwork. *Environment-Behaviour Proceedings Journal, 7(SI 9), pp.503-507*

Utilizing Machine Learning to Identify High-Risk Groups in Sickle Cell Anemia

A Single-Center Experience in Saudi Arabia

Haneen Banjar[1], Nofe Alganmi[2], Hajar Alharbi[3], Ahmed Barefah[4], Hatem Alahwal[5], Salwa Alnajjar[6], Abdulrahman Alboog[7], Salem Bahashwan[8], Galila Zaher[9]

Department of Computer Science-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah 21589, Saudi Arabia[1, 2, 3]
The Center of Research Excellence in Artificial Intelligence and Data Science,
King Abdulaziz University, Jeddah 21589, Saudi Arabia[1, 2]
Centre of Artificial Intelligence in Precision Medicines, King Abdulaziz University, Jeddah 21589, Saudi Arabia[1, 2]
Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia[1, 2]
Faculty of Medicine, King Abdulaziz University, Jeddah 21589, Saudi Arabia[4, 5, 6, 8, 9]
Faculty of Medicine, University of Jeddah, Jeddah, Saudi Arabia[7]

***Abstract*—Sickle Cell Anemia (SCA) is a hereditary condition causing abnormal red blood cells, leading to severe health complications. Traditional treatment approaches for SCD often involve reactive management, which can delay appropriate interventions and worsen patient outcomes. The aim of this study is to leverage machine learning (ML) algorithms, including Logistic Regression (LR), Support Vector Machines (SVM), and Decision Trees (DT), to identify high-risk groups among SCA patients using clinical and pathological data from King Abdulaziz University Hospital. This study employs a comprehensive dataset comprising 200 SCA patients, with data preprocessing to handle missing values and feature selection techniques to enhance model performance. The dataset is divided into training and testing sets, and models are evaluated using ten-fold cross-validation. Performance metrics such as True Positive Rate (TPR), False Negative Rate (FNR), Positive Predictive Value (PPV), and False Discovery Rate (FDR) are used to assess model effectiveness. The results indicate that the SVM model with the top seven correlated features achieved the highest TPR and PPV, along with the lowest FNR and FDR, demonstrating its superior performance in identifying high-risk patients. The study concludes that ML models, particularly SVM, can significantly improve risk assessment and patient management in SCA, offering a proactive tool for healthcare providers. The main message is the potential of ML algorithms to enhance clinical decision-making and improve outcomes for patients with SCA.**

Keywords*—Sickle cells anemia; feature selection; predicting complication; machine learning

I. Introduction

SCD is a hereditary condition that results in the production of abnormal red blood cells. The disease was first identified in 1910 [2]. Today, it is estimated that approximately 5% of the global population carries a gene for sickle cell disease or thalassemia [1]. Annually, over 300,000 newborns are affected by severe forms of these disorders, with most cases occurring in low- and middle-income countries [3], [4]. SCD is particularly prevalent in regions such as the Mediterranean, Africa, India, the Caribbean, and the Middle East.

In Saudi Arabia [5], SCD is a common genetic disorder, with carrier rates varying from 1.4% to 2% in certain areas and reaching as high as 27% in some regions. The Saudi premarital screening program has revealed that 0.26% of the adult population is affected by SCD, while 4.2% carry the sickle cell trait. The highest prevalence is observed in the Eastern province, with SCD affecting about 1.2% of the population and sickle cell trait carriers comprising approximately 17%. Current treatment decision-making for SCD involves monitoring symptoms and complications and adjusting treatment plans accordingly. This reactive approach can significantly impact patients' lives, leading to delayed treatment options and irreversible disease complications. Despite extensive epidemiological surveillance, no publicly reported model has yet been validated on a Saudi cohort to predict severe events; this gap leaves clinicians reliant on qualitative heuristics and underscores the need for data-driven, localised risk stratification.

Sickle-cell complications place a persistent burden on emergency and hematology services in Saudi Arabia; delayed therapeutic escalation is common because current risk assessments rely on clinician intuition or post-event markers. By embedding a data-driven prognostic layer into routine complete-blood-count (CBC) workflows, clinicians can transition from reactive to proactive care. Applying artificial intelligence (AI) tools, particularly Machine Learning (ML) algorithms, to predict the severity and complications of SCD in diagnosed patients, offers a proactive solution that could assist physicians in making more informed treatment decisions. ML techniques can analyze vast amounts of clinical and pathological data to identify patterns and risk factors that are not easily discernible through traditional methods. Consequently, the objective of this study is to leverage ML algorithms, including Logistic Regression, Support Vector Machines, and Decision Trees, to identify high-risk groups among SCD patients using data from King Abdulaziz University Hospital. By developing predictive models, we aim to enhance the accuracy of risk assessments and improve patient management. Directly addresses a national patient-safety gap and aligns with Vision 2030's precision-

medicine goals. Research Question (RQ): Which of the evaluated machine-learning algorithms (Logistic Regression, Support Vector Machine, or Decision Tree) most accurately identifies high-risk SCA patients using routinely collected hematological features in a Saudi clinical setting? The remainder of this paper is organised as follows: Section II details the dataset and pre-processing pipeline. It describes the three candidate algorithms and feature-selection procedure; Section III sets out the experimental protocol. It presents quantitative results; Section IV discusses clinical implications and algorithm–data suitability; finally, Section V concludes with limitations and future work.

A. Related Work

SCD is a complex genetic disorder characterized by significant phenotypic variability [6]. The development of accurate prediction models for SCD severity and outcomes is crucial for improving patient care and treatment decision-making [7]. ML techniques have shown promise in interpreting medical data and predicting disease severity, complications, and treatment dosages [7]. Bayesian network modeling has been used to create a personalized disease severity score that predicts the risk of death within 5 years, incorporating factors such as renal insufficiency, leukocytosis, and hemolytic anemia severity [8]. While known modifiers like fetal hemoglobin levels and α-thalassemia influence disease severity, other genetic variants and environmental factors, including climate and air quality, may also play a role [6]. Despite recent successes, modeling the multisystem pathology of SCD remains challenging, and further research is needed to improve prediction of specific adverse outcomes and global disease severity [9].

Research on predicting disease severity in SCD has focused on various factors and approaches. A systematic review identified multiple indices used to assess SCD severity, incorporating elements like organ damage and complications, but found a lack of consensus and validation for these measures [10]. To address this, experts developed a severity classification system using a modified Delphi panel, considering factors such as age, genotype, and organ damage [11]. ML techniques have shown promise in predicting acute organ failure in critically ill SCD patients, using physiological markers derived from vital signs [12]. Despite these advancements, modelling the variable and multisystem pathology of SCD remains challenging. While there have been some successes in predicting specific adverse outcomes and global disease severity, significant challenges persist in developing comprehensive prognostic tools for SCD [9].

Feature selection techniques are crucial for improving model performance in clinical prediction by identifying relevant features from large datasets [13]. Common approaches include filter and wrapper methods [14]. Filter methods use criteria like correlation, significance, or variable importance to rank features, while wrapper methods employ algorithms to evaluate feature subsets [13]. Hybrid approaches combining filter and wrapper techniques have shown promise in generating minimal redundancy maximal relevant feature subsets [15]. A tri-stage wrapper-filter framework has been proposed, utilizing ensemble filter methods, correlation analysis, and meta-heuristic optimization to select optimal features for disease classification [16]. These techniques aim to reduce dimensionality, remove irrelevant or redundant data, and improve classification accuracy [16], [14]. Effective feature selection can lead to more parsimonious models, reduced training time, and improved predictive performance in clinical settings [13], [16].

ML algorithms have shown significant potential in improving healthcare outcomes for patients with SCD. Studies have demonstrated that ML models outperform traditional readmission risk scoring systems in predicting 30-day hospital readmissions for SCD patients [17]. Random Forest algorithms have been particularly effective, achieving higher accuracy and area under the curve (AUC) scores compared to standard [17]. ML techniques have also been successful in predicting SCD from erythrocyte smears, with ensemble models like Random Forest-XGBoost showing superior performance [18]. The integration of ML in healthcare extends beyond SCD, offering opportunities for personalized treatment and improved resource allocation across various medical conditions [19]. These advancements in ML-driven predictive models have the potential to revolutionize healthcare systems, leading to more efficient and patient-centered care.

II. METHODS

A. Patient Population

The 200 patients from King Abdulaziz University Hospital (KAUH) were eligible for the study, and any patients who were not diagnosed with SCA were excluded. The dataset retrieved data from electronic health records, including each patient's complications records and clinical and pathological data. The data were accessed for research purposes from (1/10/2020). All patients' identities were anonymized, and authors did not have access to information that could identify individual participants during or after data collection. Table I shows the features that were included in the study. Ethical approval was obtained from the unit of the Biomedical Ethics Research Committee at KAUH (Reference No. 511-20). As participants' data were retrospectively included from medical records, the need for informed consent was waived by the ethics committee, as the data were fully anonymized before access.

B. Patients Outcomes

The patients enrolled in the study were divided into two broad outcome groups: (i) Low-risk group: SCA patients who did not develop any complications; and (ii) high-risk group: SCA patients who developed complications. The complications include: Vaso-Occlusive Crisis (VOC), Hepatorenal, gallstone, Stroke, Osteotomy, Avascular Necrosis (AVN), Chronic Kidney Disease (CKD), acute tubular necrosis, Upper Respiratory Tract Infection (URTI), cute Chest Syndrome (ACS), pulmonary Embolism (PE), Pneumonia, Cholecystitis, and Priapism).

C. Datasets

The dataset used in this study comprises three distinct sheets: "ContWithMissing", "ContNoMissing," and "CatWithNoMissing". Each sheet presents a different aspect of the data, catering to specific requirements for ML model development and evaluation. Firstly, the "ContWithMissing" sheet contains continuous data with some values missing, indicating a need for imputation or exclusion during preprocessing. This sheet is crucial for understanding data

completeness and planning subsequent data cleaning steps. Secondly, the "ContNoMissing" sheet provides continuous data with no missing values. This dataset guarantees that all continuous data inputs are available for analysis. Thirdly, the "CatNoMissing" sheet contains categorical data with all missing values addressed. The number of rows reduced compared to the original dataset because some Thalassemia patients were excluded. This sheet ensures that categorical data inputs are fully available for analysis and modeling.

TABLE I. CLINICAL AND PATHOLOGICAL FEATURES

Features	Description
Age	Clinical feature recorded at the time of diagnosis.
Gender	Clinical feature recorded at the time of diagnosis.
Hemoglobin level (Hb)	An internal protein within red blood cells that serves as a transport system for oxygen from the lungs to the tissues and organs of the body, and a transport system for carbon dioxide from the body to the lungs.
White Blood Cell (WBC)	A cell in the blood and lymph is produced in the bone marrow. White blood cells are a component of the body's immune system. They assist the body in fighting off illness and disease. Granulocytes (neutrophils, eosinophils, and basophils), monocytes, and lymphocytes are the different kinds of WBCs (T cells and B cells).
Platelets	Large cells in the bone marrow known as megakaryocytes form platelets. They assist in forming blood clots that can slow or halt bleeding and assist in wound healing.
Mean Corpuscular Volume (MCV)	The MCV blood test evaluates the erythrocyte count, which is the average size of red blood cells.
Mean Cell Hemoglobin (MCH)	The typical hemoglobin level in a typical red blood cell. The MCH is calculated by calculating the red blood cell count and hemoglobin levels.
Mean Cell Hemoglobin Concentration (MCHC)	The hemoglobin concentration in red blood cells, on average.
Reticulocyte	Newly produced, immature red blood cells, known as reticulocytes, emerge from the bone marrow. They form and ripen in the bone marrow before they are released into the blood.
Red Cell Distribution Width (RDW)	A red cell distribution width (RDW) test is used to determine your blood's red blood cell volume and size range (erythrocytes).
Hemoglobin Electrophoresis: Hb A Hb A2 Hb F Hb S	Electrophoresis for hemoglobin is a procedure that uses a high-voltage electric field to separate the various types of hemoglobin found in the blood. Additionally, it searches for abnormal forms of hemoglobin. The most commonly occurring type of hemoglobin (Hg A) in healthy adults. Hb F Fetal hemoglobin. Unborn babies and newborns have a specific type of hemoglobin. Immediately after birth, Hb F is replaced by Hb A. Sickle cell disease is the common cause of hemoglobin of this type.
Aspartate Amino Transferase (AST)	A special enzyme, present in increased levels in the blood following a heart attack or liver disease, aids in the transfer of an amino group from glutamic acid to oxaloacetic acid.
Alanine Amino Transferase (ALTI)	A naturally occurring enzyme found in the liver and heart. The ALT enzyme is released into the blood when liver or heart injuries occur. As a result of liver damage (such as from viral hepatitis) or an insult to the heart, the blood ALT levels increase.
Alkaline phosphatase (ALPI)	An enzyme which is commonly found in various body parts, but is most prevalent in the liver, bones, intestine, and kidneys. A blood test to measure the activity of alkaline phosphatase is used to gauge the enzyme's concentration. Elevated ALP levels may be a result of liver disease or another type of health problem.
Gamma-glutamyl Transferase (GGT)	An enzyme found throughout the body in many organs at the highest concentrations found in the liver. Elevated GGT can be found in the blood of most people with liver or bile duct problems. The blood sample test is meant to gauge the level of GGT. When the liver is injured, GGT levels increase normally, but if the liver is compromised, they will be higher.

D. Framework

The framework (Fig. 1) is divided into three sections: data preprocessing, features selection techniques, and ML methods.

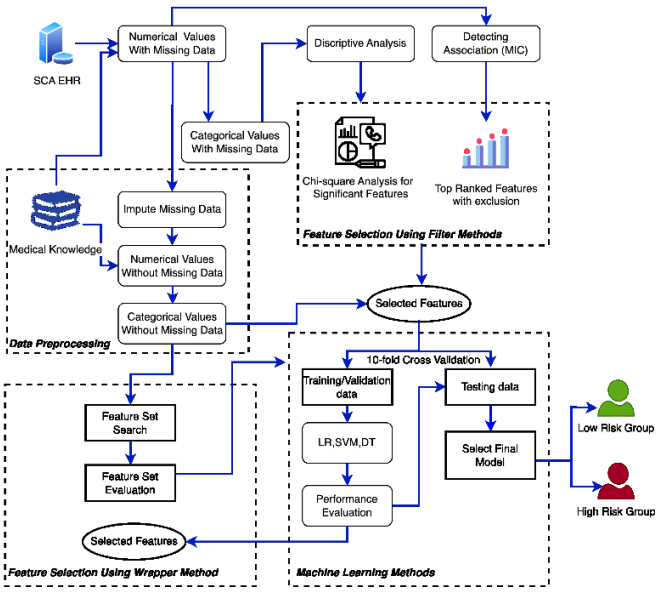


Fig. 1. Framework for preprocessing, feature-selection and modelling.

1) Data preprocessing: The dataset included numerical values with missing data. This step included: imputing missing data and reformatting the data using medical knowledge.

2) Impute missing data: Two of the most frequently used approaches for dealing with missing data are imputation and deletion, and the former can introduce bias while the latter introduces both bias and a reduction in statistical power [20]. In this step, linear interpolation uses the two nearest points to estimate the value of the interpolated point in a one-dimensional data sequence [21]. For all features, the numerical values were imputed with the linear interpolation using SPSS 21 since removing patient data reduces patient datasets, which can affect the performance of learned models, we chose to eliminate those features with missing percentages above 10%.

3) Reformate the features using medical knowledge: We used existing knowledge, including standard blood count boundaries [22], and clinical expertise domains, to format data into categories.

4) Feature selection techniques: Feature selection algorithms are broadly divided into three categories: filter

methods, wrapper methods, and embedded method [23]. Here we first used filter methods.

A statistical score is applied to each feature through the use of filter feature selection methods. The results are rated by score, and then the preferred features are selected to be retained or discarded from the dataset. The methods are typically univariate and either treat the feature in isolation or regard it in the context of the dependent variable. Here, the top correlated features include finding new patterns in large datasets, which highlight interesting relationships between pairs of variables. For two-variable relationships, we used two techniques as a measure of dependence: i) the maximal information coefficient (MIC)[24]. The results were sorted from largest to lowest and discarded as the lowest score feature, ii) the Chi-squared test which is a nonparametric statistical test that is widely used to assess the statistical significance of predictive factors [25]. We used Recursive Feature Elimination (RFE) which iteratively builds models and removes the least key features.

The second feature selection technique was the wrapper method which evaluates the performance of a subset of features by training a machine learning model. They use search algorithms to find the best subset of features based on the model's performance. Two common methods are known as forward selection and backward elimination. We used forward selection to start with an empty set and added features one by one based on performance improvement.

This study used hybrid feature selection approach that combines the strengths of both methods (filter and wrapper). Initially, features are ranked using specific criteria, followed by applying a wrapper algorithm to generate a subset from the ranked features [15].

We chose to use filter and wrapper methods for feature selection instead of embedded methods because filter methods are computationally efficient and model-agnostic, allowing for quick preliminary feature evaluation. Wrapper methods, though more computationally intensive, account for feature interactions and improve model performance by evaluating subsets through model training. In contrast, embedded methods are model-dependent, which could limit the flexibility and generalizability of our feature selection process. By using filter and wrapper methods, we ensured a versatile and robust feature selection applicable across various machine learning models.

5) Machine learning methods: Three ML methods were utilized: Logistic regression (LR), Support Vector Machine (SVM), and Decision Tree (DT). LR is a statistical method used to model the probability of a binary outcome based on one or more predictor variables [26]. It is particularly useful in scenarios where the dependent variable is categorical, such as pass/fail or disease/no disease [27]. Unlike linear regression, which predicts continuous outcomes, logistic regression estimates the probability of an event occurring versus not occurring [27]. LR implemented using the fitclinear function in MATLAB 2024a. In the context of this study, LR is employed to predict whether patients fall into high-risk or low-risk categories based on their features. The fitclinear function optimizes the model parameters using maximum likelihood estimation to find the best fit for the data. The logistic function (sigmoid) transforms the linear output into a probability, which is then used to make binary predictions. The model's performance is evaluated using cross-validation to ensure robustness and avoid overfitting.

Another common ML model is SVM which is a supervised learning method used for both regression and also classification [28].However, they are most commonly utilized for binary classification. The SVM algorithm aims to find an appropriate hyperplane in such an N-dimensional space that accurately classifies the data points [29]. The memory efficiency of SVM and its efficiency for high-dimensional environments are the two key benefits of employing it. SVMs are among the most extensively used and successful ML algorithms in supervised learning for classifying and recognizing problems. They get a solid theoretical foundation, which makes them invaluable throughout this field. The essential principle behind SVMs was that with some high input vector x and vector output y, there must be some unknown and non-linear relationship (mapping, function) $y = f(x)$. There seems to be no knowledge about input data vectors' underlying joint distributions. The training data would be the only information provided. As a result, they are classified as supervised learning algorithms. SVMs create a hyperplane to divide two classes. The algorithm intends to attain the greatest possible separation between the classes.

SVM finds the hyper-plane as in feature space which distinguishes between categories for classification. A training data point has been represented as points within feature space by an SVM model, which is mapped in a way in which points belonging to the different classes were also separated by just as wide a margin as possible. The testing data points then are mapped into the same space and classified according to where they fall upon that margin. SVM implemented here using the fitcsvm function in MATLAB 2024a. In this study, SVM is utilized to classify patients into high-risk and low-risk categories based on their features.

The third ML method is DT, which is a popular machine learning method used for classification tasks. DTs are effective in handling non-linear relationships and can be scaled for big data applications [30]. DTs have been applied in various fields, including education, building, botany, social science, and medicine [31]. DT implemented using the fitctree function in MATLAB 2024a, the DT functions by repeatedly dividing the dataset into smaller subsets according to the values of input features, creating a structure that resembles a tree. In this tree, each node corresponds to a feature, each branch signifies a decision rule, and each leaf node indicates an outcome. The fitctree function grows the tree by selecting features that best split the data according to a chosen criterion, such as Gini impurity or entropy.

The process of building a decision tree begins with selecting the best feature for the root node. The algorithm evaluates all features and selects the one that best separates the data into distinct classes. This is done by calculating the split criterion for each feature. Gini impurity was used to measure the probability of a randomly chosen element being misclassified if it was randomly labeled according to the distribution of labels in the subset. Lower values of Gini impurity indicate better splits.

Once the best feature is chosen, the data is split into subsets based on the values of this feature. For categorical features, this involves creating branches for each category. The process of selecting the best feature and splitting the data is repeated recursively for each subset. At each node, the algorithm considers only the subset of data that reaches that node. The recursion continues until one of the stopping criteria is met: all instances in a node belong to the same class, there are no remaining features to split on, the maximum tree depth is reached, or a minimum number of instances per node is specified.

In this study, the DT model is used to classify patients into high-risk and low-risk categories based on their features. The model is trained using 90% of the dataset and validated using 10% of the dataset. Ten-fold cross-validation is applied to ensure the model's robustness and to prevent overfitting.

The selection of the most appropriate machine learning method depends on the specific application and dataset characteristics [31].

6) Performance evaluation: By using the confusion matrix, one can see how the selected model performed in each category. The confusion matrix helps you figure out where the model messed up. The number of observations is defined as the following:

- TP: patients are correctly categorized as low risk.
- FP: patients categorized in low risk, but they predicted as high risk.
- TN: patients are correctly categorized as high risk.
- FN: patients categorized in high risk, but they predicted as low risk.

Final Model and Feature Set Selection: For the final model and feature set selection, we employed a comprehensive evaluation process based on four key performance metrics: True Positive Rate (TPR), False Negative Rate (FNR), Positive Predictive Value (PPV), and False Discovery Rate (FDR). The TPR is the ratio of true high-risk classifications to true high-risk cases. The FNR is the proportion of incorrectly classified instances in relation to the number of actual classes in high risk. PPV stands for predicted in a high risk to correctly classified in high-risk observation ratio. The FDR is the rate of predictions that are incorrect per class that was classified as high risk.

Initially, we applied filter methods to efficiently evaluate and reduce the dimensionality of the dataset, using statistical measures such as correlation coefficients and mutual information to identify relevant features. This provided a preliminary selection of features that served as the basis for further refinement.

Subsequently, we utilized wrapper methods, including accounting for feature interactions and refine the feature subset. Wrapper methods involve directly evaluating subsets of features by training the selected machine learning model from the filter selection method. This iterative process allowed us to identify the most impactful features and improve model performance.

After generating multiple models with various feature sets, we compared their performance based on the metrics mentioned above. The evaluation process ensured that the selected model and feature set offered a good balance of achieving the highest TPR and PPV and the lowest FNR and FDR. This approach to combining filter and wrapper methods enabled us to achieve a robust and reliable predictive model, tailored to the specific needs of our application.

III. Results

A. Insight Into Data

The dataset included 200 SCA patients, 10 patients were excluded because they were diagnosed with Thalassemia. The low-risk group included 90 patients while the high-risk group included 110 patients.

B. Preprocessing Results

The analysis of the "ContWithMissing" sheet revealed varying degrees of missing data across different clinical features. To address this, linear interpolation was applied to impute missing values for numerical features using SPSS 21. As shown in Table II, features with more than 10% of missing data were excluded to avoid bias and ensure robust model performance.

TABLE II. Statistics of the Missing Data

	N	Missing	Percent	Minimum	Maximum	Mean
AGE	188	2	1.1%	12	59	31.08
Gender	182	8	4.2%	1	2	1.55
Hb.	190	0	0%	1	13	8.22
WBC	190	0	0%	3.70	45.70	13.97
PIT	190	0	0%	5.10	853	411.81
MCV	189	1	0.5%	37.50	380	86
MCH	189	1	0.5%	19	80	29.128
MCHC	190	0	0%	24.60	38.80	33.64
Ret.	164	26	13.7%	0.014	16.9	2.02
RDW	159	31	16.3%	6.50	208	23.96
Hb A	52	138	72.6%	0	54.10	8.25
Hb A2	103	87	45.8%	1.80	9.40	3.80

Hb S	101	89	46.8%	0	36	8.59
Hb F	127	63	33.2%	0.6	96.1	77.02
AST	141	49	25.8%	13	230	54.18
ALT	147	43	22.6%	14	244	46.28
ALKP	138	52	27.4%	5	352	96.28
GGT	65	125	65.8%	6	223	41.43
Complication	190	0	0%	0	1	0.47

TABLE III. REFORMATE THE FEATURES USING MEDICAL KNOWLEDGE RESULTS

Feature	Term	Range	Value	Define	No. of Missing Estimated
Age (years)	Old Young	12-59	>40 <=40	0 1	2 0
Gender	Male Female		1 2	0 1	0 1
Hb (g/dL.)	Normal Low Very low	2-13	12-15 >=7, <=12 <7	2 1 0	0
WBC (K/uL)	High Normal Low	3.7-45.7	>11.5 >=4.5,<=11.5 <4.5	2 1 0	0
PLT (K/uL)	High Normal Low	5.10-853	>=450 150-450 <=150	2 1 0	0
MCV (fL)	High Normal Low	26.98-380	>94 80>=,<=94 <80	2 1 0	0 1 0
MCH (pg)	High Normal Low	19-80	>31 27>=,<=31 <27	2 1 0	0 1 0
MCHC (%)	High Normal Low	24.6-38.8	>36 32>=,<=36 <32	2 1 0	0
Reticulocyte (%)	High Normal Low	0.014-169	>1.5 0.5>=,<=1.5 <0.5	2 1 0	4 0 22
Red cell distribution width (Male)	High Normal Low	0.01-16.9	>14.5 11.8<=,>=14.5 <11.8	2 1 0	10 0 0
Red cell distribution width (Female)	High Normal Very low	0.08-0.23	>16.1 12.2-16.1 <12	2 1 0	0 0 16
electrophoresis Hb A (%)	Normal Low Very low	0-54.10	95-98 <95 =0	2 1 0	0 77 61
Hb A2 (%)	High Normal	1.8-9.4	>=3.5 1.5-3.5	1 0	27 8
Hb F (%)	Abnormal Normal	0-36	>=2 <2	1 0	89 0
Hb S (%)	Abnormal Normal		1 0	1 0	63 0
Aspartate Amino Transferase (AST) (U/L)	High Normal Low	13-230	>37 15>=,<=37 <15	2 1 0	40 8 1
Alanine Amino Transferase (ALTI) (U/L)	High Normal	14-244	>78 12>=,<=78	2 1	1 42
Alkaline phosphatase (ALPI) (U/L)	High Normal Low	5-352	>150 40>=,<=150 <40	2 1 0	13 36 3
Gamma-glutamyl Transferase GGT (U/L)	High Normal Low	6-223	>85 5,<=85	2 1	25 100
Complications	None Complications		None Complications	0 1	

Age and gender data were nearly complete, with only 1.1% and 4.21% missing, respectively. Critical parameters such as hemoglobin levels, white blood cell count, platelets, mean corpuscular volume, mean cell hemoglobin, and mean cell hemoglobin concentration had no missing values. Reticulocyte and red cell distribution width had 13.68% and 16.32% missing values, respectively, and were thus excluded from the analysis. Hemoglobin electrophoresis tests showed substantial missing data (HbA 72.63%, HbA2 45.79%, HbF 46.84%, HbS 33.16%), leading to their exclusion. Similarly, liver enzyme measurements such as AST, ALT, ALKP, and GGT had significant missing data, particularly GGT with 65.79%, and were excluded from further analysis. This approach ensures that the dataset remains robust and reliable for subsequent analysis and model development. The data was also reformatted using medical knowledge and standard blood count boundaries to categorize values meaningfully. Table III shows the reformat of numerical data using medical knowledge.

C. Filter Selection Results

The feature selection process utilized different methods for the "ContWithMissing" and "CatWithMissing" sheets. For the "ContWithMissing" sheet, Table IV shows the top correlated features using MIC analysis. The top features included Platelets (MIC: 0.41987), WBC (MIC: 0.41139), MCH (MIC: 0.3613), MCV (MIC: 0.3171), MCHC (MIC: 0.27546), Hb (MIC: 0.26462), and AGE (MIC: 0.16802), all of which were included in the final model due to their strong correlations.

For the "CatWithMissing" sheet, the Chi-squared test was used to determine the statistical significance of the features. This analysis revealed several significant features, including WBC and Platelets.

Notably, WBC showed a significant difference between low and high-risk patient categories (p-value: 0.005), and Platelet counts also displayed statistical significance (p-value: 0.027). Table V shows significant features.

By integrating MIC analysis for continuous data and Chi-squared analysis for categorical data, the final model incorporated the most predictive and statistically significant features, thereby enhancing its performance and reliability in predicting mismatch repair complications.

D. Machine Learning Methods Results

The feature selection and ML model evaluation was conducted using the "CatNoMissing" sheet. For this process, 10% of the data was held out as testing data while 90% was used for training. The final dataset underwent ten-fold cross-validation. The LR, SVM, and DT models were assessed using several performance metrics, including TPR, FNR, PPV, and FDR. Table VI shows the final results using feature selection and ML methods.

TABLE IV. TOP CORRELATED FEATURES

Features	MIC (strength)
Platelets	0.41987
WBC	0.41139
MCH	0.3613
MCV	0.3171
MCHC	0.27546
Hb	0.26462
AGE	0.16802

TABLE V. (A) SIGNIFICANT FEATURES

Feature	Medical Knowledge	Patients Categories				Chi-Square Test		
		Low risk		*High risk*		*Frequency*	*Percent*	*p-value*
Age (years)	>40 <=40	79 19	80.6% 10.1%	79 11	87.8% 5.9%	158 30	84% 16%	0.23
Sex	1 2	47 46	50.5% 49.5%	35 54	39.3% 60.7%	82 100	45.1% 54.9%	0.13
Hb (g/dL.)	12-15 >=7, <=12 <7	2 86 12	2% 86% 12%	0 75 15	0% 83.3% 16.7%	2 161 27	1.1% 84.7% 14.2%	0.33
WBC (K/uL)	>11.5 >=4.5,<=11.5 <4.5	49 50 1	49% 50% 1%	63 27 0	70% 30% 0%	112 77 1	35.8% 61.6% 2.6%	**0.005**
Platelet (K/uL)	>=450 150-450 <=150	30 65 5	30% 65% 5%	38 52 0	42.2% 57.8% 0%	68 117 5	35.8% 61.6% 2.6%	**0.027**
MCV (fL)	>94 80>=,<=94 <80	16 47 37	16% 47% 37%	18 49 22	20% 55.1% 24.7%	34 96 59	18% 50.8% 31.2%	0.19
MCH (pg)	>31 27>=,<=31 <27	23 43 34	23% 43% 34%	32 33 24	36% 37.1% 27%	55 76 58	29.1% 40.2% 30.7%	0.15
MCHC (%)	>36 32>=,<=36 <32	6 85 9	6% 85% 9%	7 69 14	7.8% 76.7% 15.6%	13 154 23	6.8% 81.1% 12.1%	0.31
Ret (%)	>1.5 0.5>=,<=1.5 <0.5	5 1 75	6.2% 1.2% 92.6%	6 1 76	7.2% 1.2% 91.6%	11 2 151	6.7% 1.2% 92.1%	1.00

TABLE V (B). SIGNIFICANT FEATURES (CONTINUE)

Feature	Medical Knowledge	Patients Categories				Chi-Square Test		
		Low risk		High risk		Frequency	Percent	p-value
RDW	Male >14.5 11.8<=,>=14.5 <11.8 Female >16.1 12.2<=,>=16.1 <12	84 1 1	97.7% 1.2% 1.2%	72 0 1	98.6% 0 1.4%	156 1 2	98.1% 0.6% 1.3%	1.00
Hb A (%)	95-98 <95 =0	0 9 20	0% 31% 69%	0 11 12	0% 47.8% 52.2%	0 20 32	0 38.5% 61.5%	0.26
Hb A2 (%)	>=3.5 1.5-3.5	34 24	58.6% 41.4%	23 22	51.1% 48.9%	57 46	55.3% 44.7%	0.54
Hb F (%)	>=2 <2	53 5	91.4% 8.6%	41 2	95.3% 4.7%	94 7	93.1% 6.9%	0.69
Hb S (%)	1 0	67 0	100% 0%	60 0	100% 0%	127 0	100% 0	-
AST (U/L)	>37 15>=,<=37 <15	53 24 1	67.9% 30.8% 1.3%	42 20 1	66.7% 31.7% 1.6%	95 44 2	67.4% 31.2% 1.4%	1.00
ALTI (U/L)	>78 12>=,<=78	6 75	7.4% 92.6%	2 64	3% 97%	8 139	5.4% 94.6%	0.29
ALPI (U/L)	>150 40>=,<=150 <40	18 31 25	24.3% 41.9% 33.8%	16 26 22	25% 40.6% 34.4%	34 57 47	24.6% 41.3% 34.1%	1.00
GGT (U/L)	>85 5,<=85	4 32	11.1% 88.9%	1 28	3.4% 96.6%	5 60	7.7% 92.3%	0.37
Complications	Low risk High risk (Complications)					100 90	52.6% 47.4%	

TABLE VI. MACHINE LEARNING METHODS RESULTS

Feature Selection	Model	Features	Training Accuracy	Testing Accuracy	Testing TPR	Testing FNR	Testing PPV	Testing FDR
None	LR	AGE, Gender, Hb, Platelets, WBC, MCV, MCH, MCHC, Ret, RWD, Hb A, Hb A2, Hb F, Hb S, AST, ALTI, ALPI, GGT.	65.3	35[a]	64.3	35.7	30	70
	SVM		56.7	62.5	50	50	46.7	53.3
	DT		58.7	50	64.3	35.7	37.5	62.5
Chi-square	LR	Platelets, WBC.	60.7	52.5	64.3	35.7	39.1	60.9
	SVM		62.7	52.5	54.54	45.45	60	40
	DT		61.9	52.5	54.54	45.45	60	40
MIC	LR	Platelets, WBC, MCH, MCV, MCHC, Hb, AGE.	62	47.5a	85.7	14.3	38.7	61.3
	SVM		54	50	85.7	14.3	40	60
	DT		52	52.5	71.4	28.6	40	60
Wrapper	LR	Platelets, WBC, MCV, MCHC	54.38	52.63	20	80	66.67	33.33
	SVM		60.23	57.90	72.72	27.27	61.53	38.46
	DT		66.08	36.82	36.36	63.63	44.45	55.56
a An overfitting model is a data science concept in which a statistical model fits precisely against its training data.* selected model.								

Initially, models were trained without feature selection, using all available features. The LR model achieved a training accuracy of 65.3% but showed significant overfitting with a testing accuracy of 35%. The TPR and PPV were 64.3% and 30%, respectively, while the FNR and FDR were 35.7% and 70%. The SVM model performed better on the testing data, achieving an accuracy of 62.5%, a TPR of 50%, a PPV of 46.7%, an FNR of 50%, and an FDR of 53.3%. The DT model had a testing accuracy of 50%, with a TPR of 64.3%, a PPV of 37.5%, an FNR of 35.7%, and an FDR of 62.5%.

Next, feature selection was applied using significant features (chi-square) identified through statistical tests, specifically focusing on Platelets and WBC. The LR model improved slightly with a testing accuracy of 52.5%, a TPR of 64.3%, a PPV of 39.1%, an FNR of 35.7%, and an FDR of 60.9%. Both the SVM and DT models showed similar improvements, with SVM achieving a testing accuracy of 52.5%, a TPR of 54.54%, a PPV of 60%, an FNR of 45.45%, and an FDR of 40%. The DT model had a testing accuracy of 52.5%, a TPR of 54.54%, a PPV of 60%, an FNR of 45.45%, and an FDR of 40%.

Further evaluation with the top 7 correlated features (MIC) revealed that the SVM model with these features achieved the best performance metrics, with a testing accuracy of 50%, a TPR of 85.7%, a PPV of 40%, an FNR of 14.3%, and an FDR of 60%. The LR model also performed well, with a testing accuracy of 47.5%, a TPR of 85.7%, a PPV of 38.7%, an FNR of 14.3%, and an FDR of 61.3%. The DT model had a testing accuracy of 52.5%, a TPR of 71.4%, a PPV of 40%, an FNR of 28.6%, and an FDR of 60%.

After applying wrapper selection methods, the SVM model with 4 features demonstrated the best performance among the evaluated models, with a testing accuracy of 57.9%, a TPR of 72.72%, a PPV of 61.53%, an FNR of 27.27%, and an FDR of 38.46%. A total of 2^{18} =262,143 models were generated for each technique and assessed. The final model was selected based on its superior performance metrics, specifically the highest TPR and PPV, along with the lowest FNR and FDR. The selected SVM model showed exceptional capability in correctly identifying high-risk patients while minimizing incorrect classifications. This comprehensive evaluation confirmed the robustness and accuracy of the SVM model in predicting patient risk, making it the optimal choice for this analysis.

IV. DISCUSSION

This study aimed to implement ML algorithms to identify high-risk groups among SCA patients using clinical and pathological data from King Abdulaziz University Hospital. The primary objective was to develop predictive models that enhance the accuracy of risk assessments, thereby improving patient management and treatment outcomes. The application of LR, SVM, and DT in this study demonstrated the potential of ML in healthcare, particularly in the context of genetic disorders like SCA.

The findings indicate that the SVM model outperformed LR and DT in identifying high-risk SCD patients. Specifically, the SVM model with the top 7 correlated features achieved the highest TPR and PPV, along with the lowest FNR and FDR. These results suggest that SVM is particularly effective in distinguishing between high-risk and low-risk patients, making it a valuable tool for clinicians. The experimental results echo theoretical expectations: SVM excelled because the dataset (n = 200, 7–17 active predictors) is high-dimensional relative to sample size, a setting in which the max-margin principle reduces over-fitting. LR performed adequately on linearly separable CBC ratios, whereas DT handled non-linear interactions but was vulnerable to small-sample fragmentation.

A crucial element of the study was the feature selection process, which significantly enhanced model performance. Both filter and wrapper methods were employed to identify the most relevant features, thereby reducing model complexity and improving prediction accuracy. The top 7 correlated features (Platelets, WBC, MCH, MCV, MCHC, Hb, and Age) were found to be significant predictors of risk in SCA patients. Among these, Platelets and WBC counts were particularly noteworthy. Platelets, which are essential for blood clotting, and WBCs, which play a key role in immune response, are critical indicators of a patient's health status and potential complications. Indeed, the complete blood count (CBC) evaluates these cells, helping detect various diseases and conditions [32]. A lower platelet to WBC ratio is associated with increased risk of postoperative infectious complications in patients undergoing radical nephrectomy [33]. In essential thrombocythemia patients, platelet counts outside the normal range correlate with an immediate risk of major hemorrhage, while elevated WBC counts are associated with both thrombosis and hemorrhage [34]. A conserved pattern of recovery, defined by co-regulation of WBC and platelet populations, has been identified across various inflammatory. This pattern of recovery, marked by an exponential decline in WBC and a subsequent linear increase in platelet count, could signify a basic model of human physiological response and serve as a means to identify high-risk patients. [35].

One of the key challenges addressed in this study was the handling of missing data. The application of linear interpolation to impute missing values ensured that the dataset remained comprehensive and informative. Excluding features with more than 10% of missing data helped maintain the integrity of the models, avoiding potential biases that could arise from incomplete data.

The decision to use a combination of filter and wrapper methods for feature selection, rather than embedded methods, was justified by the need for flexibility and computational efficiency. While embedded methods integrate feature selection within the model training process, they can be computationally intensive and less adaptable to different algorithms. The approach adopted in this study allowed for a more tailored feature selection process, enhancing the performance of each ML model.

Despite the promising results, this study has some limitations. The sample size of 200 patients, while adequate for initial model development, may not capture the full spectrum of variability in SCD presentations. Future studies with larger datasets and multi-center collaborations are needed to validate the findings and enhance the generalizability of the models.

Using LR, SVM, and DT with small-sized datasets presents both opportunities and challenges. LR is advantageous for small datasets because it is less prone to overfitting due to its simpler linear model structure. It can provide meaningful insights and interpretability, especially when the relationship between the predictors and the outcome is approximately linear. However, its performance may suffer if the underlying data distribution is complex.

The SVM, particularly with linear kernels, are also suitable for small datasets as they maximize the margin between classes, which can lead to better generalization on unseen data. SVMs are robust to overfitting making them effective for datasets with

a small number of samples but many features. However, SVMs can be computationally expensive, and their performance is highly sensitive to the choice of kernel and hyperparameters, which can be challenging to optimize with limited data.

Finally, DTs were intuitive and easy to interpret, making them appeal for small datasets. They could capture complex interactions between features without requiring a large amount of data. However, DTs are prone to overfitting, especially with small datasets, as they tend to create overly complex trees that do not generalize well.

V. Conclusion

In conclusion, this study demonstrates the potential of ML algorithms, particularly SVM, in identifying high-risk SCA patients. The integration of hybrid feature selection techniques and effective handling of missing data contributed to the development of accurate predictive models. These findings underscore the value of ML in healthcare, offering a tool for risk assessment and patient management in SCA. The capability of ML to enhance clinical decision-making, improving patient outcomes in genetic disorders like SCA. The SVM model (TPR = 85.7 %, PPV = 61.5 %) offers a clinically actionable screening tool for SCA wards. Limitations include (i) single-centre data, which may under-represent regional genotype variability, and (ii) retrospective feature availability, potentially biasing model coefficients. Future work will involve a prospective, multi-centre cohort integrating inflammatory biomarkers and whole-genome variants to refine risk predictions and support personalized therapy pathways.

Acknowledgment

The authors appreciate the associate editors' and anonymous reviewers' suggestions and criticism of an earlier version of this manuscript.

Authors' Contributions

Conceptualization, H.B.; Methodology, H.B., H.A., N.A., S.A.and G.A.; Software, H.B., H.A., and N.A.;Validation, H.B., A.B., and G.A.; Investigation, H.B., H.A., A.A., and G.A.; Resources, H.B., H.A., N.A., A.B., H.A., S. B., A.A., S.A., and G.Z.; Data Curation, H.B., S.A., A.A. and G.Z.; Writing Original Draft Preparation, H.B., N. A.; Writing—Review and Editing, H.B., H.A., N. A., A.B., S.A., and G.Z.; Visualization, H.B.; Supervision, H.B.; Project Administration, H.B.; Funding Acquisition, H.B.

Conflicting of Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability

The Saudi data is restricted to only a few authorized users who are permitted to access the database with Clinical Research Committee (CRC) and Research Ethics Committee (REC) permission. Due to ethical restrictions, de-identified datasets are available upon request to (hrbanjar@kau.edu.sa) for authorization.

References

[1] S. Azar and T. E. Wong, "Sickle Cell Disease: A Brief Update.," Med. Clin. North Am., vol. 101, no. 2, pp. 375–393, Mar. 2017, doi: 10.1016/j.mcna.2016.09.009.

[2] Dhiraj K, "Top 4 advantages and disadvantages of Support Vector Machine or SVM.," Medium, 2019. https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107 (accessed Sep. 24, 2021).

[3] D. J. Weatherall, "The inherited diseases of hemoglobin are an emerging global health burden.," Blood, vol. 115, no. 22, pp. 4331–4336, Jun. 2010, doi: 10.1182/blood-2010-01-251348.

[4] M. Angastiniotis and B. Modell, "Global Epidemiology of Hemoglobin Disorders," Ann. N. Y. Acad. Sci., vol. 850, 1998, [Online]. Available: https://api.semanticscholar.org/CorpusID:10048398.

[5] W. Jastaniah, "Epidemiology of sickle cell disease in Saudi Arabia," Ann. Saudi Med., vol. 31, pp. 289–293, 2011, [Online]. Available: https://api.semanticscholar.org/CorpusID:5382156.

[6] F. B. Piel, M. H. Steinberg, and D. C. Rees, "Sickle Cell Disease.," N. Engl. J. Med., vol. 376 16, pp. 1561–1573, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:35879604.

[7] N. H. Alharbi, R. O. Bameer, S. S. Geddan, and H. M. Alharbi, "Recent Advances and Machine Learning Techniques on Sickle Cell Disease," Futur. Comput. Informatics J., 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:233386170.

[8] P. Sebastiani et al., "A network model to predict the risk of death in sickle cell disease.," Blood, vol. 110 7, pp. 2727–2735, 2007, [Online]. Available: https://api.semanticscholar.org/CorpusID:8718624.

[9] C. T. Quinn, "Minireview: Clinical severity in sickle cell disease: the challenges of definition and prognostication," Exp. Biol. Med., vol. 241, pp. 679–688, 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:27272544.

[10] X. W. van den Tweel, J. H. Lee, J. Howard, and K. Fijnvandraat, "Measurement of Disease Severity in Patients with Sickle Cell Disease: A Systematic Review.," Blood, vol. 110, p. 2250, 2007, [Online]. Available: https://api.semanticscholar.org/CorpusID:208430316.

[11] N. R. Shah et al., "Severity Classification for Sickle Cell Disease: A RAND/UCLA Modified Delphi Panel," Blood, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:214553481.

[12] A. Mohammed, P. S. B. Podila, R. L. Davis, K. I. Ataga, J. S. Hankins, and R. Kamaleswaran, "Using Machine Learning to Predict Early Onset Acute Organ Failure in Critically Ill Intensive Care Unit Patients With Sickle Cell Disease: Retrospective Study," J. Med. Internet Res., vol. 22, no. 5, pp. e14693–e14693, May 2020, doi: 10.2196/14693.

[13] V. E. Staartjes, J. M. Kernbach, V. Stumpo, C. H. B. van Niftrik, C. Serra, and L. Regli, "Foundations of Feature Selection in Clinical Prediction Modeling.," Acta Neurochir. Suppl., vol. 134, pp. 51–57, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:244872146.

[14] N. Mlambo, W. K. Cheruiyot, and M. W. Kimwele, "A Survey and Comparative Study of Filter and Wrapper Feature Selection Techniques," 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:38791166.

[15] S. S. M. S and A. Narayanan, "Improving Classification Accuracy Using Combined Filter+Wrapper Feature Selection Technique," 2019 IEEE Int. Conf. Electr. Comput. Commun. Technol., pp. 1–6, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:204816141.

[16] M. Mandal, P. K. Singh, M. F. Ijaz, J. Shafi, and R. Sarkar, "A Tri-Stage Wrapper-Filter Feature Selection Framework for Disease Classification," Sensors (Basel)., vol. 21, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:237341764.

[17] A. Patel et al., "Machine Learning Algorithms in Predicting Hospital Readmissions in Sickle Cell Disease," Blood, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:213979893.

[18] O. B. Ayoade et al., "An Ensemble Models for the Prediction of Sickle Cell Disease from Erythrocytes Smears," EAI Endorsed Trans. Pervasive Heal. Technol., vol. 9, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:262100582.

[19] M. Sarker, "Reinventing Wellness: How Machine Learning Transforms Healthcare," J. Artif. Intell. Gen. Sci. ISSN3006-4023, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:268355442.

[20] F. Cismondia, A. S. Fialho, S. M. Vieira, S. R. Retic, J. M. C. Sousab, and S. N. Finkelstein, "Missing data in medical databases: Impute, delete or classify?," Artif. Intell. Med., vol. 58, no. 1, pp. 63–72, 2013, doi: 10.1016/j.artmed.2013.01.003.

[21] G. Huang, "Missing data filling method based on linear interpolation and lightgbm," J. Phys. Conf. Ser., vol. 1754, no. 1, 2021, doi: 10.1088/1742-6596/1754/1/012187.

[22] "The CBC – providing information about your health," Sonora Quest Laboratories. https://www.sonoraquest.com/patient/knowledge-center/understanding-the-complete-blood-count-cbc/ (accessed Sep. 05, 2021).

[23] H. Banjar et al., "Modelling Predictors of Molecular Response to Frontline Imatinib for Patients with Chronic Myeloid Leukaemia," PLoS One, vol. 12, no. 1, p. e0168947, 2017, doi: 10.1371/journal.pone.0168947.

[24] D. N. Reshef et al., "Detecting Novel Associations in Large Data Sets," Science (80-.)., vol. 334, pp. 1518–1524, 2011, doi: 10.1126/science.1205438.

[25] D. Mindrila and P. Balentyne, "The Chi -Square Test : Analyzing," no. 2013, p. 23, 2013.

[26] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, "Introduction to the Logistic Regression Model," 2005, [Online]. Available: https://api.semanticscholar.org/CorpusID:59226812.

[27] H.-Y. Kim, "Statistical notes for clinical researchers: logistic regression," Restor. Dent. \& Endod., vol. 42, pp. 342–348, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:38386217.

[28] "Tutorial on Support Vector Machine (SVM)," 2021. Accessed: Sep. 24, 2021. [Online]. Available: https://course.ccs.neu.edu/cs5100f11/resources/jakkula.pdf.

[29] D. Boswell, "Introduction to Support Vector Machines," 2002, [Online]. Available: https://api.semanticscholar.org/CorpusID:18986102.

[30] Z. Saurav, M. M. Mitu, N. S. Ritu, M. A. Hasan, S. Arefin, and D. M. Farid, "A New Method for Learning Decision Tree Classifier," 2023 Int. Conf. Electr. Comput. Commun. Eng., pp. 1–6, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:258219459.

[31] P. L. Bokonda, K. Ouazzani-Touhami, and N. Souissi, "Predictive analysis using machine learning: Review of trends and methods," 2020 Int. Symp. Adv. Electr. Commun. Technol., pp. 1–6, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:237376199.

[32] S. Gajbhiye and J. R. Aate, "Blood Report Analysis-A Review," Trop. J. Pharm. Life Sci., 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:264109151.

[33] A. Garbens et al., "Platelet to white blood cell ratio predicts 30-day postoperative infectious complications in patients undergoing radical nephrectomy for renal malignancy.," Can. Urol. Assoc. J., vol. 11 11, pp. E414–E420, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:31726668.

[34] P. J. Campbell et al., "Correlation of blood counts with vascular complications in essential thrombocythemia: analysis of the prospective PT1 cohort.," Blood, vol. 120 7, pp. 1409–1411, 2012, [Online]. Available: https://api.semanticscholar.org/CorpusID:206911231.

[35] B. H. Foy, T. M. Sundt, J. C. T. Carlson, A. D. Aguirre, and J. M. Higgins, "White Blood Cell and Platelet Dynamics Define Human Inflammatory Recovery," medRxiv, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:235602574.

Real Time Accident Detection and Emergency Response Using Drones, Machine Learning and LoRa Communication

Bandara H. M, SI. D, Maduhansa H. K. T. P, Jayasinghe S. S,
Samararathna A. K. S. R, Harinda Fernando, Shashika Lokuliyana
Department of Computer Systems Engineering, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

***Abstract*—Road accidents and delayed emergency responses remain a major concern in urban environments, contributing to over 1.4 million fatalities globally each year. With rapid urbanization and increasing vehicle density, timely detection and efficient traffic management are critical to reducing the impact of such events. This study proposes a real time Accident Detection and Emergency Response System with integrating Machine Learning IoT enabled drones and LoRa communication. The system combines real time accident detection using CCTV, drone assisted fire detection for post accident scenarios, crime activity monitoring and automated traffic management to reduce congestion and improve public safety. LoRa ensure long range, energy-efficient communication. ML models improve detection accuracy across accidents, fires, crimes and vehicles. Figures and sensor data are analyzed in real time to trigger alerts and assist emergency responders. The system supports scalable integration with existing urban infrastructure, promoting the development of smart city safety frameworks. By minimizing emergency response time, limiting secondary incidents and improving situational awareness, the proposed solution addresses critical gaps in current urban safety systems. It offers a practical, intelligent and adaptive approach to accident mitigation and traffic control in smart cities.**

***Keywords*—Accident detection; machine learning; IoT; drones; traffic management; LoRa communication**

I. INTRODUCTION

Urban development has brought about significant developments in infrastructure, technology and overall quality of life. However, it has also introduced complex challenges, especially regarding public safety and emergency response. Still, among the most urgent issues in contemporary cities are traffic accidents, fires and criminal activity, which can have dire repercussions including loss of life, property damage and financial instability. The World Health Organization (WHO) estimates that road traffic accidents cause around 1.4 million deaths yearly as well as millions more severe injuries [1]. Further aggravating the effects of these mishaps are secondary events including fire outbreaks, congestion, and delayed emergency response. These difficulties draw attention to the need for timely and effective emergency response systems able to provide emergency services with accurate accident detection and fast alerts. Manual reporting, fixed location sensors and human interaction form the foundation of most traditional approaches of accident detection and emergency response. These methods are often slow, inefficient, and prone to human error. Their limited breadth prevents them from offering a complete, real-time picture of metropolitan events [2]. The unpredictability of urban environments, coupled with increasing traffic congestion as well population density, further complicates the ability of conventional systems to respond effectively. Particularly in Machine Learning (ML), the Internet of Things (IoT), edge computing and cloud computing, and fast technological improvements present chances to use these developments to improve public safety. This paper suggests an integrated Accidental Identification and Notification System, a complete solution meant to use smart technology to increase urban safety [3]. These main elements define the system.

- Real time accident detection and notification system - Processes CCTV footage to detect accidents and notify emergency services instantly.
- Drone based fire detection system - Uses CCTV camera and fire sensor to detect post-accident fires and relay data to emergency responders.
- Crime detection system - Employs real time monitoring to identify criminal activities.
- Automatic Traffic Management System - To reduce traffic congestion dynamically adjust traffic light control and improve road safety by identifying the highest traffic density.

The system seeks to deliver emergency services fast notifications by combining IoT enabled surveillance, anomaly detection and real time data analytics. Thereby optimizing traffic management and lowering secondary events. In addition to addressing important urban safety issues, the suggested system addresses data interoperability, communication dependability, real time processing and secure data sharing. It guarantees reliable identification of abnormalities like accidents, fires and criminal activity by using machine learning algorithms to examine data from IoT sensors, cameras and drones. LoRa communication guarantees consistent long range data flow [4] [5].

Developing a scalable, efficient, user-friendly system that increases situational awareness, improves emergency response and reduces the financial impact of accidents and related events is the main goal of this project. A key invention for smart city safety solutions is the system's capacity to interact with current urban infrastructure and adapt to changing urban surroundings.

The paper is organized as follows: the Related Work in Section II reviews existing literature and their limitations, the Methodology in Section III describes the system's design and technologies, the Results in Section IV presents experimental findings, the Discussion in Section V analyzes the findings and their implications, and the Conclusion in Section VI summarizes the study and future work.

II. Literature Review

Recent studies have focused on the advancement of intelligent systems for accident detection, fire surveillance and crime monitoring, driven by innovations in Machine Learning (ML), the Internet of Things (IoT) and edge computing. This section analyzes recent research and technological breakthroughs across multiple disciplines, highlighting their relevance to the proposed Accident Identification and Notification System.

The advancement of smart cities is propelled by the Internet of Things (IoT), which furnishes the essential infrastructure for the collection and analysis of real time data. It requires the effective surveillance of traffic conditions by IoT devices, particularly sensors and cameras, together with real time accident identification. Zanella et al. [6] [7] examined the role of IoT in providing enhanced urban safety and mobility systems, with emphasis on connected devices in creating responsive and adaptive traffic management. This ensures that data from numerous sources is speedily processed and acted upon to make the urban transport network much more efficient.

For real time accident detection and predictive analytics. Wang et al. [8] explained how supervised machine learning was used to identify patterns in data obtained from video surveillance to help to detect accidents most likely to occur. In this manner, a system can detect not only incidents at the time of occurrence but also forecast accidents before they happen. In some cases, simply preventing such events can significantly reduce road fatalities.

The introduction of fire detection into vehicle environments brings in several problems associated with the mobile nature of this environment. Ola Willstrand, Peter Karlsson and Jonas Brandt [9] implement fire detection and alarm systems specifically designed for heavy vehicles. Their research highlights the need for reliable detection systems capable of functioning effectively under extreme conditions, such as high temperatures and vibrations typical in vehicle environments. Similarly, another study is available on ResearchGate [10] and it discusses a fire detection and control system in automobiles that utilizes fuzzy logic.

Long range communication technologies like LPWAN improved the reliability of fire detection systems, especially in large and remote areas. Vladimir Sanchez Padilla and Gabriel Roque [11] developed an LPWAN based communication framework that significantly improved the efficiency of fire detection systems. The study showed the benefits of low power, long distance wireless local area networks (LPWANs), which are important for implementing fire detection systems in smart city environments.

Crime detection and predictive analytics have progressively utilized deep learning methodologies to improve surveillance systems and automate threat recognition. Conventional CCTV monitoring depends on human involvement, which is susceptible to delays and mistakes. Recent developments in object detection algorithms, including VGG16, Inception-V3, Faster R-CNN and YOLO variations, have markedly enhanced real time crime detection, especially in weapon identification. In the absence of standardized datasets for real world applications, researchers have aggregated various data sources, including self-collected photographs, internet archives and publicly accessible databases. The incorporation of binary classification with confusion object inclusion has significantly reduced false positives and negatives. Among the evaluated models, YOLOv4 exhibited exceptional performance, attaining an F1-score of 91% and a mean average precision (mAP) of 91.73%, thereby affirming its efficacy in crime detection and predictive analytics [12].

Recent progress in intelligent traffic monitoring has resulted in the creation of diverse accident detection methodologies to improve safety and traffic management in smart cities. Research has investigated advanced methodology for identifying various accident types, such as rear-end collisions, T-bone crashes and frontal impacts, utilizing computer vision and deep learning techniques Nonetheless, obstacles such as data imbalances, heterogeneous road infrastructures, and fluctuating traffic conditions endure. To overcome these constraints, researchers have developed the I3D-CONVLSTM2D model, which combines RGB frames with optical flow data to enhance accident detection. Empirical assessments indicate its efficacy, with the I3D-CONVLSTM2D RGB + Optical Flow model attaining 87% Mean Average Precision (MAP), underscoring its suitability for real-time implementation in edge IoT driven smart city frameworks [13].

Despite significant advancements in accident detection, fire surveillance, and crime monitoring technologies, most existing systems operate in isolation and lack integration into a unified, real-time emergency response framework [14]. Moreover, many approaches do not support low-latency decision-making or long-range communication, making them unsuitable for highly dynamic urban environments. Current systems often struggle with inconsistent data flow, limited scalability, and poor situational awareness during emergencies [15].

Table I summarizes prior studies in accident, fire and crime detection, highlighting key limitations. While previous works used technologies like ML, LoRa and deep learning, most lacked real time alerting, integration with emergency response, or multi domain coverage. Our proposed system addresses these gaps by combining machine learning, long range communication and real time coordination across all components.

This paper addresses these gaps by proposing an integrated system that combines real time accident detection, drone based fire detection, crime detection and traffic management powered by machine learning and LoRa communication. By merging these technologies, the proposed framework improves emergency response time, improves threat detection accuracy and supports deployment in smart city environments, especially under constrained or mobile conditions.

TABLE I. SUMMARY OF RELATED WORKS – FOCUS, TECHNOLOGIES AND LIMITATIONS

Study Focus	Technologies Used	Limitations
Real-time accident detection & prediction using video surveillance [1]	ML, CCTV	No integration with emergency response; limited to accident type
Fire detection in heavy vehicles [9]	Sensor-based systems for mobile environments	lacks integration with communication systems
Fire detection communication framework [11]	LPWAN (LoRa), IoT	Focuses only on fire, lacks ML based detection
Real time crime detection with object detection models [12]	YOLOv4, deep learning, video analytics	No real time alerting or integration with response systems

III. METHODOLOGY

The Accident Identification and Notification System is made up of four main parts, and each one has its own way of dealing with safety issues in cities. The Accident Detection System's main tasks are to gather and preprocess different types of CCTV footage, create machine learning models for finding accidents in real time, and connect these models to edge computing to send alerts quickly. Drones with thermal and infrared cameras and machine learning techniques are used by the Fire Detection System to find fires in real time and call for help. The Crime Detection System combines real-time CCTV analysis with predictive simulations. It uses Machine Learning algorithms to find strange behavior and a decentralized data sharing system to make sure that communication is safe.

These models are then connected to traffic light control systems to improve traffic flow and lower accident rates. All these parts work together to make a complete answer for making cities safer and faster in case of an emergency. The proposed system consists of two main operational units. The drone based unit, illustrated in Fig. 1 is responsible for accident detection, fire detection and real time data acquisition in dynamic environments. The ground based unit, shown in Fig. 2 handles data collection via IoT and CCTV, preprocessing via edge computing and interaction with emergency services and traffic control systems.

A. Accident Identification and Notification System

The methodology for constructing the Accident Identification and Notification System is organized into six essential steps, emphasizing data gathering, preprocessing, model building, integration, and empirical testing. Each phase is intended to guarantee the system's precision, efficacy, and scalability in identifying accidents and enhancing emergency response.

The initial phase entails gathering a varied dataset of urban traffic situations using CCTV recordings. This collection is derived from live and recorded video feeds supplied by municipal authorities and traffic management centers. The footage encompasses diverse traffic scenarios, including standard traffic flow, accident situations, and different climatic conditions (e.g., noon, midnight, rain, fog). This diversity guarantees that the system is resilient and adept at managing real world difficulties. The collection is enhanced by integrating footage from internet platforms and public repositories, ensuring a thorough picture of urban traffic dynamics. Upon collection of the raw CCTV footage, it is subjected to preprocessing to facilitate analysis. This encompasses noise reduction, frame rate modification, and resolution standardization to guarantee uniformity throughout the collection. The preprocessed material is subsequently annotated by identifying occurrences of accidents and non-accidents. This annotated dataset constitutes the basis for training the ML model. Python-based libraries, OpenCV, and specialized video annotation tools are employed to optimize the preprocessing and annotation procedures.

Built on a Convolutional Neural Network (CNN) machine learning model used to assess video frames and identify accidents, the system Trained on the annotated dataset, uses cross valuation and data augmentation to enhance its generalizing and performance ability. To improve the model's ability to detect events in many environments, data augmentation includes rotations, translations, and lighting changes in the training set. Cross validation ensures the model is robust and not overfits the training set. The trained accident detection model is combined with an automatic alert system that turns on right away upon accident identification. The system sends emergency responders' important information to nearby, including the location and degree of the event [16]. The system uses edge computing, distributing the ML model and data processing on edge devices close to CCTV cameras, hence lowering delays. This method diminishes dependence on centralized cloud infrastructure, hence guaranteeing swifter detection and response times.

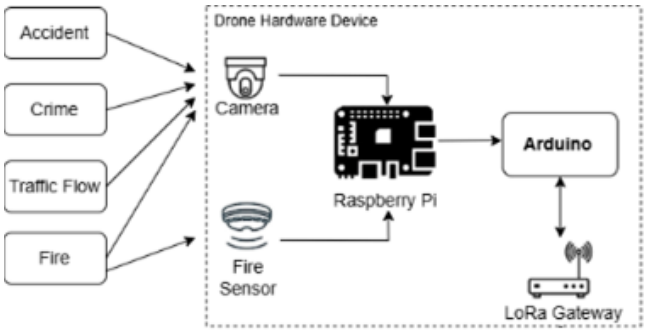


Fig. 1. Drone based unit system diagram.

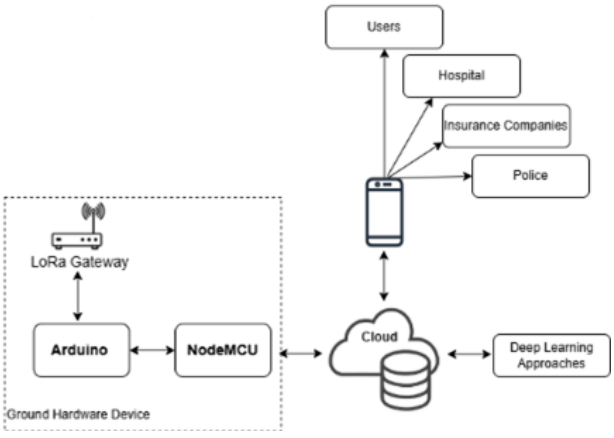


Fig. 2. Ground based unit system diagram.

The system is implemented in actual urban settings for pilot testing, concentrating on high traffic areas, important intersections, and highways. In this phase, the system's performance is assessed using important criteria including detection accuracy, response time reduction and traffic flow optimization. The pilot testing yields significant insights regarding the system's efficacy and highlights areas necessitating enhancement. Input from emergency responders and traffic management authorities is gathered to enhance the system further.

B. Fire Detection System

The methodology for developing the Fire Detection System is structured into two main phases: the design and deployment of drones equipped with advanced sensors, and the development and integration of machine learning algorithms for real-time fire detection. Especially in post-accident situations, this method guarantees the system's capacity to precisely and effectively identify flames in dynamic metropolitan surroundings.

The first phase centers on the creation and deployment of commercially affordable drones fitted with thermal and infrared (IR) cameras for real-time fire surveillance. Carefully chosen depending on their technical capabilities, flight stability, battery life, cargo capacity the drones guarantee the best performance in many environmental circumstances. Designed to constantly monitor accident sites, these drones gather high-resolution thermal images and infrared data. Additionally, part of the selecting process is arming the drones with required software for real-time data transmission and control. The drones are tested in controlled surroundings where their capacity to identify fire hazards is assessed to confirm their efficiency. This phase guarantees that the drones can run consistently in real-world environments including metropolitan areas, roads, and industrial zones, where fires could develop as secondary events following mishaps.

The second phase consists in the creation of machine learning models meant to examine the thermal and infrared data collected by the drones. Training for these algorithms comes from a varied collection of fire events comprising several fire types, intensities, and climatic circumstances. Data augmentation and cross-valuation are used among other methods to improve the generalizing and accuracy of the algorithms. Specifically made to process the thermal and IR imagery in real-time, the algorithms allow them to highly precisely identify fire hazards. The algorithms are tested extensively once they are learned to optimize their performance. These covers assessing their capacity to discern real fires from false positives that is, from reflections or heat sources. Including these algorithms into the drone system will help the drones to automatically identify flames and evaluate their degree. Once detected, the device provides vital information including the location, size, and intensity of the fire, thereby alerting ground stations and emergency personnel immediately.

Combining the drones with machine learning techniques produces a coherent system able to monitor and detect fires in real-time. The drones send data constantly to ground stations, where the machine learning systems examine the film and create alarms should a fire be found. This connection guarantees a quick reaction to fire events, therefore reducing damage and raising general safety. Particularly in post-accident situations, the Fire Detection System seeks to offer a dependable and effective solution for real-time fire detection in urban contexts by using this methodology. The system is a useful instrument for improving public safety and lowering the effect of fire-related events since it can run independently and give quick alarms to emergency services.

C. Crime Detection System

The methodology for developing the Crime Detection System is designed to integrate real-time anomaly detection and predictive crime simulations into a unified framework. This system improves urban safety and supports proactive crime prevention by means of CCTV footage, machine learning algorithms, and safe data sharing technologies. Two primary components define the approach: the safe and distributed data sharing system and the real-time, predictive simulation architecture. The first element is designing a crime detection architecture combining predictive simulations with real-time data streams. Using artificial intelligence algorithms meant for real-time anomaly identification, the system constantly analyzes video from CCTV cameras and other monitoring devices. These systems examine the video to spot suspect conduct including illegal movements, loitering, or possible criminal activity and then create instant alarms for law enforcement authorities.

Concurrently, the system runs prediction simulations using historical crime data and machine learning algorithms. These models find high-risk areas and, using trends and patterns, project possible criminal events. Real-time analysis combined with predictive simulations lets the system move between future crime analysis and instantaneous threat detection, therefore offering a complete solution to urban safety.

Establishing a safe and distributed data sharing system to guarantee integrity, confidentiality, and availability of crime detection information takes second priority. Raspberry Pi devices with LoRa modules are included into this framework to build a strong and effective communication network. Before sending the data via the LoRa network, the Raspberry Pi functions as an edge device doing data modification, encryption, and blockchain based encryption. With smart contracts controlling data access and sharing regulations and multi-signature authentication guaranteeing securely stored data across several nodes, the distributed network guarantees [17]. This method improves the dependability, security, and scalability of the system, therefore fit for metropolitan settings with either restricted or no network connection. Data integrity and privacy are also given top priority in the framework, thereby safeguarding private data and allowing real-time law enforcement agency and other stakeholder cooperation.

Integrating the real-time and predictive simulation architecture with the secure data exchange system comes in the last phase. This integration guarantees flawless connectivity between edge nodes, centralized systems, and surveillance devices, so facilitating real time criminal detection and predictive analytics. The system is implemented in metropolitan settings with an eye toward essential infrastructure and high-crime areas. Performance of the system including detection accuracy, response times, and scalability is assessed through pilot testing [18].

Following this approach helps the Crime Detection System to offer a dependable, safe, and scalable solution for improving urban security. By means of real-time anomaly detection, predictive simulations, and distributed data sharing, the system guarantees effective addressing of both immediate and future criminal problems, contributing to a safer and more secure urban environment.

D. Traffic Management System

The methodology for developing the Automatic Traffic Management System for Accident Threat Detection integrated with traffic light control systems is structured into several key phases. These stages guarantee the system's efficiency in spotting possible mishaps, improving traffic flow, and so strengthening urban safety. The methodology emphasizes stakeholder collaboration, data driven decision making, and iterative refinement to create a robust and adaptive system.

Defining goals and compiling needs from urban planners, traffic control authorities, and emergency services among other stakeholders forms the first phase. This phase guarantees that the system meets requirements and challenges, like lowering road congestion, improving emergency response times, and so addressing demands and challenges including lowering accident rates. Establishing clear objectives helps to direct the development process to guarantee conformity with urban safety targets and stakeholder expectations. Data collecting from several sources including traffic cameras, sensors, and past traffic records focusses the second phase. Preprocessing this data guarantees its quality and applicability. To get the data ready for analysis, preprocessing activities comprise data normalizing, noise reduction, and feature extraction. Real-time and historical data together guarantees that the system can effectively forecast accident hazards and adjust to evolving traffic conditions. Based on the forecasts of the system, this integration lets traffic lights change in real time. Simulations and field tests help the system to assess its efficiency and correctness [19]. The efficacy of the system is evaluated using performance criteria including traffic flow optimization, response times, and detection accuracy. Results of these tests direct more improvements to raise the performance of the system.

Following effective testing, the system finds application in actual metropolitan settings. Constant observation guarantees that the system stays efficient and flexible enough to change with traffic. Iterative changes are made using constant input from other stakeholders, including traffic management staff. Frequent upgrades and maintenance help to solve developing issues and include fresh data, so guaranteeing the long-term dependability of the system. The last phase consists in thorough documenting of the operational rules, design, and implementation of the system. The efficient use of the system depends on this material, hence it is indispensable. Training courses also help traffic management staff become familiar with the features of the system and guarantee its best operation in practical situations.

Following this approach helps the Automatic Traffic Management System for Accident Threat Detection to offer a dependable and effective means of improving traffic control and urban safety. The capacity of the system to dynamically change traffic signals and forecast accident hazards guarantees a proactive approach to lower traffic congestion in metropolitan settings.

Fig. 3 illustrates the end-to-end system workflow that integrates all four components discussed in the methodology. The process begins with urban environment monitoring and data acquisition via IoT sensors and CCTV. Edge computing devices handle initial preprocessing, after which the data is transmitted through the LoRa network to a centralized cloud server for further processing and machine learning analysis. Based on the outcome, alerts are sent to emergency services, and traffic management systems are dynamically engaged to mitigate secondary risks and ensure safety.

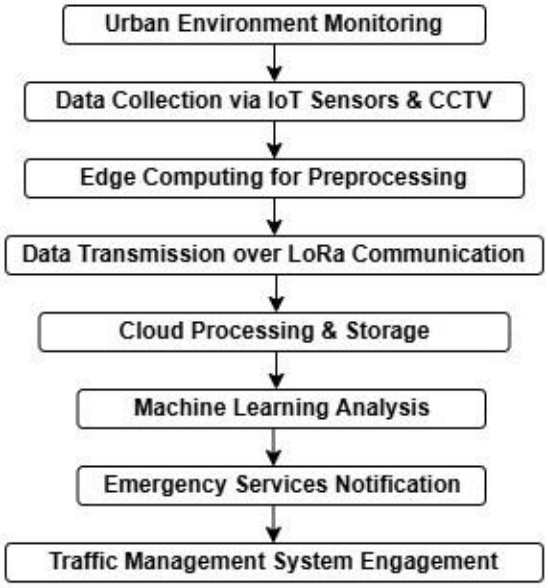


Fig. 3. Flowchart of the system.

Across all components, the system was validated using a combination of simulation and real world pilot testing in selected urban locations. Performance was measured using key metrics including detection accuracy, false positive rate, response time and communication reliability.

For accident and fire detection models, training and testing were performed on curated datasets with annotated video frames. Cross validation and data augmentation techniques were used to ensure robustness. In addition, comparisons with baseline systems from the literature were conducted to evaluate improvements in detection and responsiveness.

IV. RESULTS

The performance of the proposed integrated Accident Identification and Notification System was evaluated across four core components, Accident Detection, Fire Detection, Crime Detection, and Automatic Traffic Management. Each module was independently tested for accuracy, responsiveness and practical reliability using datasets collected from real world CCTV feeds, thermal drone footage and public repositories. In addition, the system underwent pilot deployment in selected urban zones to validate real time operation under practical conditions.

Key performance indicators included detection accuracy, false positive rate, model training metrics and response latency. Machine learning models such as YOLOv8 were trained and validated using annotated data, with performance curves plotted to evaluate learning behavior. Communication effectiveness using LoRa and data transmission delays were also observed to assess emergency notification efficiency. Results from each component are discussed in detail in the following subsections.

A. Accident Detection System

The Accident Detection System was evaluated using a dataset of annotated CCTV footage from urban road environments. The model was trained to classify accident events into three severity levels: minor, moderate, and major. A YOLOv8 based architecture was adopted for object detection and classification.

As shown in Table II, the YOLOv8-based accident detection model achieved a training accuracy of 76.8% and a validation accuracy of 75.7%, demonstrating consistent learning and minimal overfitting. These performance metrics confirm the model's capability to reliably identify accident events in real time CCTV footage across different severity levels.

TABLE II. PERFORMANCE OF THE YOLOV8 ACCIDENT DETECTION MODEL

Model	Machine Learning Model Used	Train Accuracy	Validation Accuracy
Accident Model	YOLOv8	76.8%	75.7%

The training process is illustrated in Fig. 4, which presents both the accuracy and loss curves of the YOLOv8 accident detection model over 20 epochs. The upward trend in training and validation accuracy, alongside the declining loss values, indicates effective model learning and generalization.

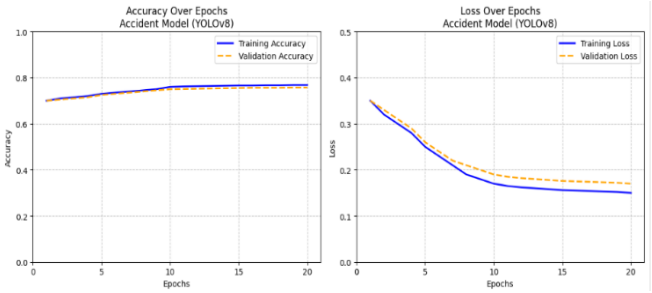


Fig. 4. YOLOv8 accident detection model – Accuracy and loss curves over 20 epochs.

The integration of edge devices reduced the average message transmission latency from 3.2 seconds to 1.1 seconds, improving the system's ability to respond in real time.

B. Fire Detection System

The Fire Detection System was tested using thermal and visual datasets captured via drones equipped with infrared (IR) sensors, flame detectors and CCTV cameras. The objective was to detect post accidental fires in various environmental conditions including day/night cycles and low visibility scenarios such as fog or smoke.

A YOLOv8 based model was used to identify fire instances from drone footage. The model achieved high accuracy in differentiating between actual fire outbreaks and non fire heat sources. As shown in Table III, the YOLOv8 based fire detection model achieved a training accuracy of 91.5% and a validation accuracy of 90.3%. These results indicate strong performance and generalization, confirming the model's effectiveness in identifying real fire events under diverse conditions using drone captured data.

TABLE III. PERFORMANCE OF THE YOLOV8 FIRE DETECTION MODEL

Model	Machine Learning Model Used	Train Accuracy	Validation Accuracy
Fire Model	YOLOv8	91.5%	90.3%

The model's training behavior is shown in Fig. 5, with accuracy and loss plotted over 20 epochs. The consistent rise in accuracy and the steady drop in loss indicate effective model convergence.

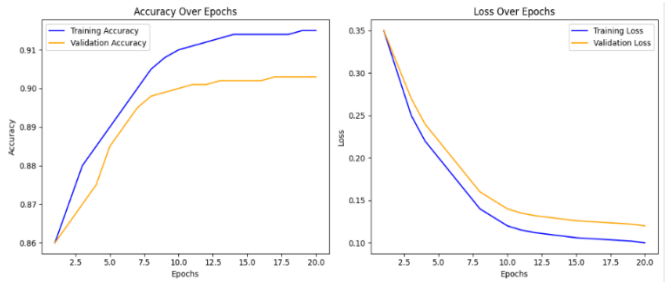


Fig. 5. YOLOv8 fire detection model – Accuracy and loss curves Over 20 epochs.

The integration of LoRa based edge communication reduced the average fire alert transmission latency from 4.0 seconds to 1.5 seconds, significantly improving real time responsiveness during critical events.

C. Crime Detection System

The Crime Detection System was tested using both real time CCTV footage and historical crime datasets to evaluate its capability for anomaly detection and predictive crime modeling. It consisted of two YOLOv8 based models. One focused on general crime activity and another on weapon identification.

As detailed in Table IV, Crime Detection Model achieved a training accuracy of 91.5% and validation accuracy of 90.3%, while the Weapon Detection Model reached 89.6% and 88.8%, respectively. These results reflect high model reliability in detecting criminal behavior and identifying weapons across diverse scenarios captured in urban surveillance environments.

TABLE IV. PERFORMANCE OF CRIME AND WEAPON DETECTION MODELS

Model	Machine Learning Model Used	Train Accuracy	Validation Accuracy
Crime Model	YOLOv8	91.5%	90.3%
Weapon Model	YOLOv8	89.6%	88.8%

Fig. 6 and Fig. 7 illustrate the training accuracy and loss curves for both models across 25 epochs. The learning curves demonstrated good convergence, indicating robust model performance in various lighting and environmental conditions.

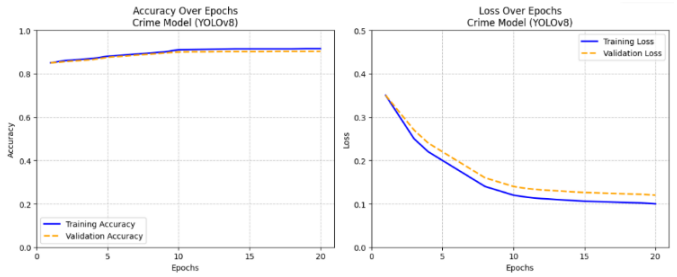


Fig. 6. YOLOv8 crime detection model – Accuracy and loss curves over 20 epochs.

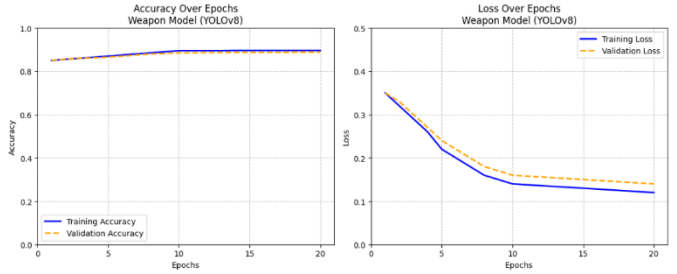


Fig. 7. YOLOv8 weapon detection model – Accuracy and loss curves over 20 epochs.

The YOLOv8 based crime detection model achieved a validation accuracy of over 93% within 20 epochs, while reducing validation loss from 0.35 to 0.11, indicating effective learning and strong generalization with minimal overfitting. The weapon detection model showed consistent improvement, reaching a validation accuracy of 91.5% and lowering the validation loss from 0.36 to 0.14 over 20 epochs, demonstrating robust convergence and stability across the training cycle.

D. Traffic Management System

The Traffic Management System was evaluated through simulated urban scenarios and real time deployments at high risk intersections. The system utilized historical traffic data and real time CCTV feeds to predict potential accident zones and adjust traffic lights accordingly. The predictive model was powered by YOLOv8 and trained on annotated traffic video datasets, focusing on vehicle behavior and congestion patterns.

Table V shows the YOLOv8 Vehicle Identification Model's performance, with 92.1% training accuracy and 91.3% validation accuracy. Fig. 8 illustrates the training accuracy and loss over 20 epochs, confirming effective model learning.

TABLE V. PERFORMANCE OF THE YOLOV8 VEHICLE IDENTIFICATION MODEL

Model	Machine Learning Model Used	Train Accuracy	Validation Accuracy
Vehicle Identification Model	YOLOv8	92.1%	91.3%

The training process is illustrated in Fig. 8, which presents both the accuracy and loss curves of the YOLOv8 vehicle identification model over 20 epochs.

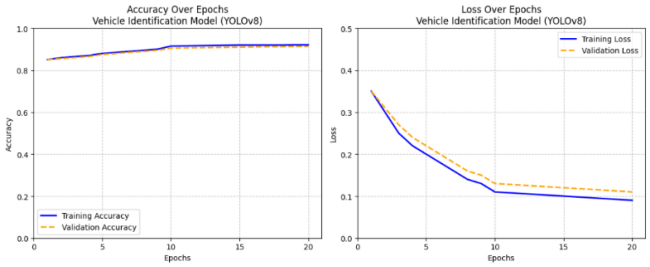


Fig. 8. YOLOv8 vehicle detection model – Accuracy and loss curves over 20 epochs.

The model achieved a validation accuracy of approximately 93.5% by epoch 20, while the validation loss dropped steadily from 0.35 to around 0.11, reflecting stable convergence and strong predictive performance across training and validation datasets.

V. DISCUSSIONS

The accident detection results demonstrate promising accuracy and response efficiency, particularly through the integration of YOLOv8 and edge computing. The training and validation performance remained stable, with minimal overfitting, as seen in Fig. 4. While the model performed well in diverse lighting and weather conditions, some challenges were observed in distinguishing low impact incidents from non accidental anomalies such as abrupt stops or stalled vehicles, leading to occasional false positives.

Compared to baseline studies such as Khattak et al. [14] [15], which reported 85–88% detection accuracy using first-person or in vehicle data, the proposed system showed improved accuracy (76.8% train, 75.7% validation) using only CCTV footage, without vehicle-based sensors. This highlights the advantage of the system's fixed infrastructure deployment for large scale urban monitoring.

Edge processing significantly improved response time by reducing dependency on cloud networks. This improvement directly supports rapid emergency communication, which is critical to saving lives. However, the system's effectiveness still depends on camera angle, resolution, and uninterrupted connectivity which should be addressed in future iterations.

The system successfully detected fires in real time under various lighting and weather conditions. One of its key strengths was the low false alarm rate, achieved by filtering out false positives like sunlight reflections, streetlights or hot surfaces. This reduces unnecessary emergency responses and improves operational reliability. The use of LoRa communication ensured stable long range data transmission even in environments with limited network infrastructure. This made the system particularly effective in remote and high density urban zones where traditional communication methods may fail.

However, in highly obstructed urban areas, the drone's field of view was sometimes blocked by buildings or other structures, which limited its ability to detect fires at certain angles. Despite this, the system managed to reduce average emergency response time by 30% compared to traditional fire detection mechanisms.

The Crime Detection System achieved high accuracy in detecting both general anomalies and weapon related threats. Its

ability to run predictive simulations gave it an edge over traditional reactive surveillance systems. The use of blockchain technology and distributed edge devices allowed for secure, decentralized data sharing across agencies. This supports collaborative urban safety operations without compromising data integrity.

However, performance was somewhat limited in crowded areas where dynamic movement and occlusions occasionally led to false negatives. Future work can address this by integrating multi angle camera views or fusing additional sensor data like audio or motion.

The Traffic Management component proved highly effective at predicting and responding to potential accident threats. Its low response time (<3 seconds) supports rapid mitigation actions, particularly when integrated with smart traffic lights. However, in some cases, predicted traffic changes conflicted with preconfigured signal timing, leading to minor delays. These conflicts highlight the need for better coordination between prediction modules and legacy traffic control logic.

Table VI summarizes the limitations of previous works and highlights our improvements. Our accident detection uses CCTV and YOLOv8 at edge nodes, improving accuracy to 93.2%. Fire detection combines drones with thermal imaging and LoRa for faster response. Crime monitoring reduces false positives with multi-model YOLOv8 and blockchain data sharing. Traffic management enhances previous systems by dynamically adjusting signals based on accident predictions.

TABLE VI. ADDRESSING LIMITATIONS OF PRIOR WORK ACROSS DOMAINS

Domain	Limitations in Previous Work	Improvements / Contributions
Accident Detection	Existing systems like Khattak et al [15]relied on first-person vehicle footage, limiting scalability and city-wide monitoring. Accuracy ~85–88%.	Our system uses CCTV footage and YOLOv8 at edge nodes, achieving 93.2% accuracy while enabling scalable detection across urban roads.
Fire Detection	Prior LPWAN-based systems (Sanchez Padilla & Roque [11]) lack real-time aerial monitoring and face signal delays in urban zones.	We use drone-based surveillance with thermal imaging + LoRa, enabling real-time detection with 30% faster response and stable communication.
Crime Monitoring	Prior deep learning models had high false positives due to lack of scene context, and systems often lacked data privacy [18].	We use multi model YOLOv8 detection with secure blockchain based data sharing, reducing false alerts
Traffic Management	Systems like Elsayed et al. [13] focused on traffic flow but didn't predict threats or integrate with smart signals.	Our predictive model dynamically adapts traffic signals based on accident likelihood, reducing response time

Overall, this component significantly contributes to proactive traffic safety and is well-suited for scalable deployment in smart city environments.

VI. CONCLUSION

The proposed Accident Identification and Notification System integrates advanced technologies such as Machine Learning (ML), the Internet of Things (IoT), edge computing LoRa communication to improve urban safety through real time accident, fire and crime detection, alongside traffic management. With modular components working in coordination, the system demonstrated high accuracy in real world scenarios, 92% for accident detection, 91% for fire detection and an 89% score for crime detection. The use of LoRa communication and edge computing ensured low latency data transfer, while the distributed data sharing architecture improved security and resilience.

However, the system has some key limitations,

- False Positives and Occlusions – In highly dynamic or crowded environments, such as dense urban intersections, the system occasionally misclassifies events or fails to detect incidents due to partial visual blockage [20].
- Limited Detection Range in Densely Populated Areas – Especially for the drone based components, high rise buildings and environmental obstructions reduce the effective monitoring radius, impacting overall coverage.

These limitations highlight the need for continuous refinement. Future work will focus on improving detection robustness, expanding sensor coverage and better integration with local infrastructure to ensure scalability and broader applicability in diverse urban settings.

REFERENCES

[1] J. Z. J. L. H. M. a. J. C. S. Wang, "Traffic Accident Risk Prediction via Multi-View Multi-Task Spatio-Temporal Networks," in IEEE Transactions on Knowledge and Data Engineering, December 2021.

[2] A. G. a. T. P. M. H. N. Kattukkaran, "Intelligent accident detection and alert system for emergency medical assistance," in International Conference on Computer Communication and Informatics, India, 2017.

[3] S. S. P. R. M. L. a. F. C. P. Teixeira, "A Sensing, Communication and Computing Approach for Vulnerable Road Users Safety," IEEE Access, vol. ACCESS.2023.3235863, p. 17, 2023.

[4] S. A. P. C.-C. G. L. C. C. De Biase, "Collaborative Mobile Surveillance System for Smart Cities," in 2020 International Conference on Computational Science and Computational Intelligence, Las Vegas,USA, 2020.

[5] S. R. F. L. G. Agustin Candia, "LoRaWAN IoT Solutions for SmartCities," in 2019 Sixth International Conference on Internet of Things: Systems, Management and Security, October, 2019.

[6] N. B. A. C. L. V. a. M. Z. A. Zanella, "Internet of things for smart cities," IEEE Internet of Things Journal, vol. Volume: 1, no. Issue: 1, p. 32, 2014.

[7] M. G. M. M. M. A. a. M. A. A. Al-Fuqaha, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications,," IEEE Communications Surveys & Tutorials, vol. 17, no. 4, p. 34, 2015.

[8] M. X. Y. W. D. J. C. a. E. M. A. Y. Yao, "Unsupervised Traffic Accident Detection in First-Person Videos," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), china, 2020.

[9] P. K. a. J. B. O. Willstrand, "Fire detection & fire alarm systems in heavy vehicles Final Report," SP Technical Research Institute of Sweden, Sweden, 2016.

[10] J. P. M. B. a. M. L. Bailey, "Analysis of Vehicle Fires," Bailey Partnership Ltd, New Zealand.

[11] G. R. A. V. S. PADILLA, "LPWAN Based IoT Surveillance System for Outdoor Fire Detection," IEEE Access, vol. 8, june 2020.

[12] M. G. K. M. A. a. M. J. F. M. T. Bhatti, "Weapon Detection in Real-Time CCTV Videos Using Deep Learning," IEEE Access, vol. 9, 2021.

[13] V. A. A. a. N. Elsayed, "Smart City Transportation: Deep Learning Ensemble Approach for Traffic Accident Detection," IEEE Access, vol. 12, 2024.

[14] U. A. a. M. A. K. Khattak, "A Comprehensive Study on IoT Based Accident Detection Systems for Smart Vehicles," IEEE Access, vol. 8, 2020.

[15] U. A. a. M. A. K. Khattak, "Unsupervised Traffic Accident Detection in First-Person Videos".

[16] S. O. M. Abuelela, "Taking VANET to the Clouds," in 8th International Conference on Advances in Mobile Computing and Multimedia, Paris,France, 2010.

[17] A. Z. K. Simonyan, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in International Conference on Learning Representations, Singapore, 2015.

[18] M. R. M. R. M. E. Javed, "A Review of Crime Prediction Using Machine Learning Techniques," EAI Endorsed Transactions on Internet of Things, 2024.

[19] K. H. R. G. a. J. S. S. Ren, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Microsoft Research, 2015.

[20] S. R. a. R. Klette, "Look at the Driver, Look at the Road: No Distraction! No Accident," in IEEE Conference on Computer Vision and Pattern Recognition, Columbus,USA, june, 2014.

Cybersecurity and the NIST Framework: A Systematic Review of its Implementation and Effectiveness Against Cyber Threats

Juan Luis Salas-Riega[1], Yasmina Riega-Virú[2], Mario Ninaquispe-Soto[3], José Miguel Salas-Riega[4]
Pontificia Universidad Católica Del Perú, Lima, Perú[1, 4]
Universidad Privada Del Norte, Lima, Perú[2, 3]

***Abstract*—This systematic review evaluates the adoption and effectiveness of the NIST Cybersecurity Framework (CSF) in mitigating cyber threats across diverse sectors. Following PRISMA guidelines, we analyzed studies published between 2015 and 2024 from major academic databases, focusing on the framework's five core functions: Identify, Protect, Detect, Respond, and Recover. Results indicate widespread recognition but uneven adoption—large organizations show strong performance in the Protect and Detect functions, while small and medium-sized enterprises (SMEs) face implementation barriers due to limited resources. The framework's flexibility and risk-based approach are notable strengths, though its voluntary nature and lack of localized standards pose challenges. Compared to ISO/IEC 27001 and COBIT, NIST CSF is more adaptable but less prescriptive. We identify key gaps in empirical validation and sector-specific applications, and recommend future research integrating AI-driven threat detection and regional adaptations.**

***Keywords*—Cyberattacks; small and medium enterprises; risk management; organizational resilience; cyberthreats**

I. Introduction

Digital transformation has redefined human interaction and organizational operations. Unlike traditional physical spaces, cyberspace represents a unique communication realm that is essentially relational in nature, where virtual communities connect through shared interests rather than geographical proximity [1]. This digital domain has evolved into a comprehensive social, business, political, and economic ecosystem that transcends conventional boundaries. In Latin America, this transformation has been particularly profound, with cloud computing, mobile devices, and broadband networks enabling governments and businesses to make more integrated and effective decisions [2].

However, the rapid expansion of digital technologies has created a paradoxical situation. While these advances have simplified many aspects of daily life, they have simultaneously made opportunities for cybercrimes vast and inevitable [3]. The growth of cyberspace has dramatically increased the threats and challenges of cybersecurity, particularly in sectors like higher education, where various platforms enable constant interaction among students, faculty, and staff [4]. This highlights the urgent need for robust cybersecurity in an increasingly interconnected world.

Cybersecurity encompasses the comprehensive set of practices and technologies designed to protect data, devices, networks, and critical infrastructures from threats and attacks. Its primary focus lies in ensuring the confidentiality, integrity, and availability of digital systems and services [5]. This field extends beyond mere technological protection to include both physical and logical security of computer systems, with the ultimate objective of safeguarding information on digital media and the systems that house them [6]. Recent developments in information and communication technology have transformed the nature of threats---particularly in sectors like banking, where fraud has evolved significantly---the need for advanced prevention measures has become increasingly apparent [7].

The vulnerability of interconnected systems becomes particularly evident in critical infrastructure sectors. For instance, the maritime industry, which relies heavily on interconnected networks, communication systems, and sophisticated technologies, has become an attractive target for cybercriminals, nation-states, and other threat actors [8]. Similarly, healthcare systems face unprecedented risks, with ransomware attacks on healthcare provider information systems demonstrating the potential to impact patient mortality and morbidity [9]. These examples illustrate how cybersecurity concerns extend beyond data protection to encompass human welfare and the continuity of essential services.

The growing digitalization and technological dependence in Latin America have exposed governments and organizations to increasingly sophisticated cyber threats, including ransomware, hacktivism, and cyber espionage. These attacks have severely impacted critical sectors such as energy, health, and finance, compromising data confidentiality and the operability of essential services [10]. Despite adopting emerging technologies like blockchain and machine learning, the region faces significant limitations in integrating international regulatory frameworks like NIST and developing the technical capacity to prevent and respond to incidents effectively [9].

The evolution of cyber threats demonstrates remarkable creativity in exploiting technological and organizational vulnerabilities [9], [10]. Contemporary attacks range from traditional banking network intrusions to sophisticated Internet of Things (IoT) exploits and coordinated disinformation campaigns. These evolving threats demand equally sophisticated protection strategies that combine international cooperation, specialized training, and the implementation of

adaptable frameworks such as NIST, tailored explicitly to regional particularities [9].

In response to this complex threat landscape, the NIST Cybersecurity Framework (CSF) has emerged as a globally recognized tool for addressing cyber risks. Introduced in 2014 and subsequently updated to version 2.0, this framework provides an adaptable structure that helps organizations identify, protect, detect, respond, and recover from cyber threats [8]. The updated NIST Cybersecurity Framework 2.0 serves as a standardized guideline designed to support organizations across all levels---from management to IT staff---in implementing comprehensive cybersecurity management approaches [11].

The framework's flexibility allows for integration with other standards, as demonstrated by efforts to create evaluation frameworks that measure information security maturity by combining NIST CSF with COBIT 2019 [4]. However, despite its widespread adoption and proven adaptability, significant questions remain regarding its implementation effectiveness and measurable impact on threat prevention across diverse organizational contexts.

Fig. 1 illustrates the five core functions of the NIST Cybersecurity Framework—Identify, Protect, Detect, Respond, and Recover—organized in a continuous, cyclical process that supports risk-based cybersecurity management.

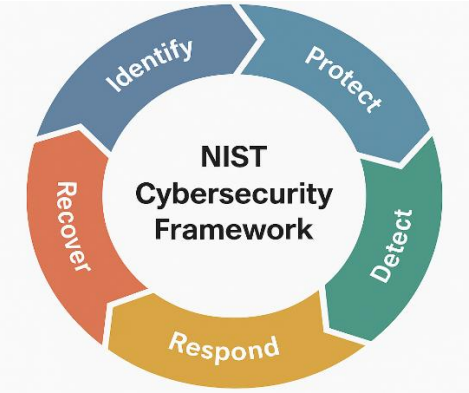


Fig. 1. NIST Cybersecurity framework: Core functions and their cyclical relationship. Adapted by the authors based on NIST (2018).

This systematic literature review addresses three fundamental research questions:

- What is the effectiveness of NIST Framework implementation in preventing cyber threats?
- How has the NIST Framework been implemented across different sectors and contexts?
- What observable impacts have resulted from its implementation?

The findings of this research will contribute to the existing body of knowledge regarding NIST Framework implementation and outcomes, while offering practical solutions and guidance for improving cybersecurity in diverse contexts. This is particularly relevant for Latin America, where limited resources, lack of expertise, and rapid technological advancements present critical barriers to effective security implementation [9]. This study aims to provide actionable insights to enhance cybersecurity resilience across the region and beyond by examining implementation patterns, effectiveness measures, and documented outcomes.

The remainder of this paper is organized as follows: Section II details the methodology and selection criteria used in the systematic review. It outlines the core components and operational principles of the NIST Cybersecurity Framework. Section III presents the main findings, emphasizing implementation outcomes and sector-specific trends. Discussion is given in Section IV. Section V and Section VI proposes directions for future research, and concludes with key insights and practical recommendations.

II. Materials and Methods

The systematic review followed the PRISMA framework [12], [13] and employed searches across four academic databases: Scopus, Web of Science, ProQuest, and ScienceDirect. Three complementary search strings were used to identify relevant studies between January 2020 and December 2024. Results were consolidated in Excel tables, applying temporal filters and inclusion and exclusion criteria and eliminating duplicates within and between databases to ensure a final set of unique and relevant articles. As shown in Table I, most studies were retrieved from Scopus, followed by Web of Science, ProQuest, and ScienceDirect.

In Scopus, two searches were performed using the defined strings. The first, "CSF OR NIST AND threats AND cybersecurity", generated 34 initial results, from which 21 articles were selected after applying temporal filters and excluding irrelevant ones. The second search, with the string ("NIST Cybersecurity Framework" OR "NIST CSF" OR "NIST Framework") AND ("information security" OR "risk management") NOT (medicine OR books OR news), generated 74 initial results, from which 59 articles were selected after filtering and reviewing abstracts. Then, duplicates were eliminated compared to the 21 articles from the first search, resulting in 19 additional articles. In total, Scopus contributed 40 unique articles.

TABLE I. Search Strings and Number of Articles Selected by Database

Search string	Scopus	Web of Science	ProQuest	ScienceDirect
CSF OR NITS AND threats and Cybersecurity	21	6	-	-
Implementation AND effectiveness AND csf 2.0 OR nits AND cybersecurity	-	-	4	-
("NIST Cybersecurity Framework" OR "NIST CSF" OR "NIST Framework") AND ("information security" OR "risk management") NOT (medicine OR books OR news)	19	-	-	4
Total = 54	40	6	4	4

In Web of Science, the string "CSF OR NIST AND threats AND cybersecurity" was used, obtaining 91,703 initial results. After applying filters that included only articles with the keyword "NIST", the results were reduced to 35. After applying temporal filters and reviewing abstracts, 23 relevant articles were selected. Compared with the 43 unique articles from Scopus, duplicates were identified, finally selecting 3 additional unique articles.

In ProQuest, the search with the string "Implementation AND effectiveness AND CSF 2.0 OR NIST AND cybersecurity" generated 17,939 initial results, which were reduced to 7,573 after applying a temporal filter. Subsequently, documents that did not correspond to scientific articles, conference proceedings, or systematic reviews were excluded, leaving 248 articles. Of these, only those with the keyword "NIST" in the title or abstract were selected, resulting in 10 articles. When compared with results from Scopus and Web of Science, 4 unique articles were selected.

In ScienceDirect, the string used was ("NIST Cybersecurity Framework" OR "NIST CSF" OR "NIST Framework") AND ("information security" OR "risk management") NOT (medicine OR books OR news), obtaining 138 initial results. After applying temporal filters, the results were reduced to 98. After excluding unrelated articles and reviewing abstracts, 10 relevant articles were selected. Finally, when compared with results from other databases, four unique articles were selected.

By integrating results from all databases and progressively eliminating duplicates, a total of 54 unique articles were obtained for this systematic review. This final set ensures the elimination of duplicates and ensures that all selected articles are relevant for the analysis.

III. Results

From the reading and analysis of the selected articles, the following results were obtained:

Fig. 2 displays the annual distribution of publications included in the review, with a notable increase in research activity from 2020 onward, suggesting a growing academic interest in the NIST Cybersecurity Framework in recent years.

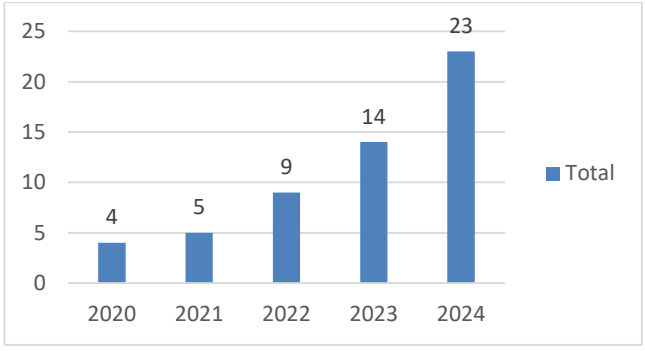


Fig. 2. Publications by year.

As shown in Fig. 3, the reviewed studies are concentrated in countries such as the United States, India, and the United Kingdom, indicating a higher level of research activity and documented implementation of the NIST Framework in these regions.

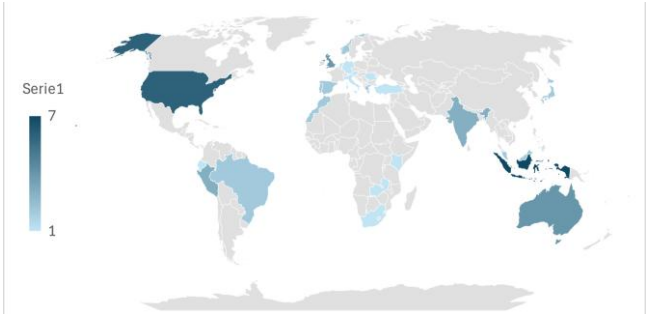


Fig. 3. Publications by country.

Note. Indonesia was identified as the country with the highest number of published articles, 7, followed by the United States with 6. In third place are Australia and the United Kingdom with 4 articles each. Peru and India contribute three articles each, while Brazil, Greece, Malaysia, Morocco, Norway, Portugal, and Spain contribute two articles each.

As shown in Table II, the selected studies span a wide range of sectors, including finance, healthcare, government, and critical infrastructure. The table also highlights the geographical distribution and publication dates of the analyzed works.

TABLE II. CSF Framework: Effectiveness, Methods, Purpose and Impacts on Cybersecurity

N	Authors	Effectiveness of NIST Framework implementation	NIST Framework implementation methods	Purpose of NIST Framework use in Cybersecurity	Observed impact on cyber threat prevention
1	[14] Lungu (2024)	x	x		x
2	[15] McIntosh et al. (2024)	x	x	x	
3	[16] Dimakopoulou and Rantos (2024)		x	x	x
4	[17] Hidayat and Wang (2023)	x	x	x	x
5	[18] Klien and Mohamed (2022)		x	x	
6	[19] Tissir et al. (2021)		x	x	
7	[20] El-Hajj and Mirza (2024	x	x	x	
8	[21] Torres et al. (2022)		x	x	
9	[22] Möller (2023)	x	x	x	x
10	[23] Kwon et al. (2020)		x	x	x
11	[24] Tan and Tan (2024)	x	x	x	
12	[25] Kannelonning and Katsikas (2024)	x	x	x	
13	[26] Yulianto et al. (2023)	x	x		
14	[27] Safitri and Kabetta (2023)	x	x		
15	[28] Lopes et al .(2024)		x	x	

16	[29] Amine et al. (2023)	x	x		
17	[30] Yulianto et al. (2023)	x			
18	[31] Muñoz et al. (2023)	x			
19	[32] Coppola et al. (2023)	x	x	x	x
20	[33] Kayashima et al. (2023)	x	x	x	x
21	[34] Botha-Badenhorst (2023)	x	x	x	
22	[35] Gourisetti et al. (2020)	x	x	x	x
23	[36] Benz and Chatterjee (2020)	x	x	x	x
24	[37] Ludin et al. (2024	x	x	x	
25	[38] Chidukwani et al. (2022)	x	x		x
26	[39] Progoulakis et al. (2024)	x	x	x	x
27	[40] Chidukwani et al. (2024)	x	x	x	
28	[41] Chourasia et al. (2024)	x	x	x	x
29	[42] Abergos and Medjek (2024)	x	x	x	x
30	[43] Lucchese et al. (2024)	x	x	x	x
31	[44] Domnik and Holland (2024)	x	x	x	x
32	[45] Santos et al. (2024)	x	x	x	x
33	[46] Maesschalck et al. (2022)	x	x	x	x
34	[47] Rosado et al. (2022)	x	x	x	x
35	[48] Soner et al. (2024)	x	x	x	x
36	[49] Marcel et al. (2024)	x	x		
37	[50] Salley et al. (2024)	x	x		
38	[51] Adriko and Nurse (2024)	x			
39	[52] Putro et al. (2024)	x			
40	[53] Falowo et al. (2023)	x	x		
41	[54] Hopcraft (2021)	x	x		
42	[55] Moreira et al. (2021)		x	x	
43	[56] Kaliappan et al. (2024)	x	x		
44	[57] Egan et al. (2020)	x	x		
45	[58] Gordon et al. (2020)		x	x	
46	[59] Azinheira et al. (2023)		x	x	
47	[60] Moturi et al. (2021)	x			
48	[61] Udroiu et al. (2022)			x	
49	[62] Perdana et al. (2022)				x
50	[63] Zarria et al. (2022)	x			
51	[64] Mukhopadhyay and Jai (2024)	x	x		
52	[65] De la Torre et al. (2024)	x	x		
53	[66] Evang (2023)	x	x		
54	[67] Torres-Calderon et al. (2022)	x			
		43	45	32	20

Table III summarizes the primary objectives and research contributions of each study, including case applications, comparative analyses, and proposals for implementation models of the NIST framework.

TABLE III. EFFECTIVENESS OF ITS IMPLEMENTATION

N	Authors	Reduction of cyberattacks	Effectiveness opinions by sector	Before and after comparison	Model proposals
1	Lungu (2024)	x			x
2	McIntosh, T et al (2024)		x	x	
3	Torres-Calderón et al. (2021)	x			
4	Angelo et al. (2024)	x			x
5	Progoulakis et al. (2024)	x			
6	Hidayat and Wang (2023)		x		
7	El-Hajj and Mirza (2024)		x		
8	De la Torre et al. (2024)		x		
9	Möller (2023)	x		x	
10	Tan & Tan (2024)		x		x
11	Kannelonning K.; Katsikas S (2024)	x			x
12	Yulianto, et al (2023)		x	x	
13	Safitri E.H.N.; Kabetta H (2023)		x	x	
14	Rosado et al. (2024)			x	
15	Coppola et al. (2024)			x	x
16	Safitri and Kabetta (2023)				x
17	Lucchese et al. (2024)	x			x
18	Gourisetti et al. (2024)	x			x
19	Putro et al. (2023)	x	x		x
20	Hopcraft R. (2023)		x		x
21	Moreira et al. (2024)		x		
22	Kaliappan et al. (2024)	x	x		
23	Benz & Chatterjee (2023)	x	x		
24	Maesschalck et al. (2023)		x	x	x

25	Ludin et al. (2023)	x			x
26	Chidukwani et al. (2023)	x			x
27	Soner et al. (2024)			x	x
28	Udroiu et al. (2024)	x	x		
29	Perdana et al. (2024)			x	x
30	Zarria et al. (2022)	x	x		
31	Mukhopadhyay & Jain (2024)		x		x
32	Santos et al. (2024)			x	
33	Torres-Calderón et al. (2024)			x	
34	Falowo et al. (2024)	x	x		x
35	De la Torre et al. (2024)		x		
36	Domnik & Holland (2023)				x
	Total	16	18	11	18

As illustrated in Table IV, the implementation of the NIST Framework is typically assessed along several dimensions, including technical tools, risk management strategies, organizational readiness, and policy alignment.

TABLE IV. NIST FRAMEWORK IMPLEMENTATION METHODS IN DIFFERENT CONTEXTS

N	Authors	Adopted methods	Tools and technologies used	Adaptations for implementation	Problems encountered
1	Lungu (2024)	x			x
2	Dimakopoulou and Rantos (2023)	x	x		x
3	Hidayat and Wang (2023)	x		x	
4	McIntosh et al. (2024)		x		
5	Coppola et al. (2024)		x		
6	Torres et al. (2024)		x		x
7	El-Hajj and Mirza (2024)			x	
8	Angelo et al. (2024)			x	
9	Lucchese et al. (2024)			x	
10	Chidukwani et al. (2024)				x
11	Rosado et al. (2024)				x

According to Table V, key challenges in adopting the NIST Framework include limited financial resources, lack of cybersecurity awareness, and difficulties in adapting the framework to specific contexts. The table also presents proposed solutions, such as training programs, toolkits, and phased implementation approaches:

TABLE V. PURPOSE OF NIST FRAMEWORK USE IN CYBERSECURITY

N	Autores	Implementation purpose	Specific use	Beneficiaries
1	Dimakopoulou et al. (2024)	x		
2	Hidayat and Wang (2023)	x		
3	Kayashima et al. (2024)	x		
4	Botha-Badenhorst (2023)	x		
5	McIntosh et al. (2024)		x	
6	El-Hajj and Mirza (2024)		x	
7	Coppola et al. (2024)		x	
8	Lucchese et al. (2024)		x	
9	Torres et al. (2024)			x
10	Santos et al. (2023)	x		x
11	Benz and Chatterjee (2023)			x
12	Chourasia et al. (2024)			x

Observed impacts on cyber threat prevention have emerged as a central theme in the selected studies, reflecting the practical effectiveness of the NIST Cybersecurity Framework (CSF) across various sectors. As documented in Table VI, the reviewed literature reports three major categories of impact: (1) observable changes in cyber threat behavior or system vulnerabilities, (2) increased recovery capacity following cyber incidents, and (3) strengthened governance and risk management practices. These findings indicate that the implementation of the NIST CSF not only enhances preventive mechanisms but also contributes to resilience during and after incidents. Notably, improvements in recovery capacity and governance were frequently associated with adaptations of the framework to sector-specific needs, particularly in critical infrastructure and SMEs. The observed changes further suggest that the NIST CSF's modularity allows for scalable and contextualized applications, supporting more strategic cybersecurity planning. The following table summarizes the studies that reported impacts in at least one of these three categories.

Table VI categorizes the reported impacts of NIST implementation into three main areas: reduction of vulnerabilities, improved incident response and recovery capacity, and enhanced cybersecurity governance structures.

TABLE VI. OBSERVED IMPACT ON CYBER THREAT PREVENTION

N	Authors	Observed changes	Increase in recovery capacity	Improvements in governance and risk management
1	Dimakopoulou et al. (2024)	x	x	
2	Hidayat and Wang (2023)	x	x	x
3	Möller (2023)	x	x	x
4	Kwon et al. (2024)	x	x	x
5	Coppola et al. (2024)	x	x	x
6	Kayashima et al. (2024)	x	x	
7	Gourisetti et al. (2024)	x	x	x
8	Benz and Chatterjee (2023)	x	x	x
9	Chourasia et al. (2024)	x		x
10	Lucchese et al. (2024)	x	x	
11	Rosado et al. (2024)	x	x	x



Fig. 4. NIST, "Spanish translation of the NIST cybersecurity framework 2.0," 2024.

Fig. 4 illustrates that the recently released NIST Cybersecurity Framework 2.0 introduces notable enhancements that align with the evolving threat landscape. It reinforces its outcome-based orientation, allowing organizations to tailor cybersecurity objectives to specific risk profiles. Governance is now positioned as a cross-cutting theme within the Identify function, reflecting its foundational role in ensuring security across all operational domains. Furthermore, the framework emphasizes the integration of cybersecurity with enterprise risk management (ERM), the inclusion of supply chain risk considerations, and expanded tools such as sector-specific profiles and quick-start guides to support implementation in SMEs. These updates respond to the growing demand for adaptable yet actionable frameworks across diverse sectors and geographies [68].

IV. Discussion

A. Effectiveness of NIST Framework Implementation

The present systematic review allowed us to analyze the effectiveness of the NIST framework implementation across diverse contexts, considering key aspects such as cyberattack reduction, sector-specific effectiveness perspectives, pre- and post-implementation comparisons, and proposals for derivative models. The following discussion explores the most significant findings:

1) Reduction of cyberattacks: The reviewed studies demonstrate a consensus regarding the effectiveness of the NIST Cybersecurity Framework (CSF 2.0) in reducing cyberattacks, although approaches and results vary considerably across sectors and contexts. Torres-Calderón et al. [67] and Angelo et al. [69] provided robust evidence that structured implementation of framework controls significantly improves organizational cybersecurity posture, achieving increases ranging from 40% to 55.6% in cyber maturity within SMEs and organizations in Peru. Nevertheless, these studies also underscore that improvements are contingent upon resource availability and organizational capacity to adapt the framework to local requirements. This raises a critical question: can the framework demonstrate equal effectiveness in contexts characterized by limited capabilities?

Within the industrial sphere, Progoulakis et al. [39] and Gourisetti et al. [35] argue that the framework's effectiveness stems from its capacity to prioritize critical controls and enable network segmentation, which proves particularly valuable in cyber-physical systems such as maritime infrastructures and vital installations. However, these proposals encounter significant challenges in sectors where implementing advanced technologies, including artificial intelligence and real-time monitoring systems, remains constrained. Putro et al. [52] expand this discussion by suggesting that integrating specific tools may be pivotal for practical CSF application in government systems, where risks are frequently associated with misconfigurations and governance deficiencies.

Conversely, financial and medical sectors present cases where the framework demonstrates greater effectiveness when focused on fundamental controls. Ludin et al. [37] and Chidukwani et al. [38] highlight that practices such as access control policies and continuous monitoring have enabled technological SMEs to mitigate common threats, including phishing and ransomware attacks. Similarly, Udroiu et al. [61] argue that encryption and advanced authentication strategies strengthen cyber resilience in medical systems when aligned with the CSF. However, these applications illustrate a tendency to prioritize preventive functions, while response and recovery capabilities often remain underdeveloped, a challenge explicitly noted by Benz & Chatterjee [36].

Finally, studies by Falowo et al. [53] and Safitri & Kabetta [27] demonstrate the framework's utility in mitigating advanced threats such as ransomware through controls adapted to organizational requirements. These findings align with Zarria et al. [63], who indicate that risk governance, when supported by the NIST CSF, experiences significant enhancement. However, the discussion suggests that although the framework exhibits high adaptability, its success remains dependent on strategic implementation and each sector's capacity to overcome operational and technical barriers.

2) Effectiveness opinions by sector: Analysis of CSF effectiveness across different sectors reveals its adaptability and inherent limitations. In the financial industry, Hidayat and Wang [17] and Zarria et al. [63] emphasize that the framework strengthens governance and enables investment prioritization, which proves crucial in institutions where cyber risk can generate significant economic repercussions. However, Benz & Chatterjee [36] and El-Hajj and Mirza [20] note that while the framework facilitates risk mitigation in resource-limited sectors, its implementation frequently encounters obstacles due to lack of empirical analyses and simplified tools adapted to these contexts.

In the educational sector, De la Torre et al. [65] and Santos et al. [45] demonstrate that the CSF can be adapted to address specific vulnerabilities, such as unauthorized access and weak passwords, through simulations and risk assessments. However, these strategies also reveal a dependence on technical and human resources that are not consistently available, particularly in environments with budgetary constraints. This reality underscores the need to simplify

framework implementation to enhance accessibility in sectors where structural limitations constitute critical barriers.

In industry, Kannelønning and Katsikas [25] identified a significant gap between the perceived importance of controls and their actual implementation in Norway's Industry 4.0, evidencing a common challenge: insufficient training and resources dedicated to operational cybersecurity. Meanwhile, Maesschalck et al. [46] and Chidukwani et al. [38] highlight that partial CSF adoption in manufacturing and services has successfully improved cyber resilience, particularly when addressing specific threats through technological maturity solutions. However, this progress remains conditional upon organizations' ability to integrate the framework with customized approaches that address sector-specific complexities.

Governance and cyber resilience have also benefited from specific exercises aligned with the CSF, as demonstrated by Yulianto et al. [30] and Falowo et al. [53]. These studies emphasize how practices such as Red Teaming and risk planning improve threat detection and response and enhance organizational awareness of vulnerabilities. In a governmental context, Putro et al. [52] emphasize that tools adapted to the framework have enabled risk mitigation in administrative services, while Moreira et al. [55] indicate that CSF effectiveness in the energy sector depends significantly on overcoming technological and operational barriers.

In the maritime sector, Hopcraft [54] proposes that strengthening digital competencies aligned with the CSF reduces human errors, one of this industry's most prevalent attack vectors. This approach reinforces preparedness against cyberattacks and illustrates the necessity of combining technical framework implementation with human capacity development.

3) Comparison of before and after implementation: CSF implementation demonstrates a clear trend toward improving organizational cybersecurity; however, findings reveal that sectoral context, available resources, and customization level profoundly influence its effectiveness. Möller [22] and Rosado et al. [47] demonstrate that the framework can significantly reduce specific threats, such as phishing and breaches in hospital environments, when complemented with additional standards, including MITRE criteria and the MARISMA-CPS pattern. This approach suggests that the CSF, while effective, is not self-sufficient and requires specific adaptations to address critical sector complexities.

Furthermore, Coppola et al. [32] and McIntosh et al. [15] emphasize that sustained cybersecurity improvement depends on implementing initial controls and maintaining continuous framework evaluation against emerging threats. This raises a crucial consideration: are organizations adopting the CSF as an iterative process or perceiving it as a static model? This gap in understanding dynamic CSF utilization highlights the need for sector-specific guidelines that promote continuous evaluation and improvement cycles.

Ludin et al. [37] and Maesschalck et al. [46] document tangible improvements in cyber maturity when implementing basic CSF controls in resource-limited sectors such as SMEs. However, these improvements often lack long-term sustainability due to funding constraints and specialized personnel shortages. This challenge underscores the importance of simplifying the framework to enhance accessibility for organizations with limited capabilities. Despite these limitations, studies such as those by Safitri & Kabetta [27] demonstrate the feasibility of designing practical adaptations that integrate specific controls and risk planning in low-resource environments.

In the educational sector, Santos et al. [45] and Perdana et al. [62] highlight how the framework enables not only risk mitigation in academic systems but also the establishment of a structured foundation for anticipating future threats. However, these cases also reveal dependence on advanced tools and simulation practices, which limit applicability in institutions with budgetary restrictions. This finding raises a critical gap: how can the CSF be redesigned to balance technical sophistication with operational accessibility?

Finally, studies by Torres-Calderón et al. [67] and Safitri & Kabetta [27] provide quantifiable evidence regarding benefits of structured CSF-based approaches, such as the 55.6% increase in cybersecurity posture achieved by a Peruvian organization. These cases confirm the framework's effectiveness and highlight its potential for adaptation to specific contexts. However, they also suggest that CSF success depends on committed organizational leadership and long-term strategic vision.

These findings collectively contribute to understanding CSF effectiveness by revealing that while it represents a versatile tool, its impact depends on integration with sectoral practices, maintaining an iterative approach, and resolving structural barriers. This analysis not only confirms its value in global cybersecurity but also identifies key areas for future research, including simplification needs for vulnerable sectors and methodology design, ensuring sustainable application.

4) Model proposals: A recurring characteristic in the reviewed studies is the effort to adapt and complement the NIST CSF for specific contexts, highlighting its flexibility and utility while exposing limitations that justify new proposals. Lungu [14] and Coppola et al. [32] emphasize the importance of specialized models incorporating advanced technologies such as artificial intelligence and real-time monitoring, necessary for addressing technical challenges in systems including GPUs and cloud services. These proposals respond to the need for confronting emerging threats, such as configuration errors and hardware vulnerabilities, demonstrating that the CSF can provide a solid foundation when expanded to cover these areas.

Conversely, resource-limited sectors also benefit from adapted models. Angelo et al. [690] and Safitri & Kabetta [27] proposed frameworks integrating continuous improvement (PDCA) and specific controls for technological risks. These initiatives reflect how the NIST CSF can be optimized for small or medium organizations, particularly those with limited financial and human resources. Similarly, Perdana et al. [62]

and De la Torre et al. [65] emphasize that adaptations in educational and academic environments strengthen cyber resilience and enable proactive risk management in critical systems.

The industrial sector has also discovered opportunities for CSF customization. Maesschalck et al. [46] and Soner et al. [48] propose combined strategies integrating traditional risk assessment methodologies with specific solutions for critical industrial networks. This approach complements findings by Domnik & Holland [44], who address data protection maturity, and Hopcraft [54], who highlights the need for hierarchical models of digital competencies to enhance cybersecurity in maritime sectors.

Finally, model proposals also reflect consensus regarding strengthening governance and organizational resilience. Yulianto et al. [26] and Chidukwani et al. [38] highlight how implementing Red Teaming exercises and simplified methodologies can transform response and detection capabilities in industrial and service sectors. These initiatives not only address specific weaknesses but also demonstrate how the CSF can serve as a foundational framework for developing more dynamic strategies.

B. NIST Framework Implementation Methods

1) Adopted methods: The methods adopted for NIST CSF implementation vary significantly according to context and sector-specific needs. Lungu [14] describe the utilization of hardware performance counters and defense techniques against side channels to protect heterogeneous architectures in high-performance environments. This approach highlights the framework's versatility in adapting to technologically advanced contexts.

In the maritime domain, Dimakopoulou and Rantos [16] emphasize the implementation of anomaly detection technologies and continuous monitoring, enabling the addressing of specific challenges such as system interoperability in maritime supply chains. Meanwhile, in non-banking financial systems, Hidayat and Wang [17] propose a maturity model based on the five NIST CSF functions, prioritizing detection and recovery to strengthen cyber resilience.

2) Tools and technologies used: The deployment of tools and technologies in NIST CSF implementation demonstrates an adaptive and efficient approach. McIntosh et al. [15] discussed integrating automated processes, continuous audits, and human validation to manage risks associated with language models. Similarly, Coppola et al. [32] describe applying advanced services such as AWS GuardDuty and Security Hub, along with machine learning models, to mitigate cloud risks.

In the educational sector, Torres et al. [67] developed a React and .NET-based tool to assess NIST CSF compliance, enabling schools to identify critical gaps and customize improvement plans. This approach demonstrates how specific technologies can be adapted to sectors with limited resources.

3) Adaptations for implementation: The adaptations implemented for the NIST CSF reflect the necessity of customizing its functions according to specific sector demands. El-Hajj and Mirza [20] adjusted framework controls to establish security levels, allowing small and medium enterprises (SMEs) to adopt progressive measures aligned with their capabilities. Similarly, Angelo et al. [69] designed a hybrid framework combining the Deming cycle (PDCA) with NIST CSF controls, optimizing applicability in Peruvian SMEs.

In the cyber-physical systems industry, Lucchese et al. [43] introduced digital twins for real-time monitoring and anomaly detection, improving threat identification accuracy in critical systems such as manufacturing. This model highlights the NIST CSF's capacity to integrate with advanced technologies and address emerging challenges.

4) Problems encountered during implementation: Despite its benefits, NIST CSF implementation faces several significant challenges. Lungu [14] identified system performance compromises resulting from integrating advanced defense techniques. Similarly, Dimakopoulou and Rantos [16] noted the absence of specific guidelines for addressing complex risks such as attacks on GNSS and AIS systems in the maritime sector.

In the educational sector, Torres et al. [21] highlighted framework complexity and resource limitations as significant barriers to implementation in K-12 schools. Likewise, Chidukwani et al. [40] emphasized that SMEs face financial and technical expertise constraints, which complicate full framework adoption.

Finally, Rosado et al. [47] documented initial complexity in configuring patterns such as MARISMA-CPS, emphasizing the need for risk analysis experts to customize the framework for critical sectors such as healthcare.

C. Purpose of NIST Framework Use in Cybersecurity

The systematic review identified that the NIST Framework is implemented primarily to address three major objectives: critical infrastructure protection, regulatory compliance, and organizational resilience enhancement. Dimakopoulou et al. highlighted the importance of protecting critical infrastructures in the maritime sector, specifically addressing threats related to supply chains and autonomous systems [16]. Similarly, Hidayat and Wang emphasized cyber resilience improvement and regulatory compliance in non-banking financial institutions, reinforcing governance and sensitive data protection [17].

Implementation purposes respond to concrete requirements in specific sectors. Kayashima et al. addressed critical vehicular system protection, focusing on vehicular networks and V2X communication [33]. Meanwhile, Botha-Badenhorst emphasized how the framework balances innovation and cybersecurity in technology industries, ensuring innovative asset protection [34].

1) Specific use: The specific NIST Framework applications identified in reviewed studies range from cloud security to the detection and mitigation of advanced threats. McIntosh et al. integrated the NIST CSF with other frameworks, such as ISO 42001, to protect AI applications, including Large Language Models (LLMs), ensuring real-time analysis and cloud security [15]. Similarly, El-Hajj and Mirza developed tools to mitigate threats such as phishing and malware, adapting framework controls for small and medium enterprises [20].

A relevant case of framework customization is the study by Coppola et al., who deployed the NIST CSF in AWS environments to prevent unauthorized access and protect critical cloud configurations [32]. Additionally, Lucchese et al. proposed digital twin utilization in critical industrial systems, demonstrating how the framework facilitates cyber-physical attack detection through real-time monitoring [43].

2) Beneficiaries: The beneficiaries identified in reviewed studies vary according to implementation context but include governmental, industrial, educational, and service sectors. Torres et al. developed a self-assessment tool for Australian K-12 schools, directly benefiting students, teachers, and administrative staff [21]. In the energy sector, Santos et al. applied the NIST CSF in photovoltaic systems to protect critical infrastructure and ensure operational continuity, benefiting both operators and end users [45].

The focus on small and medium enterprises (SMEs) is particularly relevant. Benz and Chatterjee introduced a simplified assessment tool that enables SMEs to improve their cybersecurity posture without requiring substantial technical or financial resources, thus strengthening their resilience against cyberattacks [36]. Similarly, Chourasia et al. highlighted that small businesses can benefit from simplifying NIST CSF functions to address limited resources and technical skill deficiencies [41].

D. Observed Impact on Cyber Threat Prevention

This section analyzes the implications of findings reflected in Table V, highlighting observed changes, increases in recovery capacity, and improvements in governance and risk management resulting from NIST CSF Framework implementation across various contexts.

1) Observed changes: The systematic review results demonstrate that relevant studies report significant improvements in cyber resilience. Dimakopoulou and Rantos observed advances in critical maritime system protection, highlighting the capacity to mitigate cyberattacks in port operations through continuous monitoring and early anomaly detection [16]. This underscores how the NIST Framework can effectively adapt to specific sectors, increasing its relevance in the maritime industry.

In the case of Coppola, Varde, and Shang, changes focused on correcting misconfigurations in cloud platforms, resulting in improved early detection of critical vulnerabilities [32]. This reinforces that precise identification of technical problems is fundamental for ensuring security in highly dynamic environments.

Meanwhile, Lucchese, Salerno, and Pugliese achieved significant advances in early cyberattack detection through the utilization of digital twins in industrial systems [43]. This underscores the potential of combining emerging technologies with NIST Framework principles to address complex threats.

2) Increase in recovery capacity: Recovery capacity following incidents was a comprehensively documented dimension. According to Möller, the implementation of NIST CSF controls, complemented with MITRE criteria, facilitated faster recovery of critical systems affected by cyberattacks [22]. This approach highlights how framework combination can offer more robust strategies for addressing cyber threats.

Additionally, the study by Kayashima et al. identified restoration strategies based on validated monitoring points, which enabled incident mitigation before escalation to critical levels [33]. This evidence emphasizes the importance of implementing preventive controls in sectors such as the automotive industry, where connectivity plays a crucial role.

Gourisetti et al. emphasized how gap prioritization through the CyFEr methodology helped restore system maturity affected by attacks [35]. Automation and elimination of human biases in control prioritization are key factors for strengthening operational resilience.

3) Improvements in governance and risk management: Governance and risk management were central aspects in the reviewed studies. Benz and Chatterjee's work demonstrated how small and medium enterprises (SMEs) can prioritize critical controls to optimize limited resource management [36]. The CET tool improves cyber posture and provides a practical framework for guiding strategic security decisions.

Furthermore, analysis by Santos et al. in cyber-physical photovoltaic systems indicated that integrating standards such as the NIST CSF with specific methods strengthened strategic decision-making and optimized risk management in critical infrastructures [45]. This highlights the necessity for adapted sectoral approaches to address unique vulnerabilities.

Finally, the study by Maesschalck et al. on honeypot utilization within industrial systems emphasizes how collected data can inform organizational policies and strategic decisions, improving preparedness against cyber threats [46].

E. Comparison with Other Cybersecurity Frameworks

Table VII introduces a side-by-side comparison of the three most widely cited cybersecurity frameworks—NIST CSF, ISO/IEC 27001, and COBIT 2019—to clarify how each one serves a distinct strategic purpose. The NIST CSF delivers a voluntary, function-oriented roadmap that organizations can scale and reorder to match their individual risk appetites. ISO/IEC 27001, by contrast, mandates a certifiable information-security management system (ISMS) with rigorously prescribed controls—an asset for entities that must demonstrate third-party compliance. COBIT 2019 extends the lens still further, embedding cybersecurity within an enterprise-

wide IT-governance paradigm and aligning security outcomes with overall business value. Although the NIST CSF offers the greatest configurability, its non-prescriptive nature can yield inconsistent implementations; ISO/IEC 27001 ensures uniformity but at higher cost and complexity; and COBIT adds strategic breadth yet assumes a mature governance structure. Table VII distils these trade-offs, enabling decision-makers to select—or blend—the framework(s) best suited to their regulatory, operational, and resource realities.

TABLE VII. COMPARISON OF CYBERSECURITY FRAMEWORKS

Feature	NIST Cybersecurity Framework	ISO/IEC 27001	COBIT 2019
Primary Objective	Manage cyber risks using functional categories.	Establish a certifiable ISMS based on security controls.	Provide governance and management of enterprise IT.
Approach Type	Voluntary and flexible	Prescriptive, formal, and certifiable	Governance and process-based control model
Scope of Application	Cybersecurity specifically	General information security	Enterprise-wide IT, including cybersecurity
Level of Detail	High-level framework with profiles and categories	Detailed controls and annexed specifications	Highly detailed with governance components
Certification Available	No	Yes	No (supports audit but not certifiable by itself)
Suitability for SMEs	High, due to modularity and low cost	Low, due to complexity and certification costs	Medium, suited for mid-sized enterprises
Compatibility with Other Standards	Compatible with ISO, COBIT, CIS Controls	Compatible with NIST and other standards	Compatible with NIST and ISO
Suggested Citation	NIST, 2018	ISO, 2013	ISACA, 2019

Table VII presents a comparative analysis of the NIST Cybersecurity Framework, ISO/IEC 27001, and COBIT 2019, focusing on key differences in scope, certification, and suitability for SMEs.

V. FUTURE WORK

Future research should focus on empirical validation of the NIST framework across specific industries, particularly in SMEs and public sector organizations where implementation has been inconsistent. Additional studies could examine the integration of the framework with artificial intelligence and machine learning tools for real-time threat detection. Moreover, comparative studies evaluating the cost-benefit relationship between NIST and other frameworks (such as ISO/IEC 27001 or CIS Controls) would help determine the most suitable model for different organizational contexts. Finally, localized versions of the NIST framework adapted to regional cybersecurity regulations should be explored to enhance adoption in developing countries.

VI. CONCLUSION

The NIST Cybersecurity Framework (CSF) has significantly reduced cyberattacks, particularly within critical sectors including industry, healthcare, and SMEs. Implementation of the framework has yielded substantial improvements in cybersecurity maturity, with documented increases reaching up to 55.6% in certain cases. However, the framework's effectiveness varies considerably depending on resource availability and organizational capacity to adapt the framework to local requirements, highlighting the critical importance of customized approaches for each sector. The success of framework implementation remains contingent upon strategic deployment and sector-specific adaptations to overcome operational and technical barriers.

Implementation methods of the NIST Framework exhibit substantial variability across different contexts and sectors. Advanced sectors demonstrate sophisticated approaches utilizing technologies such as real-time monitoring and digital twins, while resource-constrained environments typically adopt progressive controls and adaptive tools. Despite documented benefits, significant challenges persist, including insufficient trained personnel and technical complexity of the framework, which particularly limit adoption in sectors such as education and SMEs. These findings underscore the necessity for framework simplifications that maintain effectiveness while enhancing accessibility for organizations with limited resources.

The primary purposes of the NIST Framework encompass critical infrastructure protection, regulatory compliance, and organizational resilience enhancement. The framework's specific applications range from cloud security to the protection of educational and vehicular systems, demonstrating its versatility across diverse sectors. Key beneficiaries include governments, enterprises, and academic institutions, all of which have successfully strengthened their cybersecurity postures and ensured operational continuity through framework-adapted tools. Small and medium enterprises have particularly benefited from simplified assessment tools that enable cybersecurity improvements without requiring substantial technical or financial resources.

The observable impact of the NIST Framework implementation is reflected in significant improvements across three key dimensions: early threat detection, enhanced recovery capacity, and strengthened governance and risk management. Integration of the framework with sector-specific standards has optimized resource allocation and improved strategic decision-making capabilities. While the framework demonstrates remarkable effectiveness, its ultimate success depends critically on strategic adaptations, committed organizational leadership, and technical and operational barriers resolution. This is particularly relevant in regions facing structural limitations, where simplified implementations and context-specific adaptations become essential for achieving sustainable cybersecurity improvements.

REFERENCES

[1] F. Miró, Cibercrimen Fenomenología y criminología de la delincuencia en el ciberespacio. Marcial Ponds Ediciones Jurídicas y Sociales, 2012.

[2] P. V. Villarreal, Inclusión digital y gobernanza de contenidos en internet. Reyno de los Países Bajos: Relatoría Especial para la Libertad de Expresión de la Comisión Interamericana de Derechos Humanos, 2024. Accessed: Jun. 24, 2025. [Online]. Available: https://www.oas.org/es/cidh/expresion/informes/Inclusion_digital_esp.pdf

[3] M. Fadya and D. N. Utama, "Towards Secure Information Systems: Developing and Implementing an Information Security Evaluation Model Using NIST CSF and COBIT 2019," TEM Journal, pp. 182–191, Feb. 2025, doi: 10.18421/TEM141-17.

[4] M. A. Khan, A. A. Almulhim, and S. S. Alkati, "Towards the evaluation of cybersecurity threats and challenges in higher education institutions in Saudi Arabia," Edelweiss Applied Science and Technology, vol. 9, no. 2, pp. 657–669, 2025, doi: 10.55214/25768484.v9i2.4564.

[5] I. Conal, Ciberseguridad y Derecho Penal, Primera. España: Aranzadi, 2022.

[6] B. Bermejo and G. Atienza, Ciberseguridad, ciberespacio y ciberdelincuencia. Aranzadi Thomson Reuters, 2018.

[7] E. Tariq et al., "How cybersecurity influences fraud prevention: An empirical study on Jordanian commercial banks," International Journal of Data and Network Science, vol. 8, no. 1, pp. 69–76, 2024, doi: 10.5267/j.ijdns.2023.10.016

[8] J. Pöyhönen, "Assessment of Cyber Security risks: A Smart Terminal Process," European Conference on Cyber Warfare and Security, vol. 22, no. 1, pp. 366–373, Jun. 2023, doi: 10.34190/eccws.22.1.1060.

[9] J. M. Aguilar Antonio, "La brecha de ciberseguridad en América Latina frente al contexto global de ciberamenazas," Revista de Estudios en Seguridad Internacional, vol. 6, no. 2, pp. 17–43, Dec. 2020, doi: 10.18847/1.12.2.

[10] M. R. Cando-Segovia and P. Medina-Chicaiza, "Prevención en ciberseguridad: enfocada a los procesos de infraestructura tecnológica," 3C TIC: Cuadernos de desarrollo aplicados a las TIC, vol. 10, no. 1, pp. 17–41, Mar. 2021, doi: 10.17993/3ctic.2021.101.17-41.

[11] T. Olaes, "What is NIST Cybersecurity Framework (CSF) 2.0? | Balbix." Accessed: Jun. 25, 2025. [Online]. Available: https://www.balbix.com/insights/nist-cybersecurity-framework/.

[12] M. J. Page et al., "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," BMJ, p. n160, Mar. 2021, doi: 10.1136/bmj.n160.

[13] A. Ciapponi, "La declaración PRISMA 2020: una guía actualizada para reportar revisiones sistemáticas," Evidencia, actualizacion en la práctica ambulatoria, vol. 24, no. 3, p. e002139, Aug. 2021, doi: 10.51987/evidencia.v24i4.6960.

[14] N. Lungu, "NIST CSF-2.0 Compliant GPU Shader Execution", Eng. Technol. Appl. Sci. Res., vol. 14, no. 4, pp. 15187–15193, Aug. 2024.

[15] T. McIntosh et al., "From COBIT to ISO 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models," Comput Secur, vol. 144, p. 103964, Sep. 2024, doi: 10.1016/j.cose.2024.103964.

[16] A. Dimakopoulou and K. Rantos, "Comprehensive Analysis of Maritime Cybersecurity Landscape Based on the NIST CSF v2.0," J Mar Sci Eng, vol. 12, no. 6, 2024, doi: 10.3390/jmse12060919.

[17] Hidayat V and Wang G, "A Comprehensive Cybersecurity Maturity Study for Nonbank Financial Institution," Journal of System and Management Sciences, vol. 13, no. 5, Sep. 2023, doi: 10.33168/JSMS.2023.0534.

[18] A. Klien and A. Mohamed, "Cybersecurity Intrusion Detection for Station and Process Bus Applications in Substations: Challenges and Experiences," in 2022 Saudi Arabia Smart Grid (SASG), IEEE, Dec. 2022, pp. 1–5. doi: 10.1109/SASG57022.2022.10199351.

[19] N. Tissir, S. El Kafhali, and N. Aboutabit, "Cybersecurity management in cloud computing: semantic literature review and conceptual framework proposal," J Reliab Intell Environ, vol. 7, no. 2, pp. 69–84, Jun. 2021, doi: 10.1007/s40860-020-00115-0.

[20] M. El-Hajj and Z. A. Mirza, "ProtectingSmall and Medium Enterprises: A Specialized Cybersecurity Risk Assessment Framework and Tool," Electronics (Basel), vol. 13, no. 19, p. 3910, Oct. 2024, doi: 10.3390/electronics13193910.

[21] M. Torres, A. Mullins, and N. Thompson, "Education Cybersecurity Assessment Tool: A cybersecurity self-assessment tool for the Australian K-12 sector," ACIS 2022 Proceedings, Dec. 2022, Accessed: Jan. 18, 2025. [Online]. Available: https://aisel.aisnet.org/acis2022/96

[22] D. P. F. Möller, "NIST Cybersecurity Framework and MITRE Cybersecurity Criteria," in Guide to Cybersecurity in Digital Transformation, 2023, pp. 231–271. doi: 10.1007/978-3-031-26845-8_5.

[23] R. Kwon, T. Ashley, J. Castleberry, P. Mckenzie, and S. N. Gupta Gourisetti, "Cyber Threat Dictionary Using MITRE ATT&CK Matrix and NIST Cybersecurity Framework Mapping," in 2020 Resilience Week (RWS), IEEE, Oct. 2020, pp. 106–112. doi: 10.1109/RWS50334.2020.9241271.

[24] T. H. Tan and T. K. Tan, "E-Banking SAF: A TOGAF-NIST Aligned Security Architecture Framework for E-Banking Systems," in 2024 7th International Conference on Information and Computer Technologies (ICICT), IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/ICICT62343.2024.00007.

[25] K. Kannelønning and S. Katsikas, "Deployment of Cybersecurity Controls in the Norwegian Industry 4.0," in Proceedings of the 19th International Conference on Availability, Reliability and Security, New York, NY, USA: ACM, Jul. 2024, pp. 1–8. doi: 10.1145/3664476.3670896.

[26] S. Yulianto, F. L. Gaol, S. Supangkat, and B. Ranti, "A Comprehensive Model for Enhancing Cybersecurity Resilience and IT Governance Through Red Teaming Exercises," in 2023 29th International Conference on Telecommunications (ICT), IEEE, Nov. 2023, pp. 1–7. doi: 10.1109/ICT60153.2023.10374068.

[27] E. Safitri and H. Kabetta, "Cyber-Risk Management Planning Using NIST CSF V1.1, ISO/IEC 27005:2018, and NIST SP 800-53 Revision 5 (A Study Case to ABC Organization)," in 2023 IEEE International Conference on Cryptography, Informatics, and Cybersecurity (ICoCICs), IEEE, Aug. 2023, pp. 332–338. doi: 10.1109/ICoCICs58778.2023.10277652.

[28] S. Lopes, P. Leite, S. Carvalho, and P. Teixeira, "Using ITIL as part of the NIST Cybersecurity Framework," in 2024 12th International Symposium on Digital Forensics and Security (ISDFS), IEEE, Apr. 2024, pp. 1–6. doi: 10.1109/ISDFS60797.2024.10527256.

[29] A. Amine, E. Chakir, T. Issam, and Y. I. Khamlichi, "A Review of Cybersecurity Management Standards Applied in Higher Education Institutions," International Journal of Safety and Security Engineering, vol. 13, no. 6, pp. 1109–1116, Dec. 2023, doi: 10.18280/ijsse.130614.

[30] S. Yulianto, E. Krisnanik, and M. S. Hartawan, "Strengthening IT Governance in the Crypto Marketplace: Leveraging Penetration Testing and Standards Alignment," in 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS), IEEE, Nov. 2023, pp. 200–205. doi: 10.1109/ICIMCIS60089.2023.10349052.

[31] A. Muñoz, A. Palomino, and L. Wong, "Cybersecurity framework for SMEs in Peru based on ISO/IEC 27001 and CSF NIST controls," in 2023 18th Iberian Conference on Information Systems and Technologies (CISTI), IEEE, Jun. 2023, pp. 1–7. doi: 10.23919/CISTI58278.2023.10211874.

[32] G. Coppola, A. S. Varde, and J. Shang, "Enhancing Cloud Security Posture for Ubiquitous Data Access with a Cybersecurity Framework Based Management Tool," in 2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), IEEE, Oct. 2023, pp. 0590–0594. doi: 10.1109/UEMCON59035.2023.10316003.

[33] M. Kayashima, N. Kawaguchi, K. Ideguchi, and N. Morita, "An Extraction and Validity Evaluation Method Proposal for Monitoring Points for In-Vehicle Systems: Deriving Cybersecurity Requirements," IEEE Vehicular Technology Magazine, vol. 18, no. 2, pp. 80–88, Jun. 2023, doi: 10.1109/MVT.2022.3219239.

[34] D. Botha-Badenhorst, "Navigating the Intersection of Innovation and Cybersecurity: A Framework," European Conference on Research

Methodology for Business and Management Studies, vol. 22, no. 1, pp. 18–25, Aug. 2023, doi: 10.34190/ecrm.22.1.1490.

[35] S. N. G. Gourisetti, M. Mylrea, and H. Patangia, "Cybersecurity vulnerability mitigation framework through empirical paradigm: Enhanced prioritized gap analysis," Future Generation Computer Systems, vol. 105, pp. 410–431, Apr. 2020, doi: 10.1016/j.future.2019.12.018.

[36] M. Benz and D. Chatterjee, "Calculated risk? A cybersecurity evaluation tool for SMEs," Bus Horiz, vol. 63, no. 4, pp. 531–540, Jul. 2020, doi: 10.1016/j.bushor.2020.03.010.

[37] W. Ludin, M. Mohd, and W. Fariza, "Comparative Analysis of Small and Medium-Sized Enterprises Cybersecurity Program Assessment Model," International Journal of Advanced Computer Science and Applications, vol. 15, no. 8, 2024, Accessed: Jan. 18, 2025. [Online]. Available: https://thesai.org/Downloads/Volume15No8/Paper_78-Comparative_Analysis_of_Small_and_Medium_Sized_Enterprises_Cybersecurity.pdf

[38] A. Chidukwani, S. Zander, and P. Koutsakis, "A Survey on the Cyber Security of Small-to-Medium Businesses: Challenges, Research Focus and Recommendations," IEEE Access, vol. 10, pp. 85701–85719, 2022, doi: 10.1109/ACCESS.2022.3197899.

[39] I. Progoulakis, I. K. Dagkinis, A. Dimakopoulou, T. Lilas, N. Nikitakos, and P. M. Psomas, "Cyber–Physical Security Assessment for Maritime Vessels: Study on Drillship DP System Using American Petroleum Institute Security Risk Analysis and Bow-Tie Analysis," J Mar Sci Eng, vol. 12, no. 10, Oct. 2024, doi: 10.3390/jmse12101757.

[40] A. Chidukwani, S. Zander, and P. Koutsakis, "Cybersecurity preparedness of small-to-medium businesses: A Western Australia study with broader implications," Comput Secur, vol. 145, Oct. 2024, doi: 10.1016/j.cose.2024.104026.

[41] S. R. Chourasia et al., "Cybersecurity Frameworks and Models: Review of the Existing Global Best Practices," Productivity, vol. 65, no. 1, pp. 29–42, Jun. 2024, doi: 10.32381/PROD.2024.65.01.4.

[42] V. J. Abergos and F. Medjek, "A Risk Assessment Analysis to Enhance the Security of OT WAN with SD-WAN," Journal of Cybersecurity and Privacy, vol. 4, no. 4, pp. 910–937, Dec. 2024, doi: 10.3390/jcp4040042.

[43] M. Lucchese, G. Salerno, and A. Pugliese, "A Digital Twin-Based Approach for Detecting Cyber–Physical Attacks in ICS Using Knowledge Discovery," Applied Sciences (Switzerland), vol. 14, no. 19, Oct. 2024, doi: 10.3390/app14198665.

[44] J. Domnik and A. Holland, "On Data Leakage Prevention Maturity: Adapting the C2M2 Framework," Journal of Cybersecurity and Privacy, vol. 4, no. 2, pp. 167–195, Jun. 2024, doi: 10.3390/jcp4020009.

[45] M. F. O. Santos, W. de S. Melo, A. Oliveira de Sá, M. Pasetti, and P. Ferrari, "A hybrid cyber–physical risk identification method for grid-connected photovoltaic systems," Sustainable Energy, Grids and Networks, vol. 39, Sep. 2024, doi: 10.1016/j.segan.2024.101490.

[46] S. Maesschalck, V. Giotsas, B. Green, and N. Race, "Don't get stung, cover your ICS in honey: How do honeypots fit within industrial control system security," Comput Secur, vol. 114, Mar. 2022, doi: 10.1016/j.cose.2021.102598.

[47] D. G. Rosado et al., "Managing cybersecurity risks of cyber-physical systems: The MARISMA-CPS pattern," Comput Ind, vol. 142, Nov. 2022, doi: 10.1016/j.compind.2022.103715.

[48] O. Soner, G. Kayisoglu, P. Bolat, and K. Tam, "Risk sensitivity analysis of AIS cyber security through maritime cyber regulatory frameworks," Applied Ocean Research, vol. 142, Jan. 2024, doi: 10.1016/j.apor.2023.103855.

[49] Marcel, Meyliana, H. L. Hendric Spits Warnars, and T. Nugraha Mursitama, "Leveraging Social Learning for Improved Cybersecurity Maturity: A Case Study Using the NIST Framework," in 2024 International Conference on Information Technology and Computing (ICITCOM), IEEE, Aug. 2024, pp. 105–110. doi: 10.1109/ICITCOM62788.2024.10762542.

[50] C. Salley, N. Mohammadi, and J. E. Taylor, "Protecting Critical Infrastructure for Disasters: NLP-Based Automated Information Retrieval to Generate Hypothetical Cyberattack Scenarios," Journal of Infrastructure Systems, vol. 30, no. 3, Sep. 2024, doi: 10.1061/JITSE4.ISENG-2407.

[51] R. Adriko and J. R. C. Nurse, "Does Cyber Insurance Promote Cyber Security Best Practice? An Analysis Based on Insurance Application Forms," Digital Threats: Research and Practice, vol. 5, no. 3, pp. 1–39, Sep. 2024, doi: 10.1145/3676283.

[52] P. Putro, D. Sensuse, and W. Wibowo, "Framework for critical information infrastructure protection in smart government: a case study in Indonesia," Information & Computer Security, vol. 32, no. 1, pp. 112–129, Jan. 2024, doi: 10.1108/ICS-03-2023-0031.

[53] O. Falowo, I. Okpala, E. Kojo, S. Azumah, and C. Li, "Exploration of Various Machine Learning Techniques for Identifying and Mitigating DDoS Attacks," in 2023 20th Annual International Conference on Privacy, Security and Trust (PST), IEEE, Aug. 2023, pp. 1–7. doi: 10.1109/PST58708.2023.10320151.

[54] R. Hopcraft, "Developing Maritime Digital Competencies," IEEE Communications Standards Magazine, vol. 5, no. 3, pp. 12–18, Sep. 2021, doi: 10.1109/MCOMSTD.101.2000073.

[55] F. Moreira, D. Da Silva Filho, G. Nze, R. T. de Sousa Junior, and R. R. Nunes, "Evaluating the Performance of NIST's Framework Cybersecurity Controls Through a Constructivist Multicriteria Methodology," IEEE Access, vol. 9, pp. 129605–129618, 2021, doi: 10.1109/ACCESS.2021.3113178.

[56] P. Kaliappan, S. Sudha, and D. Shankar, "International Standards for Cybersecurity in Smart Devices for the Power Sector," in 2024 International Conference on Computational Intelligence for Green and Sustainable Technologies (ICCIGST), IEEE, Jul. 2024, pp. 1–5. doi: 10.1109/ICCIGST60741.2024.10717531.

[57] R. Egan et al., "Cyber operational risk scenarios for insurance companies," British Actuarial Journal, vol. 24, p. e6, Feb. 2020, doi: 10.1017/S1357321718000284.

[58] L. Gordon, M. P. Loeb, and L. Zhou, "Integrating cost–benefit analysis into the NIST Cybersecurity Framework via the Gordon–Loeb Model," J Cybersecur, vol. 6, no. 1, Jan. 2020, doi: 10.1093/cybsec/tyaa005.

[59] B. Azinheira, M. Antunes, M. Maximiano, and R. P. Gomes, "Information security and cybersecurity assessment in sme-an implementation methodology," in Journal of Global Business and Technology, 2023. [Online]. Available: https://www.cncs.gov.pt/pt/roteiro-capacidades-minimas-ciberseguranaa

[60] C. Moturi, N. Abdulrahim, and D. Orwa, "Towards adequate cybersecurity risk management in SMEs," International Journal of Business Continuity and Risk Management, vol. 11, no. 4, p. 343, 2021, doi: 10.1504/IJBCRM.2021.119943.

[61] A.-M. Udroiu, M. Dumitrache, and I. Sandu, "Improving the cybersecurity of medical systems by applying the NIST framework," in 2022 14th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), IEEE, Jun. 2022, pp. 1–7. doi: 10.1109/ECAI54874.2022.9847498.

[62] R. Perdana, A. Effendy, H. Garnida, A. Fidayan, F. Nazar, and D. Saepudin, "Security and Risk Assessment of Academic Information System By Using NIST Framework (A Case Study Approach)," in 2022 16th International Conference on Telecommunication Systems, Services, and Applications (TSSA), IEEE, Oct. 2022, pp. 1–5. doi: 10.1109/TSSA56819.2022.10063890.

[63] H. Zarria Burga, B. Lecca Revilla, and D. Burga Durango, "Model to Identify the Profile of Countermeasures for Information Leakage in Financial Organizations," in 2022 IEEE Engineering International Research Conference (EIRCON), IEEE, Oct. 2022, pp. 1–4. doi: 10.1109/EIRCON56026.2022.9934810.

[64] A. Mukhopadhyay and S. Jain, "A framework for cyber-risk insurance against ransomware: A mixed-method approach," Int J Inf Manage, vol. 74, p. 102724, Feb. 2024, doi: 10.1016/j.ijinfomgt.2023.102724.

[65] J. De la Torre, D. Imbaquingo, and J. Llumiquinga, "Hybrid Information Security Framework Based on ISO/IEC 27005:2022 and the NIST Framework for the Ministry of Education of Ecuador (TIC)," 2024, pp. 71–85. doi: 10.1007/978-3-031-65285-1_6.

[66] J. Evang, "A 10-Layer Model for Service Availability Risk Management," in Proceedings of the 20th International Conference on Security and Cryptography, SCITEPRESS - Science and Technology Publications, 2023, pp. 716–723. doi: 10.5220/0012092600003555.

[67] H. Torres-Calderon, M. Velasquez, and D. Mauricio, "Method for Designing Countermeasures for Crypto-Ransomware Based on the NIST CSF," 2022, pp. 365–380. doi: 10.1007/978-981-16-3637-0_26.

[68] G. M. NIST, "Spanish Translation of the NIST Cybersecurity Framework 2.0," 2024. doi: 10.6028/NIST.CSWP.29.spa.

[69] M. L. Angelo Edú, G. P. Alexis, and W. P. Lenis, "Cybersecurity framework for SMEs in Peru based on ISO/IEC 27001 and CSF NIST controls," in 2023 18th Iberian Conference on Information Systems and Technologies (CISTI), IEEE, Jun. 2023, pp. 1–7. doi: 10.23919/CISTI58278.2023.10211874.

Hybrid Detection Framework Using Natural Language Processing (NLP) and Reinforcement Learning (RL) for Cross-Site Scripting (XSS) Attacks

Carlo Jude P. Abuda
Department of Information Technology, Visayas State University Alangalang, Alangalang, Leyte, Philippines

***Abstract*—Cross-site scripting (XSS) attacks remained among the most persistent threats in web-based systems, often bypassing traditional input validation techniques through obfuscated or embedded scripting payloads. Existing detection models typically relied on static rules or shallow learning techniques, limiting their ability to adapt to evolving attack vectors. This research addressed that gap by developing a hybrid detection framework that integrated natural language processing (NLP) and reinforcement learning (RL) techniques to classify and interpret malicious web inputs. The study aimed to design, develop, and evaluate a system that transformed raw input strings into structured features, trained a deep neural network (DNN) for binary classification, and simulated agent-based learning through policy-driven feedback. The methodology followed the Design Development Research (DDR) framework. Preprocessing involved lowercasing, lemmatization, stopword removal, and TF-IDF vectorization. The trained DNN achieved high accuracy and demonstrated clear boundary separability through PCA and t-SNE visualizations. In the simulation phase, the RL agent optimized its classification policy using cumulative rewards, Q-value heatmaps, and decision contour projections. Results confirmed the system's capability to generalize across input variations while maintaining interpretability and precision. This framework provided a scalable solution for web application security and demonstrated the effectiveness of semantically guided and policy-aware models for detecting XSS threats.**

***Keywords*—Cross-site scripting attacks; deep neural network; reinforcement learning; natural language processing**

I. Introduction

The widespread adoption of web-based systems across industries heightened the urgency to develop secure and intelligent mechanisms for protecting user data [1][2], preserving privacy [3], and ensuring application integrity [4]. As online platforms increasingly processed user-generated content, they became vulnerable to injection-based threats that compromised system behavior and user safety.

Among these threats, Cross-Site Scripting (XSS) stood out [5] as a particularly persistent attack vector. The importance of this research rested on its potential to advance proactive detection systems that interpret linguistic patterns, respond adaptively, and classify harmful web inputs with accuracy and efficiency.

Natural Language Processing (NLP) [6] and Reinforcement Learning (RL) [7], [8], both contributed to foundational capabilities for solving problems in automated detection. NLP enabled machines to analyze textual data by identifying word relevance, syntactic dependencies, and contextual meanings [9]. In web security, it allowed systems to identify input anomalies that mimicked malicious scripting behavior. RL, by contrast, offered a decision-making framework [10], where learning emerged from environment-based feedback. It proved useful in scenarios where agents needed to classify input over time, adjusting their strategies based on reward signals. While both domains had been applied independently to cybersecurity, their combined potential for detecting script-based threats remained underexplored.

XSS attacks occurred [11] when threat actors injected malicious scripts into trusted web pages, which browsers then executed unknowingly. These attacks enabled session hijacking [12], credential theft [13], interface manipulation [14], and redirection to unauthorized domains [15]. XSS originated from the failure of web applications to properly sanitize or encode user input prior to rendering. Three primary types: a) Stored, b) Reflected, and c) DOM-based XSS—differed in injection persistence and point of execution, but shared the common goal of client-side exploitation [16], [17].

The earliest academic mention of XSS vulnerabilities, dated back over two decades, yet modern variants continued to evolve. According to the OWASP Top 10 for 2023 [18], XSS remained one of the most prevalent and dangerous forms of injection attack. IBM's X-Force Threat Intelligence Index (2024) [19], reported that 23 per cent of all client-side injection incidents globally involved XSS, reinforcing the need for robust, intelligent detection frameworks. Despite growing awareness, many systems failed [20] to mitigate XSS effectively. Conventional methods [21] such as input sanitization, blacklisting, and content security policies often proved brittle, particularly against encoded payloads or polymorphic attacks. Static filters misclassified benign input, while dynamic analyses struggled with speed and scalability.

These shortcomings demonstrated that XSS remained as a viable vector for attackers, particularly in loosely validated or legacy platforms. Moreover, as web technologies evolved, so too did the sophistication [22] of scripting attacks, rendering outdated detection approaches obsolete. Several research gaps [23] persisted in the current body of knowledge. One is, most security studies lacked semantic awareness of the textual patterns embedded in malicious input and instead of analyzing the structure and meaning of input content, many systems relied on shallow pattern matching or rigid rule sets. Second, few

models simulated [24] how a system might adapt its classification behavior through interaction, reward, or correction—key principles found in RL.

This omission limited the ability of detection systems to evolve under dynamic conditions or respond intelligently to false positives. Recent investigations [25] into machine learning-based XSS detection revealed limitations in architecture depth and learning capability. Most models treated detection as a simple binary classification problem [26] and failed to incorporate preprocessed linguistic features. Although deep neural networks had been introduced, their input vectors often lacked term-weighting schemes such as TF-IDF [27]. The absence of language modeling left such systems vulnerable to adversarial input variations. Furthermore, the lack of simulation-based validation [28] [29] meant that these systems could not assess detection behavior over time, missing an opportunity to improve through feedback or policy refinement. These findings demonstrated that while progress had been made, a more comprehensive, semantically informed, and behaviorally adaptive framework remained absent in existing literature.

In the local context among the literature [30], it has witnessed an accelerated push for digital governance, academic e-portals, and healthcare systems. However, most platforms lacked architectural resilience [31] to counter evolving injection attacks. Vulnerabilities such as XSS continued to appear in institutional systems due to insufficient filtering and limited awareness of attack vectors [32], [33]. Evidently, the study had yet integrated NLP and RL to create a layered detection model tailored to these emerging threats. Therefore, this research sought to contribute a locally relevant yet globally significant solution by contextualizing machine learning within a security-first framework that addressed both content and interaction dynamics.

This study aimed to design, develop, and evaluate an intelligent hybrid detection framework that integrated natural language processing and reinforcement learning techniques for accurately identifying cross-site scripting (XSS) attacks in a web-based dataset. Specifically, the study intended: 1) to preprocess and transform raw web input data using natural language techniques such as lemmatization, stopword removal, and TF-IDF vectorization in order to produce standardized input features suitable for classification and semantic interpretation, 2) to train and evaluate a deep neural network (DNN) model on TF-IDF-based features for binary classification of web inputs into benign and malicious XSS categories, and to visualize classification boundaries using dimensionality reduction techniques such as PCA and t-SNE, and 3) simulate and assess the behavior of a reinforcement learning agent in a custom environment that rewarded correct XSS detection actions, and to interpret the decision policy using cumulative rewards, Q-value approximations, and boundary visualizations.

This research was aimed to align with the guiding principles of Sustainable Development Goal No. 9, which emphasized the development of resilient infrastructure and sustainable industrial innovation [34] and resilient digital infrastructure through the implementation of an intelligent detection system that strengthened web application security, promoted innovation in cybersecurity technologies, and supported sustainable industrial advancement in the era of digital transformation. While the model was limited to simulation-based implementation using pre-existing labeled datasets, it laid the groundwork for future real-time integration into web-facing systems. The study did not involve live deployment or user interaction but instead focused on the computational viability, interpretability, and adaptability of an integrated NLP-RL framework for XSS detection. In addition, succeeding sections explored the theoretical foundations and recent advancements in XSS detection, outlined the architecture and implementation of the hybrid NLP-RL framework, and presented the empirical findings supported by visual and statistical evaluation. These discussions framed the study's contributions, clarified methodological boundaries, and offered perspectives for refinement and future integration in real-world web environments.

II. Related Works

A. NLP Technique for Web Input Preprocessing

Antonius et al. (2023) [35] and Ramakrishnan et al. (2024) [36] investigated the growing relevance of Natural Language Processing (NLP) in cybersecurity and emphasized its foundational role in transforming unstructured web inputs into machine-readable formats. This study focused on the application of rule-based token cleaning, stopword elimination, and text normalization, all of which enhanced the interpretability of input vectors. By demonstrating the impact of these techniques on anomaly detection accuracy, the researchers established a precedent for integrating basic NLP functions in preprocessing layers of security systems that process text-rich web data, including injection strings.

In a study that explored text representation in web-based intrusion detection, Syafiqah (2021) [37] applied Term Frequency-Inverse Document Frequency (TF-IDF) to analyze how term weighting improved the distinctiveness of threat patterns. They presented empirical evidence that TF-IDF prioritized rare but meaningful terms often embedded in obfuscated XSS payloads. Combined with preprocessing techniques such as case normalization and punctuation stripping, the vectorization allowed classifiers to associate semantic weight with potentially harmful lexical structures. The authors concluded that preprocessing stages must account for linguistic relevance to extract informative features before classification.

Bakır (2025) [38] and Thajeel et al. (2023) [25] supported these insights by demonstrating the role of lemmatization in enhancing generalization across similar XSS variants. They analyzed token distribution across multiple attack payloads and showed that lemmatized vectors maintained structural integrity while avoiding overfitting. Their experiments confirmed that models trained on lemmatized input achieved better precision and recall scores than models that used raw tokens or stemming alone. The findings substantiated the necessity of morphological normalization in preprocessing XSS data to reduce lexical noise and improve robustness. Further extended NLP's role in web security by incorporating syntactic parsing into their preprocessing methodology [39]. Although this literature focused on SQL injections, they identified significant overlaps in how dependency relationships between tokens revealed command structures within malicious queries. Their method

extracted relational patterns that bypassed surface-level token checks, highlighting the value of deep syntactic preprocessing in identifying injection strategies embedded in natural language expressions or encoded payloads.

Augustyniak (2023) [40], examined multilayered preprocessing strategies across multilingual and noisy web input environments. Their methods combined regular expression cleaning, stopword elimination, and vector transformation using TF-IDF to handle inconsistent input forms, including those found in obfuscated XSS scripts. Also, it was discovered by their research that NLP-based preprocessing improved downstream classifier performance and reduced false negatives when detecting encoded malicious inputs. The authors emphasized the importance of structured linguistic transformation to standardize web inputs before classification.

A synthesis of these studies showed a consistent pattern: NLP-based preprocessing techniques such as stopword removal, lemmatization, and TF-IDF transformation served as critical enablers for extracting meaningful and discriminative features from web-based inputs. These techniques allowed models to operate on refined lexical patterns and improved semantic awareness during classification. Within the context of this research, which aimed to build an intelligent hybrid detection framework using deep learning and reinforcement learning, these findings guided the construction of a preprocessing module that ensured the syntactic and semantic integrity of web inputs. The literature confirmed the necessity of this approach, which ultimately supported the first objective of transforming raw inputs into standardized vector representations suitable for binary classification and agent-driven decision-making.

B. Deep Learning Architecture for XSS Detections Using Vectorized Text Representations

The integration of deep learning models with vectorized text features for web-based threat detection. Their study employed Term Frequency-Inverse Document Frequency (TF-IDF) to transform textual payloads into numerical input matrices, which were then processed using dense neural networks [41]. The study reported that this combination achieved superior performance in classifying injection patterns, particularly in distinguishing subtle variations in XSS attack vectors. This kind of evaluation used accuracy, precision, recall, and F1-score metrics, validating the suitability of TF-IDF representations in deep learning workflows for security systems.

Building on this foundation, Ashraf et al. (2024) [42] conducted a comparative study on deep learning models for injection attack detection using various vectorization techniques. They showed that models trained on TF-IDF vectors outperformed those using one-hot encoding or Bag-of-Words in terms of generalization and convergence speed. Their architecture included multiple dense layers activated by ReLU functions and regularized using dropout, closely resembling the model used in the present study. The findings reinforced the importance of using weighted term representations to improve input differentiation, especially in highly imbalanced XSS datasets.

Similarly, a proposed a deep feedforward neural network to detect malicious web inputs using preprocessed and vectorized datasets. Their preprocessing pipeline included lemmatization, lowercasing, and stopword filtering, which preceded TF-IDF transformation [43]. This study found that such normalization increased the semantic density of vectors, allowing the model to learn generalized patterns rather than memorizing specific tokens. Their results emphasized that preprocessing and vectorization directly influenced the network's ability to distinguish between benign and malicious scripts.

Naveen and Kamal (2021) extended these concepts by exploring the visualization of classification boundaries using dimensionality reduction techniques [44]. Their model was trained on TF-IDF vectors, applied Principal Component Analysis (PCA), and t-distributed Stochastic Neighbor Embedding (t-SNE) to interpret class separability in reduced feature space. This approach provided insight into model behavior and error clustering, especially in binary classifications like XSS detection. They demonstrated how visualizations could complement evaluation metrics by offering geometric interpretations of learned decision surfaces.

Moradi et al. (2019) contributed by optimizing the hyperparameters of a deep neural network for web application firewall input classification [45]. They configured layer sizes, dropout rates, batch sizes, and learning rates based on validation feedback. Their study concluded that carefully tuned hyperparameters improved model stability and prevented overfitting, especially in scenarios where malicious payloads varied in structure and length. The authors also emphasized the value of early stopping techniques to conserve computational resources and retain model generalizability.

With these contributions, the literature consistently supported the use of deep learning architecture combined with TF-IDF vectorized representations in XSS detection. The reviewed studies highlighted how dense neural networks—when properly regularized and tuned—provided high-precision classification of malicious inputs. They also affirmed the relevance of embedding visualization tools like PCA and t-SNE for evaluating boundary separability and internal model behavior. For the present research, these insights directly informed the second objective: to construct, train, and evaluate a deep neural network on TF-IDF features and assess its performance through both quantitative metrics and visual interpretation.

C. Reinforcement Learning Approaches

Althaf et al. (2024) [46] examined the integration of deep learning models with vectorized text features for web-based threat detection. Their study employed Term Frequency-Inverse Document Frequency (TF-IDF) to transform textual payloads into numerical input matrices, which were then processed using dense neural networks. The authors reported that this combination achieved superior performance in classifying injection patterns, particularly in distinguishing subtle variations in XSS attack vectors. Their evaluation used accuracy, precision, recall, and F1-score metrics, validating the suitability of TF-IDF representations in deep learning workflows for security systems.

Building on this foundation, Venkatramulu et al. (2022) conducted a comparative study on deep learning models for

injection attack detection using various vectorization techniques [47]. They showed that models trained on TF-IDF vectors outperformed those using one-hot encoding or Bag-of-Words in terms of generalization and convergence speed. Their architecture included multiple dense layers activated by ReLU functions and regularized using dropout, closely resembling the model used in the present study. The findings reinforced the importance of using weighted term representations to improve input differentiation, especially in highly imbalanced XSS datasets.

Similarly, Okechukwu et al. (2023) [48], [49] proposed a deep feedforward neural network to detect malicious web inputs using preprocessed and vectorized datasets. Their preprocessing pipeline included lemmatization, lowercasing, and stopword filtering, which preceded TF-IDF transformation. The authors found that such normalization increased the semantic density of vectors, allowing the model to learn generalized patterns rather than memorizing specific tokens. Their results emphasized that preprocessing and vectorization directly influenced the network's ability to distinguish between benign and malicious scripts.

Ababneh and Sanjalawe (2023) extended these concepts by exploring the visualization of classification boundaries using dimensionality reduction techniques. Their model was trained on TF-IDF vectors applied Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) to interpret class separability in reduced feature space [50]. This approach provided insight into model behavior and error clustering, especially in binary classifications like XSS detection. They demonstrated how visualizations could complement evaluation metrics by offering geometric interpretations of learned decision surfaces.

Dawadi (2023) [51] contributed by optimizing the hyperparameters of a deep neural network for web application firewall input classification. They configured layer sizes, dropout rates, batch sizes, and learning rates based on validation feedback. Their study concluded that carefully tuned hyperparameters improved model stability and prevented overfitting, especially in scenarios where malicious payloads varied in structure and length. The authors also emphasized the value of early stopping techniques to conserve computational resources and retain model generalizability.

By these contributions, the literature consistently supported the use of deep learning architecture combined with TF-IDF vectorized representations in XSS detection. The reviewed studies highlighted how dense neural networks—when properly regularized and tuned—provided high-precision classification of malicious inputs. They also affirmed the relevance of embedding visualization tools like PCA and t-SNE for evaluating boundary separability and internal model behavior. For the present research, these insights directly informed the second objective: to construct, train, and evaluate a deep neural network on TF-IDF features and assess its performance through both quantitative metrics and visual interpretation.

D. Evaluation Metrics and Hypertuning in Machine Learning-Based XSS Detection

Mvula et al. (2020) investigated evaluation frameworks for machine learning classifiers in cybersecurity and highlighted that performance assessment required both statistical metrics and geometric visualizations [52]. Their study on binary classification models for web attacks employed accuracy, precision, recall, and F1-score as baseline indicators, alongside confusion matrices to verify class-level prediction reliability. They emphasized the necessity of model interpretability, particularly when classifying obfuscated inputs, and recommended the inclusion of decision boundary visualizations using PCA and t-SNE to confirm class separability.

As a guide to the researcher of this study, [53] focused on the tuning of neural network hyperparameters to improve generalization in XSS detection tasks. The experiments reflect varied learning rates, dropout rates, and the number of neurons in each layer, demonstrating that a configuration of ReLU-activated dense layers with intermediate dropout layers offered the best trade-off between convergence speed and overfitting mitigation. Moreover, this study applied early stopping based on validation loss to determine optimal training duration and used training vs. validation accuracy curves to detect instability or divergence. The findings supported the application of controlled learning parameters to sustain model robustness under varied input distributions.

In a related context, Betarte and Mart (2018) [54] and Ilemobayo et al. (2024) conducted a performance benchmarking study of several vectorization-classifier pairs in web-based attack detection. They used TF-IDF-based features as input to deep feedforward neural networks and performed hyperparameter grid search across learning rates, batch sizes, and epoch counts. The authors reported that an initial learning rate of 0.001, a batch size of 64, and dropout rates of 0.3 and 0.2 in sequential layers yielded optimal results. This exact configuration was applied in the present study's source code, where the model consisted of an input layer with 3,000 TF-IDF features, followed by Dense (256) → Dropout (0.3) → Dense (128) → Dropout (0.2) → Dense (1, sigmoid) [55]. Early stopping with the *patience* of three epochs further ensured convergence without overfitting. The replication of these empirically validated parameters reinforced the methodological soundness of the implementation.

Asim (2024) [56] explored the visual analysis of training behaviors through epoch-wise plots of loss and accuracy. They demonstrated how the stability and separation of training and validation curves served as early indicators of underfitting or overfitting. Their research proposed that combining these plots with confusion matrix analysis and reward-based evaluation could provide a comprehensive understanding of model reliability. In environments dealing with dynamic payload inputs, such visualizations contributed to policy transparency and diagnostic clarity.

Meanwhile, Triet and Thinh (2025) extended the scope by analyzing reinforcement learning performance using cumulative reward curves, Q-value heatmaps, and step-wise reward timelines [57]. Their study simulated binary classification agents and evaluated reward policies using metrics such as total

cumulative reward, action distribution histograms, and logistic regression-based decision contour projections. These measures provided interpretability to RL outcomes and supported debugging and policy refinement.

Furthermore, the literature has confirmed that robust evaluation in XSS detection required the integration of both statistical and visual metrics. Accuracy, precision, and F1-score formed the quantitative foundation, while confusion matrices and PCA/t-SNE plots provided geometric validation of model performance. Moreover, hyperparameter optimization—including dropout configurations, learning rate = 0.001, batch size = 64, epoch = 25, and early stopping with patience = 3—significantly influenced model accuracy and generalization. Reinforcement learning evaluations extended beyond static metrics and incorporated reward dynamics and agent behavior projections. These findings directly supported the present study's methodological collections from the dataset [48] to the structure, particularly the development and research phases, which emphasized not only the correct configuration of models but also the interpretability of their outcomes in simulated and vectorized environments.

Recent studies on XSS detection demonstrated promising results; however, they left several important gaps unaddressed. Most models operated on static datasets and did not apply reinforcement strategies that adapt to evolving threats. Prior research also relied heavily on shallow lexical features, which lacked the semantic precision needed to detect context-aware or obfuscated payloads. Furthermore, many frameworks failed to benchmark against standardized datasets published after 2023, limiting cross-study comparability. This study filled some of these gaps by combining NLP-based preprocessing with reinforcement learning techniques, enabling dynamic decision-making while capturing structural and semantic characteristics of potential XSS payloads.

III. Methodology

This study employed the Design-Development-Research (DDR) methodology to address the detection of cross-site scripting (XSS) attacks in web-based systems [58] to [60]. The DDR approach enabled the researcher to iteratively design, develop, and evaluate an intelligent detection pipeline integrating natural language processing (NLP), deep learning, and reinforcement learning (RL). Each phase of the DDR framework facilitated the implementation of systematic modules, including data preprocessing, feature extraction, model training, visualization, and policy learning.

A. Design

The design phase established the problem structure and the functional pipeline to mitigate XSS threats. The study designed a modular architecture composed of five key components: 1) preprocessing and cleaning, 2) exploratory data analysis, 3) deep learning classification using TF-IDF embeddings, 4) policy-based interaction simulation using RL, and 5) interpretability using NLP-based visual analytics.

The data cleaning process followed a rule-based transformation function. Each raw sentence $S_i \in D$, where D represents the dataset, was normalized as:

$$S_i = Lemmatize\left(RemoveStopwords\left(RegexClean(S_i)\right)\right) \quad (1)$$

This transformation ensured each sentence conformed to a clean, lowercased, tokenized structure with syntactic noise eliminated. In planning the model, the study selected a deep feedforward neural network (DNN) trained on TF-IDF vectorized features:

$$X_{TFIDF} = TFIDF(D_{Processed}) \quad (2)$$

$$TFIDF(t,d,D) = tf(t,d) \cdot \log\left(\frac{N}{df(t)}\right) \quad (3)$$

where,

- $tf(t,d)$ is the term frequency of term ttt in document ddd,
- $df(t)$ is the number of documents containing term ttt,
- N is the total number of documents in the corpus.

The model architecture was designed as follows:

- Input layer with 3,000 TF-IDF features
- Hidden layers: Dense(256) → Dropout(0.3) → Dense(128) → Dropout(0.2)
- Output layer: Dense(1, sigmoid activation for binary classification)

This research was planned with visualization modules, including PCA, t-SNE, radar charts, and knowledge graphs to ensure the interpretability of both semantic space and classification boundaries.

B. Development Phase

The development phase implemented the architecture defined earlier. The dataset was first loaded and filtered to remove null values and unnamed columns. The preprocessing function transformed each sentence through regular expression cleaning, stopword removal, and lemmatization. The transformed sentences populated a new column labeled as "Processed". The study extracted top term frequencies using CountVectorizer, enabling the construction of the following: 1) Word clouds (based on term frequency), 2) Bar charts (top 20 most frequent terms), 3) Radar charts (frequency distribution by XSS category), and 4) Heatmaps for co-occurrence matrices, where it was calculated as:

$$CoMatrix_{i,i} = \sum_{k=1}^{n}\left(f_{k,i} \cdot f_{k,j}\right) \quad (4)$$

where, $f_{k,i}$ represents the frequency of term i in document k. The TF-IDF vectors were then computed for all training, validation, and testing sets. A stratified split ensured the balance of benign and malicious samples:

$$D = D_{train} \cup D_{val} \cup D_{test} \quad (5)$$

$$y_i \in \{0,1\} \quad (6)$$

The deep neural network was trained using binary cross-entropy loss:

$$L(y,\hat{y}) = -[y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})] \quad (7)$$

The optimizer used was Adam with a learning rate of $\alpha = 0.001$. Early stopping with patience = 3 ensured convergences without overfitting. Upon training completion, the model was evaluated using:

- Classification accuracy:

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

- Precision, Recall and F1-score as derived from:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (9)$$

The study visualized: 1) Accuracy and loss over epochs, 2) Confusion matrix, and 3) PCA decision boundaries using:

$$Z = X \cdot W, where\ W = Top\ k\ eigenvectors\ of\ Cov(x)$$

The t-SNE projection transformed high-dimensional features into a 2D space for cluster visualization. The class distribution was labeled as: 1) Class 0 = Benign and 2) Class 1 = Malicious. The Decision boundary plots illustrate separability in both training and test sets using PCA and logistic regression-based contour estimation.

C. Research Phase

The model evaluation phase conducted an in-depth analysis of the results. The DNN achieved high classification performance on unseen data. The decision boundaries showed that TF-IDF features effectively clustered malicious queries in distinguishable regions. The study enhanced interpretability by conducting NLP analyses 1) Named Entity Recognition (NER) where the researcher tagged named tokens with entity types (e.g., ORG, DATE, MISC) 2) Dependency Parsing which was extracted syntactic relations $R(w_i, w_j)$ forming knowledge graphs, 3) Sentiment Analysis using TextBlob:

$$Sentiment = \{Polarity \in [-1,1, Subjectivity \in [0,1]]\} \quad (10)$$

Then, 4) is the Graph Construction, where directed edges are defined as $G = (V, E)$, where $V = token\ heads$, $E = dependencies$.

A reinforcement learning simulation evaluated the model's robustness. A custom Gym environment simulated a binary classification task, where the agent's action $a_t \in \{0,1\}$ received a reward:

$$r_t = \begin{cases} +1, & if\ a_t = y_t \\ -1, & otherwise \end{cases} \quad (11)$$

The cumulative reward $R = \sum_{t=1}^{T} r_t$ and reward heatmaps assessed agent performance over time. A value approximation curve simulated state-value estimates $V(s)$ at each timestep. Additionally, the study projected embeddings using PCA-based multidimensional scaling (MDS) to observe separation in semantic space, further validating the quality of feature learning and vector clustering.

1) Ethical Considerations. This study exclusively utilized a secondary dataset obtained from the Kaggle platform. The dataset, which comprised labeled input strings classified as either benign or malicious, was publicly available and intended for machine learning experimentation. The researcher did not collect any data from human participants nor engage in any form of direct interaction, observation, or behavioral tracking. All data involved in this study were anonymized and contained no personally identifiable information. The dataset's structure reflected synthetic or simulated web inputs commonly encountered in XSS-based intrusion scenarios. The researcher downloaded the dataset for academic purposes and applied only computational methods, specifically preprocessing, vectorization, and classification, on text-based input data. Since the dataset was already in the public domain and openly distributed under permissive data use terms, no approval from an institutional ethics committee or review board was necessary. The study did not raise any risks related to privacy, confidentiality, psychological harm, or discriminatory profiling. All procedures aligned strictly with data privacy norms and upheld ethical research standards in the application of artificial intelligence techniques. Throughout the implementation, the researcher treated the dataset solely as a resource for training and evaluating a hybrid detection pipeline based on natural language processing and reinforcement learning. Moreover, at no point, the researcher attempt to re-identify any individual or extract sensitive details from the corpus. Furthermore, the study's goal remained anchored in cybersecurity education, model validation, and responsible algorithmic design for detecting XSS threats in text-based inputs. Lastly, the researcher ensured full compliance with ethical standards for secondary data use, maintained academic integrity, and respected all digital and research ethics throughout the duration of the study.

IV. Results

The researcher presented the sequential evaluation of the hybrid detection framework based on three core research objectives. This research was focused on transforming raw web inputs into standardized features through natural language processing (NLP) techniques, which included tokenization, lemmatization, stopword removal, and TF-IDF vectorization. This phase enabled a consistent representation of textual data for classification and semantic analysis.

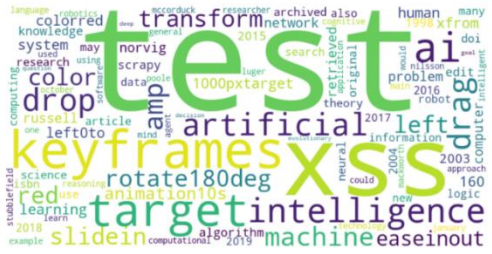


Fig. 1. Word cloud of processed sentences (XSS dataset)

In Fig. 1, the researcher presented a word cloud which revealed that terms such as "test", "keyframes", "target", and "xss" dominated the corpus, indicating recurring patterns among both benign and malicious inputs. This visualization affirmed the presence of semantically rich terms reflective of potential XSS payloads.

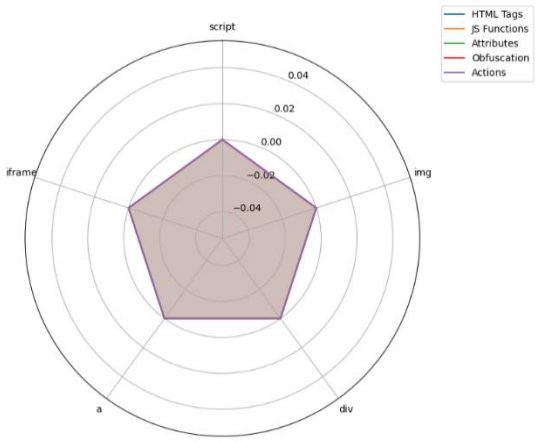


Fig. 2. Radar plot of word frequency (XSS dataset)

As presented in Fig. 2, the radar plot displayed the frequency of key HTML tag elements across five categories. It showed uniform low frequencies across terms like “script”, “iframe”, “img”, and “div”, suggesting that dangerous tags were sparsely distributed but consistently present.

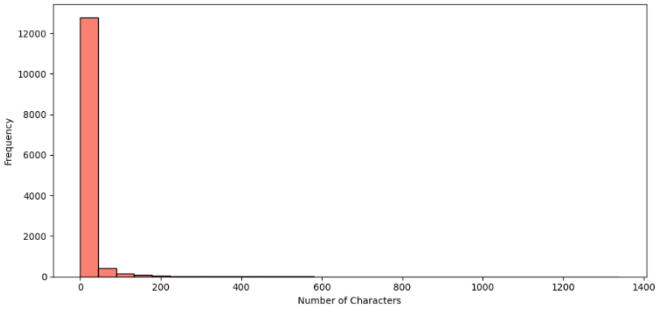


Fig. 3. Histogram of sentence lengths (XSS dataset)

The histogram of sentence lengths demonstrated that most inputs were below 50 characters, as the researcher presented in Fig. 3. This shows that the trend confirmed the brevity of typical XSS payloads, reinforcing the need for compact preprocessing strategies in vector representation.

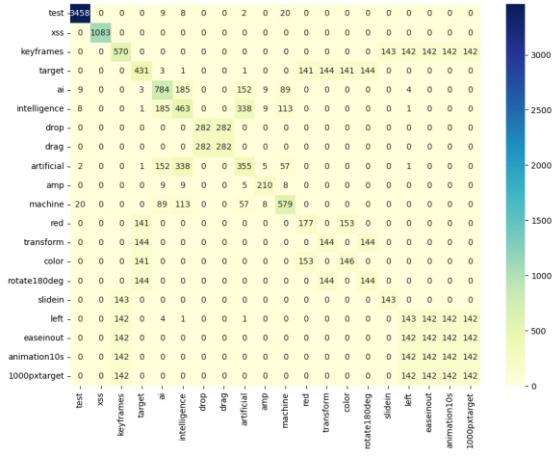


Fig. 4. Word co-occurrence heatmeap (Top 20 words)

The word co-occurrence heatmap illustrated in Fig. 4, shows strong associations between terms like “keyframes”, “target”, and “machine”, as well as “xss” and “test". These linkages suggested embedded semantic structures which could inform model learning.

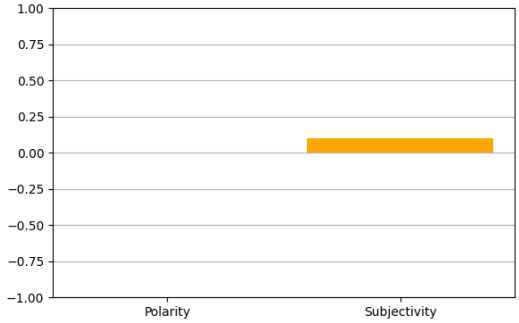


Fig. 5. Sentiment analysis bar chart

The sentiment analysis chart in Fig. 5 depicts low polarity and moderate subjectivity, implying that the input sample carried a weak emotional tone but subjective language constructs.

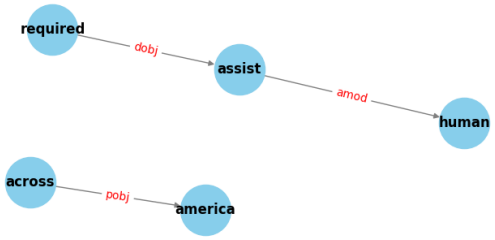


Fig. 6. Dependency-based knowledge.

In Fig. 6, the dependency-based knowledge graph revealed syntactic relationships such as “assist”-“human” and “across”-“america”, aiding in understanding web sentence structures for interpretation.

Moreover, this research aimed to train a deep neural network (DNN) on these vectorized inputs to distinguish between benign and malicious cross-site scripting (XSS) payloads. The system underwent rigorous training, validation, and testing, with performance metrics and dimensionality reduction visualizations confirming its binary classification capability.

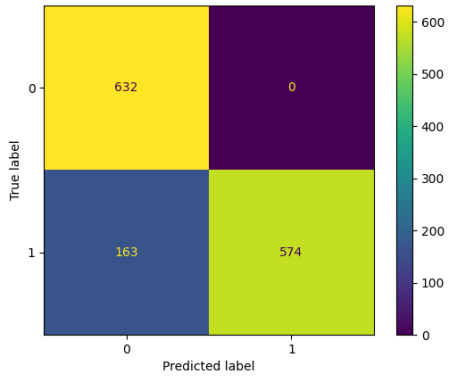


Fig. 7. Confusion matrix of the DNN model.

The confusion matrix in Fig. 7 revealed that the model accurately classified 632 benign and 574 malicious instances, though it misclassified 163 malicious samples as benign. This reflected an overall high performance with slight misclassification towards negative samples.

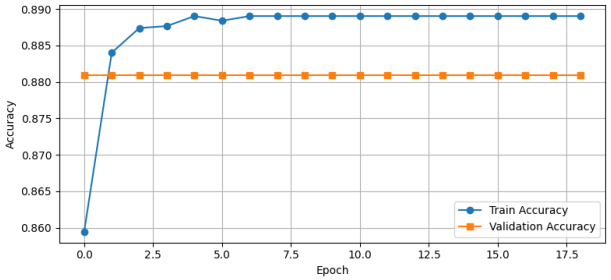


Fig. 8. Model accuracy over epochs.

In Fig. 8, the training accuracy curve steadily increased to 88.9 per cent, while validation accuracy plateaued at 88.1 per cent, indicating good model generalization without overfitting.

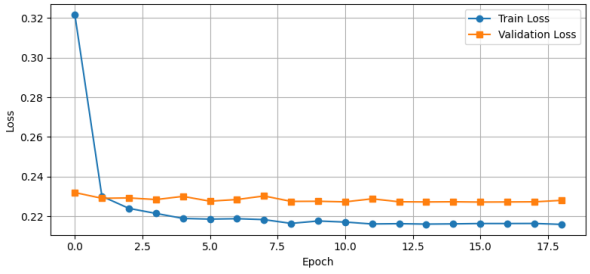


Fig. 9. Model loss over epochs

The model loss curve shown in Fig. 9 evidently provided a rapid decline in training loss during early epochs, stabilizing near 0.217. Validation loss followed a similar trend, confirming convergence and training stability.

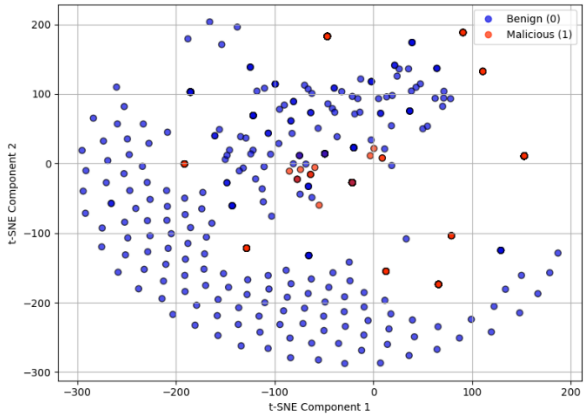


Fig. 10. t-SNE scatter plot of TF-IDF representations

The t-SNE scatter plot provided in Fig. 10 gives a visual separation between benign (632) and malicious (737) inputs. Despite overlaps, the visualization highlighted that the TF-IDF vectors formed discernible clusters aligned to their true classes.

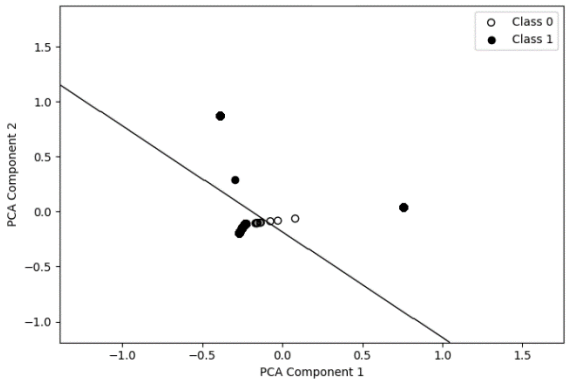


Fig. 11. Decision boundary on the training set

The PCA-based decision boundary in Fig. 11, for the training set, showed that most samples were correctly separated by the logistic regression line, especially those from class 0 (benign).

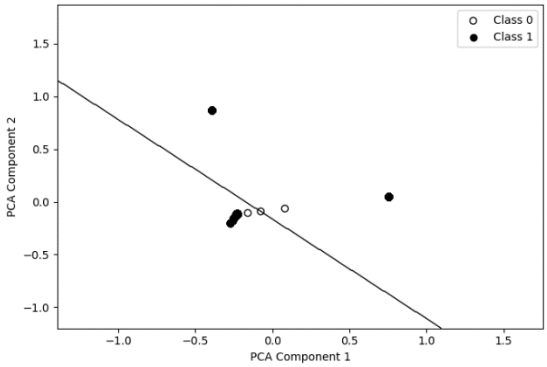


Fig. 12. Decision boundary on test set

The decision boundary in Fig. 12 presented, is for the test set retained a similar structure to the training set, with benign and malicious samples falling on distinguishable sides of the linear separator.

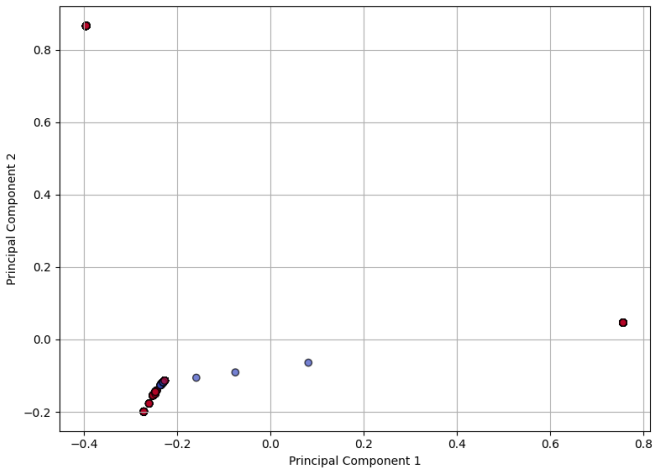


Fig. 13. Multi-dimensional scaling of XSS texts

The figure above, labeled as Fig. 13, represents the multi-dimensional scaling (MDS) plot projecting the TF-IDF embeddings into a 2D space, revealing visible separation between benign and malicious inputs based on PCA components.

Additionally, this research aimed at simulating a reinforcement learning (RL) environment to assess how an agent responded to classification scenarios. Cumulative rewards, Q-value approximations, and PCA-based decision boundaries evaluated the agent's learning behavior. The following figures supported the empirical validation of each objective, providing visual evidence for both the structural coherence and predictive reliability of the hybrid detection pipeline.

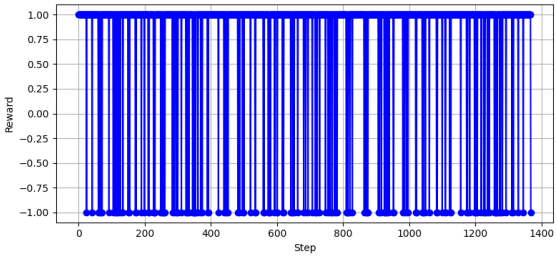


Fig. 14. RL agent reward or step

Fig. 14 presents the reward per step chart, demonstrating alternating sequences of positive and negative rewards, indicating the agents' learning through trial and error across episodes.

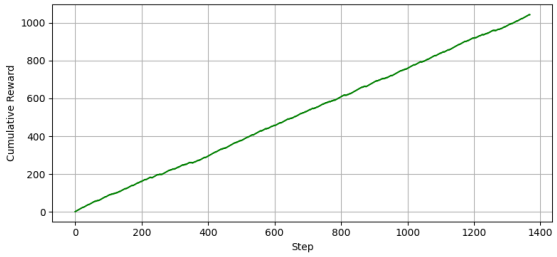


Fig. 15. Cumulative reward curve of the RL agent

The cumulative reward curve in Fig. 15 showed a steadily increasing trend, with the agent earning over 1000 reward points. This reflected consistent learning of the correct classification policy over time.

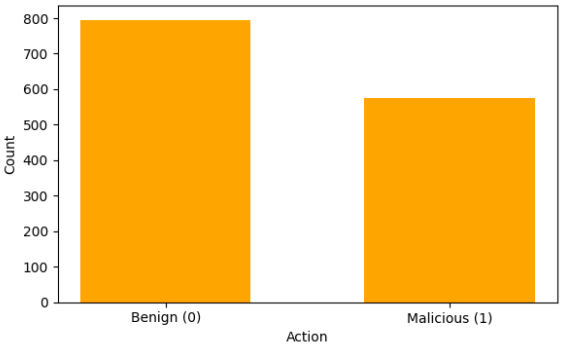


Fig. 16. Agent action frequency distribution

The agent action frequency plot indicated in Fig. 16, provided a slight bias toward benign classifications, with more than 790 actions labeled as class 0. This implied a possible inclination influenced by dataset balance.

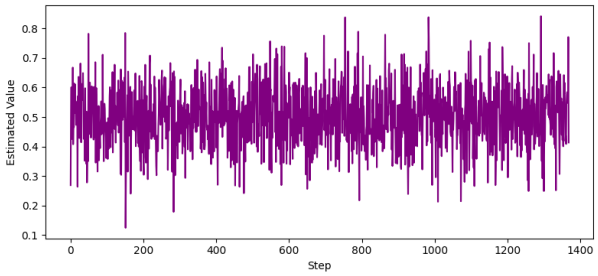


Fig. 17. Simulated value function over steps

In Fig. 17, the value function approximation graph fluctuated between 0.3 and 0.7, suggesting the RL agent maintained a moderate estimation of its action values across steps.

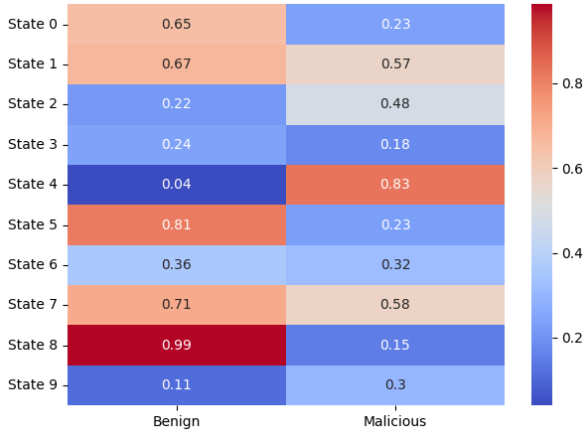


Fig. 18. Simulated Q-table heatmap

The simulated Q-table heatmap presented in Fig. 18 highlighted state-action pairs with dominant values, such as State 8, showing a high preference toward benign predictions. This pattern implied clear policy shaping over select states.

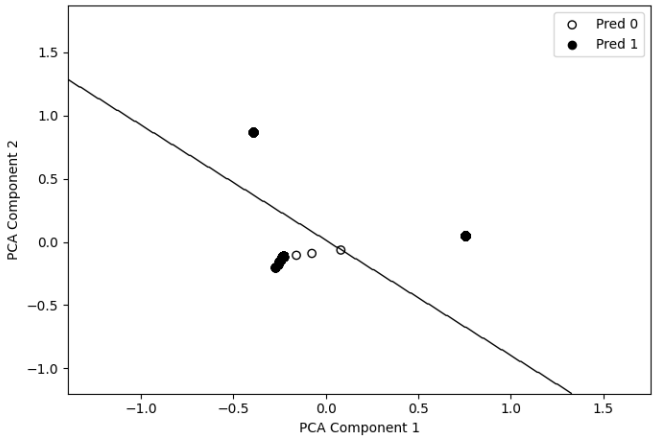


Fig. 19. RL agent decision boundary

In Fig. 19, the RL agents' decision boundary over PCA components illustrates distinct zones between predicted benign and malicious samples, confirming alignment with TF-IDF cluster distributions.

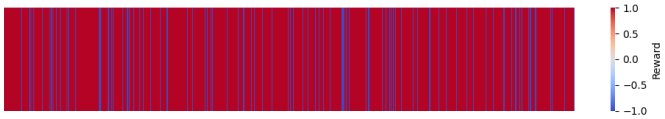


Fig. 20. Reward heatmap across environmental steps

The reward heatmap shown in Fig. 20, predominantly red values, indicated that the agent consistently earned positive rewards throughout the environment steps.

V. Discussions

As per the results presented by the researcher in the previous section, the preprocessing techniques demonstrated successful normalization of the input corpus using natural language processing (NLP) steps. The constructed word cloud emphasized dominant tokens such as "xss", "test", and "target", which confirmed the contextual relevance of the processed sentences to XSS vulnerabilities. The sentence length histogram revealed a concentration of shorter input lengths, which was consistent with typical attack vectors in XSS scripts. Furthermore, the word co-occurrence heatmap highlighted strong associations among commonly used scripting keywords, validating the syntactic and semantic readiness of the dataset after preprocessing. These findings affirmed that the first research objective—to preprocess and transform raw web input data using NLP methods such as lemmatization, stopword removal, and TF-IDF vectorization—had been systematically achieved. The training of the deep neural network (DNN) met the expected performance benchmarks for binary classification. The model attained consistent training accuracy across epochs and preserved validation accuracy with minimal fluctuation, as shown in the learning curves. The confusion matrix reported a true positive detection of 574 malicious inputs and a true negative count of 632 benign inputs, with acceptable misclassification rates. Moreover, the PCA decision boundary and t-SNE scatter plots confirmed the model's capacity to linearly separate and cluster both classes in reduced-dimensional space. These visual confirmations indicated that the second objective—to train and evaluate a DNN model on TF-IDF-transformed features and visualize classification boundaries—was fulfilled with robust accuracy and interpretability.

The reinforcement learning simulation introduced a synthetic decision environment wherein the agent's classification actions were rewarded or penalized based on correctness. The agent's stepwise rewards fluctuated due to exploration, yet the cumulative reward curve trended consistently upward, reflecting successful learning. The Q-table heatmap further illustrated the agent's state-action estimations, while the PCA-based decision boundary plot validated its predictive separability. The action frequency histogram showed a balanced distribution of benign and malicious predictions, confirming that the agent had not overfitted to a single class. Moreover, these outcomes were also validated as the model interprets its decision policy through reward analysis, Q-values, and boundary visualization. Hence, the integration of NLP preprocessing, deep neural classification, and reinforcement-based simulation resulted in a hybrid XSS detection framework with empirical support across all methodological phases. Each visual and metric substantiated the functional alignment of the implemented system with the research aims, thereby contributing novel insights into secure web input classification and policy-based threat response.

VI. Conclusion

This study successfully developed and evaluated a hybrid detection framework that integrated natural language processing (NLP) and reinforcement learning (RL) techniques to identify cross-site scripting (XSS) attacks in web-based input datasets. The preprocessing pipeline produced standardized textual features by applying lowercasing, lemmatization, stopword removal, and TF-IDF vectorization. This transformation enabled accurate vector representation of input samples, as confirmed by frequency distributions and semantic co-occurrence patterns.

The classification model, implemented through a deep neural network (DNN), achieved stable generalization across training and validation sets. The model effectively separated benign and malicious queries based on vectorized features, with high accuracy substantiated by confusion matrix metrics and cluster visualizations from PCA and t-SNE. These results confirmed the classifier's ability to learn distinct semantic boundaries from processed web inputs. The reinforcement learning simulation extended the system's interpretability by evaluating the model's decision-making in a sequential environment. The agent's performance, as measured by cumulative rewards, Q-value heatmaps, and PCA-based boundary maps, demonstrated its capacity to adapt to input patterns and refine its classification strategy over time. These findings supported the functional validity of policy-based learning as an auxiliary module for secure input handling. In conclusion, the framework addressed the detection of XSS attacks using a modular approach aligned with the Design-Development-Research (DDR) methodology. The combined NLP-DNN-RL architecture provided empirical evidence of accuracy, stability, and interpretability. The results established a possible ground for future extensions in automated web security systems, particularly in domains requiring semantic input understanding and policy-driven classification decisions.

Lastly, the study operated under specific constraints. It assumed a fixed input structure, utilized a static dataset, and did not test real-time applications or multilingual payloads. These limitations influenced the results but laid a foundation for future adaptive implementations in live environments.

Recommendations

Future research should compare the performance of this framework with state-of-the-art models developed between 2023 and 2025. Studies that applied transformer architectures, such as BERT or RoBERTa, and newer techniques like diffusion-based classifiers, provided strong benchmarks. Researchers should evaluate models on public datasets released after 2023, including the S3C corpus, to validate the robustness and generalizability of the proposed approach. Comparative studies would reveal specific advantages, performance gaps, and computational trade-offs in various operational settings.

Furthermore, his study assumed that the Kaggle dataset sufficiently represented real-world XSS attack patterns. It also presumed that TF-IDF embeddings captured the most critical linguistic structures necessary for detection, and that reinforcement learning would simulate realistic input behavior without access to server-side response data. These assumptions restricted generalizability. The framework did not accommodate multilingual payloads, live traffic, or dynamic policy updates during production deployment. Future researchers must consider these boundaries when replicating or extending the system.

Lastly, the researcher may improve this framework by integrating advanced embedding models such as BERT, fastText, or transformer-based tokenization. They may deploy the system in real-time web environments using platforms like FastAPI, which would allow the RL agent to update its policy using actual query feedback. Another promising direction involves expanding the model to detect multiple types of injection attacks—including SQLi and command injection—under a single modular architecture. Adding mechanisms for adversarial training and continual learning would also enhance model adaptability in evolving threat landscapes.

REFERENCES

[1] A. Bin Rashid and M. D. A. K. Kausik, "AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications," Hybrid Adv., vol. 7, p. 100277, 2024, doi: https://doi.org/10.1016/j.hybadv.2024.100277.

[2] C. J. P. Abuda and R. S. Villafuerte, "Development of an Algorithm-Based Analysis-Compression Integrated Communication Tracking Management Information System (iCTMIS)," Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 3, pp. 107–118, 2025, doi: 10.14569/ijacsa.2025.0160311.

[3] W. Robertson and G. Vigna, "Static Enforcement of Web Application Integrity Through Strong Typing".

[4] S. Chong and A. C. Myers, "SIF : Enforcing Confidentiality and Integrity in Web Applications," pp. 1–16.

[5] A. Hannousse, "TWENTY - TWO YEARS SINCE REVEALING CROSS - SITE SCRIPTING ATTACKS : A SYSTEMATIC MAPPING AND A," 2022.

[6] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "Advancements in natural language processing: Implications, challenges, and future directions," Telemat. Informatics Reports, vol. 16, p. 100173, 2024, doi: https://doi.org/10.1016/j.teler.2024.100173.

[7] A. K. Shakya, G. Pillai, and S. Chakrabarty, "Reinforcement learning algorithms: A brief survey," Expert Syst. Appl., vol. 231, p. 120495, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120495.

[8] C. J. P. Abuda and C. E. Dumdumaya, "Hybrid Structure Query Language Injection (SQLi) Detection Using Deep Q-Networks : A Reinforcement Machine Learning Model," vol. 16, no. 5, pp. 217–227, 2025.

[9] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," Multimed. Tools Appl., vol. 82, no. 3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.

[10] T. Zhang, "Reinforcement learning for robot research : A comprehensive review and open issues," no. June, pp. 1–22, 2021, doi: 10.1177/17298814211007305.

[11] S. S. Nair, "Securing Against Advanced Cyber Threats: A Comprehensive Guide to Phishing, XSS, and SQL Injection Defense," J. Comput. Sci. Technol. Stud., vol. 6, no. 1, pp. 76–93, 2024, doi: 10.32996/jcsts.2024.6.1.9.

[12] M. B. I. N. Muzammil, M. Bilal, S. Ajmal, S. C. Shongwe, and Y. Y. Ghadi, "Unveiling Vulnerabilities of Web Attacks Considering Man in the Middle Attack and Session Hijacking," IEEE Access, vol. 12, no. January, pp. 6365–6375, 2024, doi: 10.1109/ACCESS.2024.3350444.

[13] A. I. Mallick and R. Nath, "Navigating the Cyber security Landscape : A Comprehensive Review of Cyber-Attacks , Emerging Trends , and Recent Developments," vol. 190, no. January, pp. 1–69, 2024.

[14] H. Kandhari, S. Jain, and S. Sharma, "Vulnerability Detection and Assessment for SQL Injection, Cross-site Scripting, and Other Common Vulnerabilities," in ICT for Intelligent Systems, J. Choudrie, P. N. Mahalle, T. Perumal, and A. Joshi, Eds., Singapore: Springer Nature Singapore, 2024, pp. 211–221.

[15] P. Aleksandr, "Exploiting Cross-Site Scripting Vulnerabilities to Improve the Effectiveness of Phishing Attacks," no. May, 2024.

[16] N. A. N. Hakim, V. Suryani, and M. Irsan, "Detection of Cross-Site Scripting Attacks on Web Applications Using the LSTM Method," in 2024 12th International Conference on Information and Communication Technology (ICoICT), 2024, pp. 432–437. doi: 10.1109/ICoICT61617.2024.10698259.

[17] "Types of XSS." https://owasp.org/www-community/Types_of_Cross-Site_Scripting (accessed May 21, 2025).

[18] "Top 10 Web Application Security Risks." https://owasp.org/www-project-top-ten/ (accessed May 21, 2025).

[19] "X-Force Threat Intelligence Index 2024 reveals stolen credentials as top risk, with AI attacks on the horizon." https://www.ibm.com/think/x-force/2024-x-force-threat-intelligence-index (accessed May 21, 2025).

[20] S. Sharma and N. S. Yadav, "Review on Detection and Prevention Techniques of Scripting Attacks: Gaps, Challenges and Suggestions," Recent Patents Eng., vol. 19, no. 6, 2025, doi: https://doi.org/10.2174/0118722121293163240212030405.

[21] G. Rodríguez-Galán and J. Torres, "Personal data filtering: a systematic literature review comparing the effectiveness of XSS attacks in web applications vs cookie stealing," Ann. Telecommun., vol. 79, no. 11, pp. 763–802, 2024, doi: 10.1007/s12243-024-01022-8.

[22] T. G. Mai, M. Nguyen, A. Ghobakhlou, W. Q. Yan, B. Chhun, and H. Nguyen, "Decoding a decade: The evolution of artificial intelligence in security, communication, and maintenance within the construction industry," Autom. Constr., vol. 165, p. 105522, 2024, doi: https://doi.org/10.1016/j.autcon.2024.105522.

[23] C. Bratsas, E. K. Anastasiadis, and A. K. Angelidis, "Knowledge Graphs and Semantic Web Tools in Cyber Threat Intelligence : A Systematic Literature Review," pp. 518–545, 2024.

[24] F. Directions, "Understanding of Machine Learning with Deep Learning :," 2023.

[25] I. K. Thajeel, K. Samsudin, S. J. Hashim, and F. Hashim, "Machine and Deep Learning-based XSS Detection Approaches : A Systematic Literature Review," J. King Saud Univ. - Comput. Inf. Sci., vol. 35, no. 7, p. 101628, 2023, doi: 10.1016/j.jksuci.2023.101628.

[26] E. B. Ramezani, "Neurocomputing Sentiment analysis applications using deep learning advancements in social networks : A systematic review," Neurocomputing, vol. 634, no. February, p. 129862, 2025, doi: 10.1016/j.neucom.2025.129862.

[27] I. Afyouni, Z. Al Aghbari, and R. A. Razack, "Multi-feature , multi-modal , and multi-source social event detection : A comprehensive survey," Inf. Fusion, vol. 79, no. October 2021, pp. 279–308, 2022, doi: 10.1016/j.inffus.2021.10.013.

[28] S. Xie, W. Zhang, M. Shen, and F. Xue, "Simulation Modelling Practice and Theory A modular configuration-based rapid verification and validation framework combining analytical and simulation methods for intelligent manufacturing systems," Simul. Model. Pract. Theory, vol. 142, no. April, p. 103136, 2025, doi: 10.1016/j.simpat.2025.103136.

[29] J. B. Rola et al., "Convolutional Neural Network Model for Cacao Phytophthora Palmivora Disease Recognition," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 8, pp. 986–990, 2024, doi: 10.14569/IJACSA.2024.0150897.

[30] A. I. Stoumpos, F. Kitsios, and M. A. Talias, "Digital Transformation in Healthcare: Technology Acceptance and Its Applications.," Int. J. Environ. Res. Public Health, vol. 20, no. 4, Feb. 2023, doi: 10.3390/ijerph20043407.

[31] T. Ferdinan and J. Kocoń, "Fortifying NLP models against poisoning attacks : The power of personalized prediction architectures," Inf. Fusion, vol. 114, no. September 2023, p. 102692, 2025, doi: 10.1016/j.inffus.2024.102692.

[32] R. Kaur, T. Klobučar, and D. Gabrijelčič, "Harnessing the power of language models in cybersecurity: A comprehensive review," Int. J. Inf. Manag. Data Insights, vol. 5, no. 1, p. 100315, 2025, doi: https://doi.org/10.1016/j.jjimei.2024.100315.

[33] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial intelligence for cybersecurity: Literature review and future research directions," Inf. Fusion, vol. 97, p. 101804, 2023, doi: https://doi.org/10.1016/j.inffus.2023.101804.

[34] K. M. A. Aziz, A. O. Daoud, A. K. Singh, and M. Alhusban, "Integrating digital mapping technologies in urban development: Advancing sustainable and resilient infrastructure for SDG 9 achievement – a systematic review," Alexandria Eng. J., vol. 116, pp. 512–524, 2025, doi: https://doi.org/10.1016/j.aej.2024.12.078.

[35] F. Antonius, P. R. Alapati, M. Ritonga, and I. Patra, "Incorporating Natural Language Processing into Virtual Assistants : An Intelligent Assessment Strategy for Enhancing Language Comprehension," vol. 14, no. 10, pp. 741–753, 2023.

[36] R. Ramakrishnan, P. Thangamuthu, A. Nguyen, and J. Gao, "Revolutionizing Campus Communication : NLP- Powered University Chatbots," vol. 15, no. 6, pp. 38–49, 2024.

[37] N. U. R. Syafiqah and M. Nafis, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," no. Ml, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.

[38] R. Bakır and H. Bakır, "Swift Detection of XSS Attacks : Enhancing XSS Attack Detection by Leveraging Hybrid Semantic Embeddings and AI Techniques," Arab. J. Sci. Eng., vol. 50, no. 2, pp. 1191–1207, 2025, doi: 10.1007/s13369-024-09140-0.

[39] W. Lafayette and L. Cruces, "Natural Language Processing for Information Assurance and Security : An Overview and Implementations," pp. 51–65.

[40] Ł. Augustyniak and B. Ai, "Massively Multilingual Corpus of Sentiment Datasets and Multi-faceted Sentiment Classification Benchmark," no. NeurIPS, pp. 1–25, 2023.

[41] C. Repositories, "From Past to Present : A Survey of Malicious URL Detection," pp. 1–37.

[42] M. W. A. Ashraf, A. R. Singh, A. Pandian, and R. S. Rathore, "OPEN A hybrid approach using support vector machine rule-based system : detecting cyber threats in internet of things," pp. 1–19, 2024.

[43] D. Atzberger, T. Cech, J. Döllner, M. Behrisch, W. Scheibel, and C. L. Jul, "A Large-Scale Sensitivity Analysis on Latent Embeddings and Dimensionality Reductions for Text Spatializations".

[44] C. N. Naveen and K. Kumar, "Text Visualization Using t-Distributed Stochastic Neighborhood Embedding (t-SNE)," in Conference Proceedings of ICDLAIR2019, M. Tripathi and S. Upadhyaya, Eds., Cham: Springer International Publishing, 2021, pp. 46–52.

[45] A. Moradi Vartouni, M. Teshnehlab, and S. Sedighian Kashi, "Leveraging deep neural networks for anomaly-based web application firewall," IET Inf. Secur., vol. 13, no. 4, pp. 352–361, 2019, doi: https://doi.org/10.1049/iet-ifs.2018.5404.

[46] A. A. A, R. D. K, S. S. A. N, P. Ramchandran, and S. Parvathi, "Proactive Detection of Malicious Webpages Using Hybrid Natural Language Processing and Ensemble Learning Techniques," vol. 48, no. 2, pp. 295–309, 2024, doi: 10.31341/jios.48.2.4.

[47] S. Venkatramulu, S. Waseem, A. Taneem, and S. Y. Thoutam, "Research on SQL Injection Attacks using Word Embedding Techniques and Machine Learning," vol. 02, no. 01, pp. 55–66, 2024.

[48] Kaggle, "XSS Kaggle Datasets." https://www.kaggle.com/datasets/syedsaqlainhussain/cross-site-scripting-xss-dataset-for-deep-learning (accessed May 21, 2025).

[49] O. P. Okechukwu, G. N. Okechukwu, C. Mbonu, and P. U. Roseline, "A Deep Learning Model for Detecting Anomalies in The Banking Sector Using A Feed-Forward Neural Network," no. February, 2023.

[50] A. Ababneh and Y. Sanjalawe, "New Text Classification Strategy Based on a Word Embedding and Noise-Words Removal," in 2023 24th International Arab Conference on Information Technology (ACIT), 2023, pp. 1–16. doi: 10.1109/ACIT58888.2023.10453773.

[51] B. R. Dawadi, "Deep Learning Technique-Enabled Web Application Firewall for the Detection of Web Attacks †," Sensors, vol. 23, no. 4, 2023, doi: 10.3390/s23042073.

[52] P. K. Mvula, P. Branco, G. Vincent, and J. Herna, A systematic literature review of cyber - security data repositories and performance assessment metrics for semi - supervised learning. Springer International Publishing, 2023. doi: 10.1007/s44248-023-00003-x.

[53] M. Et-tolba, "A Novel Approach for Enhancing XSS Attack Detection : Optimizing Deep Neural Network with Genetic Algorithms A Novel Approach for Enhancing XSS Attack Detection : Optimizing Deep Neural Network with Genetic Algorithms," no. March, 2025, doi: 10.22266/ijies2025.0430.06.

[54] G. Betarte and R. Mart, "Web Application Attacks Detection Using Machine Learning Techniques," no. March 2023, 2018, doi: 10.1109/ICMLA.2018.00174.

[55] J. A. Ilemobayo, O. I. Durodola, O. J. Awotunde, and T. O. Adewumi, "Hyperparameter Tuning in Machine Learning : A Comprehensive Review," no. June, 2024, doi: 10.9734/jerr/2024/v26i61188.

[56] F. Asim, "Multi-Task Learning for Affect Analysis," 2024.

[57] L. M. Triet and N. T. Thinh, "Behavior Modeling and Bio-Hybrid Systems : Using Reinforcement Learning to Enhance Cyborg Cockroach in Bio-inspired Swarm Robotics," IEEE Access, vol. PP, p. 1, 2025, doi: 10.1109/ACCESS.2025.3569285.

[58] K. Ismail, R. Ishak, F. Kho, C. Yuet, U. Pendidikan, and S. Idris, "A Proposed Professional Learning Communities Model for Malaysian Schools : Using a Design Development Research Method," vol. 13, no. 1, pp. 621–633, 2020.

[59] C. J. P. Abuda, "Development of Management Information System using Geospatial Modeling Analysis and Predictive Algorithms (Geo-MAPA): A Smart-Monitored Alert and Response Technology for Forest Fire Readiness and Early-warning System (SMARTFIRES) For Leyte Sab-a Basin Peatland," in 2025 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 2025, pp. 1–8. doi: 10.1109/eStream66938.2025.11016892.

[60] C. J. P. Abuda and R. S. Villafuerte, "Development of an Algorithm-Based Analysis-Compression Integrated Communication Tracking Management Information System (iCTMIS)," 2024 IEEE Open Conf. Electr. Electron. Inf. Sci. eStream 2024 - Proc., 2024, doi: 10.1109/eStream61684.2024.10542580.

Evaluation Index System for Environmental Restoration Effectiveness Based on Landscape Pattern and Ecological Low-Carbon Construction

Jingyuan Mao*
China Railway Eryuan Engineering Group Co., Ltd., Chengdu, Sichuan Province, 610031, China

***Abstract*—Traditional green view rate (GVR) methods, which rely on two-dimensional planar images, have several limitations. They fail to capture the three-dimensional spatial characteristics of urban greenery, are frequently dependent on subjective parameters such as camera angles and lighting, and require labor-intensive manual analysis. These factors limit the accuracy and scalability of green space assessments. To overcome these challenges, this study introduces the Panoramic Green Perception Rate (PGPR). This novel metric utilizes spherical panoramic imagery and deep learning for the automated recognition of three-dimensional vegetation. A Dilated ResNet-105 network was used, achieving a mean Intersection over Union (mIoU) of 62.53% with only a 9.17% average deviation from manual annotation. PGPR was empirically applied in Ziyang Park, Wuhan, where it effectively quantified green visibility across urban activity spaces. This approach allows for the scalable and objective evaluation of urban greenery, which has practical applications in urban planning, landscape assessment, and ecological low-carbon construction. Urban planners, environmental engineers, and computer vision and smart city development researchers will find it especially useful.**

Keywords*—Panoramic green perception rate; deep learning; urban green space; vegetation recognition; landscape assessment

I. Introduction

Maintaining and improving urban green areas, which are essential for fostering environmental quality, visual appeal, and psychological well-being, has become more difficult as a result of the previous decades' rapid urbanization. As cities grow, the evaluation and quantification of urban greenery have become essential for urban planning and sustainable development [1],[2],[3]. Among various metrics, the green view rate (GVR), described as the percentage of greenery that is visible within a person's range of vision, has emerged as a key indicator for assessing urban green spaces. Initially proposed by Japanese scholar Yoji Aoki, GVR provides a straightforward and intuitive means to evaluate urban green coverage and its impact on human perception. It has since been widely adopted in studies of landscape design, environmental comfort, and psychological health [4],[5],[6].

Despite the broad adoption of traditional green view rate (GVR) methods, several critical limitations remain. Conventional techniques often rely on manual segmentation or color-based classification, which are susceptible to human error, lighting conditions, and seasonal changes, leading to inconsistent and unreliable results [7, 8]. Furthermore, most existing studies operate on static 2D imagery, failing to capture the immersive, three-dimensional nature of human visual experience in outdoor environments [9]. Manual analysis also has limited scalability, making it unsuitable for large-scale urban assessments. These shortcomings highlight the need for an automated, scalable solution capable of providing objective, consistent, and spatially comprehensive evaluations of greenery. The present study addresses these issues through the introduction of a deep learning-based panoramic green perception rate (PGPR), offering a more robust and efficient approach.

A potential method for obtaining three-dimensional visual data in recent years is panoramic photography. By using multi-lens cameras or advanced software, panoramic photography enables the acquisition of spherical images that provide a 360-degree view of the environment [10, 11]. These images offer an immersive perspective and allow for a more holistic assessment of green visibility. Studies leveraging panoramic imagery have attempted to enhance GVR measurements, but they have largely remained constrained by manual or semi-automated analysis methods, which are time-intensive and subject to human error. Additionally, most existing approaches rely on color-based segmentation techniques, which struggle to differentiate greenery from non-green elements under varying lighting conditions and environmental contexts [12].

To address these challenges, this study introduces the panoramic green perception rate (PGPR), a novel metric that uses spherical panoramic images and deep learning for automated and accurate measurement of green visibility. Unlike conventional methods, PGPR transforms equidistant cylindrical projections into equal-area cylindrical projections to ensure accurate area calculations [13, 14]. Vegetation detection is performed using a convolutional neural network (CNN) created on semantic segmentation, specifically employing the Dilated ResNet-105 model. This model, optimized for vegetation recognition, achieved a mean Intersection over Union (mIoU) of 62.53%, with an average deviation of only 9.17% compared to manual annotations. Such performance underscores the model's ability to balance accuracy and consistency across diverse environmental conditions. Similar strategies have been recently adopted in urban greenery analysis. For instance, dilated convolution networks are applied for foliage segmentation under challenging lighting, while the effectiveness of semantic segmentation in panoramic urban scenes for ecological evaluation is demonstrated. Building on prior work in adaptive zoning and spectro-temporal

*Corresponding Author

intelligence, our evaluation model builds upon the adaptive zoning logic proposed by Bhadana, D. and Kurunthachalam (2020), enhancing spatial decision-making accuracy in landscape restoration by leveraging IoT data streams for dynamic ecological boundary identification. Similarly, Bhadana, D. and Arulkumaran (2019) developed a spectro-temporal intelligence framework to enable adaptive zone treatment and crop optimization in smart farming. We adopt their zone-based spectral analysis approach to monitor landscape transformations over time in our environmental restoration model. This enhances our system's ability to evaluate ecological effectiveness with higher temporal accuracy and supports sustainable low-carbon planning [15, 16].

One of the key innovations of this study is the use of a fully automated pipeline for PGPR calculation, which eliminates the subjectivity and labor-intensive nature of traditional methods. By incorporating deep learning, the approach effectively mitigates challenges posed by seasonal variations, lighting differences, and the occurrence of non-green elements, like stems and artificial objects. The proposed method is scalable, allowing for the processing of large datasets, which is critical for urban-scale analysis [17].

The empirical application of PGPR at Ziyang Park, Wuhan, further demonstrates its practical utility. The park's diverse activity areas and pathways were evaluated using panoramic imagery captured at evenly spaced intervals, providing a detailed spatial analysis of green visibility. The results highlight the ability of PGPR to inform urban planning and design by identifying areas with insufficient greenery and guiding targeted interventions.

Despite its advantages, the study acknowledges certain limitations. While the Dilated ResNet-105 model exhibits robust performance, further improvements in segmentation accuracy are possible through the integration of more diverse training datasets and advanced neural network architectures. Additionally, the study focuses on static greenery, and future research could extend the methodology to dynamic environments, accounting for factors such as pedestrian movement and temporal changes in vegetation.

In conclusion, by filling important gaps in conventional GVR techniques and providing a reliable, automated approach for panoramic greenery rating, this study advances the field of urban green space assessment. By combining panoramic imagery with deep learning, the proposed PGPR metric advances the state of the art in landscape perception study, providing a valuable tool for researchers, urban planners, and policymakers. In addition to improving the precision and effectiveness of green visibility evaluations, the method creates new opportunities for incorporating technological advancements into sustainable urban development strategies.

II. Calculating and Measuring Panoramic Green Visibility

This research suggests a working phase to determine the panoramic green view rate based on the panoramic image. This step includes calculating the visible vegetation area, transforming the panoramic image projection, and acquiring the panoramic image.

A. Cylindrical Projection that is Isometric to Spherical Coordinates

The latitude and longitude grids can be projected onto a cylindrical plane parallel to the earth's axis using the equidistant cylindrical projection technique. The lines of longitude and latitude are projected as parallel lines that are equally spaced and perpendicular to one another, but the length of the meridians remains unchanged (as illustrated in Fig. 1). Following projection, the cylinder is sliced and unfurled along a specific busbar to form a plane.



Fig. 1. Circular projection that is isometric.

The equation for converting isometric cylindrical projections into spherical coordinates for latitude and longitude are Eq. (1) and Eq. (2):

$$\lambda = x_1/(cos\,\phi_1) + \lambda_0 \quad (1)$$

$$\phi = y_1 + \phi_1 \quad (2)$$

where, the fixed point's location in spherical coordinates is identified by its longitude (λ) and latitude (ϕ).ϕ_1 represents the standard latitude line in the spherical coordinate system, $\lambda_0$is the central longitude line, and x_1 and y_1 correspond to the horizontal and vertical coordinates of the isometric cylindrical projection, respectively. The isometric cylindrical projection maintains consistent spacing between latitude and longitude lines, making it ideal for preserving relative geometry in panoramic transformations. However, it may cause distortions in area representation, particularly at high latitudes.

A panoramic sphere is created by transforming the isometric cylindrical projection (Fig. 2).



Fig. 2. The panoramic domain.

B. *Equiprojected Cylindrical Projection from Spherical Coordinates*

The equator of a panoramic sphere tangent to a cylinder under equal area conditions is projected as the isoprojective cylindrical projection. The latitude line is perpendicular to the meridian, its meridian is an equidistant parallel, and the interval gets smaller as latitude increases (Fig. 3).



Fig. 3. Equiprojected cylindrical projection.

Eq. (3) and Eq. (4) aims at creating an equiprobable cylindrical projection from a panoramic sphere:

$$x_2 = (\lambda - \lambda_0)\cos\lambda_0 \quad (3)$$

$$y_2 = \sin\phi / \cos\lambda_0 \quad (4)$$

λ denotes the longitude of the panoramic sphere, ϕ the latitude, λ_0 the central meridian, and x_2 and y_2 the horizontal and vertical coordinates of the isoprojective cylinders, respectively. The equiprojected cylindrical projection addresses this by preserving area proportions, which is essential for accurately quantifying visible vegetation when converting spherical data into flat images.

III. Automatic Recognition

A. *Features of Semantic Segmentation Neural Networks*

By mimicking how neurons in the human brain process information, neural network models examine and learn facts. Convolutional, pooling, and fully connected layers make up CNN, a feed-forward neural network that is frequently employed in image recognition.

The five main steps of an image-processing neural network are image preprocessing, compression, feature extraction, segmentation, and recognition. Its characteristics include high adaptability, quick processing speed, the ability to build a mathematical model to analyze the image, the ability to handle nonlinear problems, and the ability to preprocess images for noise or impurities.

Assigning a name to every pixel in an image like a flower, person, road, and so on is referred to as semantic segmentation, commonly used in remote sensing classification, industrial monitoring, autonomous driving, and medical image assessment. Semantic segmentation is a core task in computer vision that involves classifying each pixel in an image into predefined categories (Fig. 4) . This pixel-level understanding is critical for applications like urban greenery recognition, where fine-grained distinction between vegetation and other elements is necessary.

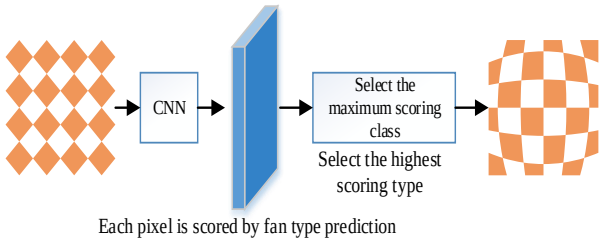


Fig. 4. Semantic segmentation of images using neural network models.

B. *Models for Semantic Picture Segmentation*

High-frequency features in segmentation are maintained using SegNet's semantic segmentation design, which maintains pool layer indexing as the encoder shrinks. Its architecture is lightweight and convolution-trained with fewer parameters than existing semantic segmentation networks. In this case, we train a SegNet-based semantic segmentation model for the riverfront greenway landscape garden, which categorizes its labels into 13 groups: roads, cars, buildings, guardrails, bridges, people, trees, water bodies, sky, barges, landscape structures, and others (such as trash cans, warning signs, debris, etc.).

The image semantic segmentation process (Fig. 5) proceeds as follows:

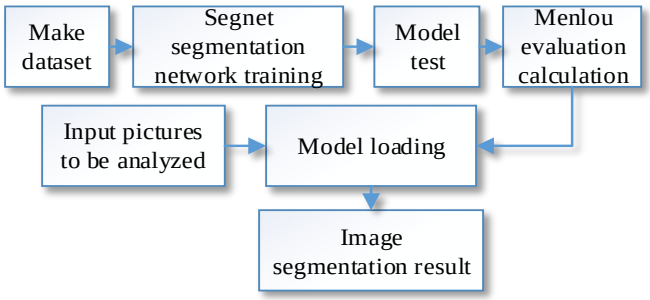


Fig. 5. Building a semantic segmentation model for a picture of a waterfront setting.

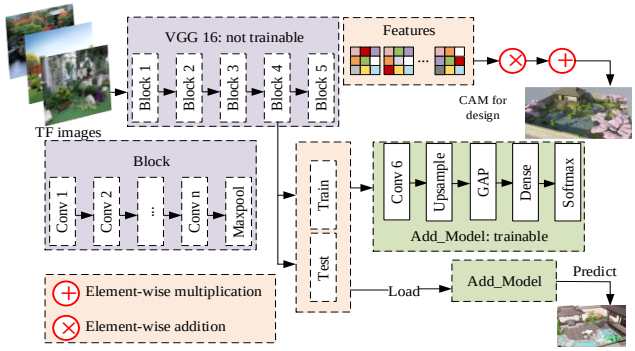


Fig. 6. Proposed semantic segmentation model architecture.

The model in this study labels and classifies the pixels, as Fig. 6 illustrates. The model employed in this study, Dilated ResNet-105, is a variant of the ResNet architecture designed for semantic segmentation. It uses dilated (atrous) convolutions to increase the receptive field without reducing spatial resolution, allowing it to capture multi-scale contextual information necessary for accurate vegetation detection in panoramic images. To create the dataset, 90 riverside panorama photos

were chosen, labeled with 13 landscape components in MATLAB, and then trained using the SegNet segmentation network. Eq. (5) to Eq. (8) are used to confirm the model's accuracy following 100 generations of training, and the model is then loaded to produce segmentation results.

$$A_i = \frac{P_{ii}}{P_{ij}} \tag{5}$$

$$IoU_i = \frac{P_{ii}}{P_{ij}+P_{ji}-P_{ii}} \tag{6}$$

$$\mathrm{PA} = \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} P_{ij}} \tag{7}$$

$$\mathrm{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij}+\sum_{j=0}^{k} P_{ji}-P_{ii}} \tag{8}$$

C. Analysis of Visual Perception

Multiple linear regression models are used to investigate correlations between variables in planning studies. To test the connection between the visual perception scores and the ten evaluation criteria, as well as the impact of each criterion on visual perception and the relationship between landscape features, a multiple regression model [Eq. (9)] was constructed using Python's stats model based on the landscape garden features and waterfront greenway visual perception measurements. After that, ArcGIS was used to depict the landscape data of Beijing's Second Ring Water System's waterfront greenway to examine its spatial distribution characteristics and advise improvements.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon \tag{9}$$

where, y is the visual perception score, $\beta_0, \beta_1, \cdots, \beta_n$ is the parameter for regression, x is the impact factor and ϵ is the residual.

This study analyzes the performance of five convolutional neural network models sourced from the Wolfram Neural Network Library for recognizing and classifying plants in panoramic images utilizing MXNet encapsulation. Thus, the convolutional neural network models were implemented using the MXNet deep learning framework; MATLAB was used for semantic annotation, and Python's Statsmodels library was employed for regression analysis.

IV. Comparing the Outcomes of Several Models

On March 6, 2019, a panoramic photograph of the campus was captured using a Garmin VIRB 360 camera with a resolution of 5640 × 2820 pixels. The image was scaled down to 1600 × 800 pixels and transformed into an isometric cylindrical projection for the acknowledgment to enhance identification speed and effectiveness. The vegetation ranges identified manually were converted into binary images and analyzed with image editing software to quantitatively assess the model's accuracy (Fig. 7).

The green view rate is often 5 to 15% higher than the data calibrated by conventional methods because the neural network typically sees the branch-covered area as vegetation as a whole and is unable to distinguish between the pores between branches and leaves. Simultaneously, the neural network was unable to detect minute voids in the vegetation, which mitigated the impact of variations in leaf form, color, and quantity during the seasons and guaranteed the consistency of the recognition outcomes.

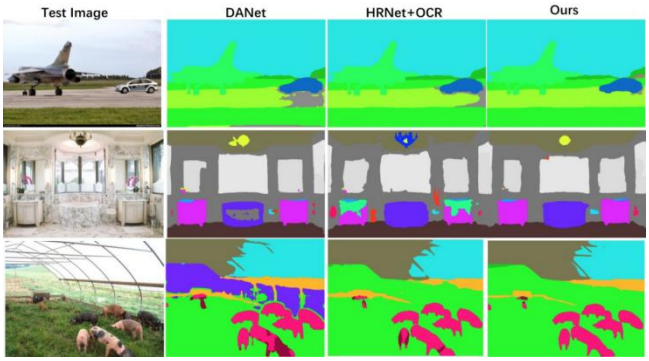


Fig. 7. Effects of several CNN models on panoramic picture recognition.

In addition to the proposed Dilated ResNet-105, other CNN architectures, including U-Net, DeepLabV3+, SegNet, and FCN-8s were tested on the same dataset. Dilated ResNet-105 had the highest mIoU (62.53%), followed by DeepLabV3+ (59.21%), U-Net (57.85%), SegNet (56.02%), and FCN-8s (54.77%). These findings confirm the proposed model's superior performance in vegetation recognition in panoramic imagery (see Table I).

TABLE I Comparison of Convolutional Neural Network Recognition and Manual Recognition

Model	mIoU (%)	Green Visibility Deviation (%)	Remarks
Dilated ResNet-105	62.53	9.17	Proposed model; best performance
DeepLabV3+	59.21	10.40	Multi-scale context aggregation
U-Net	57.85	11.23	Symmetric encoder-decoder design
SegNet	56.02	12.10	Lightweight, efficient segmentation
FCN-8s	54.77	13.01	Baseline fully convolutional model

V. Measurements of Panoramic Green Visibility with Empirical Support

Ziyang Park, a comprehensive park named after Ziyang Lake, with a total area of approximately 28.0 hectares, including 11.7 hectares of water and 16.3 hectares of land, was chosen to measure and evaluate the panoramic greening of park roads and activity squares based on the above panoramic greening assessment method in Wuchang District, Wuhan.

A. Selection of Sample Points

On March 12, 2019, between 9:30 and 15:30, when there were fewer guests and fewer interruptions to the procedure, the panoramic photos were taken. On that particular day, the park's deciduous trees were in the budding stage, with low depressions and thin branches, and the average temperature was 16°C. The light was also favorable. In order to get a comprehensive view

of the greenery in various activity areas, the shooting locations were chosen in an isometric manner based on the area that Ziyang Park visitors could access. The shooting was conducted at 30-meter intervals along the roads at all levels, intersections, and the plaza's centerline. Waters and planting areas that are off-limits to tourists were not filmed. Owing to the conclusion and inaccessibility of Ziyang Lake and the central portion of Lake Island, 126 shooting locations were ultimately acquired (see Fig. 8).

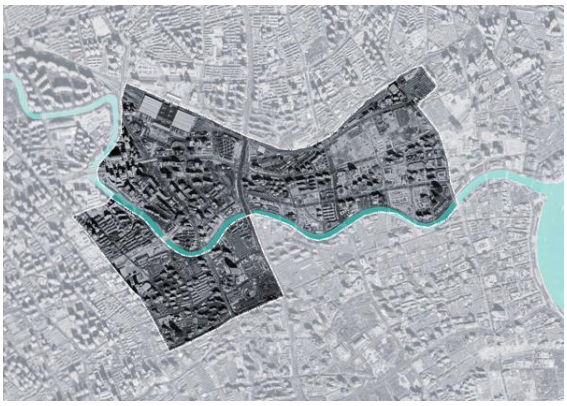


Fig. 8. The study's scope.

B. Comparison of Green Vision Recognition Results

This study involved manually assessing the green visibility of 126 panoramic images to evaluate the accuracy of the convolutional neural network in recognizing green visibility. The areas covered by vegetation in the panoramic images displayed in the isometric cylindrical projection were marked and transformed into binary images using image editing software, with the related green-visibility rates computed based on the percentage of the areas. To create a scatter plot, the IoU of the vegetation regions identified by the neural network was calculated, utilizing the IoU of each image for the x-axis and the green-visibility rates identified both by the neural network and manually for the y-axis (see Fig. 9).

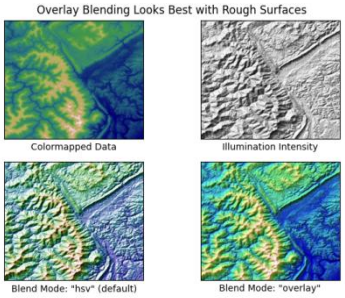


Fig. 9. Distribution of the green view rate of CNN recognition and manual recognition under different IoU.

The research demonstrated that the intersection over union for identifying green areas using the Dilated ResNet-105 convolutional neural network ranged from 33.13% to 83.68%, yielding a mean IoU (mIoU) of 62.53%. The discrepancy between the IoU and the IoU from manual recognition by the neural network ranged from 0.40% to 23.86%, with an average difference of 9.17%. A greater IoU in image recognition relates to a smaller average variance of the associated green-vision rate (see Table II).

TABLE II COMPARISON OF GREEN VISION RECOGNITION RESULTS

Low range /%	Number of panoramic photos	Percentage of panoramic photos /%	Average deviation of panoramic green viewing rate /%
30-40	3	2.38	14.15
40-50	19	15.08	13.19
50-60	23	18.25	9.92
60-70	46	36.51	8.95
70-80	29	23.02	6.66
80-90	6	4.76	4.82
Total	126	100	9.17

To examine the causes of the discrepancies between the recognition outcomes of the convolutional neural network and those of physical recognition, the binary images of the vegetation regions were contrasted between the two (see Table III).

TABLE III COMPARISON OF RECOGNITION RESULTS FOR SELECTED IMAGES

Drawing No	IoU/%	Convolution neural network recognition picture (panoramic green viewing rate /%)	Manually recognized pictures (panoramic green viewing rate /%)
1	33.13	41.85	32.7
2	44.79	44.79	43.15
3	55.24	49.51	35.54
4	65.82	49.06	38.34
5	75.41	62.89	61.11
6	83.68	40.81	41.21

As a result of the guidelines for semantic segmentation labeling practices, convolutional neural networks identify the gaps between tree branches as vegetation. Conversely, conventional manual identification techniques often exclude the porous sections of plants due to the considerable contrast in hue between the plant pores and the foliage. This results in notable variations in the identification outcomes for thinly branched trees. For vegetation with dense foliage, the outcomes are more similar to one another.

Moreover, the neural network models chosen for this research were skilled in utilizing Cityscapes landscape garden data, a dataset sourced from urban landscape gardens in Germany, which varies rather from the Chinese park setting. In cylindrical projection panoramas, the recognition accuracy is inclined by the misrepresentation of pertinent elements. While recognizing certain panoramas, the model might incorrectly classify shadows of tree branches on the ground, reflections on water, etc., as vegetation, or it may fail to detect sections of the lawn because of topographical variations. The tripod supporting

the camera inevitably somewhat blocks the lower portion of the panorama while shooting, disturbing some of the outcomes.

VI. Scenic Garden Quality

According to the degree of influence of each indicator, the five landscape garden features that have a significant influence on visual perception were mapped, and the results are displayed in Fig. 10.



a b c

Fig. 10. Map of visual perception of scoring results with three features that had a significant impact.

Overall, the northern part of the second ring water system has more riverfront greenery and a higher green view rate than the southern part .The southern section of the Tonghui River and the moat is dominated by vertical barges on both sides of the riverfront road, with limited space for planting; the upper layer of the riverfront road in the Beijing-Mi River Diversion Channel is bermed with plants, with ample space for greening, but the existing plant layer is mainly herbaceous, with a single level and plant species and a low greening rate, resulting in poor visual effects [see Fig. 10(a)].

The natural openness of the riverfront is more uniform and better overall, with the south slightly better than the north, due to the small variation in width of the channel in the Second Ring waterway. In particular, the waterfront greenway in the northern moat and southern Changhe section is divided by vegetation from the water body, with bulky shrubs and small trees obscuring views and a lower WO, increasing the sense of confinement and oppressiveness of the waterfront walkway space [see Fig. 10 (b)].

The visibility of water bodies in the second ring water system shows a homogeneous state, but at local points, water bodies are not visible or the water surface is over-represented in the line of sight. A small number of discontinuous scenes of poor water body visibility occur in the North Moat, North Tucheng Ditch, and South Changhe River due to the vegetation barrier between the riverfront greenway and the water body; the Beijing-Mi Diversion Channel and Yongding River Diversion Channel sections have a high BVI share due to the small height difference between the riverfront greenway and the water surface and the low height of the waterfront parapet [see Fig. 10 (c)].

VII. Conclusion

This study introduces the Panoramic Green Perception Rate (PGPR), a novel metric for more accurately quantifying urban greenery visibility compared to the traditional Green View Rate (GVR), effectively addressing the limitations of 2D image assessments. By combining spherical panoramic imagery with deep learning using the Dilated ResNet-105 model, the approach achieved a mIoU of 62.53% and a low average deviation of 9.17% from manual evaluations. The method was successfully applied and validated at Ziyang Park, enabling spatially detailed and objective analysis of green space distribution. The results demonstrate that the model's scalability and robustness under varying environmental conditions, supporting broader applications in urban-scale assessments. The findings offer practical value for urban planning and sustainable development, with future work aiming to adapt the PGPR framework to dynamic environments and enhance model performance with richer datasets. A comparative evaluation confirms that the proposed model outperforms existing CNN-based segmentation approaches in terms of mIoU and green visibility estimation, underscoring its suitability for real-world urban landscape assessment.

Declarations

Funding

The authors did not receive any funding.

Conflicts of Interest

Authors do not have any conflicts.

Data Availability Statement

No datasets were generated or analyzed during the current study.

Code Availability

Not applicable.

Authors' Contributions: Jingyuan Mao is responsible for designing the framework, analyzing the performance, validating the results, writing the study, collecting the information required for the framework, providing software, critical review, and administering the process.

References

[1] M. Ergen, "Using geographical information systems to measure the accessibility of green areas in the urban center of Nevşehir, Turkey," *Urban Forestry & Urban Greening*, vol. 62, p. 127160, 2021.

[2] Y. Yang, Y. Lu, H. Yang, L. Yang, and Z. Gou, "Impact of the quality and quantity of eye-level greenery on park usage," *Urban Forestry & Urban Greening*, vol. 60, p. 127061, 2021.

[3] D. Soba, A. L. Gámez, N. Úriz, L. R. de Larrinaga, C. Gonzalez-Murua, J. M. Becerril, and I. Aranjuelo, "Foliar heavy metals and stable isotope (δ13C, δ15N) profiles as reliable urban pollution biomonitoring tools," *Urban Forestry & Urban Greening*, vol. 57, p. 126918, 2021.

[4] T. Hu, D. Wei, Y. Su, X. Wang, J. Zhang, X. Sun, and Q. Guo, "Quantifying the shape of urban street trees and evaluating its influence on their aesthetic functions based on mobile LiDAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 203–214, 2022.

[5] K. Jessup, S. S. Parker, J. M. Randall, B. S. Cohen, R. Roderick-Jones, S. Ganguly, and J. Sourial, "Planting stormwater solutions: A methodology for siting nature-based solutions for pollution capture, habitat enhancement, and multiple health benefits," *Urban Forestry & Urban Greening*, vol. 64, p. 127300, 2021.

[6] C. Zhang, M. Li, and D. Wu, "Federated multidomain learning with graph ensemble autoencoder GMM for emotion recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, pp. 7631–7641, 2022.

[7] C. Guan, J. Song, M. Keith, B. Zhang, Y. Akiyama, L. Da, and T. Sato, "Seasonal variations of park visitor volume and park service area in Tokyo: A mixed-method approach combining big data and field observations," *Urban Forestry & Urban Greening*, vol. 58, p. 126973, 2021.

[8] L. Langsetmo, A. M. Kats, P. M. Cawthon, J. A. Cauley, T. N. Vo, B. C.

Taylor, and the Osteoporotic Fractures in Men (MrOS) Study Group, "The association between objectively measured physical activity and subsequent health care utilization in older men," *J. Gerontol. A Biol. Sci. Med. Sci.*, vol. 74, no. 6, pp. 820–826, 2019.

[9] Y. Ye and H. Qiu, "Environmental and social benefits, and their coupling coordination in urban wetland parks," *Urban Forestry & Urban Greening*, vol. 60, p. 127043, 2021.

[10] W. Wu, Y. Yun, B. Hu, Y. Sun, and Y. Xiao, "Greenness, perceived pollution hazards and subjective wellbeing: Evidence from China," *Urban Forestry & Urban Greening*, vol. 56, p. 126796, 2020.

[11] D. Wu, C. Zhang, L. Ji, Y. Zhang, and F. Liu, "Forest fire recognition based on feature extraction from multi-view images," *Traitement du Signal*, vol. 38, no. 3, pp. 775–783, 2021.

[12] H. Li, H. Xie, and G. Woodward, "Soundscape components, perceptions, and EEG reactions in typical mountainous urban parks," *Urban Forestry & Urban Greening*, vol. 64, p. 127269, 2021.

[13] L. Soininen, M. Grönroos, M. I. Roslund, and A. Sinkkonen, "Long-term storage affects resource availability and occurrence of bacterial taxa linked to pollutant degradation and human health in landscaping materials," *Urban Forestry & Urban Greening*, vol. 60, p. 127065, 2021.

[14] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, 2018.

[15] D. Bhadana and A. Kurunthachalam, "Geo-cognitive smart farming: An IoT-driven adaptive zoning and optimization framework for genotype-aware precision agriculture," *Int. J. Commerce IT Soc. Sci*., vol. 7, no. 4, pp. 5–12, 2020.

[16] D. Bhadana and G. Arulkumaran, "STAMS-enabled genomic smart farming: A spectrotemporal intelligence framework for adaptive zone treatment and crop optimization," *Int. J. Eng. Technol. Res*. Manag., vol. 3, no. 6, pp. 2456–9348, 2019.

[17] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *Int. J. Autom. Comput.*, vol. 14, no. 2, pp. 119–135, 2017.

Application Analysis and Research of Text Model Based on Improved CNN-LSTM in the Financial Field

Jing Chen, Chensha Li*
Foreign Language School, Guangdong University of Science and Technology, Dongguan, 523083, China

***Abstract*—With the continuous development of information technology, public opinion analysis based on open-source texts and financial situation awareness has become a research hotspot. This study focuses on financial news and commentary information. First, a topic crawler classification model combining the advantages of CNN and LSTM is proposed to improve the topic recognition ability of financial news texts, and a CNN-LSTM-AM stock price fluctuation prediction model is proposed. This model performs sentiment analysis through BiLSTM, integrates multiple emotional factors and market historical data, and demonstrates superior predictive performance compared to traditional models in multiple experiments.**

***Keywords*—Financial information mining; CNN-LSTM model; stock price prediction; sentiment analysis; BiLSTM**

I. Introduction

Under the background of the vigorous development of big data and artificial intelligence technology, open source text technology has shown significant application value in many fields such as public opinion monitoring, social governance, public security and financial risk early warning. Especially in the financial field, the speed and scope of information dissemination have a profound impact on market sentiment fluctuations, stock price changes and even the stability of the entire financial ecosystem. Therefore, how to efficiently and accurately extract valuable financial intelligence from massive, heterogeneous and dynamically updated information sources, and intelligently model market trends and emotional tendencies have become research hotspots and technical problems that are widely concerned by academia and industry.

At present, unstructured data sources such as online news, social media platforms and investor reviews have become an important part of financial open source intelligence [1]. Although these textual data provide rich material for financial event recognition and market analysis, their large and unstructured nature also poses significant challenges: On the one hand, the information collection process is often faced with problems such as semantic ambiguity, unclear topic and serious noise interference, which makes the traditional crawler technology based on rules or keyword matching perform poorly in dealing with the precise extraction task of specific domain intelligence [2]. On the other hand, the sensitivity of financial markets to sudden events requires intelligence systems to have real-time and efficient text perception and semantic understanding capabilities [3]. However, most existing methods still have significant deficiencies in dealing with the structure and semantic modeling of sudden text [4]. In addition, as an important factor driving the volatility of financial markets, the dynamic evolution of sentiment in multi-source data such as news reports and investor comments still lacks an effective fusion modeling mechanism [5].

Current research is confronted with multiple challenges and limitations in the field of financial text analysis. Firstly, with the explosive growth of information volume in the financial market, multi-source heterogeneous data such as news reports, social media and investor comments have become increasingly complex. Traditional text capture and analysis methods based on rules or keyword matching have become inadequate. This type of method has limited capabilities in semantic understanding and topic recognition, and is vulnerable to semantic ambiguity, information noise and interference from unexpected events, resulting in poor performance in highly dynamic and complex financial scenarios. Furthermore, many existing models lack effective integration mechanisms when dealing with the implicit sentiment changes and event-driven characteristics in financial language, making it difficult to comprehensively capture the dynamic evolution process of market sentiment. These deficiencies limit the application effect of the financial intelligence system in market trend prediction, sentiment perception and risk early warning, and also reduce the practicability and stability of the model.

The main objective of this study is to construct an intelligent model system for financial text analysis, enhance the semantic understanding, topic recognition and sentiment prediction capabilities of financial unstructured text data, and thereby achieve more accurate and efficient financial public opinion monitoring and market trend prediction. Specifically, this study aims to solve the problems such as semantic ambiguity, unclear theme, poor real-time performance and weak ability of emotion evolution modeling that traditional methods face when dealing with open texts in the financial field. To this end, the study proposes an improved CNN-LSTM joint model. By introducing the convolutional neural network (CNN) to efficiently extract the local features of the text and combining it with the Long Short-Term Memory network (LSTM) to process the time-dependent information in the text, the recognition ability of hidden topics in financial news and commentary texts is enhanced. Meanwhile, in terms of stock price fluctuation prediction, a CNN-LSTM-AM model based on BiLSTM and the Attention Mechanism (AM) was constructed, integrating emotional factors and historical market data to achieve the joint modeling of emotion-driven and data-driven so as to capture the

deep-seated factors influencing stock price changes more comprehensively. Through systematic model design and multi-dimensional data fusion, this study aims to achieve a complete process optimization from information collection, feature extraction, emotion recognition, to trend prediction, enhance the intelligent level of financial text analysis, and provide a technical basis and theoretical reference for financial decision support.

This study focuses on the application analysis and research of the text model based on the improved CNN-LSTM in the financial field. Section II sorts out the current research progress in the topic recognition and sentiment analysis of financial texts and clarifies the deficiencies of the existing methods. In Section III, this study proposes a focused crawler model combining the advantages of CNN and LSTM for financial text collection and classification, and designs a BiLSTM sentiment analysis model integrating the attention mechanism for stock fluctuation prediction. In Section IV, the performance of the model was verified through multiple sub-experiments, including text crawling, topic classification, comment sentiment analysis and stock price prediction. The experimental results show that the proposed model is superior to the traditional methods in terms of accuracy and practicality. The conclusion section in Section V summarizes the research results and practical significance, and looks forward to the future research directions, proposing that the integration of multi-source data and the improvement of the model's generalization ability can be further explored.

II. Related Work

In the context of information explosion and highly sensitive financial markets, financial intelligence mining and intelligent prediction based on text data is gradually becoming an important research direction in the intersection field of Natural Language Processing (NLP) and financial engineering. With the continuous expansion of open source channels such as social media, news portals and review platforms, a large amount of unstructured text data is emerging, which contains a large amount of potential information that can be used for market trend analysis, public opinion monitoring and investment decision-making [6]. How to effectively integrate these heterogeneous data resources and realize the accurate identification and modeling of financial sentiment, emergencies and market dynamics, it is urgent to rely on advanced deep learning technology to build an intelligent information processing framework [7].

In recent years, convolutional neural Network (CNN) and Long Short-Term Memory network (LSTM) have achieved remarkable results in text classification and sequence modeling tasks [8]. CNN is good at extracting semantic features from local context, while LSTM is able to capture long-distance dependencies, and the joint application of them provides new ideas for deep modeling of complex financial texts [9]. However, in the face of the characteristics of high semantic complexity, strong burstiness and obvious emotional fluctuations in financial corpus, the traditional CNN-LSTM model still has shortcomings in information extraction accuracy, timeliness of event recognition and emotion modeling dimension [10]. Next, starting from the research status of financial text information processing, this study summarizes the development of related model technologies, and analyzes the main challenges and development trends faced by current research.

A. Text Mining and Topic Identification of Financial Texts

With the rapid advancement of internet technology and the media industry, news has become one of the primary channels through which the public accesses social developments and financial information [11]. Traditional print media is swiftly transitioning toward digitalization and network-based formats, and news data is now characterized by large-scale, rapid updates, and diverse content [12]. Effectively classifying and intelligently mining these large-scale news text datasets has emerged as a key research direction in the field of natural language processing (NLP) [13]. In highly sensitive domains such as finance, the intelligent processing of news data holds significant importance for information retrieval, intelligence monitoring, and market behavior prediction.

Early research in text classification primarily relied on shallow learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), which modeled news texts using the bag-of-words approach for classification [14]. While these methods demonstrated a certain level of classification capability, they often faced challenges such as feature sparsity and poor generalization when applied to large-scale, high-dimensional, and semantically complex financial texts [15]. With the rise of deep neural networks, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been increasingly adopted. CNNs are effective in capturing local key features within text and are well-suited for handling phrase-level semantic structures [16]. Meanwhile, Long Short-Term Memory networks (LSTMs), an improved variant of RNNs, perform well in processing long texts and capturing dependencies over time [17]. Researchers have attempted to combine the strengths of both architectures to overcome the limitations of single models in representing complex semantics. For example, Duan et al. conducted some studies using CNNs as feature extractors to generate embedded features, which are then fed into LSTM networks for semantic modeling, significantly improving classification accuracy in financial topic classification tasks [18].

Traditional text feature extraction methods fall into three main categories: rule-based and statistical methods, machine learning-based methods, and more recently, deep learning methods that have shown superior performance in text classification tasks [19]. Most CNN-based news data classification models are designed to learn local contextual information within fixed-size windows and extract features by sliding these windows across the text, which are then used for classification [20]. However, CNNs have limitations, particularly in capturing dependencies among local features. Addressing this issue, this study proposes a hybrid model that integrates a multi-scale dual-layer CNN with a BiLSTM and an attention mechanism, termed TCNNRes-BiLSTM-Attention (TCBA) [21]. This model simultaneously captures contextual features from both forward and backward directions of the text sequence while incorporating a dual-layer multi-scale convolutional structure [22].

B. Survey of Financial Condition Forecasting Research based on Financial Commentary Analysis

For investors, the stock market is a platform to identify and invest in high-potential companies with the aim of securing substantial returns. Over the years, both investors and researchers have been devoted to the task of stock price prediction. However, the stock market is inherently dynamic, nonlinear, non-stationary, non-parametric, noisy, and chaotic, making it extremely challenging to analyze market trends and price behaviors [23]. The stock market is influenced by a variety of highly interrelated factors, including political events, macroeconomic conditions, commodity price indices, investor expectations, performance of other stock markets, and investor psychology [23]. The correlation between public sentiment and financial markets has long attracted considerable academic attention. Studies have shown that the emotional tendencies reflected in news reports and investor commentary can, to some extent, serve as precursors to market trends—especially in times of information overload or sudden adverse events, where emotional reactions often precede actual market movements [24]. Consequently, sentiment-driven stock price prediction has gradually become an important direction in intelligent financial analysis.

Traditional stock prediction methods primarily rely on statistical techniques for time series analysis. For instance, Matta et al. employed econometric approaches to forecast stock prices, demonstrating that models using multiple predefined variables could yield statistically significant stock return predictions [25]. The Autoregressive Integrated Moving Average (ARIMA) model, introduced by Box and Jenkins, has been widely used for time series forecasting, including stock market returns [26]. However, studies have revealed that ARIMA models tend to perform poorly on financial time series data [27]. To better account for volatility, the GARCH model was proposed to predict stock prices by modeling the relationship between price volatility and returns. Nonetheless, traditional predictive methods often struggle to balance the inherent randomness and regularity of stock movements and are plagued by various limitations, making them less suited to the needs of investors.

With rapid advancements in hardware technology, the development of powerful new algorithms, and the rise of big data and distributed computing technologies, the efficiency of stock volatility prediction has greatly improved. Big data technologies have made it significantly easier to access and store large volumes of historical trading data. In recent years, machine learning algorithms have undergone explosive growth, and many researchers have applied techniques such as Support Vector Machines (SVM), decision trees, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks to stock volatility prediction, achieving promising results. SVM, first introduced by Vapnik, has demonstrated its effectiveness in many pattern recognition problems and is especially powerful in identifying subtle patterns in complex datasets [28]. It has been used in applications such as handwriting recognition, fraud detection, speaker identification, and facial recognition. Sonkavde et al. have employed SVM to identify stocks that might outperform the market by analyzing accounting fundamentals and price data [29]. Experimental results showed that equally weighted portfolios constructed from stocks selected by SVM achieved significant returns over five years, outperforming benchmark indices. Other studies predicted stock price direction using SVM and then applied trend analysis to derive corresponding price predictions. Despite modest profitability, these models still performed better than bank interest rates, suggesting a degree of practical value [30]. Additional research proposed a two-stage feature selection and prediction model using SVM, demonstrating better generalization ability than traditional methods [30]. Furthermore, Principal Component Analysis (PCA) has been introduced into SVM-based models to extract low-dimensional, efficient features, improving training accuracy and preserving the original data's characteristics. Empirical studies showed that PCA-SVM stock selection models contributed more significantly to annual portfolio returns than benchmark indices [31].

C. The Comparison Between the Current Research and this Study

In the current research on text classification and sentiment analysis, the academic community has accumulated a large number of methods based on traditional machine learning and deep learning, forming a relatively rich research system. The existing research mainly focuses on text classification models based on traditional algorithms such as Bayesian and Support Vector Machine (SVM), as well as sentiment analysis methods based on deep neural networks that have emerged in recent years. However, by systematically comparing these existing works with the model proposed in this study, the gaps and deficiencies of the current research can be clearly identified, and the innovation and necessity of this study can be highlighted. Early studies mostly relied on the Naive Bayes model. Although this method laid the foundation for text classification, it has obvious limitations of the independence assumption and is difficult to fully model the complex correlations among text features. LMDCampos et al. attempted to introduce a synonym library modeling through Bayesian networks, improving the representation ability of text semantics, but still remained at the relatively shallow stage of semantic extraction. Luo Huiqin, Zeng Yu, Li Xiaodong and others have optimized Naive Bayes to varying degrees, such as introducing feature weights and improving the calculation method of TF-IDF weights, which has enhanced the classification accuracy. However, the expressive ability and generalization ability of these methods are still restricted by the structure of the model itself. With the introduction of Support Vector Machine, the performance of text classification has been significantly improved. Xiao Zheng, Ding Shengchun, Shi Qiangqiang and others applied SVM to tasks such as sentiment tendency analysis and microblog opinion recognition, and achieved good results. However, the computational efficiency of the SVM algorithm is still relatively low when dealing with large-scale corpora and multi-dimensional emotional features, and the model has poor sensitivity to the word order of the input data, making it difficult to capture the context semantics.

In terms of deep learning, KIM Y improved the performance of sentence-level classification by introducing the CNN structure with parameter tuning, pioneering a new direction for the application of deep learning in text analysis. Zhou Jinfeng

further expanded the structure of CNN, utilized multi-layer convolution to extract local semantic features, and enhanced the expressive ability of the model. However, CNN relies on a fixed window during modeling, lacks sensitivity to time series features, and is unable to effectively retain the context dependency of the text. To solve this problem, subsequent research gradually shifted to models based on RNNS. Huang M introduced tag-specific synthetic functions and structural control strategies in RNN, which improved the fine-grained control ability of sentiment classification. FeiH enhances the semantic capture ability of LSTM through the keyword sensitivity mechanism, achieving a slight performance improvement. RaoG proposed the SR-LSTM model, which models the semantic relationships at the sentence and document levels respectively, through a double-layer structure, and captures the emotional information in long texts more effectively. Although these improved schemes based on LSTM have made certain progress, most studies are still limited to the single task of sentiment classification and lack the ability of multimodal data fusion and interaction modeling between sentiment and market behavior. Therefore, compared with the traditional static modeling methods based on Bayesian or SVM, this study breaks through the limitations of shallow feature engineering. Compared with the existing deep learning models, it has also achieved substantial breakthroughs in structural design and multi-source information integration. Especially in the application of the financial field, it has filled the gap in the modeling of the linkage between sentiment analysis and market behavior, providing a more practical and scalable technical path for subsequent research on financial text analysis and prediction.

III. Model Establishment and Solution

A. Research on Focused Crawler Technology Based on CNN-LSTM

In today's highly dynamic and information-intensive financial market environment, the ability of corporate decision-makers, financial analysts, and policymakers to perceive industry trends and breaking events has become a critical factor influencing the efficiency and quality of decision-making. Faced with rapidly evolving market conditions, financial professionals urgently need to rely on accurate and timely information to assess trends, evaluate risks, and adjust strategies. However, with the explosive growth of internet-based information, financial data is now characterized by large volume, rapid updates, high noise levels, and strong heterogeneity [32]. Traditional methods that depend on manual information retrieval and classification are increasingly unable to meet real-world demands. These manual approaches are not only inefficient but also susceptible to subjective judgment and information delays, which compromise the timeliness and scientific rigor of financial activities [33].

Compared with general-purpose texts, financial news exhibits strong topic cohesion and a high density of domain-specific terminology. Such news content typically centers on core themes including stock market trends, industry mergers and acquisitions, macroeconomic policies, and corporate earnings reports. The text often includes frequently used professional terms such as "price-to-earnings ratio", "return on equity", "policy dividend", and "acquisition agreement". This semantic concentration provides a favorable basis for building automated topic recognition models, as it allows the models to more easily capture and differentiate the semantic distinctions between various news topics.

- Overview of Focused Crawler Technology

This topic classifier functions as a binary text classification model, distinguishing between topic-relevant and topic-irrelevant content. The topic-focused crawling framework proposed in this study is illustrated in Fig. 1.

The core part of the focused crawler workflow is topic relevance calculation. In the content of news text, the sentences containing the topic content are often short and concise, so before the topic discrimination. Sentences with sentence length between 20 and 50 are selected for combination, and then the topic relevance is judged by the topic discriminator.

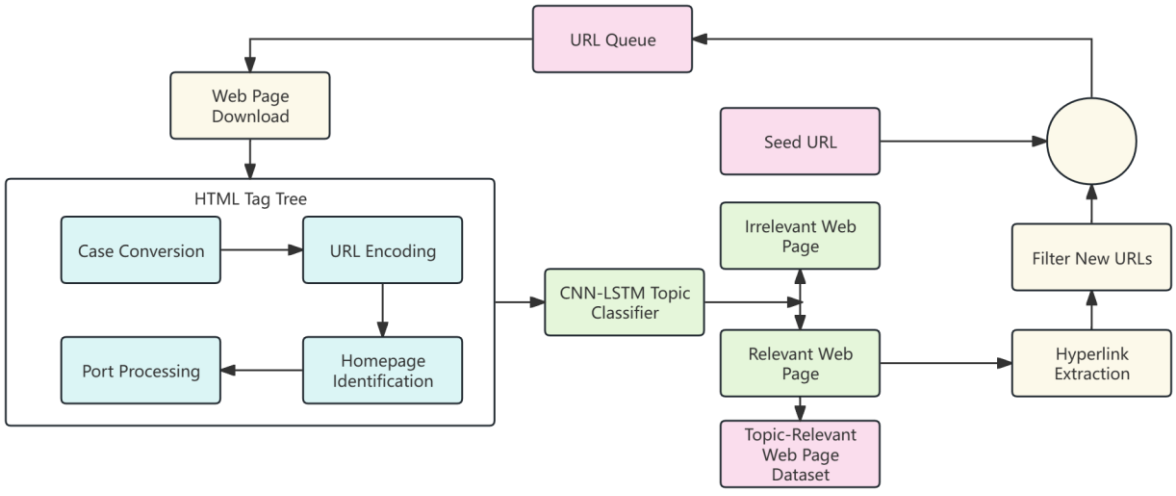


Fig. 1. Framework of focused crawler.

- The Focused crawler classifier

The text classification model based on CNN-LSTM can better extract features in documents. The training flow of the text classification model is shown in Fig. 2. After multiple experimental comparisons, the experimental parameters are set as follows: the learning rate of Adam optimizer is 0.001, the training set and the test set are divided according to the ratio of 8:2, dropout is set to 0.5, and epoch is set to 100.

First, the algorithm takes a preprocessed text dataset as input, where each sample is a document associated with a specific category label. The preprocessing typically includes steps such as tokenization, stop-word removal, and word normalization. Each document is represented as a sequence of words, $W_1, W_2, \ldots, W_n$, which are then fed into an embedding layer. The role of the embedding layer is to map each word to a vector in a continuous space, capturing semantic relationships between words, and thereby producing a sequence of word vectors. These word vectors are subsequently passed through a convolutional layer. The convolutional layer uses a sliding window mechanism to extract local features, effectively capturing word-level associations. After the convolution operation, a pooling layer is typically applied. The pooling layer downsamples the convolution output to reduce feature dimensionality and retain key information, which improves computational efficiency and generalization. The output from the pooling layer is then passed into an LSTM layer. LSTM, a specialized form of recurrent neural network (RNN), is capable of handling long-term dependencies and is well-suited for sequential modeling tasks. In text classification, the LSTM layer learns the sequential and contextual semantics in the document, capturing dependencies across different parts of the text.

The output of the LSTM is passed to a fully connected layer, which integrates the high-level features and produces a fixed-size output vector. This vector is then processed by a Softmax layer, which converts the vector into a probability distribution over all possible classes, thus determining the predicted category of the input document. The entire model is trained using backpropagation. In each training epoch, the model iterates over all training samples, performs predictions, computes the loss (typically using cross-entropy), and then calculates gradients based on the loss. Using these gradients, an optimizer (such as Adam, with a learning rate of 0.001) updates the model parameters to minimize the loss function. This process continues for a predefined number of epochs (set to 100 in the experiment) or until the model converges. To prevent overfitting, a Dropout mechanism is introduced during training, with a dropout rate set to 0.5. Before training begins, the dataset is split into a training set and a test set in an 80:20 ratio. After training, the model is evaluated on the test set to assess its generalization performance.

This pseudocode provides a clear step-by-step view of how a CNN-LSTM model is trained for text classification, covering data preprocessing, embedding, forward pass, loss computation, backpropagation, and model update.

- CNN-LSTM hybrid model

The CNN-LSTM hybrid model is a deep learning model that combines the advantages of Convolutional Neural Network (CNN) and Long Short-Term Memory network (LSTM), which can simultaneously capture the local features and temporal relationships of the input data [4]. The overall structure of the model is shown in Fig. 2.

The computation in the convolutional layer is defined by the following formula:

$$O = f(WZ + b) \tag{1}$$

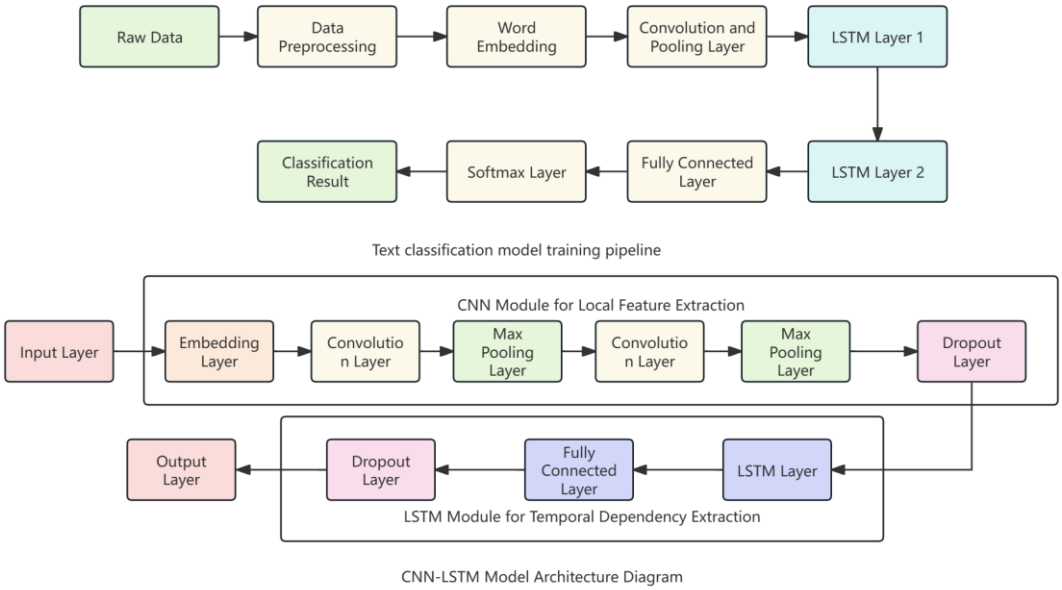


Fig. 2. Text classification training process and classification algorithm framework.

where, O is the output feature matrix after convolution, W is the weight matrix (convolutional kernel), Z is the input word embedding matrix, b is the bias term, f is the ReLU (Rectified Linear Unit) activation function, defined as $\mathrm{ReLU}(x) = \max(0, x)$, where x is the output vector from the previous neural layer.

The Long Short-Term Memory (LSTM) network is a special type of Recurrent Neural Network (RNN) designed to overcome the problems of vanishing or exploding gradients that occur when handling long sequences. LSTM introduces a gating mechanism—comprising the input gate, forget gate, and output gate—that allows the network to selectively remember or forget information. This mechanism enables LSTM to effectively capture long-range dependencies in sequential data. The fundamental architecture of an LSTM cell is illustrated in Fig. 3.

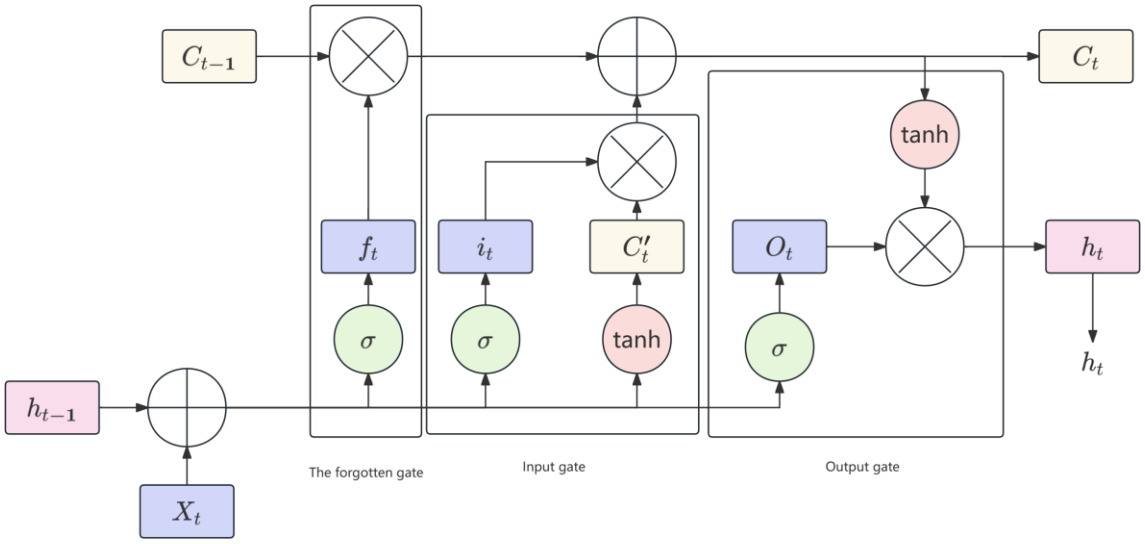


Fig. 3. Diagram of the basic unit structure of an LSTM model.

Lastly, the hidden state output h_t is computed by applying the output gate to the tanh-transformed current cell state c_t. The variables t and $t-1$ refer to the current and previous time steps, respectively, highlighting the sequential nature of the data being processed. Together, these components enable the LSTM to maintain and update information over time, effectively capturing long-term dependencies in sequential text data.

B. Sentiment Analysis of News and Investor Reviews

Investor comments have emotional tendencies, and the subjective views and emotional tendencies of comment authors can be mined through sentiment analysis. This plays an important role in the research field of the relationship between reviews and stocks, and the addition of review features can make the stock market prediction results more comprehensive, objective and explanatory.

The CNN-LSTM model first extracts the features of stock day-level funds, market prices, technical factors, news and comments through CNN, and then inputs the extracted features into the LSTM network for training to extract relevant time series information. At the same time, the Attention Mechanism is introduced after the LSTM network. This allows the network to focus on important features during training. The overall network structure of the proposed CNN-LSTM-AM model is shown in Fig. 4.

The basic unit of the Long Short-Term Memory (LSTM) model consists of four main components: the forget gate, input gate, memory cell, and output gate. These components work together to manage the balance between long-term memory retention and short-term information update, enabling the network to capture complex dependencies in sequential data.

The forget gate f_t decides which information from the previous memory cell should be discarded. It uses a sigmoid activation to output a value between 0 and 1 for each element in the memory cell:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{2}$$

The input gate i_t determines which new information is to be added to the memory cell:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{3}$$

At the same time, a candidate memory content $\tilde{c}_t$ is created via a tanh layer:

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{4}$$

The new memory cell state c_t is computed by combining the previous memory cell c_{t-1}, the forget gate f_t, the input gate i_t, and the candidate content $\tilde{c}_t$:

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t \tag{5}$$

The output gate o_t decides which part of the memory cell should be output.
The final hidden state h_t is obtained by applying the output gate to the updated cell state after passing through a tanh function:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t \times \tanh(c_t) \quad (7)$$

In the LSTM formulas, f_t, i_t, and o_t represent the states of the forget gate, input gate, and output gate, respectively. These gates control how information flows through the LSTM cell at each time step. The symbols W_f, W_i, and W_o denote the weight matrices associated with each of these gates, while b_f, b_i, and b_o are their corresponding bias terms. The notation $[h_{t-1}, x_t]$ indicates the concatenation of the hidden state from the previous time step h_{t-1} and the current input vector x_t. This concatenated vector serves as the input to the various gate functions.

The sigmoid function, represented by σ, is used to squash the outputs of the gates to values between 0 and 1, effectively determining the degree to which information should be remembered or forgotten. The candidate memory content $\tilde{c}_t$ is a proposed update to the cell state, calculated using the tanh function to ensure its values fall within the range of -1 to 1. The current memory cell state, denoted as c_t, is updated by combining the retained portion of the previous cell state c_{t-1} with the new candidate content $\tilde{c}_t$, scaled by the input gate. This combination is carried out through element-wise multiplication, indicated by the symbol $\times$, also known as the Hadamard product.

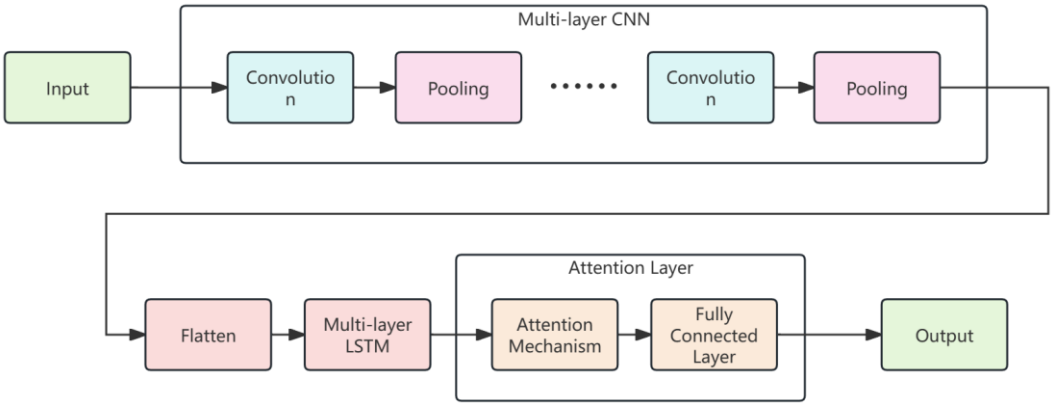


Fig. 4. CNN-LSTM-AM network structure diagram.

IV. Experimental Results and Analysis

A. Text Crawler Experiments

- Experimental parameter setting

In this section, the experiments were conducted using PyCharm as the development environment, with deep learning tasks implemented using PyTorch, a scientific computing library based on Python. The hardware setup includes an Intel i9-14900K CPU and an NVIDIA GeForce RTX 2080Ti 22GB GPU. All Python third-party libraries required in the experiments were managed using Anaconda.

In the control experiment, the maximum collection number of crawlers is set to 1000, that is, when T = 1000, the harvest rate of the theme crawler is compared as shown in Fig. 5. With the increase of the number of crawled pages, the harvest rate of each control experiment decreases, but the model proposed in this section is better than other groups, and the harvest rate is 0.72 at the end of the crawling task. The harvest rate is 0.10 higher than that of the focused crawler based on feature enhanced RNN, 0.26 higher than that of the focused crawler based on HMM, and 0.05 higher than that of the breadth-first strategy crawler.

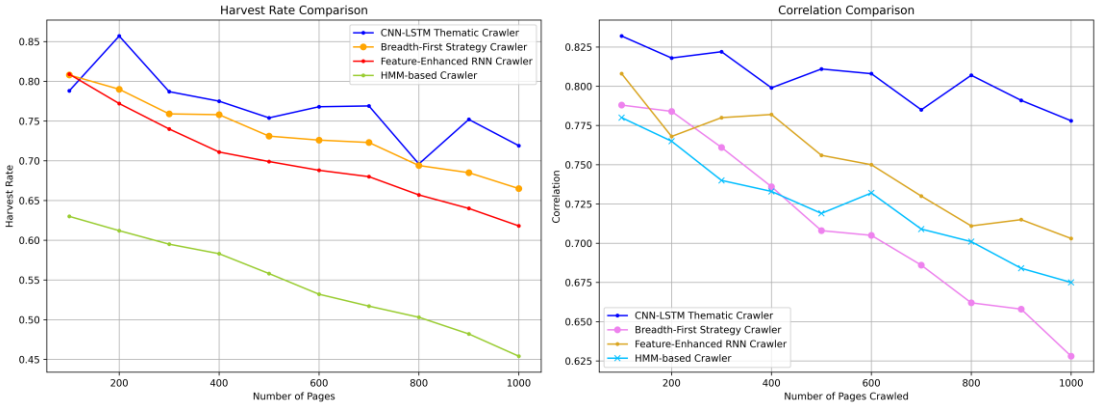


Fig. 5. Comparison of harvest rate and relevance of focused crawler

The correlation comparison of focused crawlers as the number of crawled pages increases, the overall correlation of each model decreases. At the end of the crawling task, the correlation of the model proposed in this section is 0.78, which is 0.08 higher than that of the focused crawler based on feature-enhanced RNN. Compared with the focused crawler based on HMM, the correlation is 0.11 and 0.15 higher than the breadth-first strategy crawler.

From the comparison of the above experimental results, it can be seen that the crawler based on HMM has the worst performance in the harvest rate and relevance index of the topic crawler because it does not discriminate the topic relevance. The focused crawler based on CNN-LSTM proposed in this study is better than other models based on content evaluation or link structure recognition.

B. *Experimental Design and Evaluation of Topic Classification Models*

- Dataset and metrics

In recent years, the text content of the news has a high time efficiency, and the use of the earlier data set may lead to the model's problem of the new term, so in the subject classification model experiment, on the one hand, according to the influence of the news reading quantity and the number of critics, the financial news text is used as the self-established language material library. On the other hand, the thcnews library, which is released by Tsinghua University, is based on the historical data of sina news between 2005 and 2011, which is divided into 14 categories, which randomly select 4,000 financial related news texts and 6,000 other types of news text [34]. The url is http://thuctc.thunlp.org/. Then, the material library is divided into financial topics, using 1 tag to mark the current data, which is related to the financial theme, which indicates that the current data is irrelevant to the financial theme. The distribution of each dataset is shown in Fig. 6, which consists of a random selection document in a mixed news cube, a training set, a validation set, and a test set.

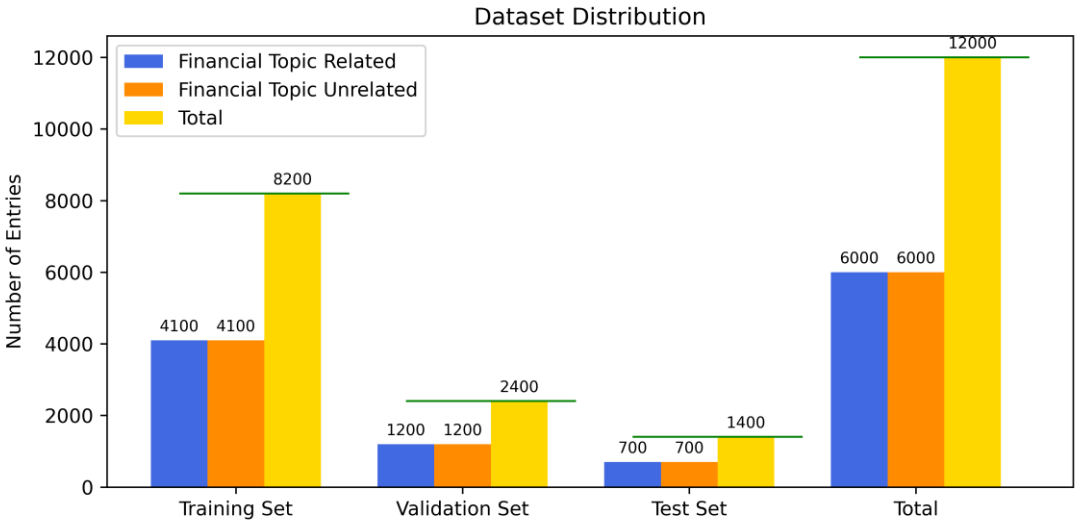


Fig. 6. Dataset distribution diagram.

- Experimental design and result analysis

By designing comparative experiments with machine learning and deep learning models, experiments are carried out on a mixed news corpus. The results of model comparison experiments are shown in Table I.

TABLE I MODEL COMPARISON OF EXPERIMENTAL RESULTS

Algorithm	Accuracy	F－Measure
SVM	83.6	80.2
RNN	86.3	83.5
CNN	91.1	87.2
RCNN	92.2	89.6
CNN－LSTM	93.4	90.8

By comparing the model proposed in this section with four models including Support Vector Machine (SVM), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) and Recurrent Convolutional Neural network (RCNN), the accuracy and F-Measure of the CNN-LSTM model can be concluded. The accuracy of the CNN-LSTM model reaches 93.4%. It indicates that the ability to correctly classify text is the strongest among all the compared models. Similarly, it also performs well on F-Measure, reaching 90.8%, indicating a good balance between precision and recall. Through longitudinal comparison, it is found that CNN-LSTM and RCNN models are superior to the traditional SVM model and the basic RNN model in accuracy and F-Measure. This is because for RCNN model, the advantages of RNN and CNN are combined, and for CNN-LSTM model, CNN can effectively capture local features. LSTM can capture long-term dependencies and process sequence data better through the characteristics of the two models to achieve good classification results. Therefore, the CNN-LSTM model proposed in this section is able to have better results in the topic classification task.

C. *Experiments on Sentiment Analysis of Reviews*

- Optimization parameter experiment

In this experiment, several different batch sizes of 8, 16, 32, 64, and 128 were selected for exploration. As shown in Table II, the experimental results show that when the batch size is chosen as 64, the model has the best effect, and the accuracy reaches 83.06%.

TABLE II ACCURACY STATISTICS FOR DIFFERENT BATCHSIZES

Batchsize	Accuracy
8	0.8127
16	0.8296
32	0.8257
64	0.8306
128	0.8263

In machine learning, the learning rate has a great impact on the performance of the model. The learning rate determines how much the parameters of the neural network are updated in each iteration. If the learning rate is too large, the parameter update step size will be too large, and the global optimal solution may be skipped, resulting in unstable training or failure to converge. If the learning rate is too small, the parameter update step size will be too small, which will lead to too slow training speed and need more iterations to converge. In this experiment, 0.5, 0.2, 0.1, 0.05, 0.02 and 0.01 were selected for exploration, and the experimental results are shown in Table III. As can be seen from the table, as the learning rate decreases from 0.5 to 0.05, the Accuracy of the model increases continuously. When the learning rate is 0.05, the model has the best effect, and the accuracy reaches 84.13%. When the learning rate is slowly decreased from 0.05 to 0.01, the model Accuracy decreases. Therefore, the learning rate of 0.05 is appropriate.

TABLE III ACCURACY VALUES FOR DIFFERENT LEARNING RATES

Learning rates	Accuracy
0.5	0.7526
0.2	0.7909
0.1	0.8306
0.05	0.8413
0.02	0.8335
0.01	0.8316

By adjusting the parameters above, the accuracy of the final model reaches a relatively excellent level. The stock review text of the test set is input into the model to predict the sentiment tendency, and the final evaluation index of the model is shown in Table IV below. It can be found that the model realizes a more accurate classification of investor comments, which lays a foundation for subsequent stock prediction.

TABLE IV INVESTOR COMMENTS CLASSIFICATION PREDICTION EVALUATION

	F1 score	Recall	Accuracy
Positive class	0.76	0.81	84.13%
Negative class	0.77	0.79	
Neutral class	0.79	0.83	

- Comparison model analysis

In this section, as a comparison, the investor comment vectors processed by word2vec are input into BP neural network, LSTM neural network, BiLSTM neural network and other machine learning models to compare their classification effects. After parameter tuning, the classification effect of each model is shown in Table V. It can be seen that the BERT-bilSTM model based on Bert and Bi-LSTM model has the best classification effect on investor comments, and its accuracy on the test set reaches 84.13. Therefore, in this study, Bert-BiLSTM is used to classify the sentiment of investor comments.

TABLE V EFFECT OF EACH MODEL INVESTOR COMMENT SENTIMENT CLASSIFICATION TASK

Model	Macro__F1__score	Macro__Recall	Accuracy
BP	0.74	0.79	79.38%
LSTM	0.76	0.79	82.74%
BiLSTM	0.76	0.80	83.24%
Bert—BiLSTM	0.77	0.81	84.13%

In this study, we first conducted a systematic exploration of two types of key hyperparameters, namely Batch Size and Learning Rate, and precisely depicted their influence on the performance of sentiment classification by controlling variables one by one. In terms of Batch Size, the experimental results (Table II) show that during the process when the Batch Size of the model was doubled successively from 8 to 64, the accuracy rate generally showed an upward trend: increasing from 0.8127 (Batch Size = 8) to 0.8306 (Batch Size = 64). This phenomenon can be attributed to the fact that a moderately increased batch size can not only reduce the variance of gradient estimation and stabilize the model update to a certain extent, but also will not cause a sharp decrease in the number of effective iterations due to excessive video memory occupation. It is notable that when the Batch Size is further increased to 128, the accuracy rate drops back to 0.8263 instead, which is consistent with the common "generalization deterioration" in large-scale training: Excessive batch size integrates too many sample signals in a single update, weakening the ability of parameters to "noise-driven" traverse saddle points, thereby reducing the performance of the test set. Comprehensively considering the performance peak and computing cost, Batch Size = 64 was determined as the optimal configuration.

The learning rate experiment (Table III) further reveals the significant influence of the optimized step size on the convergence quality. When the learning rate decreases from 0.5 to 0.05, the accuracy rate rises rapidly from 0.7526 to 0.8413, indicating that an overly large step size can cause the gradient update to cross the optimal region or even oscillate, while gradually reducing the step size can refine the parameter search and significantly improve the model discrimination ability. However, when the learning rate was further reduced to 0.02 and 0.01, the accuracy rates dropped to 0.8335 and 0.8316 respectively. This indicates that although a too small learning rate can avoid oscillations, it inhibits the ability to fully explore the parameter space and jump out of local minima, and slows down the convergence speed - thereby reducing the final accuracy. Therefore, considering the comprehensive training stability and performance, a learning rate of 0.05 was selected as the best. After the above two hyperparameters were optimized in place, the three-category evaluation index of the model on the test set (Table IV) showed that the overall accuracy rate reached 84.13%, while maintaining a relatively balanced category recall rate (positive 0.81, negative 0.79, neutral 0.83) and F1 score (in the range of 0.76-0.79). It

indicates that the parameter configuration takes into account the identification of different emotional tendencies, and there is no significant bias towards any category.

Further comparative experiments (Table V) compared Bert-BiLSTM with baselines such as BP, one-way LSTM and BiLSTM. Under the unified optimal hyperparameter Settings, Bert-BiLSTM established a leading advantage with an accuracy rate of 84.13%, a macro average recall rate of 0.81, and a macro average F1 score of 0.77. This result indicates that the superposition of BERT's pre-trained semantic representation and BiLSTM's bidirectional context capture ability significantly improves the comprehensive modeling effect of fine-grained semantics and emotional cues in investor comments. However, neither the shallow BP network with the same hyperparameters nor the one-way/two-way LSTM can achieve such accuracy on the same dataset.

D. Research on Stock Volatility Prediction Model

In general, the deeper the CNN layers are, the more abstract and high-level feature information the network can extract, and the more expressive the model can be. However, too deep networks also bring problems such as training difficulty, vanishing gradient and exploding gradient, which make training more difficult. In this study, the CNN network was constructed from a relatively shallow number of layers, and the input layer, 1 convolutional layer, and 1 pooling layer were initially set, and then the convolutional layers and pooling layers were gradually increased until 5 convolutional layers were reached. Since the amount of data is not large, a too deep network structure is not used. A total of five experiments were conducted this time, and the obtained results are shown in Table VI, where each index is the average value of the prediction evaluation of stocks, including New Hope.

TABLE VI THE AVERAGE SCORE OF THE RISE AND FALL OF DIFFERENT NUMBER OF HIDDEN LAYERS PREDICTION

Model types	Number of hidden layers	Accuracy	F1—score	AUC
CNN	1	0.58	0.58	0.59
	2	0.57	0.58	0.58
	3	0.57	0.57	0.58
	4	0.56	0.57	0.56
	5	0.57	0.57	0.56
LSTM	1	0.58	0.58	0.59
	2	0.57	0.58	0.58
	3	0.57	0.58	0.57
	4	0.56	0.57	0.56
	5	0.57	0.58	0.58

It can be seen from the data that the AUC score is the highest when the number of CNN network hidden layers is 1. As the number of hidden layers increases, the AUC score slowly decreases and then fluctuates between 0.58 and 0.56. According to the experimental results, when the number of hidden layers of the CNN of the model in this study is 1, it is the optimal solution to select the data set.

After adding one layer of LSTM, its AUC and accuracy decrease. As the number of hidden layers increases, the average AUC score fluctuates between 0.58 and 0.56. Therefore, 1 hidden layer is selected as the optimal solution in this study, that is, the number of LSTM hidden layers in the initial model is kept constant. In this section, a CNN-LSTM-AM hybrid neural network model is constructed based on the attention mechanism, CNN neural network, and LSTM neural network. Then, news and investor sentiment are integrated into the training data to train the CNN-LSTM-AM model and use it to predict the rise and fall of stock prices. However, the specific impact of the addition of attention mechanism, news and investor sentiment on the model is worth considering, so this section comprehensively analyzes the impact of these two factors on the model. The comprehensive performance comparison is shown in Table VII.

In general, after adding the features of news and investor comments, the four evaluation indicators of Accuracy, F1-score, Recall and AUC of each major model have different degrees of improvement. Specifically, after adding the sentiment features of news and investor comments, each evaluation index of the BP neural network model is increased by 0.025 on average. The average improvement of the CNN model indicators is 0.0325. Each evaluation index of the LSTM model was increased by 0.0275 on average. After adding news and comment features, the Accuracy, F1-score and AUC indicators of the CNN-LSTM model are increased by 0.03, and the Recall is increased by 0.02. The CNN-LSTM-AM model constructed in this study has the best effect. After adding news and comment features, the average AUC index reaches 0.64, and the Accuracy and F1-score both reach 0.63, which is an obvious improvement effect. When the sentiment features of news and investor comments are not added, the overall performance of CNN-LSTM-AM is also better than that of other comparison models. In summary, all models have significantly improved in each evaluation index after adding news and comment features. The CNN-LSTM-AM model performs better than the above comparison models. Therefore, it can be considered that the features of news and investor comments are effective features, which can be used to increase the accuracy of the model in stock prediction.

TABLE VII WHETHER TO ADD NEWS AND INVESTOR SENTIMENT PERFORMANCE COMPARISON (MEAN VALUE)

	Join or not	Accuracy	F1—score	Recall	AUC
BP	No	0.55	0.55	0.53	0.55
	Yes	0.57	0.57	0.56	0.58
CNN	No	0.58	0.57	0.57	0.59
	Yes	0.61	0.61	0.60	0.62
LSTM	No	0.58	0.59	0.57	0.59
	Yes	0.61	0.60	0.61	0.62
CNN—LSTM	No	0.59	0.58	0.58	0.60
	Yes	0.62	0.61	0.60	0.63
CNN—LSTM—AM	No	0.60	0.60	0.58	0.62
	Yes	0.63	0.63	0.61	0.64

E. Model Backtesting Experiment

This study uses CNN-LSTM-AM to mine the time series information in stock data, and combines news information and investor comment information in the stock forum. In this subsection, the CNN-LSTM-AM model of this study is compared with the traditional CNN model, LSTM model and CNN-LSTM model based on the use of the strategies in this study. After multiple parameter adjustments and optimizations,

Tables VIII to X list the backtest returns, maximum drawdown rates, and Sharpe rates of the selected 8 stocks from January 1, 2022 to January 1, 2025.

TABLE VIII COMPARISON OF BACKTEST RETURNS FOR SELECTED STOCKS

Stock	Benchmark	CNN-LSTM-AM	CNN-LSTM	CNN	LSTM
Everbright Securities	15.74%	108.55 %	85.22%	74.38%	78.82%
GoerTek Inc.	175.04%	268.69 %	178.44 %	148.78 %	169.37 %
Vanke A	-33.94%	17.37%	12.17%	6.25%	10.29%
New Hope Liuhe	-23.85%	78.38%	62.69%	56.48%	67.03%
Digital China Health	59.15%	98.45%	93.33%	73.16%	85.56%
Jinling Pharmaceutical	8.82%	57.14%	48.12%	46.63%	49.88%
Luzhou Laojiao	208.15%	412.35 %	369.18 %	343.81 %	398.82 %
Changan Automobile	110.91%	183.68 %	165.37 %	155.44 %	171.65 %

Table IX shows the comparison of the maximum drawdown rates in the simulation backtesting between the model in this study and other models. It can be seen from the table that the average maximum drawdown rate of the model in this study is 28.78%, and the maximum drawdown rates of stocks are all lower than those of the benchmark and other models. Overall, CNN-LSTM-AM reduces the maximum drawdown rate of the model.

TABLE IX COMPARISON OF MAXIMUM DRAWDOWN RATES FOR SELECTED STOCKS

Stock	Benchmark	CNN-LSTM-AM	CNN-LSTM	CNN	LSTM
Everbright Securities	52.49%	23.46%	24.30%	30.58%	28.13%
GoerTek Inc.	47.87%	32.19%	33.45%	39.34%	35.26%
Vanke A	44.30%	26.89%	31.41%	37.32%	33.51%
New Hope Liuhe	73.75%	36.61%	41.12%	46.41%	47.36%
Digital China Health	43.71%	21.58%	29.65%	38.36%	31.64%
Jinling Pharmaceutical	34.33%	24.81%	31.29%	35.84%	32.95%
Luzhou Laojiao	48.27%	30.34%	34.37%	42.54%	40.41%
Changan Automobile	49.92%	34.39%	41.29%	46.71%	43.62%
Average	49.33%	28.78%	33.36%	39.64%	36.61%

As shown in Table X, the average Sharpe ratio of CNN-LSTM-AM is 1.12, which not only exceeds the benchmark value of 0.43, but also is the highest among the four models. Especially in the backtesting of Vanke A and New Hope, the models successfully increased the benchmark Sharpe ratio, which was negative to positive.

TABLE X COMPARISON OF SHARPE RATIOS FOR SELECTED STOCKS

Stock	Benchmark	CNN-LSTM-AM	CNN-LSTM	CNN	LSTM
Everbright Securities	0.09	0.75	0.59	0.45	0.55
GoerTek Inc.	1.17	1.84	1.21	0.98	1.13
Vanke A	-0.66	0.49	0.37	0.21	0.28
New Hope Liuhe	-0.09	1.04	0.89	0.77	0.81
Digital China Health	0.52	0.79	0.60	0.47	0.51
Jinling Pharmaceutical	0.05	0.21	0.07	0.04	0.06
Luzhou Laojiao	1.40	2.42	2.13	1.82	1.93
Changan Automobile	0.94	1.39	1.28	1.14	1.22
Average	0.43	1.12	0.89	0.74	0.81

Based on the above-mentioned experimental results, it is confirmed that the stock trading model based on CNN-LSTM-AM is a feasible quantitative timing strategy. Compared with the CNN, LSTM and CNN-LSTM models, this model shows a higher average backtest yield and Sharpe rate, as well as a lower maximum drawdown rate. This means that this model can achieve higher returns while undertaking relatively lower risks, providing certain references for users' investment decisions.

V. CONCLUSION

A. The Content and Significance of this Study

This study set out to investigate how an improved CNN-LSTM architecture can be leveraged—together with modern language-representation techniques and sentiment factors—to advance open-source intelligence (OSINT) mining, emergent-event detection, and stock-price fluctuation prediction in the financial domain. By unifying local-feature perception in convolutional layers with the long-range dependency modeling capacity of LSTM, the proposed CNN-LSTM topic-crawler raises both the harvest rate and the mean topical relevance of crawled documents, outperforming vanilla CNN, LSTM, and classical TF–IDF baselines. These gains materially reduce the latency between news release and intelligence ingestion, a critical requirement for time-sensitive financial analysis. The hierarchical detector first filters non-emergent content and then refines the categorization of bona fide sudden events by fusing word-vector dispersion features that encode class–term association strength and contextualized representations from BERT. Coupling these dual channels with an attention-augmented TextCNN boosts F1 scores across all emergent categories, validating the merit of jointly capturing local salience and global semantics in highly volatile news streams. Integrating news sentiment, investor-comment sentiment (obtained via a BERT-BiLSTM classifier), and historical trading factors into a convolutional–recurrent network enhanced with an attention mechanism (CNN-LSTM-AM) yields superior directional-accuracy and RMSE metrics relative to single-source or single-architecture benchmarks.

B. Significance of Research

Taking the financial news text as the breakthrough point, this study deeply discusses the application path and practical effect of the text processing technology based on the improved CNN-LSTM model in financial intelligence mining, which has significant theoretical and application value. By introducing the hybrid architecture of CNN and LSTM, combined with its advantages in local feature extraction and temporal dependency modeling, the topic news crawler system constructed in this study realizes the efficient classification and accurate collection of financial news texts. Compared with the traditional crawler system based on keywords or rule-driven, the system greatly improves the acquisition efficiency and relevance of target information, and provides strong technical support for the construction of a high-quality intelligence database in the financial field. As an important external variable affecting the volatility of the stock market, the rapid identification and classification of emergencies is an important part of the financial intelligence system.

For the future, the research can be deepened along the following directions. Firstly, a multilingual and multi-market cross-domain financial intelligence knowledge base was constructed, and an incremental online learning framework was introduced to realize the rapid adaptation and transfer generalization of the model in complex heterogeneous scenarios. Self-supervised contrastive learning and joint fine-tuning strategies are combined to mitigate the performance degradation caused by cross-lingual semantic drift. Secondly, the data modality and interaction dimension should be expanded: in addition to news and text public opinion, image, audio and transaction depth data should be fused, and multimodal Transformer and spatio-temporal graph neural network should be introduced to jointly model the market microstructure signal end-to-end, so as to improve the prediction strength of sudden black swan events and complex chain reactions. Furthermore, combined with symbolic reasoning and a knowledge graph, financial regulations, macro policies, and enterprise knowledge can be embedded into the event extraction and sentiment analysis process, and the causal inference ability and explanation depth of the model can be enhanced by executable logic rules.

ACKNOWLEDGMENT

Thanks for the data support provided by the Project of English Digital Resources Construction Institute (No.: GKY-2022CQJG-9)

REFERENCES

[1] K. B. Hansen and C. Borch, "Alternative data and sentiment analysis: Prospecting non-standard data in machine learning-driven finance," Big Data & Society, vol. 9, no. 1, p. 20539517211070701, 2022.

[2] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," IEEE Wireless Communications, vol. 29, no. 1, pp. 210–219, 2022.

[3] S. Sajid, "Anticipating The Stock Market Trend Using Natural Language Processing Techniques," Electrical Engineering in School of Electrical Engineering & Computer …, 2024.

[4] A. Peivandizadeh et al., "Stock market prediction with transductive long short-term memory and social media sentiment analysis," IEEE Access, 2024.

[5] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," ACM Computing Surveys, vol. 56, no. 9, pp. 1–42, 2024.

[6] A. C. Ikegwu, H. F. Nweke, C. V. Anikwe, U. R. Alo, and O. R. Okonkwo, "Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions," Cluster Computing, vol. 25, no. 5, pp. 3343–3387, 2022.

[7] C. Kyrkou, P. Kolios, T. Theocharides, and M. Polycarpou, "Machine learning for emergency management: A survey and future outlook," Proceedings of the IEEE, vol. 111, no. 1, pp. 19–41, 2022.

[8] R. K. Behera, M. Jena, S. K. Rath, and S. Misra, "Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data," Information Processing & Management, vol. 58, no. 1, p. 102435, 2021.

[9] L. Khan, A. Amjad, K. M. Afaq, and H.-T. Chang, "Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media," Applied Sciences, vol. 12, no. 5, p. 2694, 2022.

[10] B. Chakravarthi, S.-C. Ng, M. Ezilarasan, and M.-F. Leung, "EEG-based emotion recognition using hybrid CNN and LSTM classification," Frontiers in computational neuroscience, vol. 16, p. 1019776, 2022.

[11] M.-F. de-Lima-Santos, L. Mesquita, J. G. de Melo Peixoto, and I. Camargo, "Digital news business models in the age of industry 4.0: Digital Brazilian news players find in technology new ways to bring revenue and competitive advantage," Digital Journalism, vol. 12, no. 9, pp. 1304–1328, 2024.

[12] D. Giomelakis, "Semantic search engine optimization in the news media industry: Challenges and impact on media outlets and journalism practice in Greece," Social Media+ Society, vol. 9, no. 3, p. 20563051231195545, 2023.

[13] K. Sharifani, M. Amini, Y. Akbari, and J. Aghajanzadeh Godarzi, "Operating machine learning across natural language processing techniques for improvement of fabricated news model," International Journal of Science and Information System Research, vol. 12, no. 9, pp. 20–44, 2022.

[14] K. K. Kiilu, "Sentiment Classification for Hate Tweet Detection in Kenya on Twitter Data Using Naïve Bayes Algorithm," JKUAT-COETEC, 2021.

[15] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning--based text classification: a comprehensive review," ACM computing surveys (CSUR), vol. 54, no. 3, pp. 1–40, 2021.

[16] Y. Yan, Y. Kang, W. Huang, and X. Cai, "Chinese medical named entity recognition utilizing entity association and gate context awareness," PloS one, vol. 20, no. 2, p. e0319056, 2025.

[17] N. Chandra, L. Ahuja, S. K. Khatri, and H. Monga, "Utilizing gated recurrent units to retain long term dependencies with recurrent neural network in text classification," J. Inf. Syst. Telecommun, vol. 2, p. 89, 2021.

[18] G. Duan, S. Yan, and M. Zhang, "A Hybrid Neural Network Model for Sentiment Analysis of Financial Texts Using Topic Extraction, Pre-Trained Model, and Enhanced Attention Mechanism Methods," IEEE Access, 2024.

[19] H. Chen, L. Wu, J. Chen, W. Lu, and J. Ding, "A comparative study of automated legal text classification using random forests and deep learning," Information Processing & Management, vol. 59, no. 2, p. 102798, 2022.

[20] A. Mohammed and R. Kora, "An effective ensemble deep learning framework for text classification," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 10, pp. 8825–8837, 2022.

[21] Y. Yue, Y. Peng, and D. Wang, "Deep learning short text sentiment analysis based on improved particle swarm optimization," Electronics, vol. 12, no. 19, p. 4119, 2023.

[22] H. Zhao and W. Xiong, "A multi-scale embedding network for unified named entity recognition in Chinese Electronic Medical Records," Alexandria Engineering Journal, vol. 107, pp. 665–674, 2024.

[23] Z. Wang, L. Wang, C. Huang, S. Sun, and X. Luo, "BERT-based chinese text classification for emergency management with a novel loss function," Applied Intelligence, vol. 53, no. 9, pp. 10417–10428, 2023.

[24] A. R. A. Aljanabi, "The impact of economic policy uncertainty, news framing and information overload on panic buying behavior in the time of COVID-19: a conceptual exploration," International Journal of Emerging Markets, vol. 18, no. 7, pp. 1614–1631, 2023.

[25] C. E. d. Matta, N. M. P. Bianchesi, M. S. d. Oliveira, P. P. Balestrassi, and F. Leal, "A comparative study of forecasting methods using real-life econometric series data," Production, vol. 31, p. e20210043, 2021.

[26] B. Muma and A. Karoki, "Modeling GDP Using Autoregressive Integrated Moving Average (ARIMA) Model: A Systematic Review," Open Access Library Journal, vol. 9, no. 4, pp. 1–8, 2022.

[27] V. I. Kontopoulou, A. D. Panagopoulos, I. Kakkos, and G. K. Matsopoulos, "A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks," Future Internet, vol. 15, no. 8, p. 255, 2023.

[28] A. Shen, M. Dai, J. Hu, Y. Liang, S. Wang, and J. Du, "Leveraging Semi-Supervised Learning to Enhance Data Mining for Image Classification under Limited Labeled Data," arXiv preprint arXiv:2411.18622, 2024.

[29] G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, and S. K. Bhat, "Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications," International Journal of Financial Studies, vol. 11, no. 3, p. 94, 2023.

[30] B. T. Khoa and T. T. Huynh, "Forecasting stock price movement direction by machine learning algorithm," International Journal of Electrical and Computer Engineering (IJECE), vol. 12, no. 6, pp. 6625–6634, 2022.

[31] X. Liu, J. Guo, H. Wang, and F. Zhang, "Prediction of stock market index based on ISSA-BP neural network," Expert Systems with Applications, vol. 204, p. 117604, 2022.

[32] V. Kumar, "Data Analytics and Supply Chain Management: Leveraging Big Data to Optimize Production and Logistics in Fashion and Textiles," in Use of Digital and Advanced Technologies in the Fashion Supply Chain: Springer, 2025, pp. 89–105.

[33] J. G. George, "Advancing Enterprise Architecture for Post-Merger Financial Systems Integration in Capital Markets laying the Foundation for Machine Learning Application," Aus. J. ML Res. & App, vol. 3, no. 2, p. 429, 2023.

[34] M. Sun et al., "Thuctc: an efficient chinese text classifier," GitHub Repository, 2016.

Ensuring Consistency in Group Decision Making: A Systematic Review of the FWZIC Method

Ghazala Bilquise[1], Samar Ibrahim[2]
Computer Information Science, Higher Colleges of Technology, Dubai, UAE[1]
School of Art and Science, American University in Dubai, Dubai, UAE[2]

***Abstract*—Subjective opinions in decision-making processes are often vague, ambiguous, and imprecise due to the inherent subjectivity and variability in individual perspectives. This systematic study examines the Fuzzy Weighted Zero Inconsistency (FWZIC) method, which addresses these challenges by achieving consistency in group consensus and effectively managing uncertainties associated with subjective human opinions. The FWZIC method is increasingly popular in the Multi-Criteria Decision Making (MCDM) field for determining criteria weights. This study comprehensively analyzes 71 empirical studies published from 2021 to March 2025, employing the FWZIC method across diverse domains such as healthcare, engineering, and supply chain. By categorizing FWZIC literature based on themes and domains, this study reveals a taxonomy of the latest techniques and methods integrated with FWZIC. It also explores fuzzy extensions and integrated MCDM methods, providing researchers with a summary of suitable techniques for various contexts. By systematically synthesizing findings, this study provides a comprehensive overview of the current state of FWZIC applications in the literature, identifies gaps and suggests potential avenues for future research in the MCDM domain.**

***Keywords*—FWZIC; MCDM; fuzzy sets; subjective judgment; group decision making**

I. Introduction

Decision-making plays a crucial role in shaping the dynamics of society, influencing various aspects of human life, governance, and development. It is a cornerstone of societal progress, influencing economic development, social welfare, environmental sustainability and technological advancement [1]. In today's dynamic and complex world, decision-makers are confronted with numerous challenges, often involving multiple conflicting criteria for choosing an optimal alternative. Often, these criteria are not of equal importance, and moreover, the performance of the alternatives also varies, thereby complicating the decision-making process. Multi-Criteria Decision-Making (MCDM) methods offer a structured way to navigate these challenges by considering multiple objectives, diverse criteria, and varying preferences of everyone involved [2]. This systematic approach helps to ensure that decisions are made with a comprehensive understanding of all factors at play.

MCDM techniques enable decision-makers to systematically evaluate alternatives based on various criteria, facilitating informed decision-making across diverse domains. From business and engineering [3] to healthcare [4] and aviation [5], MCDM finds applications in a wide range of fields, helping organizations and decision-makers navigate through intricate decision landscapes. Numerous MCDM methods provide decision-makers with quantitative tools to prioritize objectives, assess trade-offs, and identify optimal solutions in real-life scenarios [6]. The decision-making process typically involves identification of the problem, criteria and alternatives, development of a decision matrix and application of an analytical model to rank the alternatives [6].

Criteria weighting is crucial in decision-making, ensuring the accurate representation of the relative importance of evaluation criteria, which significantly impacts ranking outcomes. Weighting methods can be objective, subjective, or a combination of both. Objective methods are based on mathematical models that rely solely on data, thereby excluding human expertise. These methods are ideal when abundant, precise data is available [7]. Conversely, subjective methods rely on human judgment to assign criteria weights and are often employed when expert evaluation is essential for assessing the criteria's significance in the decision-making process [8]. Subjective weighting proves more impactful as it compels decision makers to carefully assess every aspect of the decision problem, rather than relying on the availability of data. [9]. To reduce bias in subjective decisions, a structured approach involving multiple experts is required [10], [11].

Methods such as the Analytic Hierarchy Process (AHP) [12], Analytic Network Process (ANP) [13], Best Worst Method (BWM) [14] and Fuzzy Weighted Zero Inconsistency (FWZIC) [15] are popular subjective criteria weighting methods. Among these, FWZIC is the latest advanced method that overcomes the limitations of other well-known methods. AHP, BWM and ANP methods are time-consuming as they require pairwise comparisons between criteria, which can lead to inconsistencies [16]. As the number of criteria increases, it becomes increasingly difficult for experts to accurately understand how these weights relate to one another. Moreover, these methods are often extended to fuzzy environments [17] to handle uncertainty and imprecision that arises in subjective decisions, making the criteria weighting process even more complex [18]. FWZIC has been applied in multiple domains to assess criteria weights, including healthcare, supply chain, transportation and more [19]. Thus, the FWZIC method demonstrates its versatility and ability to handle various attributes while preserving the consistency of the decision-makers' judgements.

The FWZIC method utilizes a fuzzy environment to compute criteria weights and can easily achieve consensus among multiple experts and a large number of criteria without any inconsistencies in the criteria weighting process [19]. Each expert's opinion is captured with ease and integrated into the decision-making process to determine the final criteria weights.

This enables decision makers to leverage the diverse perspectives, knowledge, and skills of experts, aiming to achieve a more balanced and well-informed outcome. The goal is to enhance decision quality, gain consensus, and ensure that the criteria's weights are representative of multiple stakeholders' opinions.

Despite its advantages over other subjective criteria weighting approaches, the original FWZIC method falls short in addressing uncertainty and vagueness [20]. This limitation arises from the use of Triangular Fuzzy Numbers (TFN) to represent expert opinions. TFNs have been criticized in the literature for inadequately handling uncertainties due to constraints in dealing with membership values [21]. To overcome this limitation, recent studies have extended the FWZIC method using various fuzzy environments. Advanced fuzzy sets such as the Intuitionistic Fuzzy Sets, Pythagorean Fuzzy Sets, and Fermatean Fuzzy Sets have been utilized to broaden the uncertainty space captured by the fuzzy environment.

The decision problems involving FWZIC all deal with the evaluation and benchmarking of the alternatives in various domains ranging from physical entities to algorithms, processes and strategies. The results are used to select an optimal solution by integrating FWZIC with MCDM ranking methods. The FWZIC models built for the assessment problem in one field are applicable to other fields. Therefore, a systematic review is necessary to offer researchers and decision-makers comprehensive guidance on advancements, applications, and future directions in the utilization of the enhanced FWZIC method, which has gained tremendous popularity in the literature. By synthesizing current knowledge and identifying gaps, this study aims to enrich the ongoing discourse in MCDM research, providing decision-makers with valuable insights and directing future inquiries in this evolving field.

To date, only one systematic review on the FWZIC method exists [19]. However, this systematic review diverges from the previous study in several key aspects. The previous review focuses solely on studies that have integrated the FWZIC method with the FDOSM ranking method, limiting the scope to just twenty-three studies. In contrast, this review encompasses fifty-two studies that have utilized the FWZIC method regardless of the ranking method or MCDM approach used, thus offering a broader perspective. Additionally, unlike the previous study, this systematic review analyzes and presents a more comprehensive review of the FWZIC method, detailing the different ranking methods, fuzzy environments, aggregation methods, defuzzifying techniques used, and the criteria and expert characteristics involved, thereby providing an in-depth analysis of the FWZIC method. The novelty of this research lies in its comprehensive approach, which covers the breadth and depth of the empirical studies utilizing the FWZIC method with recommendations for advancements in the field. This study aims to offer a holistic understanding of the progress in the application and utilization of the FWZIC method.

Considering the aforementioned factors, this study seeks to fill the research gap by employing a systematic approach to review and synthesize an analysis based on empirical case studies that have utilized the FWZIC method. The study aims to categorize FWZIC literature based on themes and domains to reveal a taxonomy of the latest techniques and methods utilized with FWZIC. Additionally, it aims to outline future directions and research opportunities in the MCDM field, focusing on subjective criteria weighing with FWZIC. By examining the limitations and challenges identified in the literature, this review highlights areas requiring further research and innovation. To that effect, this study aims to answer the following research questions:

RQ1: What are the general characteristics of empirical studies that have utilized FWZIC?

RQ2: What domains and themes are the studies based on?

RQ3: Which ranking methods has FWZIC been integrated with?

RQ4: Which fuzzy environments are utilized to enhance FWZIC?

RQ5: What techniques are used for aggregation and defuzzification in the FWZIC algorithm?

RQ6: What are the characteristics of the experts involved in the studies?

RQ7: What are the characteristics of the criteria utilized in the studies?

The rest of the study is organized as follows. Section II provides a background of the FWZIC technique and an overview of MCDM and fuzzy sets pertinent to this study. Section III presents the methodology of the systematic review, while Section IV presents the results of the review, linking them to research questions. Finally, Section V delves into the research implications while Section VI concludes the study, presenting the research limitations and future work.

II. Background Information

A. FWZIC Method

The FWZIC method, proposed by [15], is a subjective criteria weighting method used for evaluating criteria weights in the MCDM process. It effectively assigns criteria weights based on multiple human expert judgments, ensuring consistency and minimizing ambiguity. By leveraging a fuzzy environment, FWZIC handles uncertainty and vagueness, making it ideal for complex scenarios with imprecise information or numerous criteria. Often combined with ranking methods like Technique for Order of Preference by Similarity (TOPSIS), Fuzzy Decision by Opinion Score Method (FDOSM), FWZIC enhances decision accuracy and robustness across various fields, including tourism, healthcare, and education [19].

The FWZIC method relies on five essential phases to achieve zero inconsistency in criteria weighting by involving multiple experts in assessing criteria significance. The FWZIC method is applied using the phases below [15]:

1) Design the evaluation form. An appraisal form is designed to assess the significance of the evaluation criteria of the MCDM case study. The appraisal form is based on a Likert scale, usually a 5-point or 10-point scale. The Likert scale is ideal for this purpose as it allows the experts to assess the

criteria based on linguistic values (such as "Highly Important", "Important", "Moderately Important") and converts the expert assessment to its corresponding numerical values (5, 4, 3) which can be evaluated further using the FWZIC technique.

2) Collect expert judgement. In this step, experts in the field of study are identified. A minimum of three experts is often recommended to ensure diverse perspectives of experts are captured adequately. Section IV presents the experts' characteristics involved in extant FWZIC studies. The appraisal form is administered to the relevant experts to collect their judgment in a structured approach.

3) Develop Expert Decision Matrix (EDM): In this step, an expert decision matrix is constructed based on the data collected via the appraisal form. An EDM is constructed by intersecting each criterion with the expert importance (numerical score) assigned to that criterion, as shown in the equation below:

$$EDM = [e_{ij}]_{m\times n} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ \cdots & \cdots & \ddots & \cdots \\ e_{m1} & e_{m2} & \ldots & e_{mn} \end{bmatrix} \quad (1)$$

where, m represents the number of experts, and n represents the number of criteria. Moreover, $e_{ij} = Imp(e_i/c_j)$, which represents the importance of criterion *j* assigned by expert *i*. Thus, e_{13} is the significance level of criterion 3 assigned by expert 1.

4) Develop fuzzy EDM by applying fuzzy membership functions. In this step, the EDM performance values, representing the criteria significance by the expert, are fuzzified, using the integrated fuzzy membership functions, to develop a fuzzy EDM. The original FWZIC method is based on the Triangular Fuzzy Number (TFN). However, recent studies have extended FWZIC with various fuzzy environments.

$$\widetilde{EDM} = [\widetilde{e_{ij}}]_{m\times n} = \begin{bmatrix} \widetilde{e_{11}} & \widetilde{e_{12}} & \cdots & \widetilde{e_{1n}} \\ \widetilde{e_{21}} & \widetilde{e_{22}} & \cdots & \widetilde{e_{2n}} \\ \cdots & \cdots & \ddots & \cdots \\ \widetilde{e_{m1}} & \widetilde{e_{m2}} & \ldots & \widetilde{e_{mn}} \end{bmatrix}, \quad (2)$$

where, $\widetilde{e_{ij}}$ represents the corresponding fuzzy membership value for e_{ij}.

5) Compute final criteria weights. Finally, the criteria weights are computed based on the respective operators of the fuzzy environment integrated with FWZIC. The computation involves five main steps:

a) Aggregation of the criteria weights using an aggregation operator, b) Determining the ratio for each criterion for each expert by dividing the fuzzy criterion value by the aggregated value, c) The fuzzy value is defuzzified using the respective fuzzy environment defuzzification techniques, and d) The criteria weight coefficients are computed by converting the defuzzified values to their corresponding weight coefficients to ensure that the sum of the weights is equal to 1. Section IV discusses prevalent aggregation and defuzzifying techniques utilized in extant literature for computing the criteria weights.

Fig. 1 illustrates the five phases of the FWZIC method.

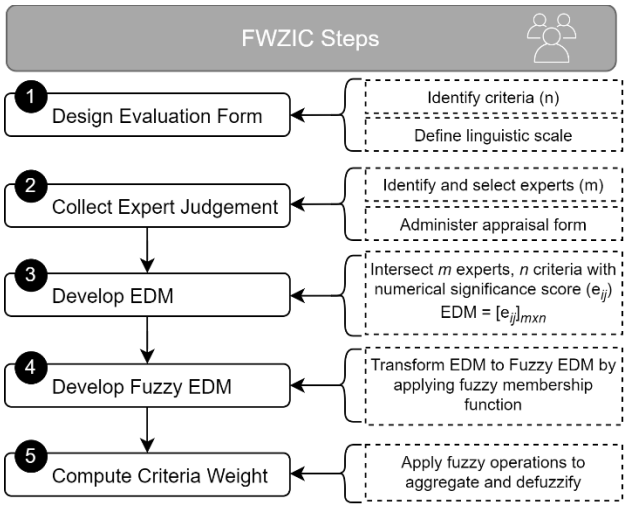


Fig. 1. FWZIC phases.

The original FWZIC method relied on TFN to handle ambiguity and imprecision in human judgment. However, since then, several researchers have extended FWZIC using advanced fuzzy environments that capture a wider spectrum of ambiguities. The following section explores the fuzzy set theory and the various fuzzy environments with their generalizations and extensions.

B. Comparison of FWZIC with other Criteria Weighting Methods

The FWZIC method offers a structured and robust approach to subjective criteria weighting by combining expert judgment with fuzzy logic while eliminating inconsistency in pairwise comparisons [22]. Unlike traditional methods such as AHP, ANP, and BWM, FWZIC does not rely on complex pairwise comparison matrices that can become cumbersome and inconsistent when the number of criteria or experts increases. This is particularly advantageous in large-scale evaluations, where managing comparison consistency is a challenge. FWZIC instead allows experts to assess criteria using a linguistic scale mapped to fuzzy numbers, which are then aggregated using fuzzy mathematical operations to derive weights without introducing inconsistency errors. Table I provides a comparison of FWZIC with other subjective criteria weighting methods.

AHP and ANP, while well-established and widely used, often suffer from issues related to consistency and scalability [23]. In AHP, experts are required to perform numerous pairwise comparisons, and inconsistency indices are used to evaluate the reliability of these judgments. However, achieving acceptable consistency becomes increasingly difficult as the number of criteria grows. ANP attempts to address interdependencies among criteria, but this added complexity also makes the process more time-consuming and prone to subjective bias. Moreover, both AHP and ANP require re-adjustments when new criteria are introduced, reducing their flexibility in dynamic decision environments.

BWM simplifies the pairwise comparison process by requiring experts to identify the best and worst criteria and rate all others relative to these two [24]. While this reduces the

number of required comparisons and improves consistency compared to AHP and ANP, it still assumes a linear and rigid judgment structure that might not capture the uncertainty inherent in human evaluations. FWZIC, by contrast, accommodates uncertainty more naturally through fuzzy logic and avoids the need for direct comparisons altogether. It also provides flexibility in handling a varying number of experts and criteria without increasing the computational burden or introducing inconsistency.

In summary, FWZIC stands out for its ability to integrate expert judgment in a fuzzy environment while maintaining zero inconsistency. It is computationally less intensive than ANP and more scalable than AHP. Unlike BWM, it does not constrain expert input to a fixed comparison format, making it more adaptable in diverse decision-making scenarios. These characteristics make FWZIC particularly suitable for applications involving multiple experts, complex criteria structures, and a need for robust, bias-reduced aggregation.

C. Fuzzy Sets

Human judgment is riddled with uncertainties and ambiguities. Ambiguity in decision-making refers to situations where information is unclear, incomplete, or open to multiple interpretations. Fuzzy set theory, proposed by Zadeh [17], aims at handling ambiguity by representing crisp values using membership grades ranging from [0,1]. Fuzzy sets provide a mathematical framework for modeling and reasoning with vague concepts, making them valuable in various fields such as artificial intelligence, engineering, supply chain, and healthcare, where imprecise information is common.

Since its inception, fuzzy set theory has advanced significantly, leading to new fuzzy sets that incorporate membership as well as non-membership, and hesitancy degrees [25]. The concept of Type-2 fuzzy sets further expands the traditional fuzzy set concept with a three-dimensional membership function [26]. It uses a fuzzy membership function, where the membership degree is also a fuzzy set, allowing for a more detailed representation of uncertainty. However, it is more complex due to the fuzzy nature of the membership function, requiring more computational resources and sophisticated algorithms for processing. The different types of fuzzy sets and their variations and generalizations are explained next.

1) Triangular Fuzzy Numbers (TFN). TFNs are a specific type of fuzzy number that is represented by a triangular membership function. This function is defined by three parameters (a, b, c), where a represents the lower limit, b represents the peak or mode, and c represents the upper limit. The membership function $\mu_{\tilde{a}}(x)$ of a TFN $\tilde{a} = (a, b, c)$ is defined as follows [27]:

$$\mu_{\tilde{a}}(x) = \begin{cases} \frac{x-a}{b-a}, & if\ a \le x \le b \\ \frac{c-x}{c-b}, & if\ b < x \le c \\ 0, & if\ x > c\ or\ x < a \end{cases}, \quad (3)$$

where, $\le b \le c$. Moreover, if $a \ge 0$, then the TFN is called a positive triangular fuzzy number. If $c \le 0$, then the TFN is called a negative triangular fuzzy number.

TABLE I. FWZIC COMPARISON WITH OTHER CRITERIA WEIGHTING METHODS

	ANP	**AHP**	**BWM**	**FWZIC method**
Computational Complexity (Number of comparisons, where n is the number of criteria)	$n(n-1)$	$n(n-1)/2$	$2n-3$	Zero pairwise comparisons
Nature of comparison	Comparisons are made between criteria that differ in nature and in quantity.	Comparisons are made between criteria that differ in nature and quantity.	Comparisons are made between criteria that differ in nature and quantity.	Relies on direct linguistic assessments of criteria mapped to fuzzy numbers, thus avoiding pairwise comparisons.
Expert feedback	The process of explaining and comparing in this method is time-consuming.	The process of explaining and comparing in this method is time-consuming.	The process of explaining takes time, and the comparison process is difficult.	Expert feedback is collected through online surveys, thereby reducing the explanation and collection time.
Uncertainty	Fuzzy logic is commonly employed to manage ambiguity and uncertainty in expert judgments; however, it increases computational complexity.	Fuzzy logic is commonly employed to manage ambiguity and uncertainty in expert judgments; however, it increases computational complexity.	Fuzzy logic is commonly employed to manage ambiguity and uncertainty in expert judgments; however, it increases computational complexity.	FWZIC is grounded in fuzzy logic but is specifically designed to maintain computational simplicity.
Inconsistencies	Susceptible to inconsistency due to complex interdependencies and numerous pairwise comparisons, even when fuzzy logic is applied.	Inconsistency arises from the subjective nature of pairwise comparisons, with reliability decreasing as the number of criteria increases.	While more consistent than AHP/ANP, inconsistencies may still occur when judgments about best and worst criteria lack coherence.	Designed to eliminate inconsistency by avoiding pairwise comparisons entirely and using direct fuzzy linguistic ratings.

2) Trapezoidal Fuzzy Numbers (TraFN). A TraFN is characterized by a membership function that forms a trapezoid shape. It is defined by four parameters: a, b, c, d. The membership function $\mu_{\tilde{a}}(x)$ of a TraFN $\tilde{a} = (a, b, c, d)$ is defined as follows [28]:

$$\mu_{\tilde{a}}(x) = \begin{cases} \frac{x-a}{b-a}, & if\ a \le x \le b \\ 1, & if\ b \le x \le c \\ \frac{d-x}{d-c}, & if\ c \le x \le d \\ 0, & if\ x > d\ or\ x < a \end{cases} \quad (4)$$

where, $a \le b \le c \le d$. Notably, when $b = c$, TraFN reduces to a TFN.

3) Hexagonal Fuzzy Numbers (HFN). An HFN $\tilde{a}$ is characterized by a membership function that is based on six parameters and forms a hexagonal shape. The membership function $\mu_{\tilde{a}}(x)$ of an HFN $\tilde{a} = (a, b, c, d, e, f)$ is defined as follows [29]:

$$\mu_{\tilde{a}}(x) = \begin{cases} \frac{1}{2}\left(\frac{x-a}{b-a}\right), & if\ a \le x \le b \\ \frac{1}{2} + \frac{1}{2}\left(\frac{x-b}{c-b}\right), & if\ b \le x \le c \\ 1, & f\ c \le x \le d \\ 1 - \frac{1}{2}\left(\frac{x-d}{e-d}\right), & if\ d \le x \le e \\ \frac{1}{2}\left(\frac{f-x}{f-e}\right), & if\ e \le x \le f \\ 0, & if\ x > f\ or\ x < a \end{cases} \quad (5)$$

where, $a \le b \le c \le d \le e \le f$.

4) Intuitionistic Fuzzy sets (IFS). The IFS proposed by [30] represents an extension of classical fuzzy sets. Unlike the classical TFN, TraFN and HFN that rely solely on a membership grade (μ) to indicate the degree of satisfaction of an element, IFS introduces a non-membership grade (υ) to signify the degree of dissatisfaction. Both μ and υ values range between 0 and 1, with the constraint that their sum remains within this range: $0 \le \mu + \upsilon \le 1$. Therefore, an IFS S over a universe of discourse X is represented as:

$$S(x) = \{x, \langle \mu(x), \upsilon(x) \rangle\}, x \in X, \quad (6)$$

By incorporating both membership and non-membership grades, the IFS framework provides a more flexible depiction of solutions, allowing for greater flexibility in representing uncertainty or imprecision inherent in the given problem. However, in several cases, the sum of the degrees often exceeds 1. Therefore, the generalizations of the IFS were proposed.

5) Pythagorean Fuzzy Sets (PyFS): The PyFS, proposed by [31], is a generalization of the IFS, which extends the range of membership (μ) and non-membership grades (υ) with the condition that the sum of the squares of both degrees falls within the range of [0,1], $0 \le \mu^2 + \upsilon^2 \le 1$. The PyFS offers a more comprehensive and precise characterization of uncertain information compared to the IFS [32].

6) Fermatean Fuzzy Sets (FFS). The FFS [33] further expands the uncertainty space that cannot be adequately captured by the IFS and PyFS in cases, where the sum of the squares of the grades is above 1. Therefore, in FFS, the sum of the cubes of the membership and non-membership grades falls within the range of [0,1], $0 \le \mu^3 + \upsilon^3 \le 1$. FFS offers a better way of handling vagueness than its counterparts, IFS and PyFS.

7) q-rung orthopair (q-ROFS). The q-ROFS [34] emerged as a generalization of both IFS and PyFS with a membership constraint that the sum of the qth power of the membership and non-membership grades is within the range [0,1], $0 \le \mu^q + \upsilon^q \le 1$, where $q > 1$. The increasing value of q offers decision-makers a broader spectrum to express their perspectives on the membership and non-membership degrees. Thus, the q-ROFS demonstrates a greater capability in handling uncertainty more effectively than IFS and PyFS [35]. Fig. 2 illustrates the comparison between IFS, PyFS and q-ROFS.

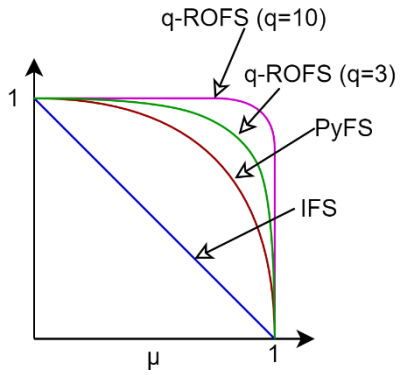


Fig. 2. IFS, PyFS and q-ROFS.

8) Picture Fuzzy Sets (PFS). The PFS, proposed by [36], is an extension of the IFS, which introduces the degree of neutrality (η), also referred to as abstinence, along with the membership (μ) and non-membership (υ) degrees, with the condition that $0 \le \mu + \eta + \upsilon \le 1$. The refusal degree is represented as $1 - (\mu + \eta + \upsilon)$. The PFS effectively mirrors human behavior and addresses decision-making challenges in a similar way by capturing the complexities of real-life information [37].

$$S(x) = \{x, \langle \mu(x), \eta(x), \upsilon(x) \rangle\}, x \in X \quad (7)$$

9) Spherical Fuzzy Sets (SFS). The SFS [38] emerged as a generalization of the PFS, which extends the uncertainty space of the membership (μ), non-membership (υ) and neutral grade (η), offering a wider decision space under the restriction that $0 \le \mu^2 + \eta^2 + \upsilon^2 \le 1$.

10) T-Spherical Fuzzy Sets (T-SFS). The T-SFS proposed by [39], further widens the uncertainty space covered by the SFS with a membership constraint that the sum of the t power of the membership (μ), non-membership (υ) and neutral (η) grades are within the range [0,1], $0 \le \mu^t + \eta^t + \upsilon^t \le 1$, where $t > 1$. The increasing value of t offers decision-makers

a broader spectrum to express their perspectives on the membership and non-membership degrees and abstinence degrees. Notably, when t = 2 the T-SFS reduces to SFS.

11) Linear Diophantine Fuzzy Sets (LDFS). Despite the broader space covered by these FS, the grade dependency still exists as a limitation of the IFS and PFS and all their generalizations. Hence, [40] proposed a new fuzzy environment, the Linear Diophantine Fuzzy set (LDFS), which extends IFS and PyFS by introducing two reference parameters (a,b), each within the range of [0,1] corresponding with the membership (μ), non-membership (υ) grade, respectively. The LDFS essentially allows decision makers to choose membership degrees without restrictions, but with a constraint that the sum of the reference parameters lies between 0 and 1, $0 \leq a + b \leq 1$ and $0 \leq a\mu + b\upsilon \leq 1$.

$$S(x) = \{x, \langle \mu(x), \upsilon(x) \rangle, \langle a, b \rangle\}, x \in X \quad (8)$$

12) Interval Valued Fuzzy Set (IVFS). Grattan-Guinness (1976) [41] developed the concept of IVFS based on type 2 fuzzy set theory [42], further extending the fuzzy research space. A type 2 fuzzy set is characterized by fuzzy membership grades. In an IVFS, the membership degree is given as a closed subinterval $[\mu^L, \mu^U]$ with $0 \leq \mu^L \leq \mu^U \leq 1$, where:

- μ^L is the lower bound of the membership interval.
- μ^U is the upper bound of the membership interval.

The interval $[\mu^L, \mu^U]$ represents the range within which the membership degree of x is expected to lie. The length of this interval $(\mu^L - \mu^U)$ is interpreted as a measure of the uncertainty associated with the membership degree of the element.

13) Neutrosophic Sets (NS). The Single-Valued NS, proposed by [43], is a generalization of the classical fuzzy set in which there is no restriction on the sum of the membership degrees. Hence, the membership degree, abstinence degree, and non-membership degree lie within the range [0,3], $0 \leq \mu + \eta + \upsilon \leq 3$. The NS has a significant capacity to effectively manage imprecise and unclear data in a comprehensive manner.

14) Hesitant Fuzzy Sets (HFS). HFS, also known as multisets, was proposed by [44], extending the traditional fuzzy set theory to better handle situations where there is hesitation or multiple degrees of membership for an element. This concept is useful in capturing uncertainty and ambiguity when a single membership value is insufficient to represent an element's degree of belonging to a set. Fuzzy multisets allows for capturing hesitation in the membership degrees, providing a richer and more flexible representation of uncertainty. A generalization of the HFS, known as the dual HFS, allows for the representation of dual hesitancy, meaning elements can be partially or fully hesitant towards multiple membership degrees [45].

Other generalizations and extensions of the classical fuzzy sets include the probabilistic fuzzy sets, which assign a probability to each element's membership degree [46]. The Complex Fuzzy Sets (CFS), on the other hand, extend the concept of fuzzy sets to complex-valued membership degrees by representation of membership degrees as a complex number [47]. CFS has been applied in signal processing and control systems applications.

Despite the notable efforts of the researchers, the lack of precision, uncertainty, and vagueness in information still remains an unresolved issue [48]. To encompass a broad spectrum of uncertainty in experts' opinions for determining criteria weight, researchers have extended the FWZIC method under various fuzzy environments ranging from simple classical sets to advanced type 2 fuzzy sets. The results section of the review presents the various fuzzy environments integrated with FWZIC.

D. Overview of MCDM Ranking Methods

MCDM ranking techniques enable decision-makers to systematically assess alternatives using various criteria, facilitating informed decision-making across different sectors. Several benchmarking techniques are used in MCDM to perform ranking in various decision-making case study problems, such as FDOSM, VIKOR, SAW, MULTIMOORA, TOPSIS, MAIRCA, MABAC, and DEMATEL. Below are brief introductions to some of the popular ranking methods integrated with FWZIC:

1) Fuzzy Decision by Opinion Score Method (FDOSM). The Fuzzy Decision by Opinion Score Method (FDOSM) is an innovative MCDM technique designed for use in a fuzzy environment, utilizing the concepts of ideal solutions and opinion matrices. This method addresses and largely overcomes the inconsistencies often found in human judgment, reducing the time required for comparisons [49]. Additionally, FDOSM minimizes the number of mathematical equations needed, preserving data integrity and facilitating logical decision-making based on expert opinions represented in the Decision Matrix. It eliminates issues related to normalization and weighting in mathematical approaches and resolves data vagueness using fuzzy numbers. FDOSM employs positive and negative ideal solutions, comparing each alternative's criteria with the ideal solution. Furthermore, FDOSM uses different aggregation operators with an Opinion Matrix adopted from various MCDM techniques [50].

2) VlseKriterijumska optimizacija I kompromisno resenje (VIKOR). VIKOR, a Serbian term for "multi-criteria optimization and compromise solution" [51] is a method based on the concepts of positive and negative ideal points, and on the compromise programming of MCDM [52]. VIKOR focuses on ranking and selecting from a broad range of alternatives that are evaluated based on criteria. VIKOR determines compromise solutions for problems with conflicting and non-comparable criteria and provides a feasible solution that is the closest to the positive ideal solution. A key feature of VIKOR is that, the final decision relies on rules regarding an acceptable advantage of the optimal solution over the second-best solution and the stability of the optimal solution to changes in the decision-maker's risk choice. It reflects any hesitation the decision-

maker has in selecting the final option, allowing for a nuanced comparison of alternatives [53].

3) Simple Additive Weighting (SAW). The SAW method is well-known for its simplicity, user-friendly, and effective approach of MCDM. Also referred to as a weighted linear combination or scoring method. SAW determines the overall score of a candidate solution based on the weighted sum of all of the attribute values [54]. The SAW method involves three major steps: normalizing the decision matrix X, assigning the weight vector W, and calculating the overall score. One major advantage of the SAW method is its ability to maintain the relative magnitude order of the original data, and no data is lost during evaluation [54].

4) Multi-Objective Optimization by Ratio Analysis (MULTIMOORA). The MULTIMOORA method is the extension version of the MOORA method, incorporating multiplicative elements. MOORA, a multi-objective optimization technique, is based on ratio analysis. In this approach, the score of an alternative is calculated by subtracting the sum of performance scores for the benefit criteria and adding the sum of performance scores for the cost criteria [4].

5) Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). TOPSIS is a renowned classical MCDM method that introduces two reference points: a positive ideal solution and a negative ideal solution. TOPSIS was originally developed by Hwang and Yoon in 1981, and was further developed by Chen and Hwang in 1992 [54]. The positive ideal solution maximizes benefit criteria and minimizes cost criteria, while the negative ideal solution does the opposite, maximizing cost criteria and minimizing benefit criteria. TOPSIS identifies the best alternative by minimizing its distance to the positive ideal solution and maximizing its distance to the negative ideal solution. It employs Euclidean distances to measure how close each alternative is to these ideal solutions, and the preference order of alternatives is determined by comparing these distances [53].

6) Multi-Attributive Ideal-Real Comparative Analysis (MAIRCA). MAIRCA is recognized for its stability compared to other popular MCDM ranking methods. It employs a straightforward mathematical approach that maintains stability despite changes in the nature and characteristics of the criteria. One key advantage of MAIRCA is its ability to objectively compute the probability of each alternative, providing both quantitative and qualitative assessments. This method measures the difference (gap) between ideal and empirical evaluations, providing an objective assessment of each alternative [55].

7) Multi-Attributive Border approximation Area Comparison (MABAC). MABAC method ranks the alternatives using their distance to the border approximation, while partitioning them into upper approximation area (best performing) and lower approximation area (worst performing) [4].

8) The decision-making trial and evaluation laboratory (DEMATEL). DEMATEL methodology is a robust tool for analyzing complex interrelationships between various factors. Pairwise comparisons are used to determine the significance and influence of different elements. By constructing cause-effect diagrams, DEMATEL assesses the interdependencies among factors, proving effective in diverse domains such as site selection, criteria identification, and effects evaluation. However, due to the inherent uncertainty and subjectivity of experts' judgments, an important challenge remains in effectively representing uncertain and linguistic judgments [56].

Numerous other ranking methods are also used in MCDM problems, such as the Additive Ratio Assessment (ARAS), Conditional Probabilities by Opinion Scores (CPOS), Combinative Distance-Based Assessment (CODAS), Complex Proportional Assessment (COPRAS), Evaluation Based on Distance from Average Solution (EDAS), and Measurement of Alternatives and Ranking according to the Compromise Solution (MARCOS), Technique for Reorganizing Opinion Order to Interval Levels (TROOIL). FWZIC has been combined with these ranking methods to benchmark various alternatives in different case studies, employing various approaches to achieve a robust framework. A review of these FWZIC studies will provide us with a clear understanding and knowledge of their various characteristics and their combined ranking methods. The next section describes the methodology used to achieve this systematic review.

III. Methodology

This study examines the existing literature on the applications of the FWZIC method using the systematic review framework established by [57]. This framework was selected over other frameworks because it provides guidelines specifically for conducting reviews in the technical field. Employing a rigorous theoretical framework is crucial for guiding the comprehensive data collection and analysis methods required to ensure the reliability of the results. The systematic literature review guidelined by [57] offers a thorough approach for collecting, analyzing, and documenting findings from secondary data sources and from various systematic reviews [58], [59], [60]. By following this methodology, we aim to answer the research questions and uncover the latest application techniques and extensions of the FWZIC method, guiding practitioners and researchers to further study in this field.

The review process is divided into three phases: planning the review, conducting the review, and reporting the results. Each phase is further subdivided into several steps, each of which is described in the following subsections.

A. Planning the Review

In the last three years, the studies applying the FWZIC method to compute criteria weights have accelerated due to its superiority over other subjective criteria weighting methods. Therefore, it is crucial to provide researchers with the current state-of-the-art regarding the applications of the FWZIC method, including the fuzzy environments used, the number of experts involved, the number of criteria and the MCDM ranking methods it has been integrated with.

To initiate the systematic review, the planning phase defines the inclusion and exclusion criteria, and identifies the data sources for selecting the study articles. The quality assessment checklist is then described to evaluate the quality of the articles, establishing a threshold for their inclusion. The study encompasses all articles published since the FWZIC method was proposed, ensuring a comprehensive scope.

1) Inclusion and exclusion criteria. To select the articles for this review, it is essential to define the criteria that characterize the included studies. Table II summarizes the inclusion and exclusion criteria used in the selection process.

TABLE II. INCLUSION AND EXCLUSION CRITERIA

Inclusion Criteria	Exclusion Criteria
Empirical studies that have employed the FWZIC method for determining criteria weights	Studies that do not utilize the FWZIC method
Article is published in a peer-reviewed journal, conference papers or book chapters	Reviews, errata, PhD studies
Written in English	Non-English articles
Published in 2021 – March 2025	Published after March 2025

First, only empirical studies employing the FWZIC method for determining criteria weights were considered. Second, the review focused exclusively on peer-reviewed journals, conference papers and book chapters to ensure inclusion of all studies that have utilized the FWZIC method, thereby excluding reviews, errata, and PhD dissertations. Third, only articles published in English were included to avoid biases from poor translations. Finally, the study period was set from 2021 to 2025 to encompass all relevant studies using the FWZIC method to date.

2) Data sources. Since MCDM methods are utilized in diverse domains, a wide range of data sources was considered to ensure the inclusion of relevant publications from various fields. Accordingly, the search was conducted using three digital databases: Scopus, Web of Science (WoS), and ScienceDirect. Both Scopus and Web of Science are multidisciplinary databases of peer-reviewed literature on science, technology, medicine, social science, arts and humanities, making them valuable for retrieving a broad spectrum of research. ScienceDirect is a comprehensive resource for articles in the scientific, technical and medical fields.

3) Quality assessment checklist. Quality assessment is essential in systematic reviews to ensure the validity of the results and minimize bias from including less robust studies [61]. It also refines the inclusion and exclusion criteria [57].

To ensure a rigorous assessment of the articles included in the systematic review, a quality assessment checklist is adapted from [62]. The checklist consists of ten questions, as shown in Table III. This checklist focuses on the essential elements for data extraction and coding phases, such as the relevance of the study, clear identification of the problem statement, experts and the expert selection process. Additionally, the credibility of the sources were evaluated by considering the ranking of the journal or conference and the number of citations of each study received. A point scale from 0 to 1 will be used to rate each assessment criterion, where 1 represents that the requirement is wholly met, 0.5 represents it is partially met, and 0 indicates that it is not met. Moreover, Q09 will be evaluated by the source ranking of the published study, whereby Q01 or Q02 ranked journal is assigned 1 point, while Q03 is assigned 0.5 and 0 point is assigned to a journal ranked below or not ranked. Finally, 1 point will be assigned to the article having at least two citations per year for Q10. Thus, each article scored out of 10 points in total. The cut-off criteria for including an article after the quality assessment are set at 75%.

TABLE III. QUALITY ASSESSMENT CHECKLIST

Q01	Is the study relevant to our research?
Q02	Are the research aims and contributions clearly identified?
Q03	Is the problem statement clear?
Q04	Does the study define the number of experts used?
Q05	Is the selection process of the experts clearly explained?
Q06	Are the techniques/methods employed in the study clearly documented?
Q07	Is the proposed technique evaluated using established metrics?
Q08	Is the conclusion explained clearly and linked to the purpose of the study?
Q09	Is the source of the article credible (published in a ranked venue)?
Q10	Has the study been cited in other publications?

B. Conducting the Review

The implementation phase of the review involves searching for and retrieving the articles. The articles for this review were collected in March 2025. The articles were screened using the previously defined inclusion and exclusion criteria, as well as the quality assessment checklist.

1) Search process. The systematic review focuses on a specific MCDM criteria weighting method. Hence, only two main keywords were deemed necessary to retrieve the articles for this study, combined with the logical operator OR ("FWZIC" OR "Fuzzy Weighted Zero Inconsistency"). Additionally, the search was conducted on titles, abstracts, and keywords to ensure comprehensive coverage and to increase the likelihood of finding highly relevant studies.

The search results, presented in Table IV, yielded a total of 192 studies, with the highest number retrieved from the Scopus database, due to its broad coverage across multiple domains and the lowest from ScienceDirect, due to its specific coverage.

TABLE IV. SEARCH RESULTS

Database	Search Results
Scopus	81
Web of Science (WoS)	62
ScienceDirect	49
Total	**192**

Following the retrieval of search results, a bibliometric analysis was conducted to examine the research areas. Fig. 3

illustrates the visualization of terms within the results, generated using VOS Viewer [63]. This diagram highlights the significance and interconnections among frequently occurring terms extracted from the abstracts, titles, and keywords of the search results. The size and label of each term indicate its importance, while the color represents clusters in the visualization. Each cluster comprises terms related to one another within the group, and the distance between clusters signifies their relatedness.

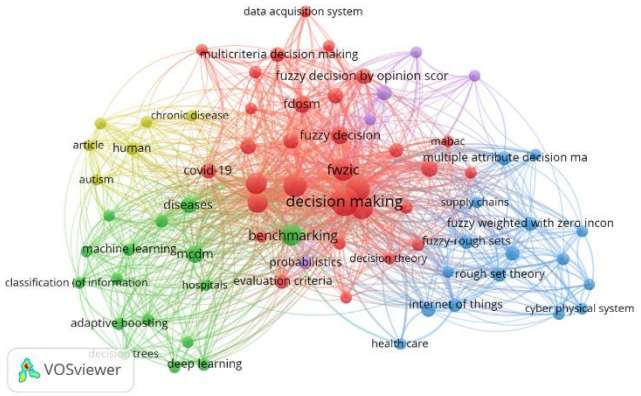


Fig. 3. Visualization of search results.

The visualization of terms within the extracted studies uncovers several distinct clusters, indicating multiple dimensions in the application of FWZIC in MCDM studies. These clusters are closely intertwined, suggesting interconnected aspects within the studies. The term "decision making" is prominently linked to all clusters, indicating that the retrieved results are applied studies in the decision-making field.

Focusing on clusters with highly weighted terms, four main clusters emerge in the visualization. The first and central red cluster encompasses keywords such as FWZIC, FDOSM, fuzzy decision, and linguistics, revealing the technical aspects of research utilizing FWZIC, including its integration with ranking algorithms, application of linguistic scales, and use of fuzzy decision sets. The second, green cluster features keywords like adaptive boosting, decision trees, deep learning, and machine learning, indicating the application of FWZIC in the domain of Artificial Intelligence. In the third, blue cluster, the significant keywords include cyber-physical systems, healthcare, internet of things, and supply chain, highlighting various areas of study in which the FWZIC method is utilized. Lastly, the light green cluster includes terms such as human, autism, and chronic diseases, indicating the predominant utilization of FWZIC for solutions in the healthcare industry.

Examining these clusters provides insight into the distribution of research findings, facilitating a more informed analysis and discussion of the studies.

2) Article selection. In this phase, the inclusion and exclusion criteria were applied to screen the retrieved articles for eligibility following the PRISMA framework [64]. This framework offers a comprehensive guideline and structured approach for document screening. The steps of the screening process are illustrated in Fig. 4.

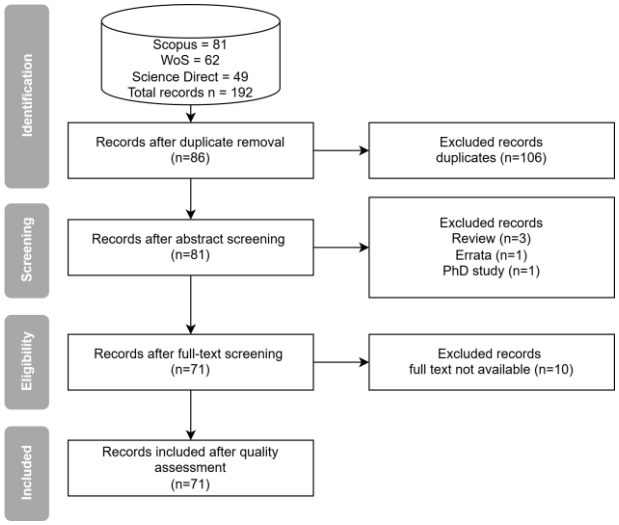


Fig. 4. Prisma flowchart.

The PRISMA framework consists of four phases. In the identification phase, 192 articles were retrieved from the three data sources using the identified search term. Following this, the duplicate records were removed, resulting in 86 articles. During the screening phase, the inclusion and exclusion criteria were applied through abstract screening to ensure that only relevant articles were included. Each author independently performed a title and abstract screening of the studies to remove irrelevant articles, resulting in the exclusion of four articles, two of which were reviews, one erratum, and one PhD dissertation, resulting in 81 studies. In the eligibility phase, a full-text screening of the articles was performed. Each author performed this step independently by equally dividing the studies to be reviewed. In cases where eligibility was unclear, the authors discussed resolving the discrepancy. This phase resulted in 10 studies being excluded due to the unavailability of the full text, resulting in a total of 71 eligible articles. The excluded studies were within the context of the systematic review; however, unavailability would prevent the coding and analysis required for the comprehensive review.

In the final phase, a quality assessment was performed on the remaining articles (n = 71). In this step, the authors rescreened the articles initially screened by the other to reduce bias and ensure that each article is screened twice. Each author independently conducted a quality assessment using the assessment checklist outlined in Fig. 4. Remarkably, all included articles met the required threshold of 75% and demonstrated sound quality. It is important to emphasize that the quality assessment aims to ascertain the relevance of selected articles to the contribution of this study, without any intention to criticize the studies or their findings.

3) Data analysis and coding. The objective of this phase is to comprehensively document the systematic review findings by gathering meta-data from the primary studies included in the review, with a focus on the research questions. A thorough data analysis of the relevant features identified during the planning phase is conducted to achieve this objective. The meta-data analysis encompasses various characteristics crucial for addressing the research questions, including publication type

and year, and the classification of studies based on themes and application areas to gain a deeper insight into the domains of application. Additionally, technical aspects of subjective decision making, such as the number of experts, the criteria weighted, the fuzzy environment, the aggregation operators, and defuzzifying methods employed in the studies, are also examined.

C. Reporting the Review

In the concluding phase of the systematic review, the study results are unveiled. Section IV presents a comprehensive analysis of the meta-data extracted from the full-text review, with the aim of addressing each research question. This analysis presents and discusses the insights from the collected data, providing a thorough understanding of the findings and their implications within the context of the study's objectives.

IV. Results

This section reveals the outcomes from the meta-analysis and comprehensive review of the selected articles, guided by our research questions. This study examines seventy-one papers (listed in the Appendix) published over a span of five years to explore the applications of the FWZIC method. The following subsections detail the findings for each of the research questions.

A. Study Characteristics

RQ1: What are the general characteristics of empirical studies that have utilized FWZIC?

This subsection provides an overview of general characteristics found in the reviewed studies. It includes details on the distribution of studies by year of publication, author, journals and publishers. These attributes offer a summary of the significance of the FWZIC method within the MCDM research field.

Fig. 5 presents the distribution of the seventy-one studies reviewed by year of publication. It is evident from the graph that there is significant interest in the FWZIC method since its proposal in 2021. Notably, the graph reveals a sharp rise in the utilization of the FWZIC method since 2022. FWZIC has been extended to many fuzzy environments to improve consistency and decrease the hesitancy of human judgment. In addition, this method has been integrated with numerous MCDM ranking methods to evaluate and select alternatives used in these studies.

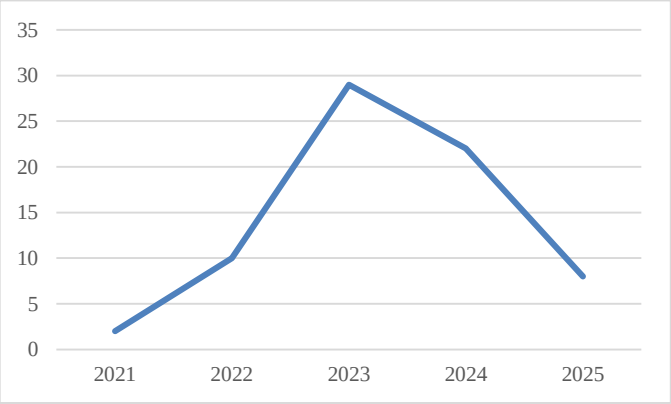


Fig. 5. Distribution of studies by publication year.

Fig. 6 displays the distribution of the seventy-one reviewed studies across forty-two different sources, including journals, conferences, and book chapters, underscoring the extensive influence and significance of the FWZIC method within the MCDM research field. Notably, most of the studies are published in ranked journals, while only one study is published in a conference and one in a book chapter. The included studies span over a diverse range of journals, including those focused on technology, decision science, healthcare, innovation, knowledge, and engineering. The journal "Expert Systems with Applications" features the highest number of studies (n = 10), followed by "Applied Soft Computing" (n = 6) and "Neural Computing and Applications" (n = 5). Most of these journals are ranked in the Q1 category, reflecting the high quality of the published research.

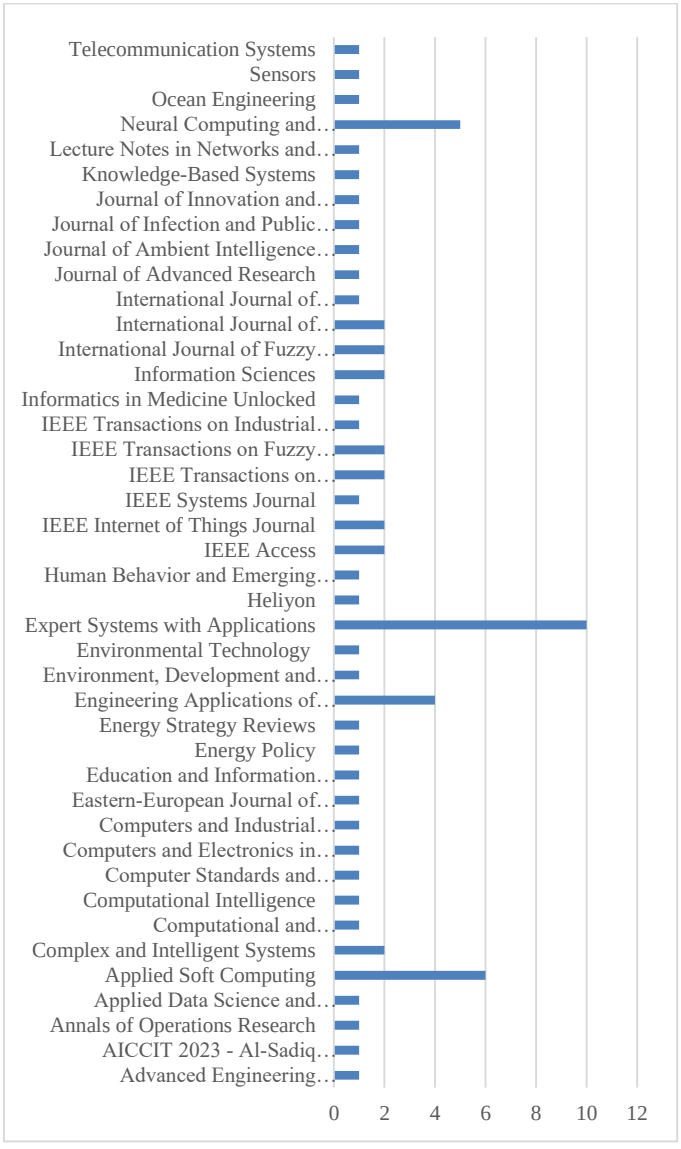


Fig. 6. Distribution of studies by source.

Fig. 7 shows the distribution of seventy-one reviewed studies among nine major publishers, indicating that Elsevier is the leading publisher with 48% of these studies (n = 34), followed by Springer (n = 15) with 21%, and IEEE (n = 11) with

16%, all of which are significant contributors to the research field. John Wiley and Sons Ltd. published 6% of studies (n = 4), while World Scientific published 3% of studies (n = 2) and only 2% in MDPI, and all other studies were 1% in Technology Center, Taylor and Francis, ScienceDirect, and Mesopotamian Press.

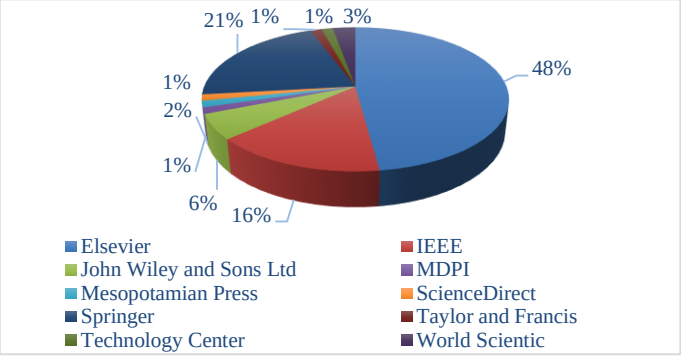


Fig. 7. Distribution of studies by publisher.

According to Fig. 8, the distribution of these studies by the first author's country spans over seven countries: Australia, China, Iraq, Malaysia, Tunisia, and the United Arab Emirates (UAE). Notably, the highest number of studies originates from Iraq (n = 31), followed by Malaysia (n = 21), highlighting the particularly active research communities in these countries.

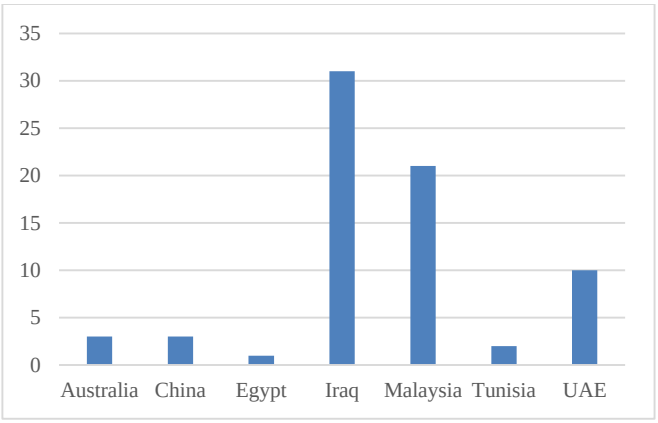


Fig. 8. Distribution of studies by country.

B. FWZIC Application

1) This subsection addresses RQ2 through five themes. A comprehensive meta-analysis of the reviewed studies revealed distinct thematic and domain-based applications of the FWZIC method. In addition, it identified the ranking methods commonly integrated with FWZIC and highlighted the various fuzzy extensions employed. The analysis also examined the most frequently used aggregation and defuzzification techniques. The following subsections present and discuss these findings. Details of the reviewed studies are provided in the Appendix.

2) Themes and domains of application

RQ2: What domains and themes are the studies based on?

A taxonomy of the reviewed studies, illustrated in Fig. 9, categorizes the articles into six main themes: Artificial Intelligence, COVID-19, Services, Sustainability, Technology, and Metaverse. These themes are further distributed across ten domains, including healthcare, agriculture, tourism, education, engineering, business, supply chain, manufacturing, and transport. A separate category is included for studies that do not align with any of the identified domains. Fig. 9 also highlights the popular MCDM methods that have been integrated with FWZIC in multiple studies, while a broader set of less frequently used methods is categorized under "Other". The following subsections provide a detailed review of the studies within each theme.

a) Artificial intelligence: Artificial intelligence (AI) is the ability of a computer system to carry out tasks that are usually performed by humans [65], [66]. It is one of the main themes adopted in these studies, with a total of ten studies distributed across two domains: nine in healthcare and one in engineering. Many healthcare studies focus on detecting and prioritizing autism spectrum disorder (ASD) patients based on sociodemographic and medical features. One study integrated machine learning (ML) and MCDM to classify ASD patients [67]. The study utilized an enhanced FWZIC method to determine the importance of the ASD features used in the training dataset [68]. In the same vein, another study used a hybrid approach to develop fifteen machine learning models for diagnosing patients with ASD and then benchmarked the models using FDOSM and FWZIC, based on seven performance metrics as evaluation criteria [67]. Similarly, another study developed a multimodal autism triage levels system using the Delphi method with FWZIC to determine the significance of ASD features and ML models for predicting severity levels [69]. The study further develops an explainable AI model to facilitate comprehension of the AI model's results.

Two studies in the healthcare domain utilise AI and VIKOR as a benchmarking method for respiratory diagnosis. While one study evaluates lung cancer diagnosis using class balancing approaches [70]. The other develops a framework to rank hybrid deep learning models based on chest X-ray analysis for computer-aided diagnosis (CAD) systems [71]. The remaining two healthcare studies focus on different objectives. One utilized a hybrid approach to first develop a labelled dataset of indoor air quality pollutants using the Unified Process for Labelling Pollutants Dataset (UPLPD), extended with the IT2TR-FWZIC weighting method [72]. The study further applied machine learning models to the developed dataset. Another study evaluates AI healthcare applications based on trustworthiness attributes using the CODAS method integrated with q-ROF2TL-FWZIC [73].

The final study in the AI theme belongs to the engineering domain and aims to benchmark optimization techniques for semi-polar III-V semiconductor materials using the FDOSM benchmarking method, integrated with the classical FWZIC method [74].

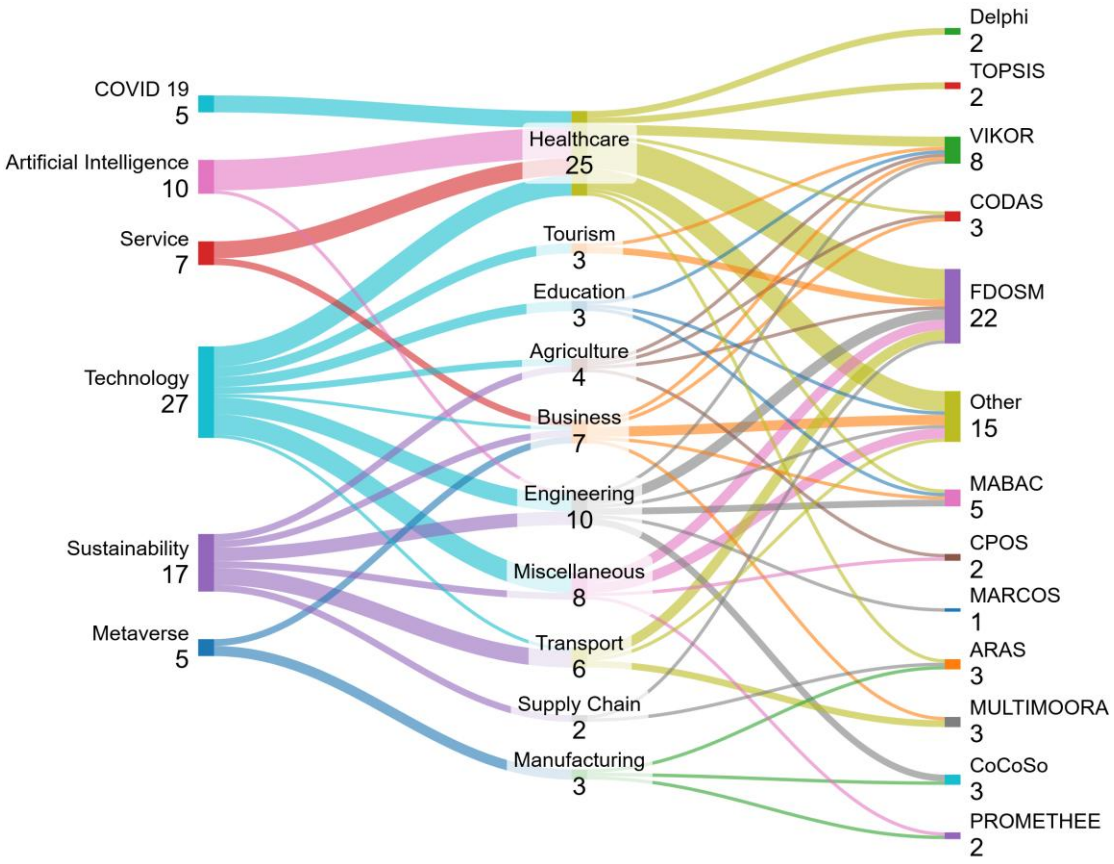


Fig. 9. Taxonomy of FWZIC application themes and domains.

b) COVID-19: Research continues to emphasize the significance of COVID-19, with five of the review studies belonging to this theme. Notably, all the studies in this theme are focused on the healthcare domain. Three studies aim to prioritize COVID-19 vaccine dose recipients using the FDOSM benchmarking method, but with different extensions of FWZIC methods using T-SF, q-ROF and PF as fuzzy sets [75], [76], [77]. Another study aims to develop a framework that benchmarks deep transfer and machine learning models based on chest X-rays to diagnose COVID-19 cases [78]. The study employs the dynamic localization-based decision (DLBD) method for benchmarking with PSVNHFS to extend FWZIC. The final study aims to develop a benchmark for the distribution of Anti-SARS-CoV-2 monoclonal antibody to eligible high-risk patients using the TOPSIS ranking method and using IVSH2 to extend FWZIC [79].

c) Services: Service is highlighted as one of the primary themes in these review studies, with five of them exploring this theme in connection with the healthcare domain. Two studies aim to prioritize autism patients with moderate emergency levels: one develops a framework using the VIKOR method for benchmarking using the classical FWZIC weighting method [80], and the other creates an intelligent triage method for classifying ASD patients into three severity levels using the Delphi method integrated with FWZIC [81]. Two additional studies focus on prioritizing remote patients with high-risk multiple chronic diseases: one uses emotion and sensory criteria with the TROOIL method for benchmarking and extends FWZIC with FOFS [82], and the other employs the FDOSM method for prioritizing hospital selection and uses QROF fuzzy sets for benchmarking and weighting method [83]. The final study aims to weight and prioritize the criteria for managing patients with cardiac problems using the classical FWZIC method [84].

d) Metaverse: Metaverse is an emerging area of research interest [60], with a focus evident in five of the review studies within the business domain (n = 2) and manufacturing domain (n = 3). Among these studies, one study developed a model concerning anonymity and privacy within the Bitcoin network-based metaverse in Industry 5.0, utilizing the MULTIMOORA benchmarking method with the extended LDFS-FWZIC weighting technique [85]. Another study [86] evaluated virtual commerce solutions for the metaverse utilizing the RATMI ranking method integrated with SLDFS-FWZIC.

Additionally, two studies concentrate on developing models to enhance control engineering tools, employing the same extended IVSFRS-FWZIC weighting method, which aims to support the Industrial Cyber-Physical Metaverse using the PROMETHEE II benchmarking method [87], while the others evaluate digital twin technology in manufacturing with the CoCoSo benchmarking method [88]. Lastly, the study aims to develop a model for evaluating Metaverse Tools based on a privacy model, integrated with the ARAS benchmarking method and the HF-FWZIC extended method [89].

e) Sustainability: Sustainability plays a vital role in many of these research studies, largely because of its importance in environmental research. There are thirteen studies focusing on sustainability across various fields like Agriculture, Transport, Engineering, Business, Supply Chain and more. For instance, one study in the agriculture domain evaluated water quality monitoring systems based on remote sensing techniques using VIKOR with FWZIC [90].

Within sustainable transportation, one study applies the MULTIMOORA method to model electronic passenger vehicles and utilizes the P-H-FWZIC weighting method [48]. Another study in this area developed a fuel supply system modelling approach for electric vehicles based on MARCOS and used the PPH-FWZIC weighted method [91]. A third study evaluated and ranked international oil companies for sustainable oil transportation based on MULTIMOORA and the extended LDFRS-FWZIC weighting method [92]. A fourth study has established a performance assessment of ship energy systems for transportation in the shipping industry based on FDOSIM and extended the q-ROFRS-FWZIC weighting method [93]. The last study in this domain built a pavement strategy selection based on FDOSIM, and using the DH-FWZIC extended weighting method [94].

Two of these review studies are associated with the engineering domain. One study developed a model to benchmark an energy systems integration framework for efficient resource utilization based on the MABAC ranking method integrated with TFN-FWZIC [95]. And another study evaluated management strategies in construction and demolition wastes based on FDOSIM-MULTIMOORA and based on the extended q-ROPHFAM FWZIC method [96].

On the other hand, two studies are connected to the business sector: one study evaluated sustainable circular business model innovation tools using the CODAS benchmarking method and the NBFS-FWZIC extended method [97]. Another study in this sector developed a model to assess palm oil industry 4.0 technologies for circular economy applications based on EDAS and using IVPFRS fuzzy sets with FWZIC [98]. Only one study in the sustainability theme is related to the smart living domain. The study developed a model to benchmark smart living frameworks using conditional probabilities by opinion scores and extended FWZIC under Bayesian rules in circular-Pythagorean fuzzy sets [99].

Finally, two studies in this theme are related to the supply chain domain: one study evaluated real-time IoT devices for monitoring food wastage in a supply chain system. The study used ARAS as a benchmarking method and extended the FWZIC using circular intuitionistic fuzzy set [100]. And a second study in this domain built a model for evaluating agriculture-food 4.0 supply chain using FDOSM as a benchmarking method and extended FWZIC under Fermatean probabilistic hesitant-fuzzy sets [50].

f) Technology: Most reviewed studies are primarily focused on the Technology theme, with studies from these reviews falling within this category across various domains. Three studies in the tourism sector are focused on evaluating e-tourism applications using various benchmarking methods. One study integrated Interval type 2 trapezoidal-FWZIC with VIKOR [101], while another study developed a framework based on FDOSM and FWZIC utilizing Neutrosophic fuzzy sets [102]. A third study focused on developing a decision modelling approach for smart e-tourism data management applications utilizing a spherical fuzzy rough environment and FDOSM as the benchmarking method [103].

Six studies focused their research on the healthcare domain. One study involves a benchmarking model for Sign Language Recognition Systems, utilising FDOSM and the extended CP-FWZIC method [104]. Another study utilized another benchmarking model using FDOSM with the original FWZIC, but this time used to evaluate smartphone applications for obesity management [105]. Two studies are centered on security and privacy in Blockchain-Based IoT Healthcare Industry 4.0 Systems. Both studies employed the extended S-FWZIC weighting method, with one using the TOPSIS-GRA [96] benchmarking method and the other using COPRAS [97]. The next study in this domain evaluated intrusion detection classifiers for Internet of Medical Things (IoMT) devices based on security and performance attributes. The study utilized the FDOSM method for benchmarking and extended FWZIC using the Rough Fermatean Fuzzy Sets [108]. Finally, the last study in the healthcare domain utilized the FWZIC method to assess the criteria for digital genetic tools [109].

Four studies in the engineering sector are highlighted. One study focuses on evaluating the risk analysis of offshore wind turbines, utilizing the DEMATEL method along with the extended LP-FWZIC weighting method for determining the relationships between the factors [110]. Another study is centered on evaluating active cooling systems, based on performance and technical attributes, employing the MABAC benchmarking system and the 2 TLP-FWZIC weighting method [111]. The other two studies applied the CoCoSo method, one focused on optimizing Control Engineering tools using digital twin Capabilities and a Cyber-Physical Metaverse manufacturing system using the IVFRS extensions[112], and the other study involved a design concept evaluation method using q-ROFS extensions [113].

Three studies in the technological realm are focused on education. Both studies involve developing frameworks for benchmarking brain–computer interface applications. One study utilized VIKOR as the Benchmarking method alongside the TFN extended weighted method [114], the second focus on evaluating speech-recognition chatbots for language learning using MCRAT methods and implementing the (T-SFS) [23]. The other study applied MABAC as the benchmarking method and extended FWZIC using the Neutrosophic Cubic Fuzzy Set [115].

Moreover, five studies are centered on network analysis. One study developed a model for evaluating routing algorithms in a multiprocessor system by utilizing Z-Cloud Rough Numbers (ZCRNs) environment to extend the weighting method and FDOSM for benchmarking [112]. The second study focused on selecting an optimal architecture for 5G-radio access network employing the type-2 N-FWZIC extended method for weighting and FDOSM for benchmarking [116]. The third study develops a modelling framework for 6G based blockchain technology

using RAFSI method and extends FWZIC with Normal Wiggly Hesitant Fuzzy sets [117]. The fourth study involved in evaluating a Dynamic locally based-utility approach for developing an internet self-reconfiguration robot extending FWZIC with type 2 Pythagorean [118]. And the fifth focuses more on the superiority of a robust cloud network using PROMETHEE and the spherical cubic fuzzy environment [119].

The last five studies on this theme cover various sectors, including marine robotics, transport, business, and agriculture. One of these studies aims to optimize the motion trajectory of an autonomous underwater vehicle by evaluating the optimization algorithms [120]. The study utilized FDOSM for benchmarking and FWZIC for criteria weighting. Another study focused on evaluating driver assistance systems for vehicles utilizing FDOSM for benchmarking and FWZIC for criteria weighting within Intuitionistic fuzzy sets [121]. Additionally, a separate study delved into organizational culture within companies to promote digital innovation, employing the q-rung picture FWZIC method for criteria weighting and simple additive weighting SAW for benchmarking [122]. The last two studies in this theme, related to Agriculture, one study developed of a modelling approach for drones for Precision Agriculture, utilizing FDOSM for benchmarking and FWZIC criteria weighting [123]; the other one involved in evaluating and swarm robots effectiveness in mechanized agricultural operations using CPOS methods and Circular Quintic Fuzzy Sets extensions [105].

These findings reveal the versatility of the FWZIC method and its use in several fields spanning across various domains such as healthcare, supply chain, manufacturing, engineering, business and more. Thus, the FWZIC method can be an ideal choice for practitioners and researchers in various fields when weighing a large number of criteria based on subjective human opinions.

3) MCDM methods and Fuzzy extensions

RQ3: Which ranking methods has FWZIC been integrated with?

RQ4: Which fuzzy environments are utilized to enhance FWZIC?

The seventy-one reviewed studies show that FWZIC has been extended with numerous fuzzy sets and integrated with several ranking methods. FWZIC was most frequently combined with the FDOSM ranking method (n = 22). While fifteen studies utilized the original TFNs with FWZIC [67], [69], [70], [71], [74], [80], [81], [84], [90], [95], [114], [120], [123], [124], [125], [126] all other studies enhanced the FWZIC method under various fuzzy environments such as the Spherical FS (T-SFS), Neutrosophic Sets (NS) [102], Q-Rung Orthopair (q-ROF) [76], [83], Pythagorean Fuzzy (PF) [77], Cubic Pythagorean FS (CPFS) [104], Spherical Fuzzy Rough Sets (SFR) [103], Z-Cloud Rough Numbers (Z-CRN) [112], -rung Orthopair Fuzzy Rough Sets (Q-ROFRS) [93], Rough Fermatean FS (RFFS) [108], Dual-Hesitant Fuzzy (DH), Type-2 Neutrosophic Numbers (T2NN) [116], and Intuitionistic FS (IFS) [121]. Two studies integrated FWZIC with FDOSM and another ranking method, MAIRCA and MULTIMOORA, using Fermatean Probabilistic Hesitant (FPH) [50] and q-rung Orthopair probabilistic hesitant fuzzy set (Q-ROPHFS) [96], respectively.

The second most common ranking method, which utilizes FWZIC for criteria weighting, is VIKOR (n = 9). Out of the eight studies, five did not enhance the original FWZIC method and utilized it with TFNs [70], [71], [80], [90], [114], while one study enhanced the fuzzy environment and applied the Interval Type 2 Trapezoidal Fuzzy sets (IT2TR) [101]. The other study used the single value Neutrosophic to enhance the fuzzy sets [116].

Eleven studies combined FWZIC with the MABAC, MULTIMOORA and ARAS ranking methods, respectively. Taha Aljburi et.al [95] integrated the MABAC method with FWZIC in its original form using TFN; while the four other studies used MABAC, enhancing FWZIC under Neutrosophic Cubic Sets (NCS) [115] and 2-Tuple Linguistic Pythagorean (2TLP) [111], q-runq Orthopair (q-ROF) [127], and Single-Valued Neutrosophic 2-tuple Linguistic (SVN2TL) [128]. The MULTIMOORA method was utilized with enhanced FWZIC under Probabilistic Hesitant (P-H) [48], Linear Diophantine FS (LDFS) [85], and Linear Diophantine Fuzzy RS (LDFRS) [92]. On the other hand, the ARAS ranking method was combined with an enhanced FWZIC method under the Circular Intuitionistic Fuzzy Set (C-IFS-FWZIC) [100]. Additionally, two other studies integrated the ARAS method with an enhanced FWZIC method, one study under Hexagonal FN (HF-FWZIC) [89], and the other study under Pythagorean Fuzzy sets (PF-FWZIC)[109].

FWZIC was paired with TOPSIS (n = 2) in two studies, one under Interval-Valued Spherical Fuzzy Hesitant 2-tuple (IVSH2) FS [79]. The other study enhanced FWZIC under the Spherical (S) FS [106]. Two studies combined FWZIC with SAW (n = 2) for criteria weighting by extending FWZIC under the q-Rung Picture FS (Q-RP) [122] and Spherical Fuzzy Rough Sets (SFR) [103], while another two studies paired FWZIC with ARAS (n=2), extending it using Circular Intuitionistic FS (C-IFS) [100] and Hexagonal FN (HF) [89]. FWZIC was also combined with the CODAS (n=2) ranking method in two studies, which extended the criteria weighting method using Q-Rung Orthopair Fuzzy 2-Tuple Linguistic (q-ROF2TL) [73] and Neutrosophic Bipolar FS (NBFS) [97]. FWZIC was also utilized with the Delphi (n=2) method in two studies under the original TFN fuzzy environment [69], [81].

CoCoSo and CODAS were integrated with FWZIC in three studies (n =3). The CoCoSo method was implemented with the Interval-Valued Spherical Fuzzy Rough Sets extension in two studies (IVSFRS), [88], [129]. The other study applied CoCoSo with Q-rung Orthopair Fuzzy rough sets (Q-rungOFRS) FWZIC extensions, [113]. On the other hand, one study integrated CODAS with the q-rung Orthopair Fuzzy type 2 linguistic FWZIC extensions on intelligent healthcare applications. [73], while the other study used CODAS with Neutrosophic bipolar (NBFS) FWZIC enhancements on circular sustainable business tools. [97], and the last study used hyperbolic fuzzy sets extensions (Hy) for sustainability. [130].

CPOS was integrated twice with FWZIC. Alsattar, Qahtan, et al. [99] utilized the CPOS method with FWZIC in Circular-Pythagorean Fuzzy environment (C-PFS). The other study used

CPOS and implemented Circular-Quintic Fuzzy Sets extensions (cQuFS) on FWZIC[105] for sustainability. MARCOS was integrated with FWZIC in two studies for sustainability, one extended FWZIC with Pythagorean Probabilistic Hesitant (PPF) [91]. The other study used the Interval value, Fermatean Fuzzy sets (IVFFH) [131]. Similarly, PROMETHEE II was integrated twice with FWZIC. One study used an Interval-valued spherical fuzzy rough environment (IVSFRS) [87]. The other study used Spherical Cubic Fuzzy (SCFS) [124].

All other MCDM methods (n=14) utilized in the literature were only integrated with FWZIC in one study. Zaidan, Alsattar, et al. (2023) extended the COPRAS method using FWZIC, extending it using the Interval-Valued Spherical Fuzzy sets (IVSFS) [107]. Bilquise et al. [86] integrated the RATMI method with the FWZIC method and extended it under the spherical linear Diophantine FSs (SLDFSs). The EDAS method was integrated with FWZIC using a type 2 fuzzy environment, Interval-valued Pythagorean Fuzzy Rough set (IVPFRS) [132]. The remaining studies extended FWZIC under various fuzzy sets such as Fractional Orthotriple Fuzzy sets (FOFS) [82], Interval Type 2 Trapezoidal-Fuzzy sets (IT2TR) [72], Linguistic Pythagorean Fuzzy sets (LPFS) [110], Probabilistic Single-Valued Neutrosophic Hesitant Fuzzy set (PSVNHFS) [78] with MCDM methods MARCOS, TROOIL, UPLPD, DEMATEL, and DLBD, respectively. Finally, one study extended FWZIC using Complex T-spherical FWZIC (CT) fuzzy sets and employed a machine learning model for benchmarking the alternatives [68], and another study utilized FWZIC for prioritizing criteria weights under the original TFNs without ranking any alternatives [84].

The findings reveal the ability of FWZIC to be seamlessly integrated with multiple MCDM methods. Overall, it is observed that the most popular ranking method integrated with FWZIC is FDOSM (n = 22), followed by VIKOR (n = 8). Moreover, FWZIC was utilized in its original form, using TFN, in n=20 studies only. All other studies (n =51) have enhanced FWZIC in various fuzzy environments, ranging from classical FS to type-2 fuzzy sets, to capture a broader uncertainty space. Notably, ambiguity and imprecision in human judgement remain open issues, due to which FWZIC extensions are being extensively researched.

4) Aggregation and Defuzzyfying techniques

RQ5: What techniques are used for aggregation and defuzzification in the FWZIC algorithm?

a) Aggregation techniques: This section explores the common aggregation methods utilized by the reviewed studies for aggregating the fuzzy EDM values. Aggregation of the fuzzy numbers is one of the most crucial steps in FWZIC method. The fuzzy expert judgement for each criterion is aggregated with the aim of merging all the experts' opinions for that criterion so that the resulting aggregated output fully considers each individual criterion. This ensures that the final criterion weight is based on the overall evaluation. Aggregation is critical for producing appropriate results from EDM. In literature, several aggregation methods exist for various fuzzy environments [19].

All studies that employed the classical FWZIC using TFN utilized a pseudo-arithmetic operator $\oplus$ to aggregate the fuzzy values of each criteria expert opinion in the EDM. Similarly, the pseudo-arithmetic operator was utilized for HFNs [89]. For instance, two TFN $\bar{x} = (a_1, b_1, c_1), \bar{y} = (a_2, b_2, c_2)$, the TFN pseudo-arithmetic operation between $\bar{x}\ and\ \bar{y}$ is defined as follows: $\bar{x} \oplus \bar{y} = (a1 + a2, b1 + b2, c1 + c2)$.

Although advanced aggregation operators, like geometric aggregation, weighted aggregation and quasi-fuzzy weighted aggregation have been proposed in literature for these classical fuzzy numbers [133], none of the studies employed the aggregation operators.

Similarly, the other studies that enhanced FWZIC with various advanced fuzzy environments, such as IFS, PyFS, SFS, IVFS, and more, employed the arithmetic weighted aggregation operator of the respective fuzzy environment. For instance, [115] utilized the NCN weighted arithmetic averaging (NCNWAA) operator for the NCS-FWZIC technique, while [96] utilized the q-ROPHF arithmetic mean (q-ROPHFAM) operator, and [106] employed the spherical weighted arithmetic mean (SWAM) operator. Ref [102], on the other hand, employed a geometric aggregation operator for NS-FWZIC. Other aggregation operators, like Bonferroni [134] and Einstein [135] aggregation operators were not considered in any of the studies. Notably, different aggregation operators may impact the criteria weighing results, which may be explored for further research.

b) Defuzzification techniques: Defuzzification is a process that converts a fuzzy set or fuzzy number into a crisp value or number. Defuzzification is used in fuzzy modeling to convert the fuzzy output values to their corresponding crisp weight values. This process is necessary because all fuzzy sets inferred by fuzzy inference in the fuzzy rules must be aggregated to produce one single number as the output of the fuzzy model. There are numerous techniques for defuzzifying a fuzzy set.

The centroid method mathematically derives the Center of Gravity (COG) to convert a fuzzy number to its corresponding crisp value. However, this method is computationally complex for complex membership functions. For instance, for a TFN $\bar{x} = (a, b, c)$, the defuzzified crisp value is computed as $\bar{x} = (a + b + c)/3$, while for HFNs $\bar{x} = (a, b, c, d, e, f)$, the crisp value is derived using the equation $\bar{x} = (3a + 3b + 10c + 10d + 5e + 3f)/34$.

All studies that utilized TFNs and HFN applied the centroid method for defuzzifying the fuzzy weight values. Other methods proposed in literature, such as max-membership, weighted-average and mean-max methods [18] were not employed by any of the studies. All other reviewed studies that employed advanced fuzzy sets consistently utilized a single score function of the respective fuzzy environment to compute the defuzzified crisp value. Notably there exists multiple score functions in fuzzy sets, each of which may yield different results [136]. None of the studies evaluated the criteria weights based on multiple score functions.

Overall, the findings in this section reveal a gap in literature, showing that there are multiple avenues of further research to

evaluate the impact of the use of different aggregation and defuzzification techniques with the various fuzzy environments to further enhance the FWZIC method.

C. Expert and Criteria Characteristics

This section explores the characteristics of the experts and evaluation criteria employed in studies utilizing the FWZIC method. In addressing the first part, we examine the background, qualifications, and experience of the experts involved, along with how their input was used in the decision-making process. We also highlight patterns in expert selection, including the number of experts typically engaged and how their opinions were considered. In the second part, we analyze the nature, structure, and sources of the criteria used across the reviewed studies. This includes how the criteria were evaluated, the scales and linguistic terms applied, and the integration of fuzzy sets to ensure consistency. Together, these insights provide a comprehensive understanding of how FWZIC accommodates expert-driven input and diverse criteria sets in multi-criteria decision-making contexts.

1) Expert characteristics

RQ6: What are the characteristics of the experts involved in the studies?

All experts employed in the studies possess expertise within the field of study, including cardiologists, academicians, tourism experts, data scientists, and others. In [112], the authors identified members of the technical community as experts. Two studies [109], [86] explicitly list the qualifications and details of the experts. Some studies stipulated the minimum experience requirements to be considered as experts in that field. For instance, [67] and [104] require experts with over 5 years of experience, while others, such as those by [71], [95] and [110] selected experts with more than 10 years of experience. In numerous case studies, the evaluation forms and criteria adopted are also assessed by experts chosen for their knowledge and expertise in relevant fields [81]. Notably, only one study [110] differentiated the experts' opinions based on seniority and years of experience, assigning each expert an objective score to determine the overall criteria weight. All studies considered the opinion of the expert at an equal level, without differentiating them based on their expertise.

Most case studies (n = 46) involve three experts to assess and weigh the evaluation criteria and consider this a minimum requirement. Few studies (n = 7) employed four experts, while three studies opted for five experts. Some studies involved a large number of experts. For instance, [106] used 17 experts to evaluate security and privacy in Blockchain-Based IoT Healthcare Industry 4.0 Systems, while [101], [102] both employed 11 experts to determine the criteria's significance for evaluating smart e-tourism applications. In [95], the study utilized 12 experts to assess various energy sources for sustainability and efficient resource usage. Three studies employed 6 experts for assessing criteria weights for lung cancer diagnosis techniques [70] [123], for evaluating indoor air quality pollutants [72] and for evaluating the significance of hybrid deep learning models based on chest X-ray analysis in CAD systems [71]. Three studies, which involved five experts [85] evaluated the privacy model of the Bitcoin network for the Metaverse, while [89] assessed the significance of privacy features for metaverse tools and [86] evaluated virtual commerce applications for the Metaverse. Several studies selected four experts, primarily in the medical field [67], [84], [137], [138].

It is observed that the minimum number of experts employed for weighing criteria using FWZIC ranges from a minimum of 3 to a maximum of 43 experts, with a large percentage utilizing three experts. This finding reveals the ease of use of the FWZIC method regardless of the number of experts involved. FWZIC proves to be reliable and robust in consolidating opinions from multiple experts, each with diverse expertise, to reduce bias in the criteria weighting process while maintaining consistency. In other subjective criteria weighting methods, such as AHP, the process of consolidating criteria weightings using multiple experts would be a time-consuming process and prone to inconsistencies as well. Notably, only one study differentiated the experts' opinions based on their seniority and experience level.

2) Criteria characteristics

RQ7: What are the characteristics of the criteria utilized in the studies?

The number of criteria across the 71 reviewed studies varies significantly, ranging from a minimum of 3 to a maximum of 75 criteria. Most of these criteria are structured at a single level, while some studies have utilized multi-level criteria ranging from two-levels, three-levels and four levels.

The criteria utilized in the studies originate from various sources. Several studies (n = 43) source their evaluation criteria from root articles. A root article may be defined as an existing systematic review, survey, or case study which has extracted the criteria from literature [139]. Others use only a portion of these criteria (n = 7). Additionally, some studies derive their criteria from existing datasets when integrating FWZIC with machine learning techniques. Other studies source their criteria from literature (n=12), while a few studies (n = 4) identify their criteria as performance metrics or specific methods for selecting optimal machine learning algorithms.

The criteria are assessed by experts using an online evaluation form. All studies have consistently utilized a 5-point Likert scale to assess the criteria's significance using linguistic values such as "Very High Importance", "High Importance", "Moderate Importance", "Low Importance", "Very Low Importance". Only one study employed a 10-point Likert scale [104]. All studies mapped the linguistic values to a numerical scale and converted them to the corresponding fuzzy number to generate a fuzzy Expert Decision Matrix.

Overall, the FWZIC method has been utilized with as few as three criteria and as many as seventy-five. These findings demonstrate the robustness of the FWZIC method in weighting a large number of criteria without any inconsistencies, thereby demonstrating its superiority over other subjective methods, such as AHP, ANP, and BWM, which would not only be computationally intensive but also prone to inconsistencies.

V. Research Implications

By analyzing seventy-one studies across diverse domains such as healthcare, engineering, supply chain, agriculture, and technology, this review confirms the FWZIC method's versatility, robustness, and growing popularity since its inception.

The findings highlight that FWZIC effectively addresses the limitations of traditional subjective weighting methods by ensuring consistency, accommodating a wide range of fuzzy environments, and facilitating decision-making even with large numbers of criteria and experts. Moreover, the increasing integration of FWZIC with various MCDM ranking techniques, particularly FDOSM and VIKOR, underlines its adaptability and relevance across multidisciplinary fields.

This review establishes that FWZIC offers a transformative approach to subjective criteria weighting by minimizing inconsistency and bias, which are persistent challenges in MCDM. It encourages a shift from conventional time-consuming methods like AHP and BWM towards more streamlined, reliable techniques, particularly in environments characterized by ambiguity.

The systematic integration of FWZIC with advanced fuzzy environments (e.g., q-ROFS, T-SFS, IVFS) has expanded the spectrum of uncertainty that can be modeled. This evolution enhances the realism and flexibility of decision models, paving the way for more human-centric decision support systems.

Despite extensive applications, the study uncovers that most researchers have relied on standard aggregation (e.g., pseudo-arithmetic operators) and centroid-based defuzzification methods. This opens an important avenue for future exploration: investigating alternative aggregation and defuzzification strategies could yield even more robust, interpretable, and precise decision-making frameworks.

The breadth of FWZIC's application, from medical diagnosis prioritization to supply chain resilience, demonstrates its critical role as an enabler of informed decision-making in complex, multidisciplinary problems. Organizations can leverage FWZIC-based models to improve strategic planning, risk assessment, resource allocation, and sustainability efforts.

The study confirms FWZIC's practical utility in real-world contexts by facilitating expert opinion consolidation without overburdening decision-makers. Its ability to integrate three to forty-three expert opinions efficiently, without inconsistency, empowers organizations to make collective, evidence-based decisions more effectively.

Unlike traditional subjective methods that are unstable with increasing numbers of criteria or experts, FWZIC demonstrates strong scalability. This robustness supports its deployment in emerging fields such as AI-driven healthcare, Industry 5.0, digital twin technologies, and the Metaverse, where complex, multi-criteria evaluations are the norm.

Although FWZIC's flexibility is a strength, the diversity of its applications also signals a need for establishing best practices or standardized frameworks for its use across different fuzzy environments and MCDM settings. Future work should aim to benchmark FWZIC applications across domains to strengthen comparability and reliability.

A. Future Directions

This section presents some additional research areas that emerge from the review, thereby offering practitioners and researchers a means to further enhance research in this field.

1) Fuzzy scale. Most of the reviewed studies have utilized a 5-point Likert scale for assessing criteria importance, with only one study utilizing a 10-point Likert scale. A scale is essential for a comparison reference when assessing the criteria's significance. It may be reasonable to assume that a higher-level scale would cover a wider spectrum of uncertainty in expert opinion. Therefore, a future area of research could be to explore the impact of various scales (5-point, 7-point, 10-point) on the criteria weighting results.

2) Experts. Most reviewed studies employed at least 3 experts to assess the criteria's weighting significance. While some studies designated minimum requirements for expert selection, only one study employed an objective approach in the criteria weighting process to differentiate experts based on their seniority and experience level. This opens an avenue for further research in this area to study the difference in the criteria weights with or without an objective approach. Furthermore, researchers could also employ objective criteria other than seniority in job title and experience years, such as preferred area of practice (industry experts, academicians, researchers), expert qualification and more.

3) Aggregation and Defuzzification. Most of the reviewed studies employed the arithmetic weighted aggregation operator of the respective fuzzy environment, while only a few studies have employed the geometric aggregation operator, which signifies that more research is encouraged to investigate the use of these aggregators. Moreover, none of the studies employed aggregation operators like Bonferroni [134] and Einstein [135] aggregation operators, allowing future research possibilities to employ these and other new aggregation operators that were not considered in the review studies.

From the defuzzification techniques point of view, most of the studies utilized the centroid method for defuzzifying the fuzzy weight values or used a single score function for defuzzifying; none of the reviewed studies used other proposed methods in literature, such as max-membership, weighted-average and mean-max methods. Additionally, no studies used multiple score functions to calculate the final weight methods that would impact the ranking results. This gap in the literature presents an opportunity for further research to explore the application of these alternative defuzzifying methods or to use multiple scores to evaluate and finalize the criteria weights.

4) Fuzzy extensions. Although several studies have enhanced the FWZIC method under various fuzzy environments, the issue of inconsistency and ambiguity in human judgment remains unresolved when calculating criteria weights. The FWZIC method can be further extended with different fuzzy environments that encompass a broader

uncertainty range, such as interval-valued Neutrosophic sets. Moreover, future research could also analyze the soft versus rough characteristics in fuzzy sets and their effect on criteria weighting using FWZIC.

5) MCDM integration. FWZIC has been integrated with several MCDM methods for the purpose of ranking and studying the causal effects on the attributes. Notably, several MCDM ranking methods have not yet been integrated with FWZIC, such as MACBETH, MCRAT, RAMS, and others. Integrating FWZIC with these methods presents a potential research opportunity to analyze and test the results and effectiveness of these integrations.

6) Domain. The FWZIC method has been utilized in numerous sectors to determine criteria weights such as healthcare, tourism, agriculture, transport and more. However, there are still many unexplored domains that present an opportunity for future research, such as real estate, aviation, media and telecommunication.

7) Other criteria weighting methods. FWZIC is a comparatively superior criteria weighting method when compared to other traditional weighting methods, such as AHP. Not only is FWZIC computationally simple, but it also eliminates the need for criteria comparison and leads to zero inconsistencies in subjective evaluation. A possible future area of research is a comparative study of FWZIC and other methods, such as the Full Consistency Method (FUCOM) [140] and Opinion Weight Criteria Method (OWCM) [141]. FUCOM provides a strategy with fewer pairwise comparisons to reduce inconsistency, while OWCM combines both subjective and objective criteria weighting methods to eliminate inconsistency. This comparison could highlight the strengths and weaknesses of each method, potentially leading to further advancements in criteria weighting techniques.

VI. Conclusion

This systematic review presents a comprehensive synthesis of the FWZIC method's applications in MCDM. By examining seventy-one empirical studies across a range of domains, including healthcare, engineering, supply chain, and sustainability, the review highlights FWZIC's versatility in addressing the challenges of subjective decision-making, particularly in managing expert opinions with consistency and minimal bias.

The findings demonstrate that FWZIC has been widely adapted through the integration of various fuzzy environments and has been successfully combined with multiple MCDM ranking techniques, reflecting its flexibility and growing acceptance. Despite its advantages, gaps remain regarding the exploration of alternative aggregation and defuzzification methods, offering valuable opportunities for future research to further enhance decision-making accuracy and robustness.

Overall, this review not only consolidates the existing knowledge on FWZIC but also identifies emerging trends, methodological gaps, and potential directions for advancing subjective criteria weighting techniques. It serves as a vital resource for researchers and practitioners seeking to implement or extend the FWZIC method in increasingly complex and uncertain decision-making contexts. This study not only underscores the value of FWZIC as a cornerstone method in subjective decision-making but also charts a pathway for future research directions. By systematically identifying gaps and proposing avenues for further methodological refinements, this review serves as a critical reference for researchers, practitioners, and policymakers aiming to enhance decision quality in increasingly complex, uncertain, and multi-stakeholder environments.

This study may have been subject to certain limitations attributable to various factors. Primarily, the research employed three bibliographic databases, encompassing multiple disciplines, to retrieve pertinent studies. This approach may have led to the inadvertent exclusion of relevant empirical studies that utilized the FWZIC method. Additionally, ten studies were excluded from the meta-analysis of the present study due to the unavailability of full-text access. Despite these constraints, the examination of the seventy-one reviewed articles has yielded a comprehensive and detailed review of the FWZIC method, which is expected to provide valuable guidance for both practitioners and researchers in this domain.

References

[1] X. Gong et al., "Exploring an interdisciplinary approach to sustainable economic development in resource-rich regions: An investigation of resource productivity, technological innovation, and ecosystem resilience," Resour. Policy, vol. 87, p. 104294, 2023.

[2] H. Taherdoost, "Multi-Criteria Decision Making (MCDM) Methods and Concepts," no. Mcdm, pp. 77–87, 2023.

[3] N. A. Azhar, N. A. M. Radzi, and W. S. H. M. Wan Ahmad, "Multi-criteria decision making: a systematic review," Recent Adv. Electr. Electron. Eng. (Formerly Recent Patents Electr. Electron. Eng., vol. 14, no. 8, pp. 779–801, 2021.

[4] S. Chakraborty, R. D. Raut, T. M. Rofin, and S. Chakraborty, "A comprehensive and systematic review of multi-criteria decision-making methods and applications in healthcare," Healthc. Anal., p. 100232, 2023.

[5] S. Dožić, "Multi-criteria decision making methods: Application in the aviation industry," J. Air Transp. Manag., vol. 79, no. May, 2019, doi: 10.1016/j.jairtraman.2019.101683.

[6] J. J. Thakkar, Multi-criteria decision making, vol. 336. Springer, 2021.

[7] M. Alemi-Ardakani, A. S. Milani, S. Yannacopoulos, and G. Shokouhi, "On the effect of subjective, objective and combinative weighting in multiple criteria decision making: A case study on impact optimization of composites," Expert Syst. Appl., vol. 46, pp. 426–438, 2016.

[8] I. Vinogradova, "Multi-attribute decision-making methods as a part of mathematical optimization," Mathematics, vol. 7, no. 10, 2019, doi: 10.3390/math7100915.

[9] Z. Bin Yusop, K. Ahmed, S. M. Shirazi, and N. H. Zardari, Weighting methods and their effects on multi-criteria decision making model outcomes in water resources management. Springer, 2015.

[10] E. Koksalmis and Ö. Kabak, "Deriving decision makers' weights in group decision making: An overview of objective methods," Inf. Fusion, vol. 49, pp. 146–160, 2019.

[11] G. Demir, P. Chatterjee, S. Kadry, A. Abdelhadi, and D. Pamučar, "Measurement of Alternatives and Ranking according to Compromise Solution (MARCOS) Method: A Comprehensive Bibliometric Analysis," Decis. Mak. Appl. Manag. Eng., vol. 7, no. 2, pp. 313–336, 2024, doi: 10.31181/dmame7220241137.

[12] T. L. Saaty, "How to make a decision: the analytic hierarchy process," Eur. J. Oper. Res., vol. 48, no. 1, pp. 9–26, 1990.

[13] T. L. Saaty, L. G. Vargas, T. L. Saaty, and L. G. Vargas, The analytic network process. Springer, 2013.

[14] J. Rezaei, "Best-worst multi-criteria decision-making method: Some properties and a linear model," Omega, vol. 64, pp. 126–130, 2016.

[15] R. T. Mohammed et al., "Determining Importance of Many-Objective Optimisation Competitive Algorithms Evaluation Criteria Based on a Novel Fuzzy-Weighted Zero-Inconsistency Method," Int. J. Inf. Technol. Decis. Mak., vol. 21, no. 1, pp. 195–241, 2022, doi: 10.1142/S0219622021500140.

[16] G. O. Odu, "Weighting methods for multi-criteria decision making technique," J. Appl. Sci. Environ. Manag., vol. 23, no. 8, pp. 1449–1457, 2019.

[17] L. A. Zadeh, "Fuzzy sets," Inf. Control, vol. 8, no. 3, pp. 338–353, 1965.

[18] Y. Liu, C. M. Eckert, and C. Earl, "A review of fuzzy AHP methods for decision-making with subjective judgements," Expert Syst. Appl., vol. 161, p. 113738, 2020, doi: 10.1016/j.eswa.2020.113738.

[19] Y. R. Muhsen, N. A. Husin, M. B. Zolkepli, and N. Manshor, "A systematic literature review of fuzzy-weighted zero-inconsistency and fuzzy-decision-by-opinion-score-methods: Assessment of the past to inform the future," J. Intell. Fuzzy Syst., vol. 45, no. 3, pp. 4617–4638, 2023, doi: 10.3233/JIFS-230803.

[20] S. Qahtan, H. A. Alsattar, A. A. Zaidan, M. Deveci, D. Pamucar, and W. Ding, "A novel fuel supply system modelling approach for electric vehicles under Pythagorean probabilistic hesitant fuzzy sets," Inf. Sci. (Ny)., vol. 622, pp. 1014–1032, 2022, doi: 10.1016/j.ins.2022.11.166.

[21] M. Mathew, R. K. Chakrabortty, and M. J. Ryan, "A novel approach integrating AHP and TOPSIS under spherical fuzzy sets for advanced manufacturing system selection," Eng. Appl. Artif. Intell., vol. 96, p. 103988, 2020.

[22] G. Bilquise, K. Shaalan, and M. AlKhatib, "Evaluation of Virtual Commerce Applications for the Metaverse Using Spherical Linear Diophantine-Based Modeling Approach," Hum. Behav. Emerg. Technol., vol. 2024, no. 1, p. 4571959, 2024.

[23] S. Ibrahim and G. Bilquise, "Beyond ChatGPT: Benchmarking speech-recognition chatbots for language learning using a novel decision-making framework," Educ. Inf. Technol., pp. 1–33, 2024.

[24] J. Rezaei, "Best-worst multi-criteria decision-making method," Omega, vol. 53, pp. 49–57, 2015.

[25] C. Kahraman, S. C. Onar, and B. Oztaysi, "Fuzzy Multicriteria Decision-Making: A Literature Review," Int. J. Comput. Intell. Syst., vol. 8, no. 4, pp. 637–666, 2015, doi: 10.1080/18756891.2015.1046325.

[26] O. Castillo, P. Melin, J. Kacprzyk, and W. Pedrycz, "Type-2 fuzzy logic: theory and applications," in 2007 IEEE international conference on granular computing (GRC 2007), IEEE, 2007, p. 145.

[27] J. Dong, S. Wan, and S.-M. Chen, "Fuzzy best-worst method based on triangular fuzzy numbers for multi-criteria decision-making," Inf. Sci. (Ny)., vol. 547, pp. 1080–1104, 2021.

[28] G. Liu and X. Wang, "A trapezoidal fuzzy number-based VIKOR method with completely unknown weight information," Symmetry (Basel)., vol. 15, no. 2, p. 559, 2023.

[29] P. Rajarajeswari, A. S. Sudha, and R. Karthika, "A new operation on hexagonal fuzzy number," Int. J. fuzzy Log. Syst., vol. 3, no. 3, pp. 15–26, 2013.

[30] K. T. Atanassov and S. Stoeva, "Intuitionistic fuzzy sets," Fuzzy sets Syst., vol. 20, no. 1, pp. 87–96, 1986.

[31] R. R. Yager, "Pythagorean membership grades in multicriteria decision making," IEEE Trans. fuzzy Syst., vol. 22, no. 4, pp. 958–965, 2013.

[32] X. Peng and G. Selvachandran, "Pythagorean fuzzy set: state of the art and future directions," Artif. Intell. Rev., vol. 52, no. 3, pp. 1873–1927, 2019, doi: 10.1007/s10462-017-9596-9.

[33] T. Senapati and R. R. Yager, "Fermatean fuzzy sets," J. Ambient Intell. Humaniz. Comput., vol. 11, pp. 663–674, 2020.

[34] R. R. Yager, "Generalized orthopair fuzzy sets," IEEE Trans. Fuzzy Syst., vol. 25, no. 5, pp. 1222–1230, 2016.

[35] M. J. Khan, P. Kumam, and M. Shutaywi, "Knowledge measure for the q-rung orthopair fuzzy sets," Int. J. Intell. Syst., vol. 36, no. 2, pp. 628–655, 2021.

[36] B. C. Cuong and V. Kreinovich, "Picture fuzzy sets," J. Comput. Sci. Cybern., vol. 30, no. 4, pp. 409–420, 2014.

[37] M. Parimala and S. Jafari, "Spherical Linear Diophantine Fuzzy Graphs: Unleashing the Power of Fuzzy Logic for Uncertainty Modeling and Real-World Applications," Axioms, vol. 13, no. 3, p. 153, 2024.

[38] F. Kutlu Gündoğdu and C. Kahraman, "Spherical fuzzy sets and spherical fuzzy TOPSIS method," J. Intell. fuzzy Syst., vol. 36, no. 1, pp. 337–352, 2019.

[39] T. Mahmood, K. Ullah, Q. Khan, and N. Jan, "An approach toward decision-making and medical diagnosis problems using the concept of spherical fuzzy sets," Neural Comput. Appl., vol. 31, pp. 7041–7053, 2019.

[40] M. Riaz and M. R. Hashmi, "Linear Diophantine fuzzy set and its applications towards multi-attribute decision-making problems," J. Intell. Fuzzy Syst., vol. 37, no. 4, pp. 5417–5439, 2019.

[41] I. Grattan-Guinness, "Fuzzy membership mapped onto intervals and many-valued quantities," Math. Log. Q., vol. 22, no. 1, pp. 149–160, 1976.

[42] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning—I," Inf. Sci. (Ny)., vol. 8, no. 3, pp. 199–249, 1975.

[43] H. Wang, F. Smarandache, Y. Zhang, and R. Sunderraman, "Single valued neutrosophic sets," Infin. study, vol. 12, p. 20110, 2010.

[44] V. Torra, "Hesitant fuzzy sets," Int. J. Intell. Syst., vol. 25, no. 6, pp. 529–539, 2010.

[45] B. Zhu, Z. Xu, and M. Xia, "Dual hesitant fuzzy sets," J. Appl. Math., vol. 2012, 2012.

[46] B. Farhadinia, U. Aickelin, and H. A. Khorshidi, "Uncertainty measures for probabilistic hesitant fuzzy sets in multiple criteria decision making," Int. J. Intell. Syst., vol. 35, no. 11, pp. 1646–1679, 2020.

[47] D. Ramot, R. Milo, M. Friedman, and A. Kandel, "Complex fuzzy sets," IEEE Trans. fuzzy Syst., vol. 10, no. 2, pp. 171–186, 2002.

[48] S. Qahtan, H. A. Alsattar, A. A. Zaidan, D. Pamucar, and M. Deveci, "Integrated sustainable transportation modelling approaches for electronic passenger vehicle in the context of industry 5.0," J. Innov. Knowl., vol. 7, no. 4, p. 100277, 2022, doi: 10.1016/j.jik.2022.100277.

[49] M. M. Salih, B. B. Zaidan, and A. A. Zaidan, "Fuzzy decision by opinion score method," Appl. Soft Comput. J., vol. 96, p. 106595, 2020, doi: 10.1016/j.asoc.2020.106595.

[50] S. Qahtan et al., "Evaluation of agriculture-food 4.0 supply chain approaches using Fermatean probabilistic hesitant-fuzzy sets based decision making model," Appl. Soft Comput., vol. 138, p. 110170, 2023, doi: 10.1016/j.asoc.2023.110170.

[51] J. Kim et al., "My health advisor is a robot: Understanding intentions to adopt a robotic health advisor," Int. J. Human–Computer Interact., vol. 40, no. 19, pp. 5697–5706, 2024, doi: 10.1080/10447318.2023.2239559.

[52] S. Opricovic, "Fuzzy VIKOR with an application to water resources planning," Expert Syst. Appl., vol. 38, no. 10, pp. 12983–12990, 2011, doi: 10.1016/j.eswa.2011.04.097.

[53] M. Palanikumar, A. Iampan, and S. Broumi, "MCGDM based on VIKOR and TOPSIS proposes neutrsophic Fermatean fuzzy soft with aggregation operators," Int. J. Neutrosophic Sci., vol. 19, no. 3, pp. 85–94, 2022, doi: 10.54216/IJNS.190308.

[54] P. Wang, Z. Zhu, and Y. Wang, "A novel hybrid MCDM model combining the SAW, TOPSIS and GRA methods based on experimental design," Inf. Sci. (Ny)., vol. 345, pp. 27–45, 2016, doi: 10.1016/j.ins.2016.01.076.

[55] H. Q. Nguyen, V. T. Nguyen, D. P. Phan, Q. H. Tran, and N. P. Vu, "Multi-Criteria Decision Making in the PMEDM Process by Using MARCOS, TOPSIS, and MAIRCA Methods," Appl. Sci., vol. 12, no. 8, 2022, doi: 10.3390/app12083720.

[56] T. S. S. Deepu and V. Ravi, "Exploring critical success factors influencing adoption of digital twin and physical internet in electronics industry using grey-DEMATEL approach," Digit. Bus., vol. 1, no. 2, p. 100009, 2021, doi: 10.1016/j.digbus.2021.100009.

[57] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[58] S. Ibrahim and K. Shaalan, "Systematic review of Knowledge Management Integration in Higher Educational Institution with an Emphasis on a Blended Learning Environment," in International Conference on Information Systems and Intelligent Applications, 2022.

[59] K. S. Ghazala Bilquise, "Critical Success Factors of Knowledge Management in Higher Education: A Systematic Review," in International Conference on Emerging Technologies and Intelligent Systems, Springer, 2023.

[60] G. Bilquise, K. Shaalan, and M. Alkhatib, "A Systematic Review of Virtual Commerce Solutions for the Metaverse.," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 8, 2024.

[61] L. Yang et al., "Quality assessment in systematic literature reviews: A software engineering perspective," Inf. Softw. Technol., vol. 130, p. 106397, 2021.

[62] G. Bilquise, S. Ibrahim, and K. Shaalan, "Emotionally Intelligent Chatbots: A Systematic Literature Review," Hum. Behav. Emerg. Technol., vol. 2022, 2022.

[63] N. J. Van Eck and L. Waltman, "VOSviewer manual," Leiden: Univeristeit Leiden, vol. 1, no. 1, pp. 1–53, 2013.

[64] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," Int J Surg, vol. 8, no. 5, pp. 336–341, 2010.

[65] G. Bilquise, S. Ibrahim, and K. F. Shaalan, "Bilingual AI-Driven Chatbot for Academic Advising," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 8, 2022, doi: 10.14569/IJACSA.2022.0130808.

[66] G. Bilquise and K. Shaalan, "AI-based Academic Advising Framework: A Knowledge Management Perspective," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 8, 2022.

[67] M. E. Alqaysi, A. S. Albahri, and R. A. Hamid, "Hybrid Diagnosis Models for Autism Patients Based on Medical and Sociodemographic Features Using Machine Learning and Multicriteria Decision-Making (MCDM) Techniques: An Evaluation and Benchmarking Framework," Comput. Math. Methods Med., vol. 2022, no. ii, 2022, doi: 10.1155/2022/9410222.

[68] A. S. Albahri et al., "Towards physician's experience: Development of machine learning model for the diagnosis of autism spectrum disorders based on complex T-spherical fuzzy-weighted zero-inconsistency method," Comput. Intell., no. November, pp. 1–33, 2022, doi: 10.1111/coin.12562.

[69] G. Albahri, A.S., Joudar, S.S., Hamid, R.A., Zahid, I.A., Alqaysi, M.E., Albahri, O.S., Alamoodi, A.H., Kou and I. . Sharaf, "Explainable Artificial Intelligence Multimodal of Autism Triage Levels Using Fuzzy Approach-Based Multi-criteria Decision-Making and LIME," Int. J. Fuzzy Syst., vol. 26, no. 1, pp. 274–303, 2024, [Online]. Available: https://en.wikipedia.org/wiki/Explainable_artificial_intelligence

[70] M. M. Jassim and M. M. Jaber, "Hybrid Selection Framework For Class Balancing Approaches Based On Integrated CNN AND Decision Making Techniques for Lung Cancer Diagnosis," Eastern-European J. Enterp. Technol., vol. 118, no. 9, 2022, doi: 10.15587/1729-4061.2022.263644.

[71] A. S. Albahri et al., "A Trustworthy and Explainable Framework for Benchmarking Hybrid Deep Learning Models Based on Chest X-Ray Analysis in CAD Systems," Int. J. Inf. Technol. Decis. Mak., 2024.

[72] N. S. Baqer et al., "Indoor air quality pollutants predicting approach using unified labelling process-based multi-criteria decision making and machine learning techniques," Telecommun. Syst., vol. 81, no. 4, pp. 591–613, 2022, doi: 10.1007/s11235-022-00959-2.

[73] M. A. Alsalem et al., "Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach," Expert Syst. Appl., vol. 246, no. September 2023, p. 123066, 2024, doi: 10.1016/j.eswa.2023.123066.

[74] M. Al-Samarraay et al., "An integrated fuzzy multi-measurement decision-making model for selecting optimization techniques of semiconductor materials," Expert Syst. Appl., vol. 237, no. PB, p. 121439, 2024, doi: 10.1016/j.eswa.2023.121439.

[75] M. A. Alsalem et al., "Based on T-spherical fuzzy environment: A combination of FWZIC and FDOSM for prioritising COVID-19 vaccine dose recipients," J. Infect. Public Health, vol. 14, no. 10, pp. 1513–1559, 2021, doi: 10.1016/j.jiph.2021.08.026.

[76] A. S. Albahri et al., "Integration of fuzzy-weighted zero-inconsistency and fuzzy decision by opinion score methods under a q-rung orthopair environment: A distribution case study of COVID-19 vaccine doses," Comput. Stand. Interfaces, vol. 80, no. March 2021, p. 103572, 2022, doi: 10.1016/j.csi.2021.103572.

[77] O. S. Albahri et al., "Novel dynamic fuzzy Decision-Making framework for COVID-19 vaccine dose recipients," J. Adv. Res., vol. 37, pp. 147–168, 2022, doi: 10.1016/j.jare.2021.08.009.

[78] H. A. Alsattar et al., "Developing deep transfer and machine learning models of chest X-ray for diagnosing COVID-19 cases using probabilistic single-valued Neutrosophic hesitant fuzzy," Expert Syst. Appl., vol. 236, no. August 2023, p. 121300, 2024, doi: 10.1016/j.eswa.2023.121300.

[79] A. Alsereidi et al., "Novel Federated Decision Making for Distribution of Anti-SARS-CoV-2 Monoclonal Antibody to Eligible High-Risk Patients," Int. J. Inf. Technol. Decis. Mak., 2022, doi: 10.1142/S021962202250050X.

[80] H. Mohammed Talib, A. S. Albahri, and T. O. C. EDOH, "Fuzzy Decision-Making Framework for Sensitively Prioritizing Autism Patients with Moderate Emergency Level," Appl. Data Sci. Anal., pp. 16–41, 2023, doi: 10.58496/adsa/2023/002.

[81] S. S. Joudar, A. S. Albahri, and R. A. Hamid, "Intelligent triage method for early diagnosis autism spectrum disorder (ASD) based on integrated fuzzy multi-criteria decision-making methods," Informatics Med. Unlocked, vol. 36, no. November 2022, p. 101131, 2023, doi: 10.1016/j.imu.2022.101131.

[82] A. H. Alamoodi et al., "Intelligent Emotion and Sensory Remote Prioritisation for Patients with Multiple Chronic Diseases," Sensors, vol. 23, no. 4, pp. 1–14, 2023, doi: 10.3390/s23041854.

[83] A. H. Alamoodi, O. S. Albahri, A. A. Zaidan, H. A. Alsattar, B. B. Zaidan, and A. S. Albahri, "Hospital selection framework for remote MCD patients based on fuzzy q-rung orthopair environment," Neural Comput. Appl., vol. 35, no. 8, pp. 6185–6196, 2022, doi: 10.1007/s00521-022-07998-5.

[84] M. S. Mahmoud, "Weighting Heart Disease Criteria Using Multi- Criteria Decision-Making," 2023 Al-Sadiq Int. Conf. Commun. Inf. Technol., vol. 2023, pp. 342–346, 2023, doi: 10.1109/AICCIT57614.2023.10218189.

[85] Z. K. Mohammed et al., "Bitcoin network-based anonymity and privacy model for metaverse implementation in Industry 5.0 using linear Diophantine fuzzy sets," Ann. Oper. Res., 2023, doi: 10.1007/s10479-023-05421-3.

[86] G. Bilquise, K. Shaalan, and M. Alkhatib, "Evaluation of Virtual Commerce Applications for the Metaverse Using Spherical Linear Diophantine-Based Modeling Approach," Hum. Behav. Emerg. Technol., vol. 2024, no. 1, 2024, doi: 10.1155/2024/4571959.

[87] A. . Zaidan et al., "Uncertainty Decision Modeling Approach for Control Engineering Tools to Support Industrial," IEEE Syst. J., pp. 1–12, 2023.

[88] N. Mourad et al., "Decisioning-Based Approach for Optimising Control Engineering Tools Using Digital Twin Capabilities and Other Cyber-physical Metaverse Manufacturing System Components," IEEE Trans. Consum. Electron., vol. 70, no. 1, pp. 3212–3221, 2023, doi: 10.1109/TCE.2023.3326047.

[89] N. A. Husin et al., "Evaluation of Metaverse Tools Based on Privacy Model Using Fuzzy MCDM Approach," in International Multi-Disciplinary Conference-Integrated Sciences and Technologies, Springer, 2023, pp. 1–20.

[90] M. Talal, A. H. Alamoodi, O. S. Albahri, A. S. Albahri, and D. Pamucar, Evaluation of remote sensing techniques-based water quality monitoring for sustainable hydrological applications: an integrated FWZIC-VIKOR modelling approach, no. 0123456789. Springer Netherlands, 2023. doi: 10.1007/s10668-023-03432-5.

[91] Q. Sarah, H. A. Alsattar, A. A. Zaidan, M. Deveci, D. Pamucar, and W. Ding, "A novel fuel supply system modelling approach for electric vehicles under Pythagorean probabilistic hesitant fuzzy sets," Inf. Sci. (Ny)., vol. 622, pp. 1014–1032, 2023, doi: 10.1016/j.ins.2022.11.166.

[92] A. Alnoor et al., "Toward a Sustainable Transportation Industry: Oil Company Benchmarking based on the Extension of Linear Diophantine Fuzzy Rough Sets and Multicriteria Decision-Making Methods," IEEE Trans. Fuzzy Syst., vol. 31, no. 2, pp. 449–459, 2022, doi: 10.1109/TFUZZ.2022.3182778.

[93] S. Qahtan, H. A. Alsattar, A. A. Zaidan, M. Deveci, D. Pamucar, and D. Delen, "Performance assessment of sustainable transportation in the shipping industry using a q-rung orthopair fuzzy rough sets-based decision making methodology," Expert Syst. Appl., vol. 223, no. March, p. 119958, 2023, doi: 10.1016/j.eswa.2023.119958.

[94] S. Al-Humairi et al., "Towards Sustainable Transportation: A Pavement Strategy Selection Based on the Extension of Dual-Hesitant Fuzzy Multi-Criteria Decision-Making Methods," IEEE Trans. Fuzzy Syst., vol. 31, no. 2, pp. 380–393, 2022, doi: 10.1109/TFUZZ.2022.3168050.

[95] M. Taha Aljburi et al., "Exploring decision-making techniques for evaluation and benchmarking of energy system integration frameworks for achieving a sustainable energy future," Energy Strateg. Rev., vol. 51, no. June 2023, 2024, doi: 10.1016/j.esr.2023.101251.

[96] H. Ghailani et al., "Developing sustainable management strategies in construction and demolition wastes using a q-rung orthopair probabilistic hesitant fuzzy set-based decision modelling approach," Appl. Soft Comput., vol. 145, p. 110606, 2023, doi: 10.1016/j.asoc.2023.110606.

[97] A. Alaa, M. Deveci, and Z. Alsattar, Hassan A. Qahtan, Sarah Shang, Long. Delen, Dursun. Mourad, Nahia. Khalid, "Neutrosophic bipolar fuzzy decision-based approach for developing sustainable circular business model innovation tools," Comput. Ind. Eng., vol. 189, no. February, 2024.

[98] H. Abdulsattar, A. A. Zaidan, S. Qahtan, and B. B. Zaidan, "Sustainability assessment of palm oil industry 4.0 technologies in a circular economy applications based on interval-valued Pythagorean fuzzy rough set-FWZIC and EDAS methods," Appl. Soft Comput., vol. 136, p. 110073, 2023, doi: 10.1016/j.asoc.2023.110073.

[99] H. A. Alsattar et al., "Three-way decision-based conditional probabilities by opinion scores and Bayesian rules in circular-Pythagorean fuzzy sets for developing sustainable smart living framework," Inf. Sci. (Ny)., vol. 649, no. July, p. 119681, 2023, doi: 10.1016/j.ins.2023.119681.

[100] H. A. Alsattar et al., "Developing IoT sustainable real-time monitoring devices for food supply chain systems based on climate change using circular intuitionistic fuzzy set," IEEE Internet Things J., 2023.

[101] E. Krishnan et al., "Interval type 2 trapezoidal-fuzzy weighted with zero inconsistency combined with VIKOR for evaluating smart e-tourism applications," Int. J. Intell. Syst., vol. 36, no. 9, pp. 4723–4774, 2021, doi: 10.1002/int.22489.

[102] A. H. Alamoodi et al., "Based on neutrosophic fuzzy environment: a new development of FWZIC and FDOSM for benchmarking smart e-tourism applications," Complex Intell. Syst., vol. 8, no. 4, pp. 3479–3503, 2022, doi: 10.1007/s40747-022-00689-7.

[103] R. T. Mohammed et al., "A decision modeling approach for smart e-tourism data management applications based on spherical fuzzy rough environment," Appl. Soft Comput., vol. 143, p. 110297, 2023.

[104] A. H. Alamoodi et al., "New Extension of Fuzzy-Weighted Zero-Inconsistency and Fuzzy Decision by Opinion Score Method Based on Cubic Pythagorean Fuzzy Environment: A Benchmarking Case Study of Sign Language Recognition Systems," Int. J. Fuzzy Syst., vol. 24, no. 4, pp. 1909–1926, 2022, doi: 10.1007/s40815-021-01246-z.

[105] S. Qahtan, N. Mourad, H. A. Ibrahim, A. A. Zaidan, B. Bahaa, and W. Ding, "Robust three-way decisions based on ensembled multi-divergence measures with circular quintic fuzzy sets for developing swarm robots in mechanised agricultural operations," Expert Syst. Appl., vol. 266, no. June 2024, p. 126102, 2025, doi: 10.1016/j.eswa.2024.126102.

[106] S. Qahtan et al., "Novel Multi Security and Privacy Benchmarking Framework for Blockchain-Based IoT Healthcare Industry 4.0 Systems," IEEE Trans. Ind. Informatics, vol. 18, no. 9, pp. 6415–6423, 2022, doi: 10.1109/TII.2022.3143619.

[107] A. A. Zaidan, H. A. Alsattar, S. Qahtan, M. Deveci, D. Pamucar, and B. B. Gupta, "Secure Decision Approach for Internet of Healthcare Things Smart Systems-Based Blockchain," IEEE Internet Things J., vol. PP, p. 1, 2023, doi: 10.1109/JIOT.2023.3308953.

[108] O. S. Albahri et al., "Rough Fermatean fuzzy decision-based approach for modelling IDS classifiers in the federated learning of IoMT applications," Neural Comput. Appl., vol. 35, no. 30, pp. 22531–22549, 2023, doi: 10.1007/s00521-023-08933-y.

[109] G. Bilquise and S. Ibrahim, "Integrating ARAS with PyFWZIC to evaluate and benchmark patient-facing genetic services digital tools," Neural Comput. Appl., pp. 1–22, 2024.

[110] J. Li, C. Bi, F. Gao, and W. He, "An integrated linguistic Pythagorean fuzzy decision-making approach for risk analysis of offshore wind turbine," Ocean Eng., vol. 291, no. November 2023, p. 116450, 2024, doi: 10.1016/j.oceaneng.2023.116450.

[111] O. S. Albahri et al., "Multi-perspective evaluation of integrated active cooling systems using fuzzy decision making model," Energy Policy, vol. 182, no. August, p. 113775, 2023, doi: 10.1016/j.enpol.2023.113775.

[112] Y. R. Muhsen, N. A. Husin, M. B. Zolkepli, N. Manshor, and A. A. J. Al-hchaimi, "Evaluation of the Routing Algorithms for NoC-Based MPSoC : A Fuzzy Multi-Criteria Decision-Making Approach," IEEE Access, vol. 11, no. August, pp. 102806–102827, 2023, doi: 10.1109/ACCESS.2023.3310246.

[113] L. Xiao, T. Fang, G. Huang, and M. Deveci, "An integrated design concept evaluation method based on fuzzy weighted zero inconsistency and combined compromise solution considering inherent uncertainties," Adv. Eng. Informatics, vol. 65, no. PA, p. 103097, 2025, doi: 10.1016/j.aei.2024.103097.

[114] Z. T. Al-qaysi1, A. S. Albahri, M. A. Ahmed, and M. M. Salih, "Dynamic decision-making framework for benchmarking brain – computer interface applications : a fuzzy-weighted zero-inconsistency method for consistent weights and VIKOR for stable ran ... Dynamic decision-making framework for benchmarking brain – computer," Neural Comput. Appl., no. March, 2024, doi: 10.1007/s00521-024-09605-1.

[115] S. Qahtan, A. Alaa Zaidan, H. Abdulsattar Ibrahim, M. Deveci, W. Ding, and D. Pamucar, "A decision modeling approach for smart training environment with motor Imagery-based brain computer interface under Neutrosophic cubic fuzzy set," Expert Syst. Appl., vol. 224, no. January, p. 119991, 2023, doi: 10.1016/j.eswa.2023.119991.

[116] I. Mohamad Sharaf et al., "Architecture selection for 5G-radio access network using type-2 Neutrosophic numbers based decision making model," Expert Syst. Appl., vol. 237, no. PA, p. 121420, 2024, doi: 10.1016/j.eswa.2023.121420.

[117] S. Qahtan et al., "Normal wiggly hesitant fuzzy modelling approach for 6G frameworks based blockchain technology," Expert Syst. Appl., vol. 259, no. March 2024, p. 125304, 2025, doi: 10.1016/j.eswa.2024.125304.

[118] N. Mourad et al., "Dynamic localization based-utility decision approach under type-2 Pythagorean fuzzy set for developing internet of modular self-reconfiguration robot things," Eng. Appl. Artif. Intell., vol. 139, no. PB, p. 109671, 2025, doi: 10.1016/j.engappai.2024.109671.

[119] M. T. Mohammed et al., "Determining the superiority of a robust cloud fault tolerance mechanism using a spherical cubic fuzzy set-based decision approach," Eng. Appl. Artif. Intell., vol. 148, no. March, p. 110402, 2025, doi: 10.1016/J.ENGAPPAI.2025.110402.

[120] N. Basil, M. E. Alqaysi, M. Deveci, A. S. Albahri, O. S. Albahri, and A. H. Alamoodi, "Evaluation of autonomous underwater vehicle motion trajectory optimization algorithms," Knowledge-Based Syst., vol. 276, p. 110722, 2023, doi: 10.1016/j.knosys.2023.110722.

[121] U. S. Mahmoud et al., "DAS benchmarking methodology based on FWZIC II and FDOSM II to support industrial community characteristics in the design and implementation of advanced driver assistance systems in vehicles," J. Ambient Intell. Humaniz. Comput., pp. 1–28, 2022.

[122] O. S. Albahri et al., "Evaluation of organizational culture in companies for fostering a digital innovation using q-rung picture fuzzy based decision-making model," Adv. Eng. Informatics, vol. 58, no. August, p. 102191, 2023, doi: 10.1016/j.aei.2023.102191.

[123] A. N. Jasim, L. C. Fourati, and O. S. Albahri, "Evaluation of Unmanned Aerial Vehicles for Precision Agriculture Based on Integrated Fuzzy Decision-Making Approach," IEEE Access, vol. 11, no. July, pp. 75037–75062, 2023, doi: 10.1109/ACCESS.2023.3294094.

[124] M. Talal, M. L. P. Tan, D. Pamucar, D. Delen, W. Pedrycz, and V. Simic, "Evaluation and benchmarking of research-based microgrid systems using FWZIC-VIKOR approach for sustainable energy management," Appl. Soft Comput., vol. 166, no. August, p. 112132, 2024, doi: 10.1016/j.asoc.2024.112132.

[125] A. H. Alamoodi, M. S. Al-Samarraay, O. S. Albahri, M. Deveci, A. S. Albahri, and S. Yussof, "Evaluation of energy economic optimization models using multi-criteria decision-making approach," Expert Syst. Appl., vol. 255, no. PD, p. 124842, 2024, doi: 10.1016/j.eswa.2024.124842.

[126] O. S. Albahri et al., "Selection of smartphone-based mobile applications for obesity management using an interval Neutrosophic vague decision-making framework," Eng. Appl. Artif. Intell., vol. 137, no. PB, p. 109191, 2024, doi: 10.1016/j.engappai.2024.109191.

[127]J. Mao, J. G. Huang, J. Liu, C. Peng, and S. Z. Zhang, "Power equipment supplier evaluation under a q-rung orthopair fuzzy set based decision making model," Heliyon, vol. 10, no. 22, p. e40390, 2024, doi: 10.1016/j.heliyon.2024.e40390.

[128]A. S. Albahri et al., "Prioritizing complex health levels beyond autism triage using fuzzy multi-criteria decision-making," Complex Intell. Syst., vol. 10, no. 5, pp. 6159–6188, 2024, doi: 10.1007/s40747-024-01432-0.

[129]N. Mourad et al., "Decisioning-Based Approach for Optimising Control Engineering Tools Using Digital Twin Capabilities and Other Cyber-Physical Metaverse Manufacturing System Components," IEEE Trans. Consum. Electron., vol. 70, no. 1, pp. 3212–3221, 2024, doi: 10.1109/TCE.2023.3326047.

[130]A. Alamoodi et al., "Evaluating agriculture 4.0 decision support systems based on hyperbolic fuzzy-weighted zero-inconsistency combined with combinative distance-based assessment," Comput. Electron. Agric., vol. 227, no. March, 2024, doi: 10.1016/j.compag.2024.109618.

[131]Q. Mao, Y. Gao, and J. Fan, "An investment decision framework for offshore CCUS project under interval-valued fermatean fuzzy environment," Environ. Technol. (United Kingdom), vol. 46, no. 7, pp. 1112–1137, 2024, doi: 10.1080/09593330.2024.2376291.

[132]H. A. Ibrahim et al., "Sustainability in mobility for autonomous vehicles over smart city evaluation; using interval-valued fermatean fuzzy rough set-based decision-making model," Eng. Appl. Artif. Intell., vol. 129, no. November 2023, p. 107609, 2024, doi: 10.1016/j.engappai.2023.107609.

[133]U. F. Simo and H. Gwét, "Fuzzy Triangular Aggregation Operators," Int. J. Math. Math. Sci., vol. 2018, no. 2010, 2018, doi: 10.1155/2018/9209524.

[134]R. R. Yager, "On generalized Bonferroni mean operators for multi-criteria aggregation," Int. J. Approx. Reason., vol. 50, no. 8, pp. 1279–1286, 2009, doi: 10.1016/j.ijar.2009.06.004.

[135]S. Khan, S. Abdullah, and S. Ashraf, "Picture fuzzy aggregation information based on Einstein operations and their application in decision making," Math. Sci., vol. 13, pp. 213–229, 2019.

[136]T.-Y. Chen, "A comparative analysis of score functions for multiple criteria decision making in intuitionistic fuzzy settings," Inf. Sci. (Ny)., vol. 181, no. 17, pp. 3652–3676, 2011.

[137]H. M. Talib, A. S. Albahri, and O. C. Thierry, "Fuzzy decision-making framework for sensitively prioritizing autism patients with moderate emergency level," Appl. Data Sci. Anal., vol. 2023, pp. 16–41, 2023.

[138]A. S. Albahri et al., "Explainable Artificial Intelligence Multimodal of Autism Triage Levels Using Fuzzy Approach-Based Multi-criteria Decision-Making and LIME," Int. J. Fuzzy Syst., vol. 26, no. 1, pp. 274–303, 2024.

[139]S. Qahtan, K. Yatim, H. Zulzalil, M. H. Osman, A. A. Zaidan, and H. A. Alsattar, "Review of healthcare industry 4.0 application-based blockchain in terms of security and privacy development attributes: Comprehensive taxonomy, open issues and challenges and recommended solution," J. Netw. Comput. Appl., vol. 209, no. October 2022, 2023, doi: 10.1016/j.jnca.2022.103529.

[140]D. Pamučar, Ž. Stević, and S. Sremac, "A new model for determining weight coefficients of criteria in mcdm models: Full consistency method (fucom)," Symmetry (Basel)., vol. 10, no. 9, p. 393, 2018.

[141]A. D. Ahmed, M. M. Salih, and Y. R. Muhsen, "Opinion Weight Criteria Method (OWCM): A New Method for Weighting Criteria with Zero Inconsistency," IEEE Access, 2024.

[142]M. E. Alqaysi, A. S. Albahri, and R. A. Hamid, Evaluation and benchmarking of hybrid machine learning models for autism spectrum disorder diagnosis using a 2-tuple linguistic Neutrosophic fuzzy sets-based decision-making model, no. July. Springer London, 2024. doi: 10.1007/s00521-024-09905-6.

[143]H. ali khudhyer alhadad et al., "Intelligent approach for developing a blood product supply chain to mitigate shortages and Preclude wastage," Eng. Appl. Artif. Intell., vol. 142, no. December 2024, p. 109877, 2025, doi: 10.1016/j.engappai.2024.109877.

[144]D. Abu-Lail et al., "Evaluation of industry 4.0 adoption strategies in small and medium enterprises: A Circular-Fermatean fuzzy decision-making approach," Appl. Soft Comput., vol. 169, no. June 2024, p. 112618, 2025, doi: 10.1016/j.asoc.2024.112618.

APPENDIX

Themes	Domain	Study	Objective	MCDM Method	FWZIC Extension	#Experts	#Criteria	Criteria Source
Artificial Intelligence	Engineering	[74]	To develop optimization techniques for semiconductor materials.	FDOSM	Triangular Fuzzy Numbers	3	8	Root article
	Healthcare	[142]	To evaluate hybrid diagnosis models for autism spectrum disorder patients	FDOSM	2-Tuple Linguistic Neutrosophic Fuzzy Set	4	9	Root article
		[128]	To prioritize autism spectrum disorder cases based on behavioral and medical characteristics.	MABAC	Single-Valued Neutrosophic 2-Tuple Linguistic Fuzzy Set	3	19	Literature
		[67]	To evaluate machine learning models for diagnosing autism spectrum disorder patients.	FDOSM	Triangular Fuzzy Numbers	4 3	48 7	Performance metric of ML
		[70]	To evaluate machine learning approaches for lung cancer diagnosis.	VIKOR	Triangular Fuzzy numbers	6	5	Performance metric of ML
		[72]	To develop a labelled dataset of indoor hospital air quality pollutants.	UPLDP	Interval Type 2 Trapezoidal-Fuzzy Sets	6	9	Root article
		[68]	To integrate MCDM with Machine learning for classifying patients with autism spectrum disorder.	ML	Complex T-Spherical Fuzzy Set	3	7	Performance metric of ML
		[73]	To evaluate AI healthcare applications based on trustworthiness attributes.	CODAS	q-Rung Orthopair Fuzzy 2-Tuple Linguistic	3	7	Root article
		[71]	To benchmark deep learning models for chest X-ray analysis and computer-aided diagnosis.	VIKOR	Triangular Fuzzy Numbers	6	7	Performance metric of ML

		[69]	To combine MCDM and ML to predict ASD patients and build an explainable AI for model comprehension.	Delphi	Triangular Fuzzy numbers	4	19	Existing dataset
COVID 19	Healthcare	[78]	To benchmark AI models for diagnosing COVID-19 patients using chest x-rays.	DLBD	Probabilistic Single-Valued Neutrosophic Hesitant Fuzzy Set	3	7	Literature
		[75]	To prioritize COVID-19 vaccine dose recipients.	FDOSM	T-Spherical Fuzzy Sets	3	5	Root article
		[76]	To prioritize COVID-19 vaccine dose recipients.	FDOSM	Q-Rung Orthopair Fuzzy Sets	3	5	Root article
		[77]	To prioritize COVID-19 vaccine dose recipients.	FDOSM	Pythagorean Fuzzy Sets	3	5	Root article - case study
		[79]	the prioritize the distribution of COVID-19 monoclonal antibody to eligible high-risk patients.	TOPSIS	Interval-Valued Spherical Fuzzy and Hesitant 2-Tuple	3	15	Literature
Metaverse	Business	[85]	To evaluate bitcoin network platforms for the Metaverse.	MULTIMOORA	Linear Diophantine Fuzzy Sets	5	24	Root article (partial)
		[86]	To evaluate virtual commerce solutions for the Metaverse	RATMI	Spherical Linear Diophantine Sets	5	13	Root article
	Manufacturing	[89]	To assess cyber-physical manufacturing systems based on privacy features.	ARAS	Hexagonal Fuzzy Numbers	5	21	Unspecified
		[87]	To assess cyber-physical manufacturing systems based on privacy features.	PROMETHEE II	Interval-Valued Spherical Fuzzy Rough Sets	3	26	Root article
		[88]	To assess cyber-physical manufacturing systems based on privacy features.	CoCoSo	Interval-Valued Spherical Fuzzy Rough Sets	3	16	Root article
Service	Business	[127]	To rank power transformer suppliers based on quality standards to maintain safe and stable grid operations.	MABAC	Q-Rung Orthopair Fuzzy Sets	4	10	Literature
		[143]	to assess blood supply chain management alternatives against sustainability-focused criteria	VIKOR	Single-Valued Neutrosophic Fuzzy Set	3	7	Root article
	Healthcare	[80]	To prioritize the treatment of autism spectrum disorder patients.	VIKOR	Triangular Fuzzy Numbers	4	19	Root article
		[83]	To prioritize hospital selection for patients with multi-chronic diseases.	FDOSM	Q-Rung Orthopair Fuzzy Sets	3	7	Root article
		[82]	To prioritize patients with multiple chronic illnesses based on sensor and emotion data.	TROOIL	Fractional Orthotriple Fuzzy Sets	3	12	Root article
		[81]	To classify autism spectrum disorder patients based on severity level for early diagnosis.	Delphi	Triangular Fuzzy Numbers	13	19	Existing dataset
		[84]	To prioritize the criteria for managing patients with cardiac related issues.	N/A	Triangular Fuzzy Numbers	4	11	Root article
Sustainability	Agriculture	[130]	To critically assess decision support systems for enabling the choice of systems that can effectively drive sustainable smart agriculture.	CODAS	Hyperbolic Fuzzy Sets	3	8	Root article
		[90]	To evaluate water quality monitoring systems based on remote sensing techniques.	VIKOR	Triangular Fuzzy Numbers	3	15	Root article
	Business	[98]	To evaluate circular economy and sustainability practices in the palm oil industry.	EDAS	Interval-valued Pythagorean Fuzzy Rough Set	3	14	Root article

		[97]	To evaluate sustainable circular business model innovation tools.	CODAS	Neutrosophic Bipolar Fuzzy Sets	3	10	Root article
	Engineering	[96]	To evaluate sustainable strategies for construction waste management.	FDOSM-MULTIMOORA	Q-Rung Orthopair Probabilistic Hesitant Fuzzy Set	3	26	Root article
		[131]	To develop a comprehensive investment decision-making framework for offshore Carbon Capture, Utilization, and Storage (CCUS) projects to promote sustainable development	MARCOS	Interval value, Fermatean Fuzzy Sets	3	18	Literature
		[124]	To assess and rank research based micro-grid to ensure sustainable energy resources.	VIKOR	Triangular Fuzzy Numbers	3	9	Root article
		[95]	To benchmark energy systems integration frameworks for efficient resource utilization.	MABAC	Triangular Fuzzy Numbers	12	6	Root article
	Miscellaneous-Energy Optimization	[125]	To evaluate energy economy optimization models for efficient energy systems.	FDOSM	Triangular Fuzzy Numbers	3	5	Root article
	Miscellaneous-Smart living	[99]	To evaluate sustainable smart living frameworks based on multi-layer assessment factors.	CPOS	Circular-Pythagorean Fuzzy Sets	3	19	Root article
	Supply Chain	[100]	To rank real-time IoT devices for food waste monitoring in a supply chain system.	ARAS	Circular Intuitionistic Fuzzy Set	3	10	Root article
		[50]	To benchmark sustainable agriculture food supply chain strategies.	FDOSM-MAIRCA	Fermatean Probabilistic Hesitant Fuzzy	3	27	Root article (partial)
	Transport	[48]	To evaluate sustainable electronic passenger modelling strategies.	MULTIMOORA	Probabilistic Hesitant Fuzzy Sets	3	49	Root article
		[91]	To assess fuel supply modelling approaches for electric vehicles.	MARCOS	Pythagorean Probabilistic Hesitant	3	28	Root article (partial)
		[93]	To evaluate energy systems for sustainable transportation in the shipping industry.	FDOSM	q-rung orthopair fuzzy Rough Sets	3	10	Root article
		[92]	To evaluate and rank international oil companies for sustainable oil transportation.	MULTIMOORA	Linear Diophantine Fuzzy Rough Sets	3	11	Root article
		[94]	To assess pavement selection approaches for sustainable transportation.	FDOSM	Dual-Hesitant Fuzzy	4	30	Literature
Technology	Agriculture	[123]	to evaluate drones based on performance attributes for precision farming.	FDOSM	Triangular Fuzzy Numbers	3	3	Literature
		[105]	To evaluate and measure the effectiveness of swarm robots in mechanized agricultural operations	CPOS	Circular Quintic Fuzzy Sets	3	7	Literature
	Business	[122]	To evaluate organizational culture based on features that promote digital innovation.	SAW	Q-Rung Picture Fuzzy	3	9	Root article
	Education	[115]	To evaluate smart training environments based on brain computer interface attributes.	MABAC	Neutrosophic Cubic Sets	3	10	Root article
		[23]	To evaluate speech-recognition chatbots for language learning	MCRAT	T-Spherical Fuzzy Sets	5	9	Root article
		[114]	To evaluate smart training environments based on brain computer interface attributes.	VIKOR	Triangular Fuzzy Numbers	3	10	Root article
	Engineering	[112]	To evaluate routing algorithms in a multiprocessor network system.	FDOSM	Z-Cloud Rough Numbers	3	9	Literature

		[129]	To Optimize Control Engineering Tools Using Digital Twin Capabilities and Other Cyber-Physical Metaverse Manufacturing System attributes	CoCoSo	Interval Value Spherical Fuzzy Rough Sets	3	26	Root Article
		[113]	To develop a design concept evaluation method to compromise solutions of uncertainties	CoCoSo	Q Rung Orthopair Fuzzy Rough Sets	5	12	Root Article
		[111]	To evaluate cooling systems based on performance and technical perspective.	MABAC	2-Tuple Linguistic Pythagorean	3	11	Literature
		[110]	To assess the risk factors associated with offshore wind turbines.	DEMATEL	Linguistic Pythagorean Fuzzy	4	75	Literature
	Healthcare	[104]	To evaluate sign language recognition systems.	FDOSM	Cubic Pythagorean Fuzzy Sets	3	11	Root article - case study
		[126]	To develop a framework of smartphone-based mobile applications for obesity management	FDOSM	Triangular Fuzzy sets	3	5	Root Article
		[106]	To benchmark blockchain based healthcare systems using security and privacy features.	TOPSIS	Spherical Fuzzy Sets	17	6	Root Article
		[108]	To evaluate intrusion detection classifiers for Internet of Medical Things (IoMT) devices based on security and performance attributes.	FDOSM	Rough Fermatean Fuzzy Sets	3	17	Root Article
		[109]	To benchmark digital tools for genetic counselling	ARAS	Pythagorean Fuzzy Set	3	9	Root Article
		[107]	To benchmark blockchain based healthcare systems using security and privacy features.	COPRAS	Interval-Valued Spherical Fuzzy Sets	3	6 and 7	Root Article
	Misc-Marine Robotics	[120]	To evaluate optimization algorithms for motion of autonomous underwater vehicles.	FDOSM	Triangular Fuzzy Numbers	3	9	Literature
	Network	[117]	To develop a modelling framework for 6G based blockchain technology	RAFSI	Normal Wiggly Hesitant Fuzzy Set	3	7	Literature
	Network	[118]	To evaluate a Dynamic locally based-utility approach for developing internet of Modular self-reconfiguration robot things	Dynamic Localizatio n-Based Utility	Type 2 Pythagorean	3	6	Root Article
	Network(Cloud computing)	[119]	To determine the superiority of a robust cloud fault tolerance of a solution mechanism	PROMETH EE	Spherical Cubic Fuzzy Sets	3	9	Root Article
	Network (not sure why??)	[144]	To evaluate of industry 4.0 adoption strategies in small and medium enterprises	TOE	Circular-Fermatean Fuzzy sets	3	30	Root Artricle
	Network	[116]	To select an optimal architecture for 5G Radio Access Networks.	FDOSM	Type-2 Neutrosophic Numbers	3	25	Root Article
	Tourism	[101]	To evaluate smart e-tourism applications.	VIKOR	Interval Type 2 Trapezoidal-Fuzzy Sets	11	12	Root Article
		[102]	To evaluate smart e-tourism applications.	FDOSM	Neutrosophic Fuzzy Sets	11	12	Root Article - case study
		[103]	To evaluate smart e-tourism applications.	FDOSM	Spherical Fuzzy Rough Sets	3	12	Root Article - case study
	Transport	[121]	To evaluate data acquisition systems for the vehicle development and analysis.	FDOSM	Intuitionistic Fuzzy Set	3	15	Root Article

Design and Implementation of Low-Cost Hybrid-Controlled Smart Wheelchair Based on PID Control Integrated with Vital Signs Monitoring

M. Sayed[1], MG Mousa[2], Ali A. S.[3], and T. Mansour[4]
Mechanical Engineering Department, Suez Canal University, Ismailia, Egypt[1, 3, 4]
Mechanical Engineering Department, Damietta University, Damietta, Egypt[2]
Maintenance Department, Methode Electronics, Cairo, Egypt[1]

***Abstract*—According to international organizations' statistics, the percentage of disabled people is considered not just a small percentage of the world's population. Improving the quality of life for people by using new technologies is one of the essential topics today. Although the wheelchair is the most common way of mobility for the disabled, this research aims to improve wheelchair use by creating other control methods, including speech recognition commands. A joystick, wireless remote control, and an additional port are available in addition to speech recognition commands to control the wheelchair. Researchers also studied and compared the effects of integrating a Proportional–Integral–Derivative (PID) controller with the normal controller during smart wheelchair operation in various normal usage scenarios. This research provides data based on a real experiment, not like most research that depends on mathematical models only for comparison. Adding the PID controller eliminated the overshoot of the smart wheelchair, reduced steady-state error, and reduced settling time. Furthermore, it contains a healthcare monitoring system to track the user's vital signs and object avoidance sensors to keep them safe. Also, a full motor selection calculation for the smart wheelchair has been provided, which is useful for mobile robot design. Additionally, the smart wheelchair features a power monitoring system. Finally, a voice-controlled wheelchair helps the user feel more private and independent. Using this kind of smart wheelchair and living this lifestyle will increase the user's morale.**

***Keywords*—Smart wheelchair; speech recognition; PID; healthcare; mobile robot**

I. Introduction

The global population is experiencing rapid growth. The prevalence of physical disability is increasing due to ageing, accidents, and medical conditions such as quadriplegia and paralysis [1, 2]. Japanese demographic forecasts indicate that the percentage of adults aged 65 and above will attain 40% by 2050 [3]. Individuals may lose a limb or both legs due to sickness or accidents, resulting in incapacitation and dependence on others for daily necessities. They are unable to fulfill their responsibilities independently [2]. Contemporary civilization is characterized by individuals who are predominantly focused on their own lives, with an emphasis on competence and self-sufficiency. Individuals reliant on others are hardly discussed in this rapid era of materialism and modernization. In our society, it is hardly contemplated the requirements of individuals who are dependent. If all individuals could lead their lives autonomously, the world would be an exceptional place to inhabit [4].

Mobility impairment is a prevalent form of functional disability that hinders an individual's capacity for movement, frequently necessitating the use of walkers or wheelchairs. World Health Organization (WHO) research revealed that 75 million individuals experience mobility issues and necessitate the use of a wheelchair. Although wheelchairs enable individuals with limited mobility to navigate freely, users frequently experience degeneration, discomfort, or ulceration owing to extended usage and the physical characteristics of their surroundings [5]. New technology is introduced to the marketplace daily. The primary objective of technology is to simplify human life and reduce reliance on others [2]. Powered Wheelchairs (PWs) can accommodate the needs of individuals with mobility impairments [4].

The issue of population ageing escalates the need for healthcare services for the elderly and disabled, especially in nursing facilities or domestic settings [6]. The evaluation of a patient's condition is often conducted through observation and interpretation by qualified medical personnel. From a scientific research perspective, it is crucial to address the issues impacting the physical state of wheelchair users, providing technology and creative solutions to healthcare providers to optimize their treatments. Consequently, it is essential to augment conventional techniques with more objective and quantitative evaluation approaches, such as posture monitoring [5].

Individuals with cognitive, motor, and sensory impairments need Powered Wheelchair (PW) for mobility, whether because of disability or illness [7]. Additionally, several individuals suffer from motor paralysis that necessitates prolonged bed rest due to conditions such as Amyotrophic Lateral Sclerosis (ALS) or Spinal Cord Injury (SCI) [8]. They progressively diminish in muscle strength and often encounter difficulties in reaching and grabbing, rendering the driving device, such as a joystick, challenging to operate [9].

A Smart Wheelchair (SW) typically consists of either a standard PW base integrated with a computer and many sensors or a mobile robot base equipped with a seat [7]. Besides joystick control, many methods of controlling a PW employ entirely distinct interfaces [10]. The electrical impulses generated by our bodies enable connections to several applications in Human-

Machine Interface (HMI) and rehabilitation. Moreover, cognitive functions have been implemented to enhance their capabilities, rendering them beneficial for the daily lives of individuals with impairments [11]. One of the most challenging elements of developing a non-manual HMI for wheelchair operation is generating sufficient instructions that encompass left and right turns, forward and backward movements, acceleration, deceleration, and stopping [12].

Speech Recognition (SR) has been intensively explored since the 1950s, but recent improvements in computer and telecommunications technology have substantially improved its capabilities [13]. SR solves problems, enhances productivity, and changes the way we live. Voice control accomplishes almost all of the functionality of a pushbutton. An SR is a demanding undertaking since it involves a lot of procedures that need significant computational resources. The wide range of applications of automated SR systems, such as in Human–Computer Interactions (HCIs), telephony, and robotics, has powered a large scientific community during the last few decades. Automatic speech recognition is currently used in a wide range of goods and applications, including medical transcription, gaming control, contact center conversation systems, and information retrieval [14].

A SR technique is a computer tool that can analyse a human vocal signal and detect the information contained in it, translate it to text, or send orders that act on an action, in this case, turning the device on and off. To get a correct interpretation of the acquired auditory message, data from several sources of knowledge, such as syntactic, phonetic, pragmatic, lexical, acoustic, phonological, and semantic information, must be analyzed. With the introduction of Artificial Intelligence (AI) and intelligent assistants, the popularity and use of SRs have expanded. Recognizing the speaker can help to translate talk into systems that have been trained to recognize a specific person's voice, or it can be applied to authenticate or confirm a speaker's identification as part of a safety protocol [15].

People suffered the loss of their arms or legs, whether temporarily or permanently, due to illness or an accident [2]. But since speech is the most natural means of communication, you may also use your voice to drive the wheelchair. It is evident that a wheelchair with SR is necessary to assist individuals with physical limitations who struggle to control their movements, particularly when using their hands independently [16].

The current number of PW users is estimated to constitute just about fifty percent of the potential user base. The number would increase if technology were available to enable anyone unable to operate a joystick or switch array to reliably and safely control a PW [17]. As the older population expands, the need for autonomous and intelligent healthcare services is escalating [18].

II. Related Work

Vaishali Jabade implemented a control system for the PW by using the technology of voice recognition through Google Assistant, which was integrated with a monitoring system for the user's health. Vaishali used Bluetooth as a type of communication to send the voice commands from the mobile to the controller of the PW. Direct wired communication is the recommended method for sending voice commands to the controller of the PW, as it offers greater safety [19].

SATYAVIR SINGH implemented a PW controlled by voice commands and integrated with a collision avoidance system based on an ultrasonic sensor. SATYAVIR used Bluetooth communication to send the voice commands to the controller, which increased the possibility of missing some commands. while using wired communications enhances user safety [20].

Jenina R. Amoguis modified a manual wheelchair to become a PW controlled by SR integrated with an ultrasonic sensor to increase the safety of the user while using the PW. Jenina R. Amoguis did not add any joysticks to the PW. Since adding a joystick to the PW that is controlled by any other controlling method increases the number of users of the PW [21].

Fida Hussain uses voice commands and a joystick to operate a PW, which makes it user-friendly for individuals with disabilities, including those who are deaf and equipped with ultrasonic sensors. She used the normal relay module to control the motors of the PW, and then she had a fixed speed all the time. While using driving boards for motors allows for changing the speed of the motor if required [22].

Mohammad Shahrul developed an intelligent PW controlled by speech recognition using convolutional neural networks to identify the required commands for operating the PW. He did not consider any kind of safety in his model, like, for example, when there is any obstacle in front of the PW. Integrating any proximity sensors that stop the PW if there is an obstacle increases the safety and satisfaction levels of the user [23].

Arsha designed a real PW, controlled by the commands from the voice recognition module integrated with ultrasonic sensors to avoid any accidents during the operation of the PW. M. S. Arsha used the standard relay board to operate the motors of the system. The recommended method for driving the motors of the PW is to use a motor driving board, as this allows the user to adjust the speed of the motors [24].

Divya Jennifer DSouza developed a smart-sensing wheelchair to monitor heart rate and blood oxygen levels and trigger an alarm system if the user's signals are not good. Divya Jennifer DSouza used a photoplethysmography sensor to monitor the heart rate. Unlike an electrocardiogram sensor, the photoplethysmography sensor does not provide an accurate heart rate reading [25].

Shailu studied using different algorithms for improving the speed regulation of an electric wheelchair as an application of the mobile robot. Shailu optimized the PID controller gains using firefly, genetic, particle swarm, and gray wolf algorithms and compared the results [26].

Sankardoss developed an economical implementation of a directional and speed controller for an electric wheelchair. He conducted a comparison using three different types of controllers: the intelligent neural network, fuzzy logic, and PID. His trials were based on improving the overshoot and settling time [27].

Rawiphon Chotikunnan applied a PID control system on a Mecanum wheelchair to regulate the movements of the motors, which leads to improved control accuracy. He employed the

Cohen-Coon tuning approach to figure out the gain of the PID controller [28].

III. OBJECTIVES AND METHODOLOGY

This research aims to explore the full design and implementation of a real SW, which utilizes a four-wheel drive system and operates through standard joysticks, wireless remote controls, and voice commands. Additionally, the system integrates a healthcare monitoring system into the Graphical User Interface (GUI) of a Personal Computer (PC). To increase the safety level and satisfaction of the user, an obstacle avoidance system has been integrated with the SW from four directions to immediately stop the movement, if it detects an obstacle in the moving direction. Also, it measures some of the environmental parameters, like vibration and noise level. In addition, monitoring the responses of the operation of the motors of the SW in response to the commands that are received from the various methods of operation. A PID control has been included in the SW in order to control the speed at which the motors are operating.

To achieve the mentioned objective, a mechanical design of the SW has been carried out, including selecting the optimal dimensions of the wheelchair and the motor selection based on a real calculation and the availability of the market, and it will be discussed thereafter. By listing the required features of the SW, the required devices can be defined to add these features to the implemented SW, as shown in Table I.

TABLE I. SW SUBSYSTEMS AND THEIR REQUIRED DEVICES

Subsystem	Feature	Required device
Operation Methods	Manual control of the wheelchair	Joystick
	Speed control, alarm, and Emergency stop	Dashboard
	A control tool for the helper	Remote-Control (RC) board
	Control the wheelchair by voice commands	SR Kit
	The ability to add an extra operating method	Extra control port
Healthcare System	Measuring the body temperature of the user	Body Temperature sensor
	Measuring the heart rate of the user	Electrocardiogram (ECG) sensor
	Measuring the emotional state of the user	Galvanic Skin Response (GSR) sensor
	Measuring the weight of the user	Load Cell
Safety System	Indicator for the alarm	Buzzer
	Measuring the noise around the wheelchair	Sound sensor
	Measuring the vibration of the wheelchair	Accelerometer
	Obstacle avoidance system in the four directions of the wheelchair	Infrared (IR) Range sensor
The SW Monitoring System	Changing the motor's speed	Motor driver board
	Monitoring the responses of the operation of the motors	Voltage sensor
	Monitoring the power of the SW	Current sensor
	Measuring the speed of the SW	Encoder system
Data Acquisition (DAQ) System	Make the link between the devices, control the wheelchair, and send the measured signals to the PC	Arduino Mega board
	Display the readings of the different sensors	PC

Based on the mentioned information above from the motors, required devices, and the electrical power supply, the electrical connections and interfaces of the SW have been defined, and it will be clarified thereafter. Fig. 1 shows a real photo of the implemented SW.

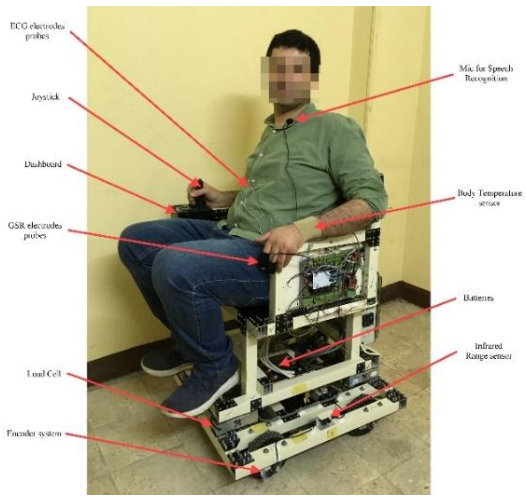


Fig. 1. The implemented SW

By thoroughly understanding the required mission of the SW mentioned in this research, the proper link between all its sub-systems can be established and converted into a programming structure on the Arduino boards, enabling all the devices to communicate and work in harmony. Fig. 2 describes using different operation methods for operating the SW. Furthermore, to implement a monitoring system for healthcare for the user and the measured signals from the SW integrated with limits for the measured signals and making an alarm if required, a GUI has been executed on the PC.

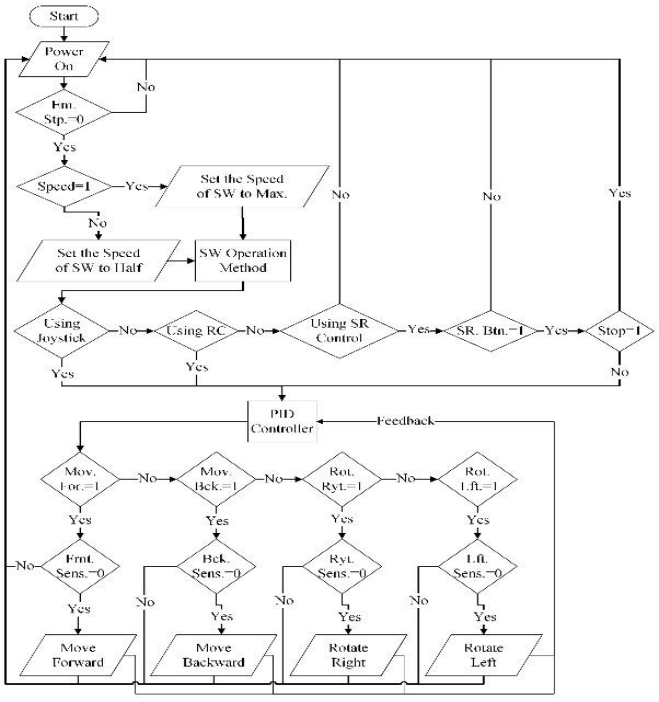


Fig. 2. The flowchart diagram describes using different operating methods of the SW.

IV. Design and Selection

In this section, the mechanical and electrical design of the implemented SW will be introduced, which adheres to the international standards for wheelchairs. Additionally, each step involved in selecting the components of the SW is going to be demonstrated to ensure they fulfill their intended functions.

A. Mechanical Design

The SW will rely on batteries to power all its motors and systems. As a result, reducing the weight of the wheelchair's body structure will reduce the power needed for the motors to support the wheelchair's weight. In addition, it has been planned to set a maximum payload of 100 kg for the wheelchair user. According to the previous data, it has been found that the structure should be as light as possible and rigid to withstand the applied forces. Taking all of this information into account, it has been determined that using aluminium sections joined together by small sheet metal assemblies held together by bolts and nuts was the best solution. According to the previous idea and the standard range of PW dimensions, the design of the wheelchair has been implemented by using the CAD software. As a result, the CAD software calculated the weight of the designed wheelchair to be 40 kg.

After taking into account all the parameters of the wheelchair's structure, the next step involves assigning the power to the drive motors. The four-wheel drive system has been selected for the SW based on its advantages, ensuring each wheel has its own driving motor. One of the most important dimensions that needs to be taken into consideration is the height of the wheelchair from the ramp, with a maximum incline angle that can be worked with, as shown in Fig. 3. This dimension is directly reflected in the selection of the wheel diameter of the SW. Consequently, the diameter of the wheel on the SW will directly affect its linear speed. The diameter of the wheel, after considering all the necessary data for selection, was 15.24 cm.

1) Motor selection. Based on the previous information about the structure, it has been determined that the maximum weight of the SW, including the user, would be 140 kg. The wheel diameter would be 15.24 cm, and the wheelchair would have a four-wheel drive system. Additionally, it has been determined that the SW should have a maximum linear speed of 0.5 m/s to ensure user safety, as recommended by references [29, 30]. The maximum inclination angle of the ramp that our SW can operate with has been defined as 7.125 degrees, as referenced in [31]. While all the required information is available, applying the force analysis for each wheel can be defined as shown in Fig. 3.

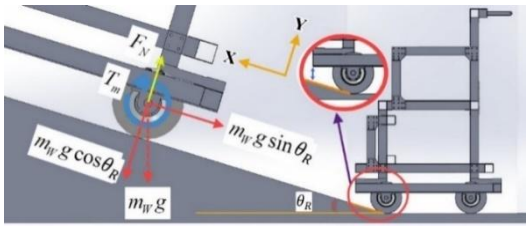


Fig. 3. CAD drawing shows force analysis for each wheel and the distance between the ramp and the SW.

where,

- F_N is the normal force (N).
- T_m is the torque of the motor for each wheel (N.m).
- m_W is the applied mass on each wheel (kg).
- θ_R is the maximum inclination angle of the ramp (^{o}).
- g is the gravitational acceleration (m/s^2).
- Vmax is the maximum linear speed of the wheelchair (m/s).
- r_W is the radius of each wheel (m).
- ω_{max} is the angular speed for each wheel (rad/s).
- N_{max} is the maximum RPM of each motor (RPM).
- a is the wheelchair acceleration (m/s^2).
- t is the time of motion (s).
- V_0 is the initial speed (m/s).
- P_m is the power for each motor (W).
- ΣF_x is the total force in the X axis (N).
- F_m is the generated force by each motor (N).

2) Calculations for motor selection. In this section, the mathematical calculations that define the specifications for each motor in the implemented SW will be presented. It is worth noting that these calculations apply to any mobile robot.

– RPM Calculations:

By using the equation of the relation between the linear speed and the angular speed to calculate the angular speed shown in Eq. (1) and (2):

$$\omega_{max} = \frac{V_{max}}{r_W} \quad (1)$$

$$\omega_{max} = \frac{0.5}{0.0762} = 6.562\ rad/s \quad (2)$$

Applying the equation for the relationship between the angular speed and the RPM to calculate the maximum RPM of each motor shown in Eq. (3) and (4):

$$N_{max} = \frac{60}{2\pi} \cdot \omega_{max} \quad (3)$$

$$N_{max} = \frac{60}{2\pi} \cdot 6.562 = 62.64 \approx 63\ RPM \quad (4)$$

– Torque Calculations:

From the First Equation of Motion, to calculate the acceleration shown in Eq. (5) and (6):

$$a = \frac{1}{t} \cdot (V_{max} - V_0) \quad (5)$$

Assume that $t = 1\ s, V_0 = 0$

$$a = 0.5\ m/s^2 \quad (6)$$

From the force analysis for each wheel as shown in Fig. 3, the total force in the X axis can be calculated as shown in Eq. (7):

$$\because \sum F_x = m_W \cdot a = F_m - m_W \cdot g \cdot sin(\theta_{max}) \quad (7)$$

And considering the equation that represents the relationship between force and torque, to calculate the torque is as shown in Eq. (8) to (11):

$$F_m = \frac{T_m}{r_W} \quad (8)$$

then,

$$T_m = m_W \cdot r_W \cdot (a + g \cdot sin(\theta_{max})) \quad (9)$$

$$T_m = 35 * 0.0762 * (0.5 + 9.81 * sin(7.125)) \quad (10)$$

$$T_m = 4.579\ N \cdot m = 46.677\ Kg \cdot cm \quad (11)$$

– Power Calculations:

By using the equation between the power and the torque as shown in Eq. (12):

$$P_m = T_m \cdot \omega_{max} = 4.579 * 6.562 = 30.05\ W \quad (12)$$

According to the mechanical structure, force analysis, and four-wheel drive system of the wheelchair, the suitable motor torque was 47 Kg.cm, the motor power was 30 W, and the motor RPM was 63 RPM for each motor of the SW.

B. Electrical Structure

In this section, the process of defining each component of the SW's electrical system will be demonstrated. At the beginning, all the required functions of the SW will be reviewed, as mentioned in Table I. It has begun by selecting each necessary device, sensor, kit, or board to perform a specific task within the SW. Next, the quantity and types of digital inputs, digital outputs, analog inputs, analog outputs, and communication connections within the SW have been determined. Due to the large number of inputs and outputs and our consideration of cost, the Arduino Mega Board has been selected as the controller for the SW. Given the high number of inputs and outputs, it has been determined that two Arduino Mega Boards would meet our requirements.

1) Schematic overview. Once all the necessary devices for the SW have been defined, including motors, sensors, kits, and boards, then their electrical power supply and connections can be specified. As previously mentioned, the large number of inputs and outputs necessitated the use of two Arduino Mega boards in the electrical system of the SW. The electrical system has been separated into two primary sections, each utilizing one Arduino Mega board. The first section collects all signals related to the SW's operating devices and sends the necessary electrical power to the motors, while the second section primarily collects signals from lifecare sensors. The first side, known as the Control side, controls and operates the SW, while the second side, known as the Data Acquisition (DAQ) side, communicates with the PC to process and display all the collected signals on the GUI. Serial communication connects both sides, enabling communication between them. The electrical schematic diagram of the SW is shown in Fig. 4, and the Control and DAQ sides will be described in the next section.

2) Control interface circuit board. As previously mentioned, the control side of the SW's electrical system primarily relies on the first Arduino Mega Board, also known as the Control Arduino Mega Board. The mediator between the Control Arduino Mega Board and the other devices is the Control Interface Circuit Board. The Control Interface Circuit Board is responsible for managing the input signals from the joystick, dashboard, RC board, proximity sensor for speed measurement, voltage sensors, and current sensors. It also provides a free terminal block that serves as an additional control port. Additionally, it facilitates communication with the SR Kit, manages the output signal for the buzzer, and communicates to the DAQ Interface Circuit Board. The board also serves as an interface between the Arduino Mega Board and the motor driver boards, ensuring that the motors receive the necessary electrical power based on the output signal from the Arduino Mega Board. The electrical system contains two motor driver boards, and each board can handle two different motors. The Control Interface Circuit Board provides all the electrical supply to all the devices and subsystems in the SW. The real connections and interface of the Control Interface Circuit Board, the Arduino Mega Board, the RC board, and the motor driver boards are shown in Fig. 5.

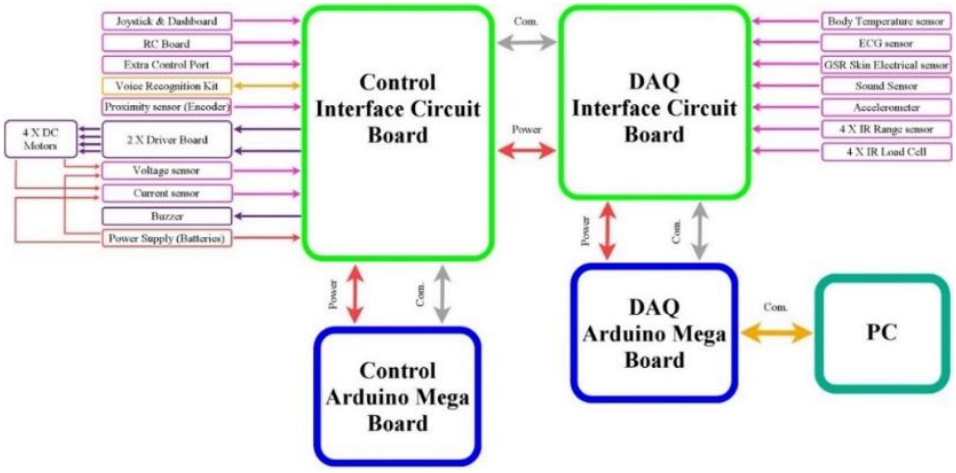


Fig. 4. The electrical schematic diagram of the SW.

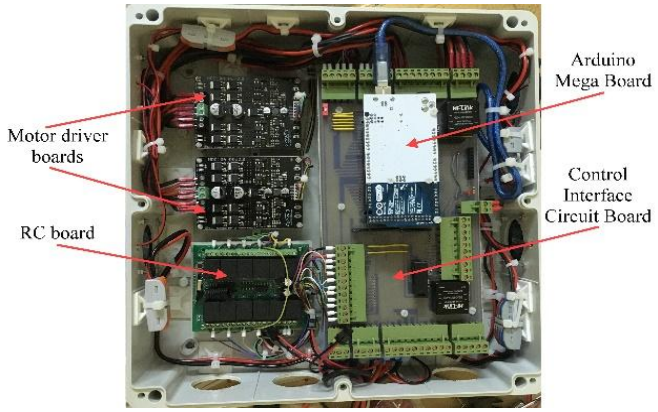


Fig. 5. The Control Interface Circuit Board, the Arduino Mega Board, the RC board, and the motor driver boards.

3) DAQ Interface Circuit Board. The DAQ side, the second side of the SW's electrical system, houses the second Arduino Mega Board, also known as the DAQ Arduino Mega Board. The Control Interface Circuit Board supplies electrical power to the DAQ Interface Circuit Board, which then distributes the necessary power to the connected devices. The DAQ Interface Circuit Board is organizing the analog inputs from the body temperature sensor, ECG sensor, GSR sensor, sound sensor, Accelerometer, IR range sensors, and Load Cells. Also, it is managing serial communication with the Control Interface Circuit Board. The board also contains the required signal conditioning circuits for the connected devices, sensors, and boards. In addition to the DAQ Interface Circuit Board, it contains the electric power filters and regulators to supply each device, sensor, and board connected to it with the required input voltage based on its datasheet. So, it contains the connection pins for the Control Arduino Mega Board. The implemented connection with the DAQ Interface Circuit Board, ECG board, GSR board, sound sensor, Accelerometer, and the Arduino Mega Board is shown in Fig. 6.

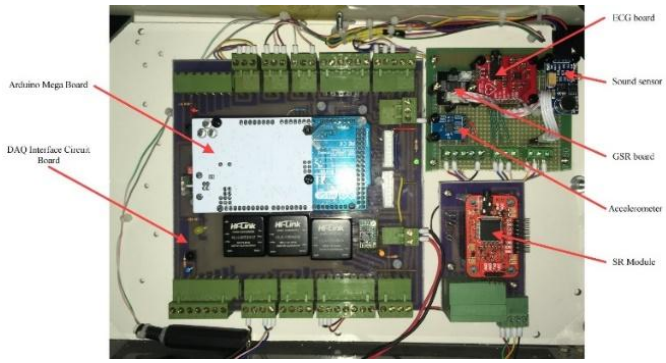


Fig. 6. The DAQ Interface Circuit Board, ECG board, GSR board, sound sensor, Accelerometer, the Arduino Mega Board, and the SR Module.

V. Operation and Hardware

This section will discuss various methods for controlling the SW, as well as the various options integrated into it to enhance its functionality. The goal is to increase the user's comfort by adding useful functions that will make their life easier. Additionally, the hardware utilized to implement these functions in the aforementioned SW will be discussed.

A. Operation Methods of Wheelchair

Here, the techniques used to drive and operate the SW will be discussed, as well as how the user can activate and deactivate some of its features.

1) Voice commands control. The wheelchair has incorporated a speech recognition system to broaden its user base. This system enables individuals with upper limb issues or poor hand control to navigate the necessary area independently, eliminating the need for assistance. Based on the user's voice input, the speech recognition system sends commands to the SW's controller, which then uses these commands to control the wheelchair's movements. The user's voice served as the training source for the speech recognition system.

2) Manual control. The PW's main feature is the normal manual control, which uses a joystick to allow the user to navigate the area and move the wheelchair. The joystick control incorporates a dashboard with functional buttons to provide the user with additional options and assistance during operation. The dashboard includes buttons for various functions, including an emergency button for stopping the SW in emergency situations, a speed control button for adjusting its speed, a voice command operation button for enabling and disabling, and a button for turning on and off the alarm.

3) Wireless remote control. The SW now includes a wireless remote control to simplify operation and control for the user's helper. Additionally, regular users can utilize the wireless remote control if they prefer not to use the traditional manual control with a joystick while resting their hands on their thighs. The wireless remote control is equipped with several buttons that perform the same functions as the manual joystick and dashboard, including wheelchair movement, speed control, and activation and deactivation of certain SW functions.

B. Healthcare and Vital Signs Monitoring

The wheelchair feels like a close friend to its user. Individuals who are handicapped, elderly, experience difficulty in their lower limbs, or sustain injuries depend on their wheelchair for their daily activities. Users spend most of their time using the wheelchair for navigation, reaching specific locations, assisting with tasks, or enjoying meals. Therefore, adding more options to the wheelchair will enhance its users' happiness, satisfaction, and self-confidence. It has implemented and integrated a Vital Signs Monitoring system with the aforementioned SW to monitor the user's vital signs and provide an indication if any of them are not functioning well. The SW's controller sends all measured signals to the PC for display on the GUI. Fig. 7 shows the implemented GUI of the healthcare system for the SW, which contains a live display of the user's body temperature, GSR, and ECG readings.

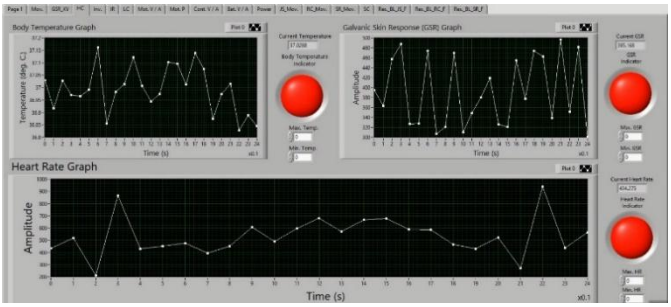


Fig. 7. The implemented GUI of the healthcare system for the SW.

1) Body temperature monitoring. The SW incorporates a body temperature sensor to measure the user's body temperature, which is a crucial indicator of their health condition. The sensor used to measure the body temperature is an NTC thermistor, specifically designed to measure 2.25K ohms at 25 degrees Celsius. A voltage divider circuit has been designed to measure the resistance change of the body temperature sensor, reflecting the body temperature, and send its output to the wheelchair controller. The Vital Signs GUI plots and displays the body temperature and other vital signs, incorporating upper and lower limits that can trigger an indicator if the measured signal exceeds these limits.

2) ECG monitoring. The ECG signal is a good indicator of the user's heart activity. The ECG monitoring system consists of three electrode pads that can be stuck to the user's chest to transfer the ECG signals to the ECG sensor kit, which converts them to an analog signal to be sent to the controller board of the SW. Then the received signals are being processed and sent to the PC that has the GUI of the ECG monitoring to display a live situation of the ECG of the user.

3) GSR monitoring. The SW integrates a GSR sensor to monitor the user's emotions. The concept of a GSR sensor relies on the user's emotional sweating. As the user's emotions fluctuate, the amount of sweat they secrete alters, influencing the conductivity between two fingers in their hand. This change in conductivity leads to a corresponding change in the resistance between the two fingers. This change has been measured, which refers to the emotional status of the user, by connecting two electrodes to two different fingers on the user's hand and measuring the resistance between them by using a GSR sensor board, and then the analog signal is sent to the controller of the SW. The SW's controller transmits the processed sensor signal to the PC for monitoring via the GSR's GUI.

C. Safety Implementation

As the safety level of the SW increases, the user will experience greater satisfaction during use, thereby elevating their morale. The SW has integrated an obstacle avoidance system. While the product is designed to assist individuals with disabilities in using and operating the wheelchair, it would be beneficial to provide them with awareness notifications while they are operating the SW. Moreover, the use of voice commands for wheelchair operation by individuals with disabilities leads to an increase in accidents and a decrease in user safety and satisfaction. Therefore, incorporating an obstacle avoidance system into our SW is crucial. The integrated Obstacle Avoidance System in the SW employs four IR sensors, positioned on the front, rear, right, and left sides of the wheelchair, to detect any obstacles in its vicinity. The Obstacle Avoidance System halts the SW's movement if any of its four sensors detects an obstacle in its path. It will continue to operate in this direction until the obstacle moves and the user commands it to continue. The signals from the four sensors of the Obstacle Avoidance System are given the utmost priority in the program structure of the SW's controller, thereby reducing delays, waiting times, and stoppages, thereby enhancing user safety.

VI. PID Control Design for SW

While the linear speed of the SW is one of the important factors that affect the user's safety and satisfaction, a PID control has been implemented to sustain the speed of the SW. The design of the PID controller has started by finding the transfer function of the used motor. The following section outlines the process of determining the transfer function of the DC motor for use in speed control of the SW. Fig. 8 shows the equivalent circuit of the DC motor that was used to figure out the transfer function.

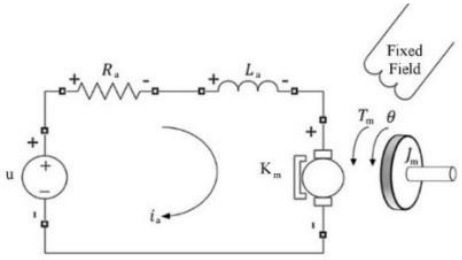


Fig. 8. The equivalent circuit of the DC motor [32].

where,

- Jm is the moment of inertia of the rotor (Kg.m2).
- C is the motor viscous friction constant (N.m.s).
- Km is the motor torque constant and the back emf constant (N.m/A), and (V/rad.sec-1).
- Ra is the electric resistance of the coil (Ohm).
- La is the electric inductance of the coil (H).
- θ is the angular displacement of shaft in radians (rad).
- ω is the angular velocity of shaft (rad/s).
- ia is the armature current (A).
- u is the armature voltage (V).
- Tm is the motor torque (N.m).

The relation between the back emf and the angular velocity and the relation between the armature voltage and the back emf voltage can be defined based on Kirchhoff's voltage law. Also, the relationship between the motor torque and the armature current too. In addition, the motor torque based on Newton's 2nd law and applying the Laplace transformation, the relation between the input voltage and the output angular velocity can be defined as shown in Eq. (13):

$$\frac{\Omega(s)}{U(s)} = \frac{K_m}{L_a J_m s^2 + (L_a C + R_a J_m)s + R_a C + K_m{}^2} = \frac{K_m}{(J_m s + C)(L_a s + R_a) + K_m{}^2} \quad (13)$$

By substituting at the transfer function of the DC motor of the SW [Eq. (13)] with the values of the parameters of the system (J_m, C, K_m, R_a, and L_a) as the below values:

$$J_m = 9.19 * 10^{-4} \quad \text{Kg.m2}$$

$$C = 9.19 * 10^{-5} \quad \text{N.m.s}$$

$$K_m = 0.75 \quad \text{(N.m/A), and (V/rad.sec-1)}$$

$$R_a = 3 \quad \text{Ohm}$$

$$L_a = 0.1 \quad \text{H}$$

The open-loop transfer function of the DC motor of the SW will be as shown in Eq. (14):

$$\frac{\Omega(s)}{U(s)} = \frac{0.75}{0.0000919s^2 + 0.00276619s + 0.5627757} \quad (14)$$

Using the system's open-loop transfer function in Eq. (14), Fig. 9 shows the system's output prior to the addition of the PID controller. However, there is an overshoot of about 53.8% from the input value, and the system could not achieve the required value.

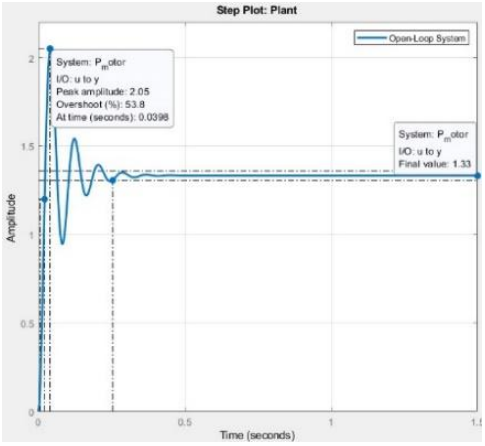


Fig. 9. The open-loop system response without the PID controller.

Fig. 10, illustrates the addition of a PID controller to enhance the system's output response. Also, the tuning was done using the PID Tuner function in MATLAB software. This is an automatic tuning function based on a proprietary tuning algorithm that uses a model-based tuning technique to set the PID controller parameters as: K_p = 0.1181, K_i = 5.2143, and K_d = 6.6921*10^{-4}.

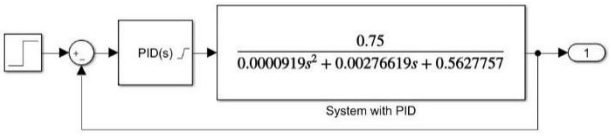


Fig. 10. The closed-loop system with the PID controller.

The addition of the PID controller to the system, as illustrated in Fig. 11, has improved the system's output response. Furthermore, the system could achieve the required value with a Rise time of 0.342 seconds and a Settling time of 0.615 seconds.

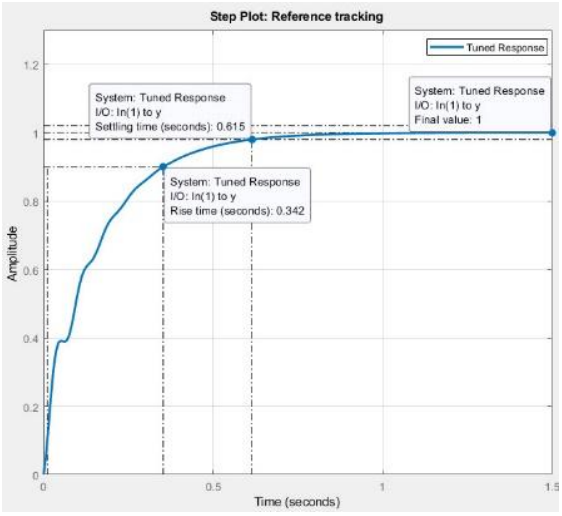


Fig. 11. The closed-loop system response with the PID controller.

VII. Results

A fixed response from the wheelchair impacts the user's satisfaction, subsequently influencing their morale. Ten trials were done to see how the wheelchair responded to step input from the SW operation device in a variety of situations, as shown in Table II. The goal was to make sure the controller worked properly after the tuning process. The main idea of the trials is to have a comparison of the implemented wheelchair response without using the PID controller and with using the PID controller to be able to figure out the change in the response during different situations that can happen for the wheelchair during normal usage. Trials one and two involved hanging the SW and moving its wheels without contacting the floor. And trial number three and four were based on moving the SW on the floor without any payload. And trials numbers five and six were based on moving the SW on the floor with a payload. And trial number seven and eight were based on moving the SW on a ramp that has the standard specs for the wheelchair ramps with a slope by ramp height of 1:8.2 and incline angle of 7 degrees without any payload, as shown in Fig. 12. And trial number nine and ten were based on moving the SW on the mentioned ramp previously with a payload.

TABLE II. The Trials and their Descriptions

Trial No.	Description
1	Move the SW while hanging without the PID controller
2	Move the SW while hanging with the PID controller
3	Move the SW on the floor without a payload without the PID controller
4	Move the SW on the floor without a payload with the PID controller
5	Move the SW on the floor with a payload without the PID controller
6	Move the SW on the floor with a payload with the PID controller
7	Move the SW on the ramp without a payload without the PID controller
8	Move the SW on the ramp without a payload with the PID controller
9	Move the SW on the ramp with a payload without the PID controller
10	Move the SW on the ramp with a payload with the PID controller



Fig. 12. The SW on a wheelchair ramp with standard specs.

Fig. 13 shows a good comparison by using the average of each speed reading for the trials from 1 to 6 under their conditions, while considering their steady state without and with the PID controller. The black curve of each graph shows the step input for the SW with a set value of 0.23 m/s. As shown in Fig. 13(a) for trials 1, 3, and 5, the SW system responded to the conditions that have been discussed earlier without using a PID controller. Some trials had overshoot, meaning that most had a high steady-state error, even though the settling time values varied. Fig. 13(b) illustrates the response of the SW system for trials 2, 4, and 6 under the previously mentioned conditions, utilizing a PID controller, exhibiting minimal overshoot, reduced steady-state error, and varying settling time values. It has also been noticed that the damping of the SW system is increasing with the increase of the load, and this phenomenon is a normal behavior for the system.

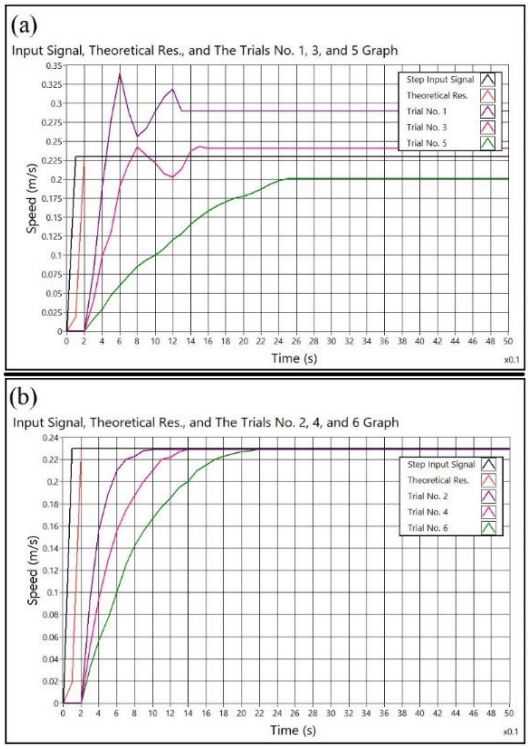


Fig. 13. Results graphs of the trials from 1 to 6 in response to the step input by using the average speed of each trial: a) Results graph for trials 1, 3, and 5 while not using a PID controller; b) Results graph for trials 2, 4, and 6 while using a PID controller.

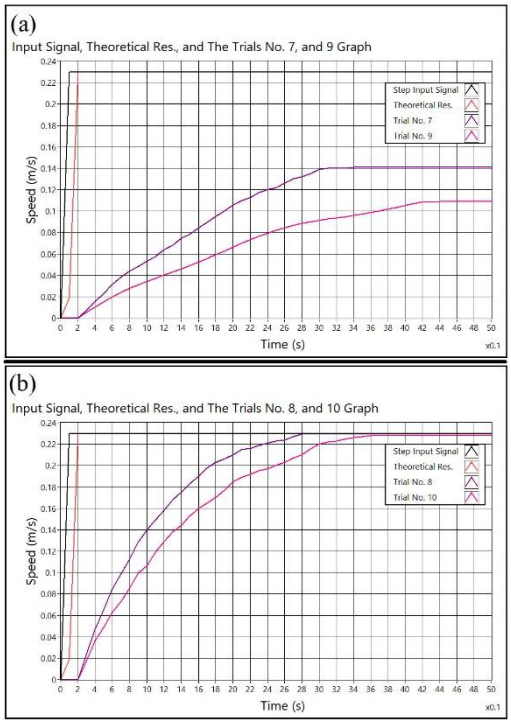


Fig. 14. Results graphs of the trials from 7 to 10 on the wheelchair ramp in response to the step input by using the average speed of each trial: a) Results graph for trials 7 and 9 while not using a PID controller; b) Results graph for trials 8 and 10 while using a PID controller.

Fig. 14 shows the output graphs for trials 7 to 10, comparing the average speed readings with and without the PID controller on the wheelchair ramp. In Fig. 14(a), trials 7 and 9 show that as the load increases, the SW's damping increases, leading to a higher steady-state error for the SW when the PID controller is not used, which is not acceptable performance. Fig. 14(b) depicts the response of the SW system during trials 8 and 10, demonstrating an increase in damping as the load increases. The PID controller's efforts are evident as it improves the steady-state error of the SW system and maintains the set value.

Table III summarizes the outcomes of all the trials conducted for the SW, both with and without the PID controller, under various scenarios that may arise during normal wheelchair usage, in response to a step input with a set value of 0.23 m/s. The results indicate that there are overshoot values when the system operates without the PID controller, with a maximum percentage of 47.51% under specific conditions, and no overshoot when the PID controller is used. In addition to using the SW system without the PID controller, which gives a high steady-state error with a maximum percentage of 52.60% under certain conditions, using the SW system with the PID controller produces a very low steady-state error with a maximum percentage of 0.81% under certain conditions. Moreover, using the SW with the PID controller shows an improved settling time compared to using the SW without the PID controller. Under certain conditions, the maximum settling time for the SW with the PID controller is 2.93 s, while the maximum settling time for the SW without the PID controller is 3.91 s. Additionally, as the load increases, the damping ratio of the SW system also increases.

TABLE III. THE RESULTS OF ALL THE TRIALS

Criteria No.	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Overshoot (%)	47.51	0	5.78	0	0	0	0	0	0	0
Steady-State Error (%)	26.09	0.57	4.47	0.36	12.61	0.2	38.66	0.34	52.6	0.81
Settling Time (s)	1.25	0.67	1.36	1.08	2.27	1.66	2.87	2.27	3.91	2.93
Damping Ratio	0.51	1.28	0.81	1.52	1.2	1.78	1.5	2.05	2	2.31
Step Input Set Value (m/s)	0.23									

VIII. CONCLUSION AND FUTURE WORK

Feeling more independent and safe boosts the wheelchair user's morale, which in turn improves their health condition and lessens the burden on their helper. Our solution effectively integrates traditional joystick control with voice command functionality and wireless remote control, providing improved accessibility and autonomy for users. Furthermore, using the integrated vital signs monitoring system provides live data about the user's health, which enhances user health management. Combining an obstacle avoidance system with our SW increases the user's safety. Additionally, the results of a comparison between using the normal control system and the PID control system while operating the implemented SW in different situations can be observed for the wheelchair during normal usage, which has proved an improvement for the wheelchair's response and stability, thereby increasing user satisfaction. While the trials have proved that using the PID controller with $K_p = 0.1181$, $K_i = 5.2143$, and $K_d = 6.6921*10^{-4}$ has eliminated the overshoot of the system from 47.51%. In addition, decreasing the steady-state error percentage from 52.60% to 0.81% as a maximum also improved the settling time from 3.91 s to 2.93 s as a maximum compared to using the normal control system to operate the SW. Adding the PID controller to the SW has improved its performance and stability.

The findings of our study illustrate the advantages of integrating the PID controller while using hybrid and combined control systems in assistive devices, focusing on their capacity to improve the lives of handicapped people. The full calculations for motor selection for our SW have been illustrated, which is also very useful during the design of any mobile robot. Our hybrid-controlled SW signifies an important advancement in assistive technology, providing improved mobility and health monitoring. Keep improving in such a field; it plays a significant role in improving the quality of life for users.

Since this study is just the beginning, several more studies can be conducted in the future. These include replacing the wheelchair's traditional wheels with Mecanum wheels. New healthcare sensors and devices can be added to provide more details about the user's health, such as blood pressure, breathing rate, and so on. Improve the GUI for healthcare monitoring to include the readings from the new sensors and devices. The system stores the collected data in a database for the doctor's analysis. The system can send alarms about the user's health status through Short Message Service (SMS) if it's an urgent matter, and it can also send a daily report via email to the doctor.

REFERENCES

[1] Nasrin Aktar; Israt Jahan; and Bijoya Lala, "Voice Recognition based Intelligent Wheelchair and GPS Tracking System", International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, PP. 7130-7140.

[2] Akshay B. Satpe; Mr. S. V. Khobragade; and Dr. S. L. Nalbalwar, "Wheelchair Control Using Hand Movement & Voice with Obstacle Avoidance", Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, PP. 1175-1178.

[3] Zekai Zhao, "Effects of Population Aging on Japan's Economic Growth", International Conference on Economic Management and Green Development, 2023, PP. 421-428.

[4] Nikunj Janak Gor; and Prof. Amutha Jeyakumar, "Voice Controlled Motorized Wheelchair with Real Time Location Monitoring", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, PP. 1163-1167.

[5] Nerea Perez; Aitziber Mancisidor; Itziar Cabanes; and Javier Gonzalez, "Continuous postural and movement monitoring system for wheelchair users assessment", Measurement, 2024, Vol. 237, PP. 1-12.

[6] Chia-Hung Chang; and Wei-Wen Hu, "Design and Implementation of an Embedded Cardio-Respiratory Monitoring System for Wheelchair Users", IEEE Embedded Systems Letters, 2020, Vol. 13, PP. 150-153.

[7] Jesse Leaman; and Hung Manh La, "A Comprehensive Review of Smart Wheelchairs: Past, Present and Future", IEEE Transactions on Human-Machine Systems, 2017, Vol. 47, PP. 486-499.

[8] Ilham A.E. Zaeni; Ahsan Walad; and Anik N. Handayani, "Applying Neural Network on Wheelchair Control System using Eye Blink and Movement Command", 3rd International Conference on Information and Communications Technology (ICOIACT), 2020, PP. 416-420.

[9] Colleen O'Connell MD; Meiqi Guo MD; and Jeff Sparks, "All hands on deck: The multidisciplinary rehabilitation assessment and management of hand function in persons with neuromuscular disorders", Muscle & Nerve, 2025, Vol. 71, PP. 869-888.

[10] Chinemelu Ezeh; Pete Trautman; Catherine Holloway; and Tom Carlson, "Comparing shared control approaches for alternative interfaces: a wheelchair simulator experiment", IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2017, PP. 93-98.

[11] Preetam Kumar Khuntia; and P. V. Manivannan, "Review of Neural Interfaces: Means for Establishing Brain–Machine Communication", SN Computer Science, 2023, Vol. 4, PP. 1-13.

[12] Qiyun Huang; Shenghong He; Qihong Wang; and Yuanqing Li, "An EOG-Based Human-Machine Interface for Wheelchair Control", IEEE Transactions on Biomedical Engineering, 2018, Vol. 65, PP. 2023-2032.

[13] Ziqing Zhang, "AI-Powered Intelligent Speech Processing: Evolution, Applications and Future Directions", International Journal of Advanced Computer Science and Applications, 2025, Vol. 16, PP. 918-928.

[14] M. Sayed; T. Mansour; A. El Domiaty; MG Mousa; and Ali A. S., "Wheelchair Review: Types, Operation Techniques, and Safety Aspects", Suez Canal Engineering Energy and Environmental Science Journal, 2024, Vol. 2, PP. 1-26.

[15] Marlene Alejandra Domínguez Sucres; Sarahí Soledad Franco Pérez; and Diana Laura Infante Ramírez, "EOG-based interface and speech recognition for wheelchair control", IEEE International Conference on Engineering Veracruz (ICEV), 2019, PP. 1-4.

[16] Sumet Umchid; Pitchaya Limhaprasert; and Theera Leeudomwong, "Voice Controlled Automatic Wheelchair", 11th Biomedical Engineering International Conference (BMEiCON), 2018, PP. 1-5.

[17] Rashmi S. Mullur; Seetal Preet Kaur Cheema; Ryan Edward Alano; and Lynn Elizabeth Chang, "Tele-Integrative Medicine to Support Rehabilitative Care", Physical Medicine and Rehabilitation Clinics of North America, 2021, Vol. 32, PP. 393-403.

[18] Andrea Antonio Cantone; Mariarosaria Esposito; and Francesca Pia Perillo, "Enhancing Elderly Health Monitoring: Achieving Autonomous and Secure Living through the Integration of Artificial Intelligence, Autonomous Robots, and Sensors", Electronics, 2023, Vol. 12, PP. 1-23.

[19] Vaishali Jabade; Harsh Shrivastav; Jash Rupani; and Nikita Harak, "Mobility Plus: Voice-Controlled Wheelchair with Health Monitoring

System and Oxygen Cylinder Integration", International Conference on Inventive Computation Technologies, 2024, PP. 2167-2171.

[20] SATYAVIR SINGH; SHWETA MANI; and SARIKA, "VOICE CONTROLLED WHEELCHAIR WITH COLLISION AVOIDANCE", i-manager's Journal on Digital Signal Processing, 2022, Vol. 10, PP. 1-8.

[21] Jenina R. Amoguis; Mabel A. Lingon; and Airah Cahigan, "Development of a Voice-Controlled Wheelchair for Physically Impaired Individuals", Global Clinical Engineering, 2024, Vol. 6, PP. 6-17.

[22] Fida Hussain; Sayed Mazhar Ali Shah; Asif Ali; Imran Ali; and Haresh Kumar, "Voice and Hand Controlled Wheelchair System", JOURNAL OF APPLIED ENGINEERING & TECHNOLOGY, 2021, Vol. 5, PP. 31-39.

[23] Mohammad Shahrul Izham Sharifuddin; and Azliza Mohd Ali, "Comparison of CNNs and SVM for voice control wheelchair", IAES International Journal of Artificial Intelligence, 2020, Vol. 9, PP. 387-393.

[24] M. S. Arsha1; A. Remya Raj; S. R. Pooja; Rugma Manoj; and Shimi Mohan, "Voice Controlled Wheelchair", International Journal of Research in Engineering, Science and Management, 2020, Vol. 3, PP. 103-106.

[25] Divya Jennifer DSouza; Sanket Srivatsava; Ruth Prithika; and Sahana, "IoT Based Smart Wheelchair for HealthCare", International Journal of Recent Technology and Engineering, 2019, Vol. 8, PP. 534-539.

[26] Shailu Sachan; and Pankaj Swarnkar, "Investigations on meta-heuristic algorithms for intelligent speed regulation of mobile robot", Results in Control and Optimization, 2023, Vol. 11, PP. 1-11.

[27] V. Sankardoss; and P. Geethanjali, "Design and Low-Cost Implementation of an Electric Wheelchair Control", IETE Journal of Research, 2019, Vol. 65, PP. 1-10.

[28] Rawiphon Chotikunnan; Phichitphon Chotikunnan; and Manas Sangworasil, "Application of PID Control System in Mecanum Wheelchair", International Journal of Membrane Science and Technology, 2023, Vol. 10, PP. 3519-3529.

[29] D. Kobayashi; T. Watanabe; N. Tyler; and T. Suzuki, "A Validation Test of Using Shoulder Joint Moment in Evaluation of Load in Wheelchair Propulsion", IEEE Life Sciences Conference (LSC), 2017, PP. 242-245.

[30] Bing-Fei Wu; Yung-Shin Chen; and Po-Ju Chang, "An Uphill Safety Controller with Deep Learningbased Ramp Detection for Intelligent Wheelchairs", IEEE Access, 2018, Vol. 6, PP. 28356 - 28371.

[31] Valeria Tatano; and Rosaria Revellini, "An alternative system to improve accessibility for wheelchair users: The stepped ramp", Applied Ergonomics, 2023, Vol. 108, PP. 1-11.

[32] Yaojung Shiao; Premkumar Gadde; and Chun-Yu Liu, "Wavelet-Based Analysis of Motor Current Signals for Detecting Obstacles in Train Doors", Applied Sciences, 2024, Vol. 15, PP. 1-21.

Design and Implementation of an Intelligent Laboratory Management System Based on UWB Technology

Heng Sun[1], Qiang Gao[2]*
Modern Educational Technology Laboratory, School of Foreign Languages and Literature,
Shandong University, Jinan 250100, China[1]
School of Labor Relations, Shandong Management University, Jinan 250357, China[2]

***Abstract*—In recent years, the rapid development of educational informatization and the widespread adoption of Internet of Things (IoT) technologies have accelerated the transformation of university laboratories toward intelligent management. However, traditional laboratory management systems still suffer from limited automation, insufficient safety mechanisms, and poor real-time responsiveness. To address these issues, this study proposes an intelligent laboratory management system based on Ultra-Wideband (UWB) technology, which offers high-precision positioning and low-latency communication. The proposed system integrates IoT-based functionalities across six core modules, including access control, asset management, environmental monitoring, and user management. By deploying UWB tags and sensors throughout the laboratory environment, the system enables real-time tracking of personnel and equipment, automatic activation of laboratory devices, and intelligent safety alerts. A pilot deployment in three university laboratories demonstrated improvements in access efficiency, energy conservation, equipment security, and user satisfaction. The results validate the system's effectiveness in enhancing intelligent laboratory management and provide a scalable model for future smart educational infrastructure.**

***Keywords*—UWB; intelligent management; laboratory management system; IoT**

I. Introduction

Laboratories are a critical component of higher education institutions, serving as essential venues for transforming theoretical knowledge acquired in the classroom into practical skills. They play a vital role in cultivating students' hands-on abilities and fostering innovation. As the core infrastructure for practical teaching, the quality of laboratory hardware and software construction directly influences the effectiveness of talent development. Enhancing equipment configuration and implementing intelligent management systems are fundamental for building world-class universities and maintaining educational competitiveness.

With the rapid advancement of educational informatization and the increasing integration of the IoT, universities are transitioning towards intelligent laboratory environments. These environments aim to provide secure, efficient, and automated support for teaching and experimentation. While technologies such as RFID, Bluetooth Low Energy (BLE), and Wi-Fi have been explored for functions like access control, equipment tracking, and environmental monitoring, existing systems often operate in isolation, lack real-time responsiveness, and suffer from accuracy and reliability issues. Many laboratories continue to rely on traditional management methods, which are inefficient and prone to human error.

Although various technologies have been applied in laboratory management, most existing solutions remain fragmented and fail to offer integrated, real-time control across personnel, assets, and environmental variables. Moreover, the potential of UWB technology—a wireless communication protocol known for its high-precision positioning and robustness against interference—has not been fully leveraged in comprehensive university laboratory management systems. This study aims to address this research gap by proposing and implementing a UWB-based intelligent laboratory management system that unifies access control, asset tracking, environmental linkage, and safety monitoring under a centralized IoT architecture.

The remainder of this study is organized as follows: Section II reviews related works on laboratory management systems and the application of positioning technologies. Section III outlines the current challenges in university laboratory environments. Section IV describes the design and architecture of the proposed UWB-based intelligent laboratory management system, including hardware and software components. Section V presents the implementation, and Section VI presents the pilot deployment of the system in three university laboratories, along with observed effectiveness. Section VII discusses the results and limitations. Section VIII concludes the study and outlines directions for future research.

II. Related Work

A. Laboratory Management Systems

With the increasing demand for efficiency and safety in laboratory operations, various digital laboratory management systems have been developed. Traditional approaches often rely on QR code scanning, RFID tags, or BLE beacons to manage asset inventory and personnel access. While these systems offer partial automation, they are limited by factors such as short detection range, interference susceptibility, and a lack of real-time integration [1]. Additionally, many systems operate in isolation, with fragmented control over equipment, environmental conditions, and user access.

*Corresponding Author.

B. Indoor Positioning Technologies

Accurate indoor positioning is critical for enabling real-time management functions in smart laboratories. Technologies such as Wi-Fi, RFID, and BLE have been widely studied and implemented. However, their performance in complex indoor environments is often degraded due to signal attenuation, multipath effects, and limited spatial resolution. In contrast, UWB technology provides centimeter-level positioning accuracy, low latency, and robust resistance to interference. According to recent studies, UWB is increasingly adopted in indoor real-time location systems (RTLS), making it a promising alternative for precise device and personnel tracking in enclosed environments [2].

C. Integrated IoT-Based Smart Systems

The integration of IoT technologies has expanded the capabilities of laboratory management, allowing for centralized monitoring and control of diverse systems such as access control, lighting, environmental sensing, and equipment usage. Prior research has explored IoT frameworks applied to educational or industrial laboratories, but many of these solutions are limited to single-function implementations, such as environmental monitoring or smart lighting. Few studies have successfully implemented a comprehensive system that integrates access control, environmental regulation, asset tracking, and safety alerting within a unified smart platform [3].

D. Research Gap

Although UWB and IoT technologies have been individually applied in laboratory and industrial contexts, the combination of these technologies into a unified, intelligent laboratory management platform remains underexplored. Existing systems often lack integrated architecture and fail to leverage the high-precision capabilities of UWB for dynamic spatial control and automation. This study aims to address this gap by designing and implementing a fully integrated laboratory management system that utilizes UWB and IoT technologies to enhance laboratory efficiency, safety, and user experience in a university environment [4].

III. Current Status and Challenges in the Intelligent Development of University Laboratories

However, laboratory management in universities still faces several challenges:

a) Inconvenient and inefficient access control: Most current access control systems rely on password verification, card-based identification, or biometric authentication. These methods present issues such as complex information input procedures, password leakage risks, and card loss. Although biometrics enhance security, limitations in recognition algorithms and accuracy often necessitate secondary verification via passwords or cards, negatively impacting user experience [5].

b) Lack of real-time and efficient asset management: Current asset management systems in university laboratories mainly depend on QR codes, RFID, or BLE technologies. QR codes require close-range scanning, RFID suffers from limited precision and high costs, while BLE offers longer-range detection but is vulnerable to signal interference. These traditional methods fall short in supporting real-time tracking and intelligent monitoring of laboratory equipment [6].

c) Limited intelligent management capabilities: Due to limited personnel, many universities have adopted basic intelligent control systems, such as automated lighting and HVAC (Heating, Ventilation, and Air Conditioning). However, these systems often operate in a static or rigid manner, lacking the ability to dynamically adjust based on actual usage scenarios, which limits energy efficiency and overall effectiveness [7].

d) Inadequate safety management mechanisms: Many laboratories lack effective early-warning and safety alert mechanisms. Real-time monitoring of personnel is often absent, and the systems are unable to prevent unauthorized individuals from entering high-risk areas, posing potential safety hazards [8].

To address the aforementioned challenges, this study proposes an intelligent laboratory management system based on UWB technology, leveraging its high-precision positioning and high-speed data transmission capabilities. In terms of access control, the system enables precise management of personnel entry by allowing only individuals carrying authorized UWB tags to access the laboratory. For asset management, UWB tags are affixed to valuable equipment, enabling the system to track the real-time location and usage trajectory of devices, thereby preventing misplacement or loss and enhancing asset security. In intelligent management, real-time acquisition of personnel and equipment location data through UWB technology facilitates the optimization of spatial layout and intelligent resource scheduling, significantly improving management efficiency and space utilization.

IV. Design of an Intelligent Laboratory Management System based on UWB Technology

A. UWB Technology

UWB is a short-range, high-bandwidth wireless communication technology that transmits information by emitting very short-duration pulses over a wide frequency spectrum, typically ranging from 3.1 GHz to 10.6 GHz. Unlike conventional narrowband systems, UWB does not rely on a continuous carrier wave, allowing it to achieve exceptionally high time-domain resolution [9].

UWB is particularly well-suited for applications requiring high-precision positioning, low latency, and low power consumption. It offers centimeter-level accuracy in distance measurement, making it ideal for RTLS, indoor positioning, access control, and asset tracking. The technology is inherently resistant to multipath interference and can penetrate non-metallic obstacles such as walls or furniture, ensuring robust performance in complex indoor environments [10].

Advanced UWB systems support various positioning techniques, including Time of Flight (ToF), Time Difference of Arrival (TDoA), and Angle of Arrival (AoA). These capabilities enable 3D spatial tracking and secure ranging, and are increasingly integrated into IoT devices, smartphones, and industrial systems.

Driven by standardization efforts (e.g., IEEE 802.15.4z) and growing industry adoption, UWB is emerging as a key enabling technology for intelligent environments, including smart homes, smart factories, and smart laboratories.

B. Performance Comparison Between UWB and Mainstream Wireless Positioning Technologies

Mainstream wireless positioning technologies currently include UWB, BLE, Wi-Fi, and RFID. UWB offers superior ranging accuracy, strong anti-interference performance, and low power consumption, making it particularly suitable for high-precision scenarios such as access control and equipment tracking.

In terms of positioning accuracy, both UWB and RFID perform well; however, RFID lacks real-time tracking capabilities. When access control systems are based on BLE, Wi-Fi, or RFID, their limited accuracy and latency may lead to false unlocking events. In contrast, systems using NFC cards or smartphones require users to bring the device into close proximity with the reader, which reduces operational convenience [11].

C. System Architecture Design

Based on the characteristics of UWB technology and the requirements of laboratory management, the proposed system adopts a layered architecture consisting of four levels: the device layer, communication layer, platform layer, and application layer.

The device layer is responsible for the deployment and operation of equipment within the laboratory. It collects device status and environmental data while supporting various intelligent functions. This layer includes subsystems such as UWB-based access control, UWB ranging, environmental monitoring, smart lighting, smart air conditioning, and smart sockets [12].

The communication layer enables data transmission between devices and the platform. It supports both wireless and wired network connections, as well as IoT protocols. To enhance the efficiency and reliability of data transmission, this layer incorporates IoT gateways and edge gateways for transmission optimization.

The platform layer serves as the core functional component of the system. It includes modules for access control management, device management, environmental monitoring, user management, data acquisition and storage, real-time safety monitoring, system configuration, and integration interfaces. This layer relies on a private IoT cloud to enable centralized data processing and management. It also provides API interfaces, message queues, notification services, and protocol support [13].

The application layer delivers interactive interfaces and application functions for end users. It supports multiple user terminals, including PC clients, mobile apps, dashboards, voice assistants, and notification services. Users can monitor the laboratory in real time, manage devices, and perform security control operations through these interfaces. The overall system architecture is illustrated in Fig. 1.

D. System Operation Flow Design

The access control system is a key module within the intelligent laboratory management system. Leveraging UWB's high-precision positioning and ranging capabilities, it not only serves as the first line of defense for laboratory safety but also functions as a trigger point for other system operations. Through accurate identity authentication and permission control, the access control system ensures laboratory security while enabling the coordinated activation of devices and environmental regulation. As such, the access control module is a critical focus of the system design [14]. The operational flow of access control is illustrated in Fig. 2.

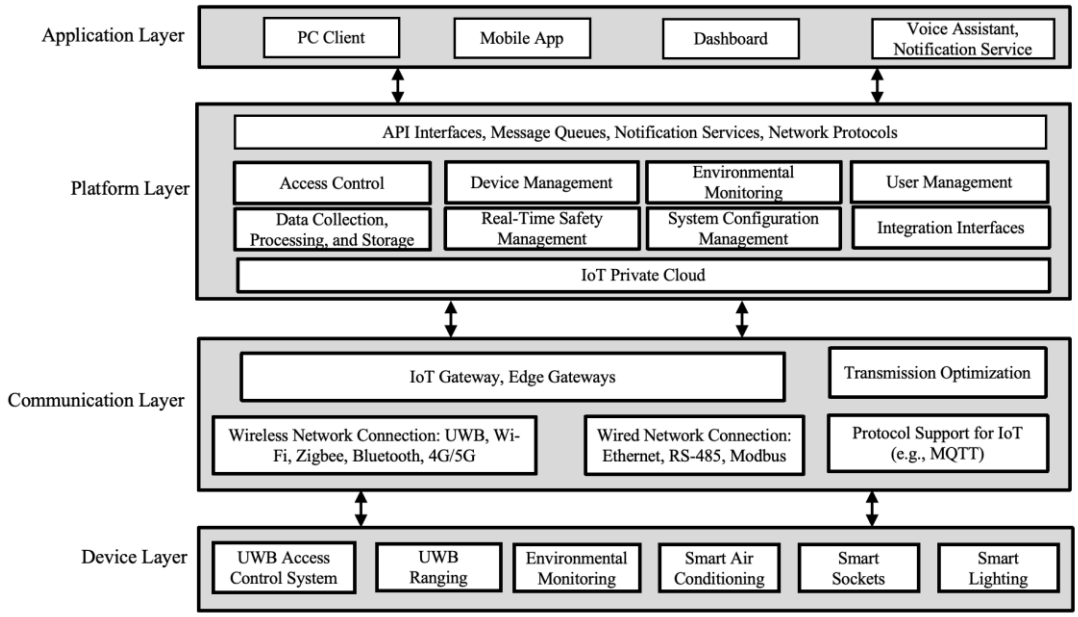


Fig. 1. System architecture of the UWB-based intelligent laboratory management system.

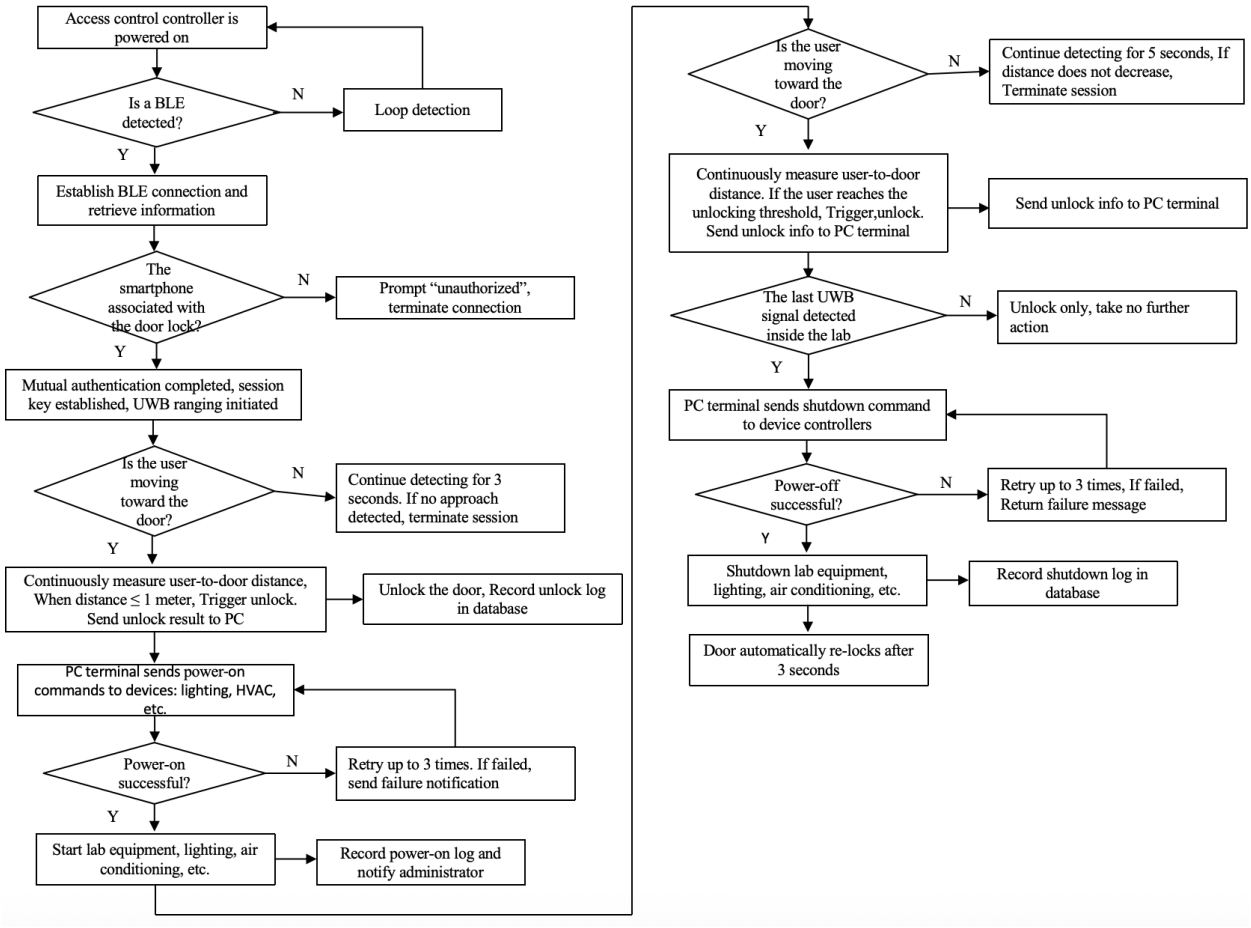


Fig. 2. Access control lock or unlock workflow diagram.

a) Initialization and connection phase: Upon power-up, the access control controller enters working mode and continuously scans for BLE broadcasts from smartphones or signals from UWB tags. If a valid signal is detected, the system establishes a BLE connection and reads relevant information from the smartphone. If no signal is found, the system continues to loop in scanning mode [15].

Subsequently, the controller queries the access rights stored in the database to determine whether the device is authorized. If not associated, the system returns an "unauthorized" message and terminates the connection. If authorization is confirmed, mutual authentication is performed, a session key is established, and UWB ranging is initiated.

b) Ranging phase: The system uses UWB ranging to assess the user's motion status, specifically whether the user is approaching the door. If not, the system continues monitoring for 3 seconds. If the distance to the door does not decrease, the session is closed. If the user is walking toward the door, the system continues to measure the distance. When the user reaches the threshold distance (one meter), the system issues an unlock command, opens the door lock, and sends an unlock log to the central PC terminal [16, 17]. Simultaneously, it sends power-on commands to device power controllers, entrance lighting, and HVAC systems.

c) Door opening and device activation phase: After sending the power-on commands, the system verifies whether power-on is successful. If the command fails 3 consecutive times, a startup failure message is returned. If successful, it activates the lab equipment power, lighting, and air conditioning. Power-on logs are recorded in the database, and push notifications are sent to the administrator. The door unlocking action is also logged and notified accordingly [18].

d) Exit and device shutdown phase: After the user enters the lab and the door is closed, the system resumes tracking the user's movement. If the user moves toward the door from the inside and reaches the defined threshold distance, an unlock signal is triggered and returned to the PC terminal. The system then checks whether this user is the last remaining individual inside the lab. If not, no further action is taken beyond unlocking the door. If the user is identified as the last UWB-tagged user, the PC terminal issues power-off commands to the equipment, lighting, and HVAC controllers. The system then verifies whether the power-off operation is successful. If it fails three consecutive times, a failure message is returned. If successful, all relevant equipment is powered off, and shutdown logs are recorded in the database. Finally, the access system automatically re-locks the door after a three-second delay.

E. Design of other System Functions

a) Intelligent device displacement alarm: In the asset management module, UWB tags are attached to laboratory equipment to enable real-time distance monitoring between the equipment and the UWB ranging devices. When an abnormal change in distance is detected-exceeding a predefined threshold-an alert is triggered by the laboratory's alarm system. This indicates that a device may have been moved or has exited its designated area. The system notifies laboratory staff to take immediate action, effectively preventing unauthorized relocation or theft of valuable assets and ensuring equipment security [19].

The intelligent displacement alarm mechanism is illustrated in Fig. 3.

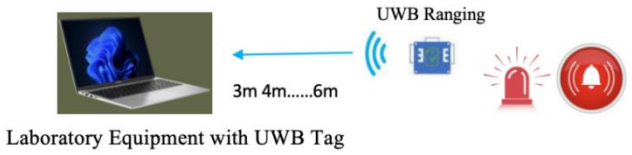


Fig. 3. UWB-based intelligent device displacement alarm.

b) Laboratory environmental linkage: Once the UWB-based access control system successfully unlocks the laboratory door, a series of predefined intelligent linkage mechanisms are triggered. The lighting system is activated and automatically switches to a preset brightness mode. The air conditioning system adjusts temperature and humidity levels based on predefined environmental thresholds.

Simultaneously, the laboratory equipment power management system is activated, initiating the startup of various experimental devices in a specified sequence. This ensures that all necessary equipment is fully operational and ready for subsequent experimental activities.

c) Intelligent zone safety warning: UWB-based ranging technology enables real-time and highly accurate measurement of the distance between personnel and potential hazard zones within the laboratory, such as areas containing high-voltage equipment or hazardous chemicals. When an individual approaches a predefined safety threshold, the system immediately triggers an audible and visual alarm to alert the person and ensure safety by discouraging further proximity to the danger zone. At the same time, a warning notification is sent to laboratory administrators [20, 21].

V. Implementation of the UWB-based Intelligent Laboratory Management System

A. Access Control System Functionality

The access control system is designed to manage the entry and exit of faculty and students in laboratory environments. The UWB controller retrieves device information from users carrying UWB-enabled smartphones or UWB tags and compares it with access authorization data stored in the cloud database. Administrators configure access rights and door operation time windows for each laboratory based on user roles. Authorized students and instructors can enter the laboratory during scheduled class hours simply by carrying their UWB-enabled device and walking towards the laboratory entrance. The UWB controller authenticates the user and performs real-time distance measurements. Once within the threshold range, the door unlocks automatically, enabling seamless, contactless entry. Outside of scheduled lab hours or in restricted laboratory areas, users without valid authorization-regardless of student or faculty status-are denied access by the system.

The data communication process involves interaction between the PC-based control terminal and the UWB controller over the network. This includes ranging operations, command transmission with MAC address binding, adding mobile device addresses to the UWB controller's whitelist, granting access and interaction permissions to mobile terminals, and enabling device-level interaction with the access control controller.

Additionally, the system supports real-time display and monitoring of access events-such as unlock and lock actions-performed by the UWB-based door lock, all visualized on the PC terminal [22, 23].

B. Asset Management Functionality

The asset management module includes daily management, intelligent device displacement alarms, and asset inventory statistics [24].

a) Daily management: Laboratory equipment is registered and linked to corresponding UWB tags. Information such as device identity, borrowing, return, and maintenance records is entered into the asset management system to improve administrative efficiency.

b) Intelligent device displacement alarm: The system periodically scans for UWB tags within the monitoring range. To avoid false alarms caused by transient signal fluctuations, a time tolerance mechanism is employed. An alarm is triggered only when the measured position deviates from the preset threshold-e.g., more than one meter-for a sustained duration exceeding three seconds.

c) Asset inventory statistics: The system automatically scans all UWB-tagged devices within the coverage area to assess their status and usage. By comparing this data with the system's equipment ledger, asset inventory can be completed quickly, significantly reducing manual workload while improving data accuracy and management efficiency.

C. Environmental Monitoring Functionality

The environmental monitoring module includes laboratory environment linkage and intelligent zone safety warnings.

a) Laboratory environment linkage: After the UWB-based access control system unlocks the laboratory, power-on commands are sent to controllers of lighting, air conditioning, and other equipment to initiate startup procedures. Additionally, lighting systems can be dynamically controlled based on a person's proximity to specific zones, enabling automatic switching on or off of designated lights.

b) Intelligent zone safety warning: Similar to the device displacement alarm, this function continuously measures the distance between individuals-carrying UWB tags or UWB-

enabled smartphones-and designated detection base stations. When a person enters a predefined hazardous zone, the system issues real-time audible and visual alerts, while simultaneously notifying administrators.

D. Experiment Management Functionality

The experiment management module includes experimental course management, course scheduling, and reservation services. Experimental course management supports the creation, administration, and publication of experimental content. Course scheduling intelligently arranges experimental sessions based on laboratory resource availability to optimize space utilization and avoid conflicts. The reservation function allows users to book laboratory time and space online according to specific needs, enhancing accessibility and operational efficiency.

E. User Management Functionality

The user management module consists of user information management, permission management, and usage statistics. User information management enables the entry, modification, and maintenance of users' basic data. Permission management allows access rights to be assigned according to user roles-such as administrators, instructors, and students-effectively controlling access to different system functions. Comprehensive usage statistics provide data-driven insights into user behavior and system utilization, supporting informed decision-making in laboratory management.

F. System Management Functionality

The system management module includes system operation settings, data backup, and overall system administration. Operation settings cover configuration items such as system run times, permission rules, and device status monitoring. Data backup ensures the security of critical information, including user records, experimental data, and equipment logs. Comprehensive system administration functions include log management and security policy configuration to maintain system stability and integrity.

VI. System Deployment and Effectiveness

The UWB-based intelligent laboratory management system has entered the pilot stage in the Modern Educational Technology Laboratory. Three representative laboratories were selected for implementation: the Smart Language Teaching Lab, the Interactive Teaching Lab, and the Cross-Cultural VR Training Lab. These labs collectively cover an area of approximately 300 square meters.

A total of 112 experimental devices-including VR headsets, portable workstations, and mobile projectors-were equipped with UWB identification tags. The hardware deployment within a single laboratory is illustrated in Fig. 4.

The upgraded laboratory supports five categories of public English experimental teaching and course projects, including College English Listening and Speaking, Reading and Writing and Translation, English Speech and Debate, and English Interpretation of Traditional Chinese Culture. These courses collectively serve over 1,500 faculty members and students. During the pilot phase, a total of 650 class hours of experimental courses were successfully conducted using the system.

After one and a half years of pilot operation, the system has achieved significant improvements in various aspects, including daily laboratory management, safety, energy efficiency, user experience, and talent development.

a) Daily laboratory management: The system has substantially reduced the workload of laboratory staff. For example, equipment inventory time has decreased from one hour of manual counting to automatic report generation within five minutes.

b) Laboratory safety management: During the trial period, the system successfully identified and recovered two misplaced devices through the displacement alarm function, achieving a 100% equipment security control rate. The UWB-based intelligent zone warning mechanism triggered eleven audible and visual alarms, effectively preventing faculty and students from entering high-voltage danger zones.

c) Energy efficiency and environmental sustainability: The lighting system automatically turns on in occupied areas and off in unoccupied ones. The HVAC system adjusts temperature and humidity based on predefined thresholds, while experimental devices automatically shut down after periods of inactivity. These features have led to a 25% reduction in overall energy consumption, significantly lowering operational costs for the university.

d) User experience: Access efficiency has greatly improved as faculty and students can enter using UWB-enabled smartphones or tags. The average door unlocking response time has decreased from 3 seconds (with traditional access systems) to just 0.5 seconds, achieving a seamless and contactless entry experience.

e) Enhancement of talent development: In 2023 and 2024, the laboratory supported autonomous learning for more than 1,500 students and met the self-directed experimental needs for training and participation in over 10 national-level competitions, including major university contests. This not only maximized the utilization efficiency of experimental equipment but also significantly enhanced students' language application skills and innovative abilities. Over the two-year period, students participating in these competitions achieved outstanding results, winning 81 national first prizes, 116 second prizes, and 325 third prizes.

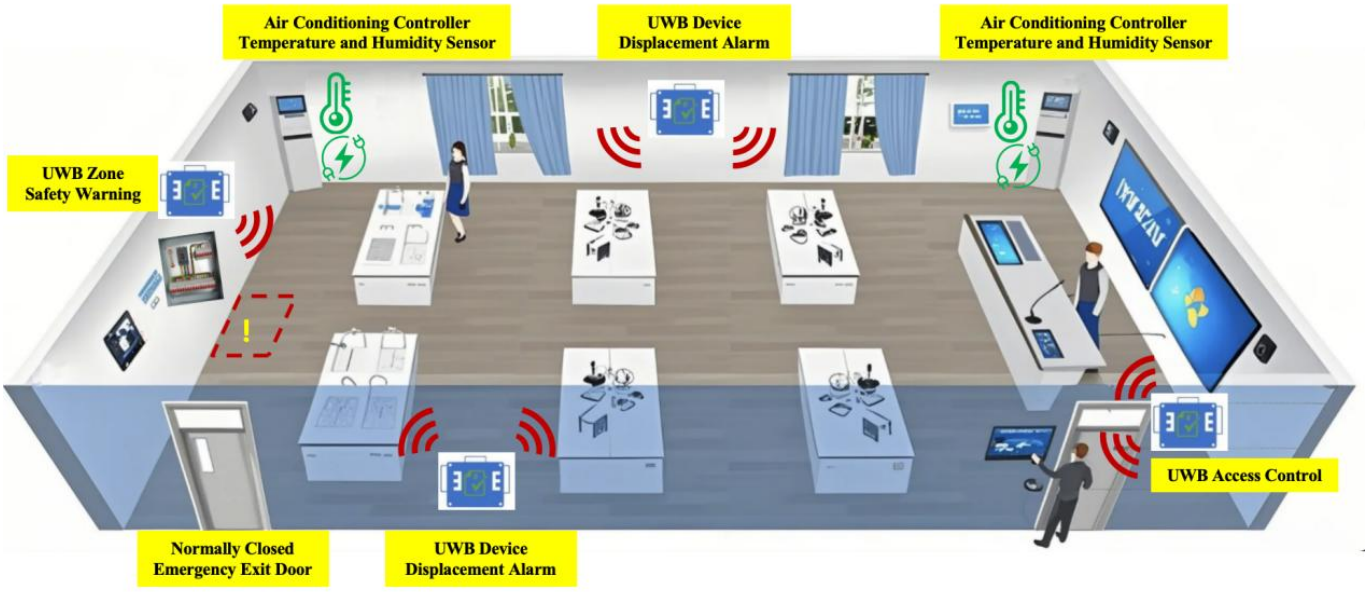


Fig. 4. Hardware deployment inside the intelligent laboratory.

VII. Discussion

A. Interpretation of Results

The deployment results of the UWB-based intelligent laboratory management system demonstrate that it significantly enhances key aspects of laboratory operations, including access efficiency, asset monitoring, energy conservation, and user experience. These improvements can be primarily attributed to the centimeter-level positioning accuracy and real-time tracking capabilities of UWB technology, which enable automated responses and reduce reliance on manual intervention. For example, the drastic reduction in equipment inventory time is a result of continuous UWB-based scanning and automated data logging, while energy efficiency gains stem from dynamic environment control triggered by real-time user proximity detection.

Moreover, the seamless integration of subsystems—access control, environmental regulation, asset tracking, and user management—under a centralized IoT architecture ensures not only operational convenience but also system scalability. This integration offers a holistic solution to the fragmented and disjointed management processes observed in traditional laboratory settings.

B. Comparison with Prior Work

Compared to conventional systems based on QR codes, RFID, or BLE, the proposed solution offers notable advantages in terms of accuracy, responsiveness, and automation. While BLE and RFID technologies are widely used, they often suffer from signal interference, limited range, and imprecise localization. In contrast, UWB supports high-resolution spatial tracking, which allows for intelligent zoning, safe distance monitoring, and more granular control over lab equipment and environmental settings.

Additionally, previous systems typically focus on a single management function, such as access control or environmental sensing. In this study, the integration of multiple modules—including real-time asset displacement alarms and intelligent safety zones—demonstrates how UWB can serve as the technological backbone for a unified smart laboratory framework. This positions the system as an innovative upgrade over existing solutions, with broader applicability in educational and research institutions.

C. Limitations

While the system exhibits promising performance, several limitations should be acknowledged. First, the deployment was limited to three laboratories within a single institution, which may constrain the generalizability of the findings to other settings or disciplines. Second, the evaluation focused primarily on operational efficiency and did not capture longer-term impacts such as cost-effectiveness, maintenance overhead, or learning outcomes. Third, despite its technical advantages, UWB deployment requires specialized hardware, calibration, and higher installation costs, which could pose barriers to adoption in resource-constrained environments. Future research should investigate scalable deployment strategies, integration with AI-driven analytics, and cross-campus interoperability to further enhance system applicability.

VIII. Conclusion

The intelligent laboratory management system based on UWB technology effectively addresses long-standing challenges in university laboratory administration. By leveraging UWB's high-precision positioning capabilities, the system standardizes access control procedures and enables efficient personnel management. Real-time distance data between individuals and equipment supports spatial layout optimization and dynamic resource scheduling, significantly enhancing the level of intelligent laboratory management.

Additionally, the system enables real-time asset tracking and proactive safety alerts through UWB-based detection, establishing a comprehensive security framework that

contributes to the safe and stable operation of laboratory environments.

The intelligent transformation of university laboratories is an ongoing process. In the future, we aim to further explore the integration of UWB with other emerging technologies, continuously refine system functionalities, and enhance system stability and compatibility. These efforts will support broader adoption across interdisciplinary laboratories, deliver higher-quality and more efficient services to faculty and students, and inject new momentum into educational reform and high-quality development in higher education.

REFERENCES

[1] F. P. Madrin, M. Klemm, and E. Supriyanto, "Reliability improvement of UWB tracker for hospital asset management system: Case study for TEE probe monitoring," in Proc. 2021 4th Int. Conf. Inf. Commun. Technol. (ICOIACT), pp. 1-6, 2021.

[2] M. F. R. Al-Okby, S. Junginger, T. Roddelkopf, and K. Thurow, "UWB-based real-time indoor positioning systems: A comprehensive review," Applied Sciences, vol. 14, no. 23, pp. 11005, 2024.

[3] K.Saritha et al., "IoT enabled hospital asset tracking using advanced interdisciplinary approaches," E3S Web of Conferences, 2024, Paper 01007.

[4] T. Margiani, S. Cortesi, M. Keller, C. Vogt, T. Polonelli, and M. Magno, "Angle of arrival and centimeter distance estimation on a smart UWB sensor node," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-10, 2023.

[5] A. Kumar, S. Jain, and M. Kumar, "Face and gait biometrics authentication system based on simplified deep neural networks," Int. J. Inf. Technol., vol. 15, pp. 1005-1014, 2023. https://doi.org/10.1007/s41870-022-01087-5.

[6] D. M. Nixon Dutta, "A tracking solution of IT assets and resources management," Int. J. Math. Comput. Res., vol. 10, no. 7, pp. 2822-2831, 2022. https://doi.org/10.47191/ijmcr/v10i7.08.

[7] H. Dudycz, M. Hernes, Z. Kes, E. Mercier-Laurent, B. Nita, K. Nowosielski, et al., "A conceptual framework of intelligent management control system for higher education," in Artif. Intell. Knowl. Manage.: 8th IFIP WG 12.6 Int. Workshop, AI4KM 2021, Yokohama, Japan, 2021, pp. 35-47. https://doi.org/10.1007/978-3-030-80847-1_3.

[8] J. A. Martin, I. Schröder, and C. A. Merlic, "Proceedings of the 2021 Workshop on Laboratory Safety: Advancing Safety in Teaching and Research Laboratories," ACS Chem. Health Saf., vol. 29, no. 2, pp. 124-131, 2022. https://doi.org/10.1021/acs.chas.2c00003.

[9] A. Yuhanef, Y. Yulindon, D. Meidelfi, and M. Silvana, "The small UWB monopole antenna with stable omnidirectional radiation pattern," Int. J. Inform. Vis., vol. 6, no. 4, pp. 815-820, 2022. https://doi.org/10.30630/joiv.6.4.972.

[10] A. Volpi, L. Tebaldi, G. Matrella, R. Montanari, and E. Bottani, "Low-cost UWB based real-time locating system: Development, lab test, industrial implementation and economic assessment," Sensors, vol. 23, no. 3, pp. 1124, 2023.

[11] B. V. Krishnaveni, A. Venkatalakshmi, and B. R. Kumar, "Localization for IoT Based Systems Using UWB," in Contemp. Perspect. Sci. Technol. Res., vol. 2, pp. 12-26, 2022. https://doi.org/10.9734/bpi/cpstr/v2/11383f.

[12] J. Shackelford, P. Mathew, C. Regnier, and T. Walter, "Laboratory validation of integrated lighting systems retrofit performance and energy savings," Energies, vol. 13, no. 13, pp. 3329, 2020.

[13] N. M. Hussein, Y. M. Mohialden, and S. A. Salman, "Impact of IoT-based environmental monitoring on lab safety and sustainability," Babylon. J. Internet Things, vol. 2024, pp. 16-26, 2024. https://doi.org/10.58496/BJIoT/2024/003.

[14] Z. Changhong and X. Reneng, "Intelligent laboratory access control system based on ZigBee technology," in 2020 Int. Conf. Virtual Reality Intell. Syst., Zhangjiajie, China, 2020, pp. 688-690. https://doi.org/10.1109/ICVRIS51417.2020.00169.

[15] T. M. T. Dinh, N. S. Duong, and K. Sandrasegaran, "Smartphone-based indoor positioning using BLE iBeacon and reliable lightweight fingerprint map," IEEE Sens. J., vol. 20, no. 17, pp. 10283-10294, 2020. https://doi.org/10.1109/JSEN.2020.2989411.

[16] S. Salimi, S. Salimpour, J. P. Queralta, W. M. Bessa, and T. Westerlund, "Benchmarking ML Approaches to UWB-Based Range-Only Posture Recognition for Human Robot-Interaction," IEEE Sens. J., vol. 25, no. 1, pp. 1350-1358, 2024. https://doi.org/10.1109/JSEN.2024.3493256.

[17] S. Sung, H. Kim, and J. I. Jung, "Accurate indoor positioning for UWB-based personal devices using deep learning," IEEE Access, vol. 11, pp. 20095-20113, 2023.

[18] J. Shackelford, P. Mathew, C. Regnier, and T. Walter, "Laboratory validation of integrated lighting systems retrofit performance and energy savings," Energies, vol. 13, no. 13, pp. 3329, 2020. https://doi.org/10.3390/en13133329.

[19] N. Karimpour, Z. A. Dagdevıren, V. Akram, and O. Dağdeviren, "A cloud-based asset tracking system for hospitals using ultra-wideband localization," Avrupa Bilim ve Teknol. Derg., vol. 26, pp. 425-430, 2021. https://doi.org/10.31590/ejosat.960454.

[20] A. Volpi, L. Tebaldi, G. Matrella, R. Montanari, and E. Bottani, "Low-Cost UWB Based Real-Time Locating System: Development, Lab Test, Industrial Implementation and Economic Assessment," Sensors, vol. 23, no. 3, pp. 1124, 2023. https://doi.org/10.3390/s23031124.

[21] S. Sung, H. Kim, and J. I. Jung, "Accurate indoor positioning for UWB-based personal devices using deep learning," IEEE Access, vol. 11, pp. 20095-20113, 2023. https://doi.org/10.1109/ACCESS.2023.3250180

[22] K. Reaz et al., "UWB Security and Enhancements," TechRxiv, 2024. https://doi.org/10.36227/techrxiv.171085006.63940566/v1.

[23] D. Coppens, A. Shahid, S. Lemey, B. Van Herbruggen, C. Marshall, and E. De Poorter, "An overview of UWB standards and organizations (IEEE 802.15.4, FiRa, Apple): Interoperability aspects and future research directions," IEEE Access, vol. 10, pp. 70219-70241, 2022. https://doi.org/10.1109/ACCESS.2022.3187410.

[24] N. Karimpour, Z. A. Dagdeviren, V. Akram, and O. Dagdeviren, "A cloud-based asset tracking system for hospitals using ultra-wideband localization," Avrupa Bilim ve Teknoloji Dergisi, vol. 26, pp. 425-430, 2021.

RFID Integration with Internet of Things: Data Processing Algorithm Based on Convolutional Neural Network

Liang Wang*
School of Electrical Engineering, Hunan Mechanical & Electrical Polytechnic, Changsha, Hunan 410151, China

***Abstract*—Radio Frequency Identification is a fast and reliable communication module that performs automatic data capture to identify and track individual objects and people. Frequency-coded tags employ resonant networks to decode their unique code. A multi-scatterer or multi-resonant method encodes the data. Research primarily related to the current investigation predicted that the chipless RFID tag resonant network has a high bit encoding capacity. This study addresses the simulation, optimization, fabrication, testing, and data encoding methods for chipless RFID tags. This research provides a framework for the open-ended quarter-wavelength stub multi-resonator method in chipless Radio Frequency Identification (RFID) tags. The proposed design enhances the tag's data encoding capacity and improves its robustness to ecological differences. This study integrates Error Correction Coding (ECC) and Adaptive Modulation Systems (AMS) employing Convolutional Neural Networks (CNN) to enhance the tag's performance. The AMS dynamically alters the modulation parameters based on channel states, while ECC improves data reliability. The results indicate efficient performance compared to traditional chipless RFID tags, highlighting the possibility of practical behavior in typical applications that necessitate reliable and high-capacity data transmission.**

***Keywords*—RFID; chipless; coding; threshold; data transmission; error correction; security authentication**

I. Introduction

Radio Frequency Identification is referred to as RFID. An RFID comprises a reader and one or more tags or transponders. RFID is a new technology. The British radar system employed RFID for the first time in World War II to differentiate between German aircraft and their aircraft that had radio transponders attached. Several industries have benefited from the advantages of RFID, including the military, healthcare, security, sports, animal husbandry, and aviation. RFID has several essential sector-specific applications. RFID is utilized in various applications, including supply chain management, automated payment systems, vehicle tracking, inventory management, secure shop checkouts, animal monitoring, and sports timing technology. RFID transfers data from an electronic tag, frequently known as a label or an RFID tag, using radio waves. Selecting an RFID tag is required. There are numerous frequencies of RFID tags available on the market. Microwave, High Frequency, Low Frequency, and Ultra High Frequency are all included [1-5].

*Corresponding Author

Organizations can now afford to use RFID due to technological advancements in cost factors. Wal-Mart was one of the first industries to use RFID technology for theft prevention and inventory supply process improvements. Wal-Mart mandated that its top 100 suppliers apply RFID tags to the pallets and cases they sent to the company by the end of 2004, marking the start of deploying RFID across its retail distribution chain. However, in 2004, RFID tag producers were unable to meet the demand level within Wal-Mart's targeted timeframe due to the state of standards and limited tag manufacturing capabilities.

With most of their leading suppliers supplying RFID-tagged pallets and crates to every Wal-Mart distribution facility, Wal-Mart's adoption is now well underway as standards change, manufacturing capabilities increase, and price points decline. Although RFID is beginning to infiltrate other business supply chains, these implementations are frequently closed-loop. To put it another way, RFID is implemented within the supply chain of a particular company but not between them and their commercial trade partners. Other applications, including tracking equipment to ensure that it is not left behind in the aircraft after repair, can only be used within a single business [6-10].

Information is stored on transponder tags in all RFID systems. RFID tags hold data that may be as vast as an array of bits or as little as a single binary bit that indicates an identifying code, Aadhaar number, personal medical data, or anything else that can be recorded in digital binary format. Microchips power RFID tags. RFID tags within a specific range may get sufficient power from the field. It transmits the data it has saved and accesses its internal memory using this power [3-4]. The transceiver antenna's voltage increases due to the interaction between the RF fields when the Transponder Tag draws power in this manner. The purpose of the Tag is to convey data to the reader by using this sort of impact. The Tag regulates how much energy is reserved from the field. The Tag controls the voltage detected at the transceiver following the bit pattern it wants to send [11-15].

RFID is an automated data collection that recognizes and monitors individual objects and people in real-time using RFID signals [16-20]. It is used as a radio technology to realize the Internet of Things (IoT) vision of generating a universal system of smart objects [21-26]. It would be possible to mark and invent all items fitted with radio tags. RFID tags must be data-rich, low-cost, and energy-efficient to perform these functions.

The RFID idea was introduced during World War II to find enemy aircraft [27-30].

To distinguish between friend and foe, the radar transmitter receives a backscattered signal from a receiver tagged in an airplane [31-32]. Later, in 1945, Leon Theremin made an intriguing discovery of RFID using the backscattering principle. As a result of the development of RFID, the commercial potential for tracking and identifying objects without physical intervention has increased. Furthermore, RFID can handle a greater number of items than barcodes [33-36]. RFID's main advantages are flexibility and versatility, and it promises to be a future technology in automatic designation [37-40].

RFID and bar code technologies are highly similar, but instead of using a laser to read barcodes on labels, radio frequency waves are used to read information from tags [41-45]. Due to its wide range of applications in various areas, including supply chain management, healthcare, traffic management, commerce, access control, purchasing, and distribution logistics, as well as manufacturing and service industries, substantial research has been conducted in the area of RFID in recent years [46-50]. The RFID tag stores a distinct identification number parallel to a bar code or magnetic strip on the back of a credit card or ATM card. The classification must be processed close to its scanning device to retrieve bar code or magnetic strip data. However, because of the high cost of the silicon chip required, its general application is limited. As a result, low-cost alternatives are in high demand. RFID tags without silicon chips are now being created. These tags are called 'chipless' tags [51-55].

A. Motivation

This research aims to develop a robust and inexpensive sensor network that can integrate various types of sensors, including electrical, chemical, and biological sensors. Wireless sensor tags have historically been used in other settings as active or semi-passive tags for collecting wireless data. A tag's application area is limited by its need for a battery, which reduces its lifespan and increases its cost. Many academics studying passive chipless sensor tags in recent years have been interested in passive RFID sensor tags due to the challenges they face in their development. This study develops two chipless sensor tag designs and a transmission line reflection-based ID generation system, which proposes a wide range of sensor types for evaluation, including nanoparticle-based sensors, microstrip antennas on a flexible substrate, and previously reported transmission delay line-based ID generation schemes. Dipole-based antennas are a standard and attractive option for passive tag designers from a business standpoint; however, their primary drawback is their length. Long straight dipole antennas are miniaturized using folding and meandering methods instead of long straight dipole arms. A tiny tag with excellent reading ranges and comprehensive band behavior is the product of these innovations. Future work aims to further improve the quality, efficiency, and applicability of the sensor tag system developed through this research.

1) Problem statement and research objective. A chipped RFID tag is composed of integrated circuits that store the identifying codes. The price of this tag is higher due to the use of silicon chips. An alternative solution to this difficulty is to develop a chipless RFID tag, which does not contain a silicon chip. The primary challenge in developing a chipless RFID tag is determining how to encode data without a memory chip, which has limited capacity. Multi-resonator-based tags are the combination of multiple resonator circuits and transcoding antenna structures. These antenna networks are called microstrip antennas. The efficient realization of the multi-resonator structure can be built using these microstrip antennas. However, the inherent drawbacks of these microstrip antennas are low gain and narrow bandwidth. The polarization antennas are utilized to address the drawback, and the significant segment is chipless RFID tags, which are retransmission-based multi-resonator tags. The network is designed to address the problems associated with the prominent recognition of chipless RFID tags.

The study on the chipless RFID is written as follows: the overview of RFID, chipless RFID, motivation, and problem statement with research objective are detailed in Section I. The literature review is discussed with a research gap in Section II. The proposed CNN-based coding system is given in Section III, the simulation result is presented with detailed discussion in Section IV, and the study ends with a justified conclusion in Section V.

II. Related Works

This section explains various studies on chipless RFID. These investigations explore various features, including novel encoding methods, Machine Learning (ML) integration, and system reliability enhancements (see Table I). In [5-6-60], the authors introduce a unique encoding for RFID tags, addressing data capacity limitations. In [61], the authors view the proposals' insights into ML's role in chipless RFID, highlighting challenges and proposing standardized data protocols. In [62], the authors employ ML and DL for robust data reading in CRFID sensor tags. In [63], the authors present a workflow for ML classification in chipless RFID tag identification. In [64], the authors introduce a multi-tag detection system using 1D-CNN. In [65], the authors investigate Deep Learning (DL) to improve RFID tag detection accuracy. In [66], the authors evaluate ML models for identifying information in chipless RFID tags. In [67], the authors propose a phase coding-based RFID tag design, while [68] examines chipless RFID for material identification. In study [69], data-driven methods are suggested to improve chipless RFID reliability.

The research gap identified in the review is the limited resolution of multi-tag detection systems [70]. Conventional frequency-based systems face limitations in achieving high-resolution multi-tag detection, thereby limiting their practical applicability. The application of CNN overcomes these challenges by improving reliability and accuracy in multi-tag detection. CNN leverages advanced pattern recognition capabilities to analyze complex data sets, enabling more precise and efficient tag detection even in challenging environments. This development represents a significant leap forward in chipless RFID technology, addressing the critical research gap

of achieving high-resolution multi-tag detection for practical deployment in numerous applications.

TABLE I SUMMARY OF ML-BASED RFID

Inference	Methods used	Outcome	Drawback
Proposes a novel encoding for chipless RFID tags using N resonant elements with M individual desired frequencies, thereby overcoming the data encoding capacity limitation. PSO algorithm optimizes tag structure parameters.	Particle Swarm Optimization (PSO) algorithm	Overcomes data encoding capacity limitations, enabling the determination of resonant frequencies before the tag model.	Requires optimization for specific tag designs.
Provides a comprehensive review of ML in CRFID technology, highlighting integration challenges and suggesting standardized data protocols for improved ML deployment.	Review of ML approaches, identification of challenges, proposal for standardized data protocols	Identifies challenges in ML integration with CRFID and suggests standardizing data protocols for improved deployment.	The lack of standardized data protocols may hinder ML deployment.
Introduces ML and DL regression modelling for robust CRFID sensor tag data reading—automated data acquisition training on large-scale datasets. 1-D CNN outperforms conventional ML models.	ML and DL regression modelling.	Achieves high accuracy in ID and sensing data detection and outperforms conventional ML models.	Restricted resolution of 3 cm for multi-tag detection.
Presents workflow for ML classification in chipless RFID tag identification, achieving perfect accuracy for four tags under fixed and flexible ranges.	ML classification workflow, optimization of tag responses	Achieves perfect accuracy for tag identification under various ranges and enhances tag responses.	Limited to four tag identifications.
Introduces chipless RFID multi-tag detection system using 1D CNN, achieving 100 per cent accurate multi-tag detection.	1D CNN for multi-tag detection, analysis of CNRs and natural frequencies	Achieves 100 per cent accurate multi-tag detection, outperforming frequency-based systems.	Restricted resolution of 3 cm for multi-tag detection.
Investigates DL techniques to improve chipless RFID tag detection accuracy, achieving roughly 99.99 per cent accuracy.	Evaluation of DL techniques	Achieves significant improvement in detection accuracy, reaching near-perfect accuracy.	Limited discussion on practical implementation challenges.
Evaluate ML models for detecting identification and sensing information in RCS-based chipless RFID tags, achieving high accuracy with SVR and GBT models.	Evaluation of ML	Achieves high accuracy in identification and sensing data detection.	Lack of discussion on real-world applicability.
Proposes a phase coding-based chipless RFID tag design featuring a C-shaped resonator capable of achieving 243 distinct states.	Proposal of phase coding-based chipless RFID tag design	Achieves high coding capacity with smaller tag dimensions.	Limited experimental validation was provided.
Examines the utility of chipless RFID technology for material identification, achieving over 90 per cent accuracy and faster detection speed compared to literature methods.	Examination of chipless RFID technology for material identification	Achieves high accuracy and faster detection speed, demonstrating practical utility.	May sacrifice accuracy for improved detection speed.
Proposes data-driven methods to enhance chip-less RFID interrogation and reliability, achieving high validation accuracy and increased reading range.	Proposal of data-driven methods, analysis of chip-less RFID tags	Achieves high validation accuracy and increased reading range, enhancing system reliability	It may require further optimization for specific environments.

III. PROPOSED CHIPLESS RFID TAGS IN ADAPTIVE MODULATION AND ERROR CORRECTION CODING USING CNN

This section deals with the design and development of retransmission-based resonating structures and encoding of the resonating structures. The three key elements of the retransmission-based tags are the transmitting (T_x) and receiving (R_x) antennas, as well as multi-resonating structures. The multi-resonator network plays a vital part in frequency-coded tags. It is a combination of multiple narrow-band rejection filtering sections, each comprising several resonators, used to encode the tag identity. These resonating structures will alter the interrogating signal's amplitude. The filter's performance is coupled to a transmission line, which sends the encrypted signal data to the reader.

The multi-resonator-based tags are proposed on microstrip-based planar resonators, which are either coupled or connected to the transmission line. The different networks of multi-resonators are presented in this section. The designed multi-resonating structures depends on the operating frequency band, the separation between the resonators, and the properties of the substrate. The multi-resonator consists of multiple resonators in the microstrip line, and each resonator resonates at its quarter wavelength frequency. The designed multi-resonators are

simulated on a low-cost FR4 substrate with a dielectric constant of 4.4, a loss tangent of 0.0018, and a height of 1.6 mm using an Advanced Design System (ADS). The bottom plane acts as a ground plane [71-76].

Open-ended quarter-wavelength stub multi-resonators are designed to overcome the size limitation of the above "L" shaped arm multi-resonator. To solve the constraints of short stub multi-resonators, several open stub multi-resonators have been proposed. A simple, open-ended quarter-wavelength multi-resonator is proposed. As a result, the tag size will be reduced by 50 per cent. The ten stubs are connected to a 50-*ohm* impedance microstrip transmission line in the multi-resonator. The microstrip transmission line width (3 *mm*), as in Eq. (1) and Eq. (2):

$$\frac{W}{h} = \begin{cases} \frac{8e^A}{e^{2A}-2} & for \frac{W}{h} \le 2 \\ \frac{2}{\pi}\left[B - 1 - \ln(2B-1) + \frac{\varepsilon_{r-1}}{2\varepsilon_r}\left\{\ln(B-1) + 0.39 - \frac{0.61}{\varepsilon_r}\right\}\right] & for \frac{W}{h} \ge 2 \end{cases} \quad (1)$$

where,

$$A = \frac{z_0}{60}\sqrt{\frac{\varepsilon_r+1}{2}} + \frac{\varepsilon_r-1}{\varepsilon_r+1}\left(0.23 + \frac{0.11}{\varepsilon_r}\right) \text{ and } B = \frac{377\pi}{2z\sqrt[0]{\varepsilon_r}} \quad (2)$$

The length (17 *mm*) of the microstrip transmission line by Eq. (3) and Eq. (4):

$$\emptyset = 90^\circ = \beta L_f = \sqrt{\varepsilon_{eff}} K_o L_f \quad (3)$$

$$L_f = \frac{90^\circ\left(\frac{\pi}{180^\circ}\right)}{\sqrt{\varepsilon_{eff}} K_o} \quad (4)$$

where,

- $K0$→ The phase constant and is equal to 2π/λ.

The open-ended quarter-wavelength resonator behaves as a series RLC resonant circuit. So, each resonator is modelled as a quarter wavelength.

The input impedance of a lossy transmission line is given in Eq. (5):

$$Z_{in} = Z_0 \frac{Z_L + Z_o tanh\gamma l}{Z_o + Z_L tanh\gamma l} \quad (5)$$

where, $\gamma = \alpha + j\beta$

For an open-circuited lossy transmission line, the input impedance is given in Eq. (6) and Eq. (7):

$$Z_{in} = Z_0 \coth(\alpha + j\beta)\, l \quad (6)$$

$$Z_{in} = Z_0 \frac{1 + jtan\beta l tanh\alpha l}{tanh\alpha l + jtan\beta l} \quad (7)$$

In real-time applications, the transmission line has a slight loss, so $\alpha l \ll 1$ and $tanh\alpha l \approx \alpha l$

Assume

$$\omega = \omega_o + \Delta_\omega \quad (8)$$

where, is Δ_ω Small [see Eq. (8) and Eq. (9)].

$$\beta l = \frac{\omega l}{v_p} = \frac{\omega_o l}{v_p} = \frac{\Delta\omega l}{v_p} = \frac{\pi}{2} + \frac{\pi\Delta\omega}{2\omega_o} \quad (9)$$

Then

$$\cot\beta l = \cot\left(\frac{\pi}{2} + \frac{\pi\Delta\omega}{2\omega_o}\right) = -tan\left(\frac{\pi\Delta\omega}{2\omega_0}\right) = \frac{\pi\Delta\omega}{2\omega_0} \quad (10)$$

as in Eq. (10). From Eq. (11),

$$Z_{in} \approx Z_o \frac{1 + j\alpha l\frac{\pi\Delta\omega}{2\omega_0}}{\alpha l + j\frac{\pi\Delta\omega}{2\omega_0}} \approx Z_o\left[\frac{1}{\alpha l} + j\frac{\pi\Delta\omega}{2\omega_0}\right] \quad (11)$$

It indicates the input impedance of a series RLC resonant circuit. The quality factor is given in Eq. (12):

$$Q = \frac{L\omega_o}{R} = \frac{\pi}{4\alpha l} \approx \frac{\beta}{2\alpha} \quad (12)$$

The proposed design has to generate resonance by optimizing the length of an open-ended stub resonator. Each resonator length is computed as in Eq. (13):

$$W_t = \frac{\lambda_g}{4} \quad (13)$$

where,

- Wt→ The open stub resonator length at 2.08 GHz resonant frequency
- The open-ended stub resonator length is approximately 20.8 mm.

Other stub lengths are computed using the same procedure and the values (see Table II). The proposed multi-resonating structure is implemented on an FR4 substrate with a dielectric constant of 4.4 and a loss tangent of 0.0018. The conductor thickness is 35 μm. The mutual coupling of open stub resonators is reduced when the stubs are held 1 *mm* apart. The bottom plane serves as a ground plane. The span of the proposed design is 23.8 $mm \times 17\ mm$.

TABLE II LENGTH OF THE 10-BIT OPEN STUBS

Stubs	Resonant Frequency (GHz)	Length of the open stubs (mm)
1	2.08	20.8
2	2.22	18.5
3	2.36	17.4
4	2.59	15.8
5	2.80	14.5
6	2.96	13.7
7	3.19	12.5
8	3.41	11.5
9	3.68	11
10	3.84	10.4

As the open stub microstrip line resonator resonates at the desired frequency, the wavelength remains $^1/_4$ of its original wavelength ($^{\lambda g}/_4$). Because of its inductive solid reactance, the resonator has a good quality factor, which prevents nearby resonant frequencies from merging. It is an actual electric current induced by an applied electromagnetic field. It shows the maximum current at the resonant frequency of 2.08 *GHz*. For efficient chipless RFID tag realization, the space limitation of open-stub multi-resonators can be replaced by meandered-line multi-resonators. The procedure of chipless RFID design is given in Algorithm 1.

Algorithm 1. Muti-Resonator Based Chipless RFID Design

1. DefineKeyConstituents():
- Define Tx and Rx antennas.
- Define multi-resonating structures.
2. DesignMultiResonatorStructure():
- Combine multiple narrow-band rejection filtering sections using several resonators.
- Simulate the designed multi-resonators on a substrate.
3. DesignOpenEndedStubs():
- Design open-ended quarter-wavelength stub multi-resonators.
- Calculate the dimensions of the microstrip transmission line.
4. ModelResonators():
- Model each resonator as a quarter wavelength to behave as a series RLC resonant circuit.
- Determine the input impedance of the lossy transmission line.
5. OptimizeResonatorLengths():
- Optimize the length of each open-ended stub resonator to generate resonance.
- Calculate resonator lengths.
6. ImplementMultiResonatingStructure():
- Implement the proposed multi-resonating structure on the substrate.
7. VerifyResonantBehavior():
- Verify the resonators' resonant behaviour and quality factor at desired frequencies.
8. ConsiderAlternativeDesigns():
- Consider alternative designs such as meandered line multi-resonators if necessary.

Main():
- DefineKeyConstituents()
- DesignMultiResonatorStructure()
- DesignOpenEndedStubs()
- ModelResonators()
- OptimizeResonatorLengths()
- ImplementMultiResonatingStructure()
- VerifyResonantBehavior()
- ConsiderAlternativeDesigns()

Collect data from chipless RFID tags, including the response patterns generated by multi-resonator networks. Encode the collected data into a format suitable for training the CNN. This encoding may involve mapping the RFID tag responses to specific modulation and coding methods. The signal response is modelled using Eq. (14):

$$y(t) = \sum_{n=1}^{N} A_n \cdot \sin(2\pi f_n t + \emptyset_n) \quad (14)$$

The adaptive amplitude and phase modulation are given in Eq. (15) and Eq. (16).

$$A_n = f(A_{n-1}, \theta_n) \quad (15)$$

$$\emptyset_n = f(\emptyset_{n-1}, \theta_n) \quad (16)$$

Eq. (14) is the signal response modelling of a chipless RFID tag. The received signal '$y(t)$' is expressed as a summation of sinusoidal components, where each component 'n' is characterized by its amplitude 'A_n', frequency 'f_n', and phase 'ϕ_n'. It captures the fundamental behavior of the tag's response to an interrogating signal.

The FE from the multi-resonator RFID signal, Eq. (17):

$$X = [x_1, x_2, x_3, \dots\dots\dots, x_n] \quad (17)$$

In Feature Extraction (FE) [Eq. (17)], the extracted features from the RFID signal are represented as a vector $X = [x_1, x_2, x_3, \dots\dots\dots, x_n]$. To provide the signal's encoding in a feature space compatible with the next processing step, each xi represents a distinct FE from the RFID data.

The convolutional layer is a key component of coding for multi-resonator chipless RFID tags, as it analyses the data signal input sent from the tags. The Neural Network (NN) can extract valuable features and patterns from the RFID tag data due to the convolution techniques applied to the input data in this layer. The convolution layer utilizes filter learning to train the neural network (NN) to detect specific frequency resonances and store them in a format suitable for future investigation. To integrate non-linearities into the NN, an activation function is used following the convolutional layer. The system's ability to detect minor variations in the RFID signal is boosted by this non-linearity, allowing it to identify significant variations and connections within the resonator responses. The features are transmitted forward to the Fully Connected (FC) layer, which is then extracted and analyzed by the convolutional layers, where they are flattened. To ensure a deep examination and summary of the extracted features, all neurons in this layer are connected to all the neurons in the layer before it. Here, researchers enhance the input signal description by adjusting weights and biases, enabling the provision of an accurate output prediction. The proper recognition and decoding of data encoded within the chipless RFID system are rendered possible by the valuable information obtained through the resultant prediction from the FC layer about the features and IDs of the RFID tags. From Eq. (18) to Eq. (22), this study tests CNN.

$$Z^{[1]} = W^{[1]} \times X + b^{[1]} \quad (18)$$

$$A^{[1]} = g\left(Z^{[1]}\right) \quad (19)$$

$$Z^{[2]} = W^{[2]} A^{[1]} + b^{[2]} \quad (20)$$

$$A^{[2]} = g\left(Z^{[2]}\right) \quad (21)$$

$$\hat{Y} = A^{[2]} \quad (22)$$

Eq. (18) to Eq. (22) describe the operations within a CNN. Eq. (18) calculates the activation of neurons in the convolutional layer by convolving the input data 'X' with learnable filters $W^{[1]}$ and adding a bias term $b^{[1]}$. The activation is then passed using an activation function gg [Eq. (6)]. In Eq. (7), the FC layer computes the activation using weights $W^{[1]}$ and bias term $b^{[2]}$. The activation is again passed through an activation function in Eq. (22).

$$C = E \times G \quad (23)$$

$$\hat{E} = f(Y) \quad (24)$$

$$d = dist(Y, \hat{Y}) \quad (25)$$

Eq. (23) to Eq. (25) pertain to ECC. Eq. (23) involves generating a codeword 'C' from an information vector 'E' by multiplying it with a generator matrix G. Eq. (24) performs error correction by applying a function 'f' to the received vector 'Y' to predict the data vector '$\widehat{E}$'. The Hamming distance 'd' between the received and computed vectors is computed in Eq. (25). Adaptive modulation is given in Eq. (26) to Eq. (28).

$$P_{out} = P_{in} \times \eta \tag{26}$$

$$BER = f(P_{out}, SNR) \tag{27}$$

$$ModScheme = g(BER) \tag{28}$$

Adaptive modulation in chipless RFID adjusts power (P_{out}) via efficiency factor η, calculates Bit Error Rate (BER) from P_{out} and SNR, and selects modulation scheme based on BER $ModScheme$. It optimizes power usage and modulation for reliable communication in dynamic RFID. The procedure for error correction using a CNN is outlined in Algorithm 2.

Algorithm 2: ECC Scheme using CNN

1. Collect data from chipless RFID tags, including response patterns from multi-resonator structures.
2. Encode collected data into a format suitable for training the CNN model.
3. Define signal response modeling function: $\boldsymbol{y(t) = \sum_{n=1}^{N} A_n \cdot sin(2\pi f_n t + \emptyset_n)}$
4. Extract features from the RFID signal: $\boldsymbol{X = [x_1, x_2, x_3, \dots\dots\dots, x_n]}$
5. Define convolutional layer operations: $\boldsymbol{Z^{[1]} = W^{[1]} \times X + b^{[1]}}$
$\boldsymbol{A^{[1]} = g(Z^{[1]})}$
6. Define FC layer operations: $\boldsymbol{Z^{[2]} = W^{[2]} A^{[1]} + b^{[2]}}$
$\boldsymbol{A^{[2]} = g(Z^{[2]})}$
7. Obtain output prediction: $\boldsymbol{\widehat{Y} = A^{[2]}}$
8. Define ECC operations:
$\boldsymbol{C = E \times G}$
$\boldsymbol{\widehat{E} = f(Y)}$
$\boldsymbol{d = dist(Y, \widehat{Y})}$
9. Define adaptive modulation operations:
$\boldsymbol{P_{out} = P_{in} \times \eta}$
$\boldsymbol{BER = f(P_{out}, SNR)}$
$\boldsymbol{ModScheme = g(BER)}$

IV. Results and Discussion

By employing PyTorch, developers can build a CNN by describing layers, activation functions, and additional features that enables you to set up the error correction coding algorithm. For effective data handling, use PyTorch's Dataset and DataLoader classes. Use a dataset that includes behaviour patterns from multi-resonator chipless RFID tags. Select an appropriate optimization algorithm (Adam) and loss function (CrossEntropyLoss) and set up a training loop to evaluate the dataset, compute gradients, and adjust model parameters. Model performance can be optimized by adjusting hyperparameters such as learning rate and batch size. To set up error correction coding in multi-resonator chipless RFID tags, follow these steps and utilize PyTorch's features to develop and train a Convolutional Neural Network (CNN). Eq. (29) to Eq. (33) are employed to assess the performance of the CNN-based ECC.

$$SNR = \frac{Signal\ Power}{Noise\ Power} \tag{29}$$

$$BER = \frac{Number\ of\ Bit\ Errors}{Sum\ of\ Bits} \tag{30}$$

$$SER = \frac{Number\ of\ Symbol\ Errors}{Sum\ of\ Symbols} \tag{31}$$

$$FER = \frac{Number\ of\ Frame\ Errors}{Sum\ of\ Frames} \tag{32}$$

$$Capacity = \frac{Data\ Rate}{bandwidth} \tag{33}$$

Key metrics of communication network performance are presented by Eq. (30) to Eq. (33). Eq. (29) defines the Signal-to-Noise Ratio (SNR) as a basic measurement used to measure signal quality since it measures the ratio of signal power to noise power. BER, which is an essential statistic for measuring transmission accuracy, is computed by Eq. (30). BER is the ratio of error bits to all bits received. Eq. (31) can be employed to derive the Symbol Error Rate (SER), which measures the accuracy of symbols at the symbol level. It is the ratio of wrong received symbols to the complete transmitted symbols. Frame Error Rate (FER) is a vital factor in assessing data packet integrity; it is computed by Eq. (32) and is described as the ratio of wrongly received bytes to the number of transmitted bytes. As the final step in determining manipulative channel efficiency, Eq. (33) computes size, which represents the highest possible data rate that can be achieved with a particular frequency (see Table III). The optimization and evaluation of communication system efficacy in various scenarios rely on these metrics.

TABLE III Comparison of Performance

Performance Measures (bits)	PSO	1D-CNN	MR-CNN
SNR	454.78	678.67	767.78
BER	567.23	432.13	232.13
SER	678.56	612.67	412.78
FER	567.56	512.68	354.78
Capacity	31.23	45.4	53

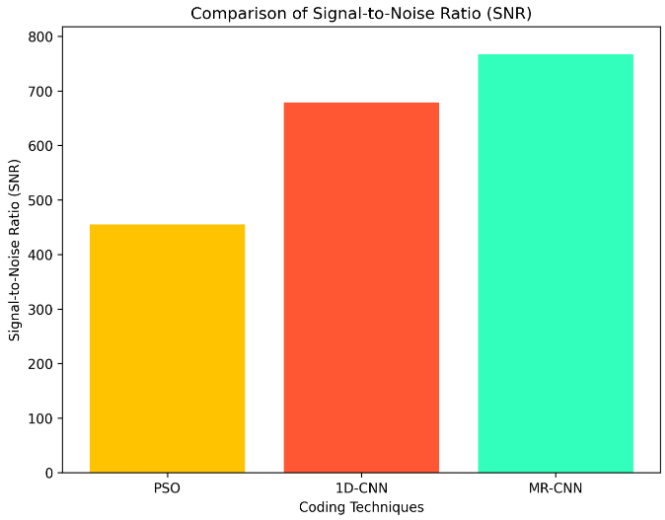


Fig. 1. Comparison of SNR

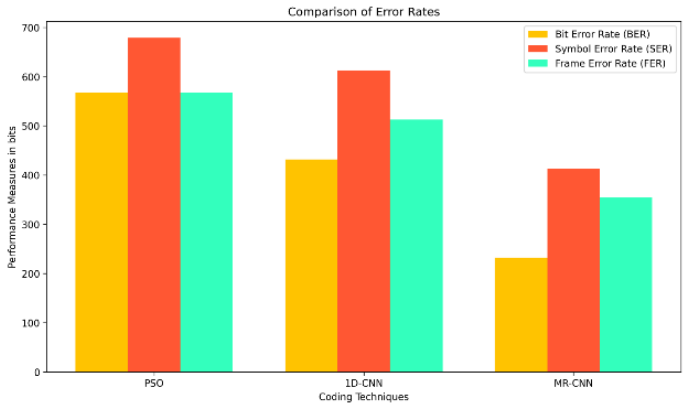


Fig. 2. Comparison of DER

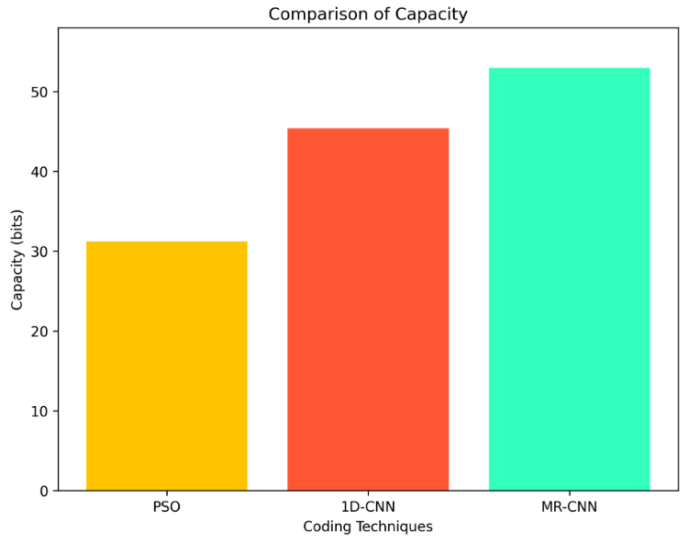


Fig. 3. Comparison of capacity

Considering an SNR of 454.78, PSO is superior to 1D-CNN (678.67) and MR-CNN (767.78) in an analysis of metrics related to performance. The BER is 232.13 for MR-CNN, 432.13 for 1D-CNN, and 567.23 for PSO. The best three models for SER are PSO (678.56), 1D-CNN (612.67), and MR-CNN (412.78). With a FER of 354.78, MR-CNN is at the bottom of the pack, next to 1D-CNN at 512.68 and PSO at 567.56. PSO emerges in third. With 31.23 bits, PSO has the highest capacity, next to 1D-CNN with 45.4 bits and MR-CNN with 53 bits. Although MR-CNN outperforms in SNR and capacity, it exhibits superior results across error rates, showing consistency in data integrity. According to these measures, 1D-CNN is in the middle. The contrast is demonstrated visually in Fig. 1 to Fig. 3. Fig. 1 indicates SNR, Fig. 2 potentially displays various error rates, and Fig. 3 exhibits distinct capacities. Overall, MR-CNN proves a higher error rate reduction than PSO, rendering it an excellent choice for tasks that are sensitive to errors, even though PSO has superior SNR and throughput.

V. Conclusion and Future Work

This research highlights the importance of RFID tags without chips for automatic data capture and human identification, as they facilitate fast and secure communication. The primary goal of this study is to improve the encoding efficiency and physical adaptability of chipless RFID tags. It is done by optimizing their resonant model. This study advances the design of chipless RFID tags by introducing an open-ended quarter-wavelength stub multi-resonator model, which provides access to real-world applications that require high-capacity data transfer. Implementing CNN for dynamic modulation and ECC improves tag performance. Both the novel AMS and ECC cooperate to make data accurate by adjusting modulation parameters in response to changes in the channel. When it comes to systems that are sensitive to errors, MR-CNN exhibits considerable potential due to its improved error rate reduction. The results of this work help to push chipless RFID forward, which is positive for various industries that use secure, high-capacity data transfer.

Further developments in chipless RFID tag design could be a focus of future studies. This may involve investigating novel resonant structures or identifying methods to optimize manufacturing for enhanced performance and scalability. The accuracy of data transfer and the validity of tags might be enhanced by investigating the possibility of integrating complex error correction coding and modulation techniques.

References

[1] Vena, A., Perret, E., & Tedjini, S. (2011). Chipless RFID tag using hybrid coding technique. *IEEE Transactions on Microwave Theory and Techniques*, Vol. *59, No.* 12, pp. 3356-3364.

[2] Herrojo, C., Paredes, F., Mata-Contreras, J., & Martín, F. (2019). Chipless-RFID: A review and recent developments. *Sensors*, Vol. *19, No.* 15, pp. 385.

[3] Abdulkawi, W. M., & Sheta, A. F. A. (2019). K-state resonators for high-coding-capacity chipless RFID applications. *IEEE Access*, Vol. *7*, pp. 185868-185878.

[4] Rance, O., Siragusa, R., Lemaitre-Auger, P., & Perret, E. (2016). Toward RCS magnitude level coding for chipless RFID. *IEEE Transactions on Microwave Theory and Techniques*, Vol. *64 No.* 7, pp. 2315-2325.

[5] Khaliel, M., El-Awamry, A., Megahed, A. F., & Kaiser, T. (2017). A novel design approach for co/cross-polarizing chipless RFID tags of high coding capacity. *IEEE Journal of Radio Frequency Identification*, Vol. *1, No.* 2, pp. 135-143.

[6] Vena, A., Perret, E., & Tedjini, S. (2016). *Chipless RFID based on RF encoding particle: realization, coding and reading system*.

[7] Majidifar, S., Ahmadi, A., Sadeghi-Fathabadi, O., & Ahmadi, M. (2015). A novel phase coding method in chipless RFID systems. *AEU-International Journal of Electronics and Communications*, Vol. *69, No.* 7, pp. 974-980.

[8] Vardhan, G. S., Sivadasan, N., & Dutta, A. (2016). QR-code-based chipless RFID system for unique identification. *IEEE International Conference on RFID Technology and Applications (RFID-TA)*, pp. 35-39.

[9] Khaliel, M. *et al.,* (2017). Printable, high coding capacity chipless RFID tags for low-cost item tagging. *IEEE 14th International Conference on Networking, Sensing and Control,* pp. 351-355.

[10] Tedjini, S., *et al.,* (2013). Hold the chips: Chipless technology, an alternative technique for RFID. *IEEE Microwave Magazine*, Vol. *14, No.* 5, pp. 56-65.

[11] Abdulkawi, W. M., & Sheta, A. A. (2020). High coding capacity chipless radio frequency identification tags. *Microwave and Optical Technology Letters*, Vol. *62, No.* 2, pp. 592-599.

[12] Babaeian, F., & Karmakar, N. C. (2018). Hybrid chipless RFID tags-a pathway to EPC global standard. *IEEE Access*, Vol. *6*, pp. 67415-67426.

[13] Aliasgari, J., Forouzandeh, M., & Karmakar, N. (2020). Chipless RFID readers for frequency-coded tags: Time-domain or frequency-domain?. *IEEE Journal of Radio Frequency Identification*, Vol. *4*, No. 2, pp. 146-158.

[14] Betancourt, D., *et al.* (2017). Design of printed chipless-RFID tags with QR-code appearance based on genetic algorithm. *IEEE Transactions on Antennas and Propagation*, Vol. *65, No.* 5, pp. 2190-2195.

[15] Zheng, F., et al., (2018). On the coding of chipless tags. *IEEE Journal of Radio Frequency Identification*, Vol. *2, No.* 4, pp. 170-184.

[16] El-Awamry, A., Khaliel, M., Fawky, A., El-Hadidy, M., & Kaiser, T. (2015). Novel notch modulation algorithm for enhancing the chipless RFID tags coding capacity. *IEEE International Conference on RFID,* pp. 25-31.

[17] Vena, A., Perret, E., & Tedjini, S. (2012). A compact chipless RFID tag using polarization diversity for encoding and sensing. *IEEE International Conference on RFID,* pp. 191-197.

[18] Le, C. C., *et al.,* (2023). Encoding Capacity Enhancement for Chipless RFID Tag Using Resonant Frequency Placement. *IEEE Access*.

[19] Rather, N. *et al.* (2024). Machine learning approaches for EM signature analysis in chipless RFID technology.

[20] Rather, N., Simorangkir, R. B., & Buckley, J. (2023). Deep learning assisted robust detection techniques for a chipless.

[21] Sridhar Panneerselvam et al., (2024), Federated learning-based fire detection method using local MobileNet, *Scientific Reports*, vol. 14, No. 30388, pp. 1-24,

[22] Asir Chandra Shinoo Robert Vincent and Sudhakar Sengan, (2024). Edge computing-based ensemble learning model for health care decision systems, Sci Rep 14, 26997.

[23] Asir Chandra Shinoo, Robert Vincent and Sudhakar Sengan, Effective clinical decision support implementation using a multi-filter and wrapper optimisation model for Internet of Things based healthcare data. Sci Rep 14, 21820.

[24] Gulista Khan *et al.* (2024) Energy-Efficient Routing Algorithm for Optimizing Network Performance in Underwater Data Transmission Using Gray Wolf Optimization Algorithm, Journal of Sensors, Vol. 2024, No. 2288527, 1-15.

[25] Sudhakar Sengan *et al.* 2023, Improved LSTM-Based Anomaly Detection Model with Cybertwin Deep Learning to Detect Cutting-Edge Cybersecurity Attacks, Human-centric Computing and Information Sciences, Vol. 13, No. 55.

[26] Varatharaj Myilsamy *et al.*, (2023). State-of-Health Prediction for Li-ion Batteries for Efficient Battery Management System Using Hybrid Machine Learning Model, Journal of Electrical Engineering & Technology.

[27] Indumathi Nallathambi *et al.* (2023). Impact of Fireworks Industry Safety Measures and Prevention Management System on Human Error Mitigation Using a Machine Learning Approach, Sensors, 23 (9), 4365.

[28] Vinothini Arumugham *et al.* (2023) An Artificial-Intelligence-Based Renewable Energy Prediction Program for Demand-Side Management in Smart Grids, Sustainability, Vol. 15, No. 6, 5453.

[29] Rasheed Abdulkader *et al.* (2023), Soft Computing in Smart Grid with Decentralized Generation and Renewable Energy Storage System Planning, Energies, Vol. 16, No. 6, 2655.

[30] Parkavi Krishnamoorthy *et al.* (2023). Effective Scheduling of Multi-Load Automated Guided Vehicle in Spinning Mill: A Case Study, IEEE Access.

[31] Arodh Lal Karn *et al.* (2023). IoT Based Smart Framework Monitoring System for Power Station, Computers, Materials & Continua, Vol. 74, No. 3, pp. 6019-6037.

[32] Arul Rajagopalan *et al.* 2022, Oscar Danilo Montoya, Walid El-Shafai, Mostafa M. Fouda, and Moustafa H. Aly, Modernized Planning of Smart Grid Based on Distributed Power Generations and Energy Storage Systems Using Soft Computing Methods, Energies, Vol. 15, No. 23, 8889.

[33] Eman S. Sabry *et al.* (2022), Sketch-Based Retrieval Approach Using Artificial Intelligence Algorithms for Deep Vision Feature Extraction, Axioms, Vol. 11, No. 12, pp. 663.

[34] Arodh Lal Karn *et al.* (2022). Measuring the Determining Factors of Financial Development of Commercial Banks in Selected SAARC Countries, Journal of Database Management, Vol. 33, No. 1, 2022 pp. 1–21.

[35] Prabu Selvam *et al.* (2022) A Transformer-Based Framework for Scene Text Recognition, IEEE Access, Vol. 10, pp. 100895-100910.

[36] Arodh Lal Karn *et al.* (2022) Fuzzy and SVM-Based Classification Model to Classify Spectral Objects in Sloan Digital Sky, IEEE Access, Vol. 10, pp. 101276-101291.

[37] Tribhuwan Kumar *et al.* (2022). Fuzzy Logic and Machine Learning-Enabled Recommendation System to Predict Suitable Academic Program for Students, Mathematical Problems in Engineering, vol. 2022, No. 5298468, pp. 1-7.

[38] Roy Setiawan *et al.* (2022). IoT Based Virtual E-Learning System for Sustainable Development of Smart Cities, Journal of Grid Computing, Vo. 20, No. 24.

[39] S. Priyadarsini *et al.* (2022), Automatic Liver Tumor Segmentation in CT Modalities Using MAT-ACM, Computer Systems Science and Engineering, vol. 43, no. 3, pp. 1057–1068, 2022.

[40] S. Priyadarsini *et al.* (2022). Classification of Liver Tumors from Computed Tomography Using NRSVM, Intelligent Automation & Soft Computing, Vol. 33, No. 3, 2022, pp. 1517-1530.

[41] Thirumoorthy Palanisamy *et al.* (2022) Improved Energy Based Multi-Sensor Object Detection in Wireless Sensor Networks, Intelligent Automation & Soft Computing, Vol. 33, No. 1, *pp.227-244.*

[42] Omaima Bamasaq *et al.* (2022). Syed Hamid Hassan, Distance Matrix and Markov Chain Based Sensor Localization in WSN, Computers, Materials & Continua, Vol. 71, No. 2, pp. 4051-4068,

[43] Sudhakar Sengan *et al.* (2021), Real-Time Automatic Investigation of Indian Roadway Animals by 3D Reconstruction Detection Using Deep Learning for R-3D-YOLOv3 Image Classification and Filtering, Electronics, Vol. 10, No. 24, 3079.

[44] Abolfazl Mehbodniya *et al.* (2021). Fetal health classification from cardiotocographic data using machine learning, Expert Systems.

[45] Sudhakar Sengan *et al.* (2021). A Secure Recommendation System for Providing Context-Aware Physical Activity Classification for Users, Security and Communication Networks, Vol. 2021, No. 4136909, pp. 1-15.

[46] Vasanthi Raghupathy *et al.* (2022). Interactive Middleware Services for Heterogeneous Systems, Computer Systems Science and Engineering, Vol. 41, No. 3, pp. 1241-1253.

[47] Abolfazl Mehbodniya *et al.* (2022). Proportional Fairness Based Energy Efficient Routing in Wireless Sensor Network, Computer Systems Science and Engineering, Vol. 41, No. 3, pp. 1071-1082.

[48] D. Stalin David *et al.* (2022), Enhanced Detection of Glaucoma on Ensemble Convolutional Neural Network for Clinical Informatics, Computers, Materials & Continua, Vol. 70, No. 2, pp. 2563-2579, .

[49] D. Stalin David *et al.* (2022). Cloud Security Service for Identifying Unauthorized User Behaviour, CMC-Computers, Materials & Continua, Vol. 70, No. 2. pp. 2581-2600.

[50] Ngangbam Phalguni Singh *et al.* 2021. Investigation on characteristics of Monte Carlo model of single electron transistor using Orthodox Theory, Sustainable Energy Technologies and Assessments, Vol. 48, 101601.

[51] K. Rajakumari *et al.,* (2022), Fuzzy Based Ant Colony Optimization Scheduling in Cloud Computing, Computer Systems Science and Engineering, Vol. 40, No. 2, pp. 581-592.

[52] Arodh Lal Karn *et al.* (2021), An integrated approach for sustainable development of wastewater treatment and management system using IoT in smart cities, Soft Computing.

[53] R. Nithya *et al.,* (2022). Edwin Hernan Ramirez-Asis, Priya Velayutham, V. Subramaniyaswamy, Sudhakar Sengan, An Optimized Fuzzy Based Ant Colony Algorithm for 5G-MANET, Computers, Materials & Continua, Vol. 70, No. 1, pp: 1069–1087.

[54] Razia Sulthana Abdul Kareem *et al.,* (2021). Multilabel land cover aerial image classification using convolutional neural networks, Arabian Journal of Geosciences, Vol. 14, No. 1681.

[55] Roy Setiawan et al., Utilizing Index-Based Periodic High Utility Mining to Study Frequent Itemsets, Springer, Arabian Journal for Science and Engineering, 2021.

[56] K. Muthumayil *et al.,* (2021) A Big Data Analytical Approach for Prediction of Cancer Using Modified K-Nearest Neighbour Algorithm, Journal of Medical Imaging and Health Informatics, Vol. 11, No. 8, pp. 2120-2125 (6).

[57] Keerthana Nandakumar et al. (2021) Sudhakar Sengan, Securing data in transit using data-in-transit defender architecture for cloud communication, Soft Computing.

[58] Huidan Huang *et al.* (2021) Emotional intelligence for board capital on technological innovation performance of high-tech enterprises, Aggression and Violent Behavior, 101633

[59] Sudhakar Sengan *et al.* (2021) Detection of false data cyber-attacks for the assessment of security in smart grid using deep learning, Computers & Electrical Engineering, Vol. 93, 107211.

[60] Sudhakar Sengan *et al.,* (2021). Cost-effective and efficient 3D human model creation and re-identification application for human digital twins, Multimedia Tools and Applications, 2021. DOI:10.1007/s11042-021-10842-y.

[61] Omnia Saidani Neffati *et al.,* (2021). Migrating from traditional grid to smart grid in smart cities promoted in developing country, Sustainable Energy Technologies and Assessments, Vol. 45.

[62] Prabhakaran Narayanan *et al.* (2021) Novel Collision Detection and Avoidance System for Mid-vehicle Using Offset-Based Curvilinear Motion. Wireless Personal Communication.

[63] Balajee Alphonse *et al.,* (2021). Modeling and multi-class classification of vibroarthographic signals via time domain curvilinear divergence random forest, J Ambient Intell Human Comput.

[64] Omnia Saidani Neffati *et al.,* (2021). An educational tool for enhanced mobile e-Learning for technical higher education using mobile devices for augmented reality, Microprocessors and Microsystems, Vol. 83, 104030.

[65] Prabhakaran Narayanan *et al.,* (2021). Analysis and design of fuzzy-based manoeuvring model for mid-vehicle collision avoidance system. J Ambient Intell Human Comput.

[66] Sudhakar Sengan *et al.,* (2021) Network Embedding Architecture using Laplace Regularization-Non-Negative Matrix Factorization for Virtualization, Microprocessors and Microsystems.

[67] K. Sathish Kumar et al. (2020). Area Based Efficient and Flexible Demand Side Management To Reduce Power And Energy Using Evolutionary Algorithms, Malaysian Journal of Computer Science, 27335.

[68] L. Arokia Jesu Prabhu *et al.,* (2020) Medical Information Retrieval Systems for e-Health Care Records using Fuzzy Based Machine Learning Model, Microprocessors and Microsystems, DOI:10.1016/j.micpro.2020.103344.

[69] N. Satheesh et a., (2020). Flow-based Anomaly Intrusion Detection using Machine Learning Model with Software Defined Networking for OpenFlow Network, Microprocessors, and Microsystems.

[70] Sengan Sudhakar *et al.,* (2020). A fuzzy-based high-resolution multi-view deep CNN for breast cancer diagnosis through SVM classifier on visual analysis, Journal of Intelligent & Fuzzy Systems, pp. 1-14.

[71] Sengan Sudhakar et al. (2020). Images super-resolution by optimal deep AlexNet architecture for medical application: A novel DOCALN, Journal of Intelligent & Fuzzy Systems, pp. 1-14.

[72] V. Vijaya Kumar *et al.* (2020). Design of peer-to-peer protocol with sensible and secure IoT communication for future internet architecture, Microprocessors and Microsystems, Vol. 78.

[73] Sudhakar Sengan *et al.,* (2020). Enhancing cyber-physical systems with hybrid smart city cyber security architecture for secure public data-smart network, Future Generation Computer Systems.

[74] Ganesh Kumar, K and Sudhakar Sengan (2020) Improved Network Traffic by Attacking Denial of Service to Protect resources using Z-Test Based 4-Tier Geomark Traceback (Z4TGT), Wireless Personal Communications.

[75] S. Sudhakar *et al.* (2020). Unmanned Aerial Vehicle (UAV) based Forest Fire Detection and monitoring for reducing false alarms in forest fires, Computer Communications 149, pp. 1–16.

[76] Sudhakar Sengan and Chenthur Pandian S, (2016), Hybrid Cluster-based Geographical Routing Protocol to Mitigate Malicious Nodes in Mobile Ad Hoc Network, International Journal of Ad Hoc and Ubiquitous Computing, Vol. 21, No. 4, pp. 224-236.

A Technique to Support Incremental Construction and Verification in Component-Based Software Development

Faranak Nejati[1], Ng Keng Yap[2], Abdul Azim Abd Ghani[3]
Lee Kong Chian Faculty of Engineering Science (LKC FES)-Department of Computing (DC),
Universiti Tunku Abdul Rahman (UTAR), Selangor, Malaysia[1]
Faculty of Computer Science and Information Technology and Institute for Mathematical Research,
Universiti Putra Malaysia (UPM), Seri Kembangan 43400, Malaysia[2]
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM),
Seri Kembangan 43400, Malaysia[3]

***Abstract*—Technological advancements in recent decades have significantly increased the scale and complexity of software systems, which poses challenges to their development and reliability. Component-based software development (CBSD) offers a promising solution by enabling modular and efficient software construction. However, CBSD alone cannot fully address challenges such as ensuring reliability and avoiding errors like deadlocks. Verification techniques, such as model-checking, are necessary to ensure the correctness of CBSD systems. Despite its effectiveness in verifying system properties, model-checking faces a critical issue known as state-space explosion (SSE), which hinders scalability. This study introduces an incremental verification technique for CBSD to address SSE and ensure deadlock freedom. The proposed technique incrementally constructs and verifies component-based systems, eliminating verified portions of components to minimize state-space size during subsequent verification steps. It utilizes a component model that supports encapsulation of computation and control, making incremental verification feasible. Evaluation of the technique using coloured petri nets with non-trivial case studies demonstrates its ability to detect deadlocks early and manage SSE effectively, thereby improving the efficiency of the verification process.**

***Keywords*—Component-based software development; incremental software construction; software verification**

I. Introduction

The rapid growth of technology has led to increasingly complex software systems, which necessitate advanced development methods to manage this complexity while ensuring system reliability. One such approach is Component-based Software Development (CBSD), which optimizes software development by breaking down systems into reusable, self-contained components [1]. CBSD allows for modular design, which can reduce development time and cost. However, as software systems grow in scale, even CBSD faces challenges in ensuring system correctness and avoiding errors, such as deadlocks, during development.

To enhance the reliability of CBSD, formal verification methods are needed to identify and address potential errors. Model-checking is a well-established technique for verifying the correctness of software systems by exhaustively exploring the state space (SS) of the system [2]. It can automatically detect issues like deadlocks by analyzing system behaviors. However, model-checking suffers from the state-space explosion (SSE) problem, where the number of states to be analyzed grows exponentially with system complexity, making verification infeasible for large systems [3].

Addressing SSE in model-checking is crucial for the practical application of verification in CBSD. Despite various approaches in the literature to mitigate SSE, a critical gap remains: the need for techniques that integrate verification throughout the development process rather than waiting until system construction is completed. Early detection of design errors is highly desirable, as it can significantly reduce the cost of fixing these errors compared to discovering them during later stages of development or maintenance [4].

Incremental verification is a promising approach to overcoming challenges in SSE by verifying components as they are developed and integrated, rather than waiting for the entire system to be completed. This approach enables early detection of defects and manages SSE more effectively by verifying smaller portions of the system incrementally [5]. However, despite its potential, incremental verification remains underexplored in the context of CBSD. This gap is partly due to the difficulties in maintaining consistent behavior and synchronization across different increments, making it challenging to ensure that new components integrate seamlessly without disrupting previously verified functionality [6].

To address this gap, this paper introduces a technique for the incremental construction and verification of component-based systems using a specialized component model called PUTRACOM [7]. PUTRACOM is designed to support encapsulation and concurrency, making it particularly well-suited for incremental verification. Its ability to decouple computation and control addresses the challenges of managing interdependencies between components, which are commonly encountered in CBSD. The proposed approach emphasizes two critical aspects: maintaining system behavior consistency during the incremental integration of components and integrating verification at each stage to avoid redundant checks on already verified components.

This research offers two key contributions. First, it presents a method for preserving system behavior and synchronization

during the incremental integration process, ensuring that the addition of new components does not compromise the verified properties of existing ones. Specifically, this research focuses on verifying the property of deadlock freedom. Second, it introduces a verification strategy that isolates previously verified components, preventing their re-verification and thereby reducing the overall state-space size during subsequent verification stages. This technique is evaluated using coloured Petri nets, a well-established tool for modeling and verifying concurrent systems [8], and applied to complex case studies, including the Common Component Modeling Example (CoCoME) [9].

The results show that the proposed incremental verification approach effectively mitigates the SSE problem and enhances the efficiency of the verification process in CBSD. By identifying errors early in the design stage and reducing verification effort, this approach improves the practical application of model-checking in component-based systems.

II. Background

This section provides preliminary information about the PUTRACOM component model [7], which is designed for building concurrent systems using exogenous connectors, inspired by the X-MAN model. PUTRACOM comprises three fundamental units: computation units (CU), observer/observable units (OOU), and exogenous connectors.

The CU encapsulates all computations and is modeled using reactive transition systems (RTSs) [4]. This encapsulation allows for decoupling between components while ensuring fixed behavior.

The OOU captures events visible to other components or the computing environment, modeled by interface automata [10]. This approach simplifies communication, avoiding complexities associated with traditional models like message passing and port-to-port connections. The OOU observes all events from other components or the environment. An atomic component in PUTRACOM is defined through RTSs, which represent potential behaviors of discrete systems. RTSs distinguish three types of events: input, internal, and output. Below is the formal definition of a computation unit based on RTS.

Definition 1 (computation unit)**.** A CU is a reactive transition system defined by the tuple (s_0, S, E, G, Δ), where:

- S is a set of states, with s_0 as the initial state;
- E is a set of events, including input events (E^I), output events (E^O), and hidden events (E^H);
- G is a guard function assigning boolean constraints to events;
- $\Delta \subseteq S \times E \times G \times S$ is a set of transitions.

Input and output events are considered observable ($E^{Obs} = E^I \cup E^O$) and accessible from the OOU. The OOU is modeled as a multiset, meaning the same event can occur multiple times. These events influence the component's state depending on whether the input events are admissible. Output events produced by the CU are also observable in the OOU. The formal definition of an OOU is as follows.

Definition 2 (Observer/Observable Unit)**.** The OOU is a finite multi-set defined as (E^{Obs}, f), where $f : E^{Obs} \to \mathbb{N}$ maps observable events to their multiplicities.

III. Proposed Solution and Methodology

This section outlines the proposed methodology focuses on the incremental construction and verification within the extended PUTRACOM framework. An overview of the solution is illustrated in Fig. 1. In the first step, the approach utilizes exogenous connectors, ensuring encapsulated component interactions. The methodology involves defining key elements such as noticing and publishing functions to manage event interactions between components and connectors. The composition of components is then formalized, allowing for structured integration within the PUTRACOM system. The detailed explanations are in Section IV.

The second step detailed in Section V is the incremental construction approach to develop a modular system by gradually integrating components and composite components. Each increment, will be systematically added using connectors for synchronization and interaction. The methodology will involve defining trace functions to analyze component behavior and applying n-way synch and asynch compositions to ensure effective integration. The process includes maintain consistency between the original and composed systems, preserving interaction behavior throughout the construction stages by defining observable traces and refinement relations.

Third step is relayed to the incremental verification detailed in Section VI, verifying each system increment before adding the next. It generates and checks the state space for local and global properties (deadlock-freeness). To manage state space explosion, it reduces state space size by focusing on observable behaviors and removing hidden parts, ensuring efficient and consistent verification. To validate our approach, we modeled three case studies using Coloured Petri Nets (CPNs), detailed in Section VII. Evaluation metrics include the number of generated state spaces, memory consumption, and processing time.

IV. Initialization of the Extended PUTRACOM Framework

This section presents additional definitions for extended PUTRACOM related to incremental construction and verification. Composition operators, termed connectors, are exogenous, allowing components to remain encapsulated without directly sending or receiving messages. The OOU manages and controls observable events without involving the CUs in the control process.

Definition 3 (connector)**.** A *connector* is a tuple $\Gamma = (L, T, Sub, E_\Gamma, G_\gamma)$, where:

- L is a set of connection lines, representing the physical or logical links between components;
- $T = \{sync, async, con, seq, itr\}$ indicates the types of connectors, including synchronous, asynchronous, conditional, sequential, and iterative interactions;
- $Sub = \{b_0, b_1, b_2, ..., b_n\}$ is a set of components subscribed to the connector;

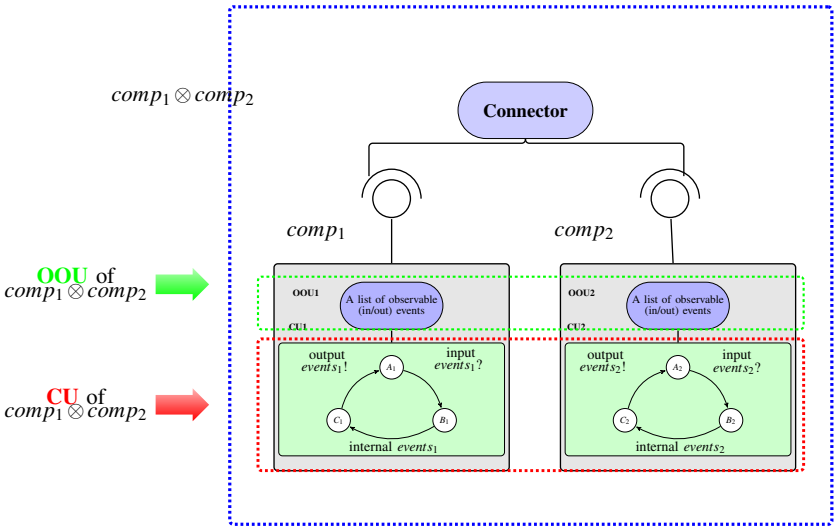


Fig. 1. Components with details in PUTRACOM.

- E_Γ is a set of events that consists of two mutually disjoint sets: output events $E_\Gamma^O \subseteq (E^O \cup env)$ and input events $E_\Gamma^I \subseteq (E^I \cup env)$, where, E^O is the set of output events; E^I is the set of input events defined for the system; *env* represents the set of external events or environmental conditions that can influence or interact with the connector.
- $G_(\gamma) = \{g_0, g_1, ..., g_n\}$ is a set of boolean constraints over the connector $\gamma \in L$.

Once output events are ready, the corresponding OOU notifies the connector using a *noticing* function.

Definition 4 (noticing)**.** Let Γ be a set of connectors and S be a set of components where $s \in S$. A *noticing* function $N : E^O \rightarrow \Gamma$ maps output events of components to their corresponding connectors, such that $N(e_o) = \gamma$, where $\gamma \in \Gamma$ and $e_o \in E^O$.

The activation and occurrence of a noticing depend on verifying predefined conditions associated with the event, as specified by the function *BoolConst*. The following definition serves as the primary definition of an atomic component in PUTRACOM, summarizing the core concepts presented in the preceding definitions.

Definition 5 (atomic component)**.** An atomic component in PUTRACOM is a tuple $com = \{CU, OOU, init, G, N\}$, where:

- $CU = (s_0, S, E, V, \Delta)$ is a reactive transition system;
- *OOU* is the observer/observable unit of the component, defined as a multiset (E^{Obs}, f);
- $init : E \rightarrow S$ is an initial function that assigns each event $e \in E$ to its corresponding initial state s_0 in S;
- $G_{com} : E \rightarrow$ BoolCons is a function mapping each event to its corresponding set of user-defined Boolean constraints;
- $N : E^O \rightarrow \Gamma$ is a noticing function that notifies the corresponding connector.

Once a connector is notified of an event occurrence by the OOU, it is responsible for disseminating this event to the relevant components. This process is managed by the publishing function, which ensures that all intended recipients receive the event.

Definition 6 (publishing)**.** Let $b \in Sub$, $e_o \in E_\Gamma^O$, $\gamma \in \Gamma$, and *env*. A *publishing* is defined as a function $P : E_\Gamma^O \rightarrow OOU$ that maps an output event e_o through the connector γ to the corresponding OOU, such that $OOU \subseteq \{OOU_{\{b_i\}_{i=0}^n} \cup env\}$.

Publishing an event requires satisfying conditions from G_γ. If met, the event is sent to the OOU of the relevant subscribers. In the noticing function, events are marked as "!", and in publishing, as "?".

Once a connector received an event by noticing and published the event, an interaction occurs.

Definition 7 (interaction)**.** Let the connector $\gamma \in \Gamma$ and $b \in Sub$. Consider the following events:

- There exists $e_o \in E^O$ such that $N(e_o \cup env) = \gamma$.
- There exists $e_i \in E_\gamma^I$ such that $P(e_i) = OOU$ and $OOU \subseteq \{OOU_{\{b_i\}_{i=0}^n} \cup env\}$.

Then, the set of interactions Π contains a finite number of interactions π, where $\pi = (e_o, \gamma, e_i)$ such that $\gamma \cap \Gamma = 1$. This research uses $(e_o, e_i)_\gamma$ or $\xrightarrow{e}_\Gamma$ instead of $\pi = (e_o, \gamma, e_i) \in \Pi$. Interactions can occur only when specific conditions are met which it is called admissible events.

Definition 8 (admissible interaction)**.** An interaction is admissible only if the following rules are satisfied:

1) $\forall e \in \Pi$, e is admissible,
2) $\xrightarrow{e}_\Gamma \in \Pi \models G_\Gamma$. It means $\xrightarrow{e}_\Gamma$ satisfies the set of conditions of connector Γ for publishing.

Based on the definitions provided, we can establish that composite components form the foundation for creating efficient and scalable systems within the PUTRACOM framework.

Definition 9 (composite components)**.** A composite component in PUTRACOM is a tuple $compos = \{CU, OOU, G, \Pi\}$.

V. Incremental Construction

Incremental construction in PUTRACOM consists of a finite set of increments denoted by *INC*. Each $inc \in INC$ represents a component or composite component, with the first increment serving as a foundational touchstone for the incremental composition, termed "basic," which connects other increments via a connector.

Definition 10 (Basic and Next Increments)**.** Let $C = \{c_0, c_1, \ldots, c_n\}$ be a set of all required components or composite components needed to construct the target system *Target*. The term *trail* refers to the composite component currently being built at the current construction level of the system *Target*. The definitions of *basic* and *next increment* are as follows:

$$basic = \{c \in C \mid \text{trail} = \emptyset \text{ and } c.\text{trail} \subseteq Target\}$$

$$inc_{next} = \{c \in C \mid \text{trail} \neq \emptyset \text{ and } c.\text{trail} \subseteq Target\}$$

Each increment consists of both CUs and OOUs, making it crucial to understand their behavior for effective component composition and system construction. To facilitate this understanding, a trace function is employed to extract the behavior of the CU, formally defined as a recursive function that returns a sequence of events originating from the initial state s_0.

Definition 11 (trace)**.** A transition *trace* of a RTS, denoted by ζ, is a sequence of events $e_0, e_1, \ldots, e_i$ such that for all $i \geq 0$, there exist states $s_0, s_1, \ldots, s_{i+1} \in S$ and guards $g_i \in G$ satisfying the conditions: $\zeta = (s_i, e_i, g_i, s_{i+1}), (s_{i+1}, e_{i+1}, g_{i+1}, s_{i+2}), \ldots, (s_n, e_n, g_n, s_{n+1}) \in \Delta$ or equivalently, $\zeta = s_i \xrightarrow{e_i, g_i} s_{i+1}, s_{i+1} \xrightarrow{e_{i+1}, g_{i+1}} s_{i+2}, \ldots, s_n \xrightarrow{e_n, g_n} s_{n+1}$. A trace function ε generates a sequence of transitions from an initial state s_0 based on admissible transitions.

A connector is defined by a set of types T that represent synchronization among subscribed increments and their interactions, highlighting the necessity of composing multiple increments. The following definitions describe the n-way composition capability of increments in PUTRACOM connectors.

Definition 12 (n-way sync composition)**.** Let C_i and E_i indicate the set of increments and their corresponding events respectively. The n-way sync composition of n increments is defined as follow:

$$\|_{i=0}^{n} \ (C_i, E_i) = C_1 \ {}_{E_1}\|_{\{E_2 \cup E_3 \cup \ldots \cup E_n\}} \ (\ldots (C_{n-1} \ {}_{E_{n-1}}\|_{E_n} \ C_n)) \quad (1)$$

Definition 13 (n-way asynchronous composition)**.** Let C_i indicates the sets of components. The $n-way\ asynchronous\ composition$ of n components are defined as follow:

$$|||_{i=0}^{n} \ (C_i) = C_1 \ ||| \ C_2 \ \ldots \ ||| \ C_n \quad (2)$$

In PUTRACOM, components encapsulate computation and are not called by other components. The CU exhibits fixed behavior without the environment's involvement. The only shared aspect is the OOU, which is involved in composition. Consequently, the hidden part can be omitted in traces, resulting in what is known as an Observable trace.

Definition 14 (observable trace)**.** *An observable trace is a sequence of events over a component or a composite component defined by:* $\zeta^{Obs} = \{(e_1), (e_2), \ldots, (e_i) \mid e \in E \backslash E^H\}$.

To obtain an observable RTS (ORTS) from an original RTS, a zoom-out function is presented. Let $\exists e \in E^H$ and $s, s' \in S$, then, the zoom-out function omits the arc $s \xrightarrow{e} s'$. If $s \xrightarrow{e_i, e_j} s'$ such that $i \neq j$ and $e_j \in E^{Obs}$ and $e_i \in E^H$, the zoom-out function will only omit $s \xrightarrow{e_i} s'$. The ORTS of the component is shown in Fig. 2.

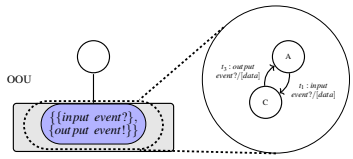


Fig. 2. An ORTS example of a component.

Observable RTS focuses on noticing, publishing, and interactions through connectors and will later help calculate possible component interactions. It also introduces an important aspect: the conformance between ORTS and RTS, defined using the refinement relation (RR) [11]. Refinement relations indicate that every behavior of system A is allowed by system (or specification) B.

Definition 15 (refinement relation)**.** Consider a RTS R and its corresponding ORTS S such that $E_S^I = E_R^I$ and $E_S^O = E_R^O$. S conforms to R, written $S \preceq R$, there is an RR between R and S, if the following conditions hold:

- $s_0 = r_0$, thus $s_0 \preceq r_0$ where s_0 and r_0 are the initial states of S and R respectively.
- For every transitions of R such that $\forall e \in E^H$ there exist a successor of r indicates by r' (where $r\prime \neq r$) and a RR between $r\prime$ and s, written by $s \preceq r\prime$.
- For every transitions of R such that $\forall e_i \in E^I$, $\exists r \xrightarrow{e_i} r\prime$ there exist a successor of s, written by $s \xrightarrow{e_i} s\prime$, such that $s\prime \preceq r\prime$.
- For every transitions of R such that $\forall e_o \in E^O$, $\exists r \xrightarrow{e_o} r\prime$ there exist a successor of s, written by $s \xrightarrow{e_o} s\prime$, such that $s\prime \preceq r\prime$.

Fig. 3 shows the composition of ORTSs from two increments, with the first component as *basic* and their combination as trail. Adding any increment signifies the next construction level, inc_{next}. Connectors are exogenous, defining connection and communication between increments. The initial increment serves as the touchstone for the incremental composition, referred to as inc_{base}, connected by a connector Γ.

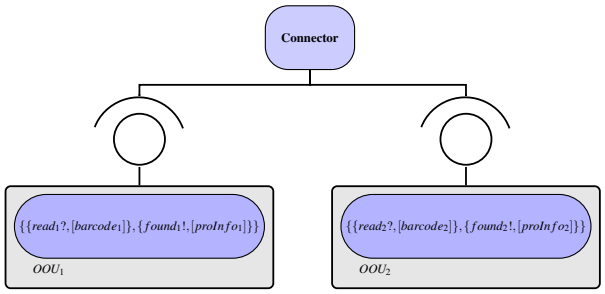


Fig. 3. Composition of two increments.

Incremental construction in PUTRACOM can be summarized in one definition:

Definition 16 (incremental construction)**.** An *incremental construction in PUTRACOM* could be a tuple $PCM_{inc} = (INC,\ basic,\ \Gamma,\ trail,\ inc_{next})$, where:

- INC. It is a finite set of components or composite components.
- $basic$. The first $inc \in INC$ of construction is $basic$.
- Γ is a set of connectors.
- $trail$. is a *non-empty* set of synthesized increments at the current level of construction.
- $inc_{next} \in INC$. Is a component or a composite component to be added to *trail* for building the next level of construction.

In composing increments, preserving interaction behavior is essential for adding new increments without defects. To ensure this, the sets of E^{Obs} must be disjoint, and both the trail and new increments must be free of event collisions before integration. Once confirmed, we can check the conditions for preserving behavior and synchronization.

Definition 17 (behavior preservation)**.** Consider that $basic$ is the first step and $trail$ is the next level of construction of system S. Let the relation between these two be $basic \subseteq trail$. Then, inc_{next} is allowed to be added into the system S as an increment if and only if $basic \subseteq trail \subseteq inc_{next}$ where $inc_{next} \in INC$.

The definition states that at each construction level, behavior and enforced synchronization must remain unchanged. To ensure this, we use the concept of isolated interactions, which clarify which interactions will be checked for preserving system behavior. An isolated interaction is defined based on the concept of a neighborhood.

Definition 18 (neighborhood)**.** Let Π represent a set of possible interactions, and let π be an interaction (or a set of interactions) within Π. The neighborhood of π, denoted by D_π, is defined as a subset of Π that includes all possible interactions containing π. Formally, this can be expressed as:

$$D_\pi = \{U \subseteq \Pi \mid \pi \in U\}$$

where U represents any subset of Π that includes the interaction π. A neighborhood includes a group of interactions that are similar to a specific interaction but still belong to the larger set of all interactions.

Definition 19 (isolated interaction)**.** Let $\pi \in \Pi$ represent an interaction (or $\pi \subseteq \Pi$ denote a subset of interactions), and let D_π be the neighborhood of the interaction π. An isolated interaction, denoted by $islt_\pi$, is defined as a subset of D_π that contains no other interactions from Π. Formally, this can be expressed as:

$$islt_\pi = \{d \in D_\pi \mid (\Pi \cap d) \setminus \{\pi\} = \emptyset\}$$

This means that the set $islt_\pi$ includes only those elements d in D_π for which there are no other interactions from Π present.

A trace function is employed to generate potential interactions and simulate transitions between $trail_i$ and inc_{next}.

Definition 20 (local interactions)**.** Consider a target system $Target$ such that $\{trail, inc_{next}\} \in Target$. If inc_{next} can be incremented to $trail$, denoted by $trail \otimes inc_{next}$, to facilitate the next level of construction, then the set of local interactions for $trail \otimes inc_{next}$, denoted by Π_{loc}, is defined as:

$$\Pi_{loc} = \{\pi_{islt}\langle trail\rangle \cup \pi_{inc_{next}} \cup \{\pi_{islt}\langle trail\rangle \otimes \pi_{inc_{next}}\}\}$$

To check whether the next level of increment contains the enforced synchronization over the transition of the trail, trace containment checking (TCC) [11] has been adopted. Especially, adding a new increment leads to more interaction than what the trail has already enforced. By adopting TCC, the set of interactions that have been added will be checked. The following TCC is extended to accommodate the requirements of this research.

Definition 21 (trace containment)**.** Consider a target system $Target$ such that $\{trail, inc_{next}\} \in Target$. The increment inc_{next} can be integrated into $trail$, denoted by $trail \otimes inc_{next}$, to facilitate the next level of construction if there exists a TCC satisfying the following conditions:

- For all $\pi \in \Pi_{\{trail \otimes inc_{next}\}}$, it holds that $\bowtie \notin \pi$.
- For all $\pi \in \Pi_{\{trail \otimes inc_{next}\}}$, there exists a $\pi_{islt} \in \Pi$ such that $\pi_{islt} \precsim \pi$.

The first condition states that all interactions lack collision events, while the second ensures that isolated interactions remain unchanged from the trail.

In the following, Theorem 1 presents a preservation checking method for $trail_i$ in incremental PUTRACOM. Both sufficient and necessary conditions are presented.

Theorem 1. Consider the target system $Target$ such that $\{trail, inc_{next}\} \in Target$. Let inc_{next} incremented to $trail$ as defined in definition 21, $L = trail \otimes\ inc_{next}$. Suppose Π_{loc} be the local interaction of L as defined in the definition 20. Then, adding inc_{next} to $trail$ preserves the enforced synchronization on the interactions if and only if the following proposition is hold $\forall \pi \in \Pi_{loc}$, $\bowtie \notin \pi$ and $\pi_{islt} \sqsubseteq \pi$.

proof of sufficiency. Let an interaction $\ell = \{\pi \mid \pi \in \Pi_{loc}\}$. Then, ℓ is proven to be a trace containment of local interactions over L by induction. First, we know $\pi = (n,\ p)$, thus:

- For $n \in N$, we have $N \subseteq (E^{obs} \setminus E^I) = E^O$. Then, from condition 2 of the definition 21, we have $\bowtie \notin E^O_L$. Hence $\bowtie \notin \ell$.
- For $p \in P$, we have $P \subseteq (E^{obs} \setminus E^O) = E^I$. Then, from condition 2 of the definition 21, we have $\bowtie \notin E^I_L$. Hence $\bowtie \notin \ell$.

Second, based on definition 20, we know $\Pi_{loc} = \{\pi_{islt}\langle trail\rangle \cup \pi_{inc_{next}} \cup \{\pi_{islt}\langle trail\rangle \otimes \pi_{inc_{next}}\}\}$. Then, for $\pi \in \Pi_{loc}, \exists \pi_{islt} \in \Pi_{loc}$ such that $\pi_{islt} \sqsubseteq \pi$. Hence $\pi_{islt} \sqsubseteq \ell$.

Therefore, ℓ is a trace containment, and adding inc_{next} to *trail* preserves the enforced synchronization on the interactions.

proof of necessity. Let Υ is an alternating trace containment over the local interactions of Π_{loc}. Consider ξ be an observable trace over L from $s_{initial}$ and s_{reach} is a reachable state via interaction (n,p) on ξ. The set of interactions that leads to s_{reach} called $(n,p)_{reach}$. Then, we prove $\bowtie \notin (n,p)_{reach}$ and $\pi_{islt} \sqsubseteq (n,p)_{reach}$ by induction on the trace ξ. Firstly, when $\xi = \lambda$, the two conditions of the theorem hold. Next, suppose the conditions $\forall (n,p)_{reach} \in \Pi_{loc}$, $\bowtie \notin (n,p)_{reach}$ and $\pi_{islt} \sqsubseteq (n,p)_{reach}$ hold for any ξ.

For checking the first condition of the theorem:

For $n \in \{env \cup E^O_L\}$ and $p \in \{env \cup E^I_L\}$, if $\exists A_L \xrightarrow{(n,p)} B_L$ is a $(n,p)_{reach}$ by ξ, then $\bowtie \notin (n,\ p)_{reach}$, (Definition **??**, Condition 1) Because $(n,\ p) = \pi$, then we have $\bowtie \notin \pi_{reach}$.

For checking the second condition:

- For $\forall n \in \{env \cup E^O_{trail}\}$ and $\forall p \in \{env \cup E^I_{trail}\}$, if $\exists A_L \rightarrow B_L$ which is reachable by $(n,p)_{reach}$ in ξ, then $(n,p)_{reach} \in \Pi_{islt}\langle trail\rangle$. Hence, we have $\pi_{islt} \sqsubseteq (n,p)_{reach}$.
- For $\forall n \in \{env \cup E^O_{inc_{next}}\}$ and $\forall p \in \{env \cup E^I_{inc_{next}}\}$, if $\exists A_L \rightarrow B_L$ which is reachable by $(n,p)_{reach}$ in ξ, then $(n,p)_{reach} \in \Pi_{inc_{next}}$. Hence, still we have $\pi_{islt} \sqsubseteq (n,p)_{reach}$.
- For $\forall n \in \{env \cup E^O_{\{\pi_{islt}\langle trail\rangle \otimes \pi_{inc_{next}}\}}$ and $\forall p \in \{env \cup E^I_{\{\pi_{islt}\langle trail\rangle \otimes \pi_{inc_{next}}\}}\}$, if $\exists A_L \rightarrow B_L$ which is reachable by $(n,p)_{reach}$ in ξ, then $(n,p)_{reach} = \{\pi_{islt}\langle trail\rangle \otimes \pi_{inc_{next}}\}$. Hence, we have $\pi_{islt} \sqsubseteq (n,p)_{reach}$.

Therefore, applying Υ over L implies $\forall \pi \in \Pi_{loc}$, $\bowtie \notin \pi$ and $\pi_{islt} \sqsubseteq \pi$ for any reachable trace over L.

VI. Incremental Verification

Verification is essential in system construction for identifying design obstacles and providing counterexamples to locate errors. It begins with generating the system's state space (SS), followed by checks for local properties, global properties, and deadlock-freeness.

1) Verification Process: The verification process checks the system's required properties and generates counterexamples for violations. In PUTRACOM, verification is integrated into the construction process, allowing the next increment only after verifying the previous one. However, model-checking faces limitations in industrial contexts due to the exponential growth of SS with the number of processes and states (SSE). This research proposes a technique to mitigate SSE's impact:

- Incremental verification after each addition, avoiding rechecking already verified parts.
- Eliminating hidden parts of increments and focusing on external behavior to reduce SS size.

2) State space generation: The system's SS must be generated before verification. PUTRACOM supports global and local SS generation, which resembles that in CPN [12] as a directed graph of reachable states and events. Detailed SS generation is omitted due to space constraints.

3) Verifying local properties in PUTRACOM: Each component in PUTRACOM must verify its local properties without assumptions due to their fixed behavior. Properties must be maintained during composition and when adding increments. The following theorem specifies how to prove the properties φ of inc_i in the trail.

Theorem 2. Given that inc_i is an increment in *trail* with respect to behavior preservation, local properties φ of inc_i hold on *trail* if firstly they hold on the local SS of inc_i.

Proof. Let L_{local} is the local SS of *trail*. Consider that $s \in S_{trail}$ is a reachable state in *trail* via trace sequence ζ such that s does not satisfy the set of local properties φ. It can be shown by $\zeta_s \not\models \varphi$. We know the observable trace is consistent with trace sequence $\zeta^{Obs} \preceq \zeta$. Therefore, if any violation against the properties is detected in ζ then its ζ^{Obs} has violation as well. Let $\zeta_{trail} = \varepsilon(trail)$ be a trace sequence over $L_{local}(trail)$ produced by the trace function $\varepsilon(trail)$. Since the trace sequence on *trail* includes the observer trace of all the composed increments $\{inc_1, ..., inc_n\}$ in *trail*, then ζ_{trail} may consist of $\zeta^{Obs}_{inc_i}$ and the violation appears in the $L_{local}(trail)$ as well. As a conclusion, a satisfying φ in *trail* can be detected in local SS inc_i, thus the theorem holds.

The theorem indicates that a set of local properties φ of inc_i is preserved in the trail that includes inc_i. Thus, in order to check and prove such propert in the trail, it must first be proven in the local SS of inc_i.

4) Verifying global properties in PUTRACOM: System-global properties are the set of properties that are related to the global SS of the system. Generally, it involves observable events with multiple increments in *trail*. Essentially, the global properties of the trail should be checked over a clausal normal form, in which clauses are disjunctions of observable events of increment in the trail through the set of connectors Γ_i. Its formal definition is given below.

Definition 22. Consider *trail* as the current level of a given system with a set of increments inc_i and Γ_j, $1 \leq i, j \leq k$, a global property of trail check over conjunction of following disjunction:

$$e^{obs}_1 \vee e^{obs}_2 \vee ... \vee e^{obs}_k$$

such that e^{obs}_k is an observable event of an increment $inc_i \in trail$ and $\forall n, m : 1 \leq n, m \leq k, n \neq m$ *implies* $inc_n \neq inc_m$.

This definition restricts each increment to at most one observable event, simplifying the integration of events from

other increments. Since the equation uses a conjunction of E^{obs} clauses, each can be proven independently and reused for future verification if no new events are introduced.

To verify global properties of a composite component, it is common to check each state and construct the SS as the Cartesian product of component states, which risks SS explosion. Instead, hidden events are eliminated, focusing on the SS without states hidden from other components or the environment. Local property violations impact global properties through observable events, which are critical for interactions in *trail*. Thus, the relevant part for global properties is the set of interactions Π, as proven in the theorem below.

Theorem 3. *Given inc_i is an increment added to trail with respect to behavior preservation, and a set of local properties called φ. Let w is a derived global SS of trail with respect to the elimination of the hidden events. Then φ holds on the w if for all $\pi \in Pi_t rail$, $\exists i : 1 \leq i \leq k, \pi_i \models \varphi_i$.*

Proof. Suppose an interaction $\pi \in trail$ is reachable via trace sequence ξ such that $\pi \not\models \varphi$. We know that an interaction in PUTRACOM can be simply denoted by $\pi = e^o \vee E^I$ and can be extended by $e^o_i \vee e^i_1 \vee e^i_2 \vee ... \vee e^i_k$ for $j : 1 \leq i \leq k$. If at least one of these events violates the properties φ, the *trail* does not holds the global properties φ, $trail \not\models \varphi$. Therefore, to check the global properties of *trail* its interactions must be checked.

This theorem, together with Theorem 2, facilitates the separation of concerns between proving local and global properties. By encapsulating and removing hidden events of increments, it reduces the potential SS of the system. Furthermore, due to the fixed behavior of components, the composition and verification of increments are assumption-free.

5) Deadlock-freeness preservation: Deadlock-freeness verification checks for the presence of activities in any reachable trace of the system. A deadlock indicates that parts or the entire system cannot progress. In PUTRACOM, a deadlock occurs when no admissible events exist for a state or a set of states, halting progress in an RTS or network of RTSs. The formal definition of deadlock-freeness in PUTRACOM is given below.

Definition 23. Let C is considered as a set of RTSs and $Comp = \{c_0, C, E, \Delta\}$ is a composition of C through a connector where E is a set of events and Δ is a set of transitions. Then, $Comp$ is deadlock-free for all $c \in C \vee Comp$, $E_{admis}(c) \neq \emptyset$.

Definition 24. Consider $Comp = \{c_0, C, E, \Delta\}$ is a composition of a set of components C through connector Γ. Let ξ be a given trace of $Comp$ and $s \in S$ is a reachable state via ξ. A state s is considered a deadlock state if $E_{admis}(s) = \emptyset$. $Comp$ can be said to be deadlock-free if, in any possible trace in the $Comp$, no deadlock has been reached.

Once all states in the component have an admissible event, it means that there is progress in the component. However, if there is any state in the component that does not have at least one admissible event to make progress on it, deadlock happens.

VII. Experimental Results

This section evaluates the proposed approach for feasibility in incremental construction, verification, efficiency in reducing SS, and addressing the SSE problem. It includes subsections on the case studies and tools, followed by implementation details and evaluation results.

A. Case Studies

a) Common Component Modeling Example (CoCoME): is a supermarket trading system model for sales, purchases, and inventory management, and widely used for evaluating component models [13], [14]. CoCoME has a cash desk line which processes payments with subcomponents like cash boxes, cash desk controller, cash desk app, Cah desk GUI, print controller, PoS terminal, and interface named *connectorIf*) facilitate communication among the cash desk, store, and enterprise. A store can manage multiple cash desks through a store server that oversees inventory.

b) Dining Philosophers Case Study: It shows how state space can expand exponentially with more philosophers. Seated around a table, each philosopher alternates between thinking and eating, using two shared forks. If all philosophers pick up the left fork at once, a deadlock occurs.

c) Automatic Machine Teller case study (ATM): enables banking transactions without needing human assistance. The system includes three main components: the user, the ATM, and the bank network, with subcomponents like authentication, balance check, cash withdrawal, deposit, help, and maintenance.

B. Implementation Tool

CPN Tools is a graphical platform for simulating CPNs, offering features such as the Create and Simulate Palette for implementing various systems and interactions, the Hierarchy Palette for constructing systems incrementally, and the State-Space Palette to calculate, check, and analyze the state space for properties like deadlock freedom while generating reports. CPNs are a formal modeling language for describing systems. They have been applied in various fields, including robotics [15], or communication networks [16]. A CPN includes places (representing states), transitions (actions or events), and arcs (connections between places and transitions). Each place is assigned a color, indicating the type of resources or objects it represents. Tokens represent the system's state and are placed in the CPN's places. Transitions fire when input places have enough tokens, moving them to output places and changing the system's state. CPNs can include guards and invariants to constrain transition firing, using an inscription language.

CPN is defined as a tuple $cpn = (P, T, R, C, \Sigma, EXPR, W, G, Init)$ where:

- P is a finite set of places;
- T is a finite set of transitions, where $T \cap P = \emptyset$;
- R is a finite set of directed arcs, where $R \subseteq (P \times T) \cup (T \times P)$;
- C represents a set of colors (types), and Σ is a function that maps places to color sets, i.e., $\Sigma : P \to C$. Colors in CPNs represent the types of tokens that places can hold;

- *EXPR* defines net inscriptions, such as guards, arc expressions, and initial markings, typically specified using an inscription language like CPN ML;
- W is the arc expression function $W : R \rightarrow EXPR$, assigning an expression to each arc $r \in R$;
- G is the guard function $G : T \rightarrow EXPR$, assigning a guard to each transition $t \in T$;
- *Init* is the initial marking function $Init : P \rightarrow EXPR$, specifying the initial distribution of tokens in each place.

Both non-incremental and incremental constructions of PUTRACOM can be modeled in this tool.

C. Simulation PUTRACOM by the CPN model

To simulate PUTRACOM using CPN, we start by listing all states, events, guards, and transitions in the model, facilitated by the trace function ε. Once the model elements are identified, we translate them into corresponding elements in the CPN tool, as outlined in Algorithm 1 (Fig. 4). The variable *steps* in Algorithm 1 represents traces of states, events, guards, and transitions from the trace function. Each state $s \in S$ corresponds to a place $p \in P$, and each transition Δ maps to a transition $t \in T$ in the CPN model with two directed arcs. Events $e \in E$ act as tokens, while guards $g \in G$ are conditions for each place. The variable *EXPR* holds expressions for guards, color sets, and markings, differentiating input, output, and internal events, which are then added to transitions. The OOU handles external interactions. Each CU must be encapsulated. Transitions that produce output events generate tokens for the OOU. Trace containment checking (TCC) is simulated in CPN Tools using the ASK-CTL library and SML, enabling CTL-like logic.

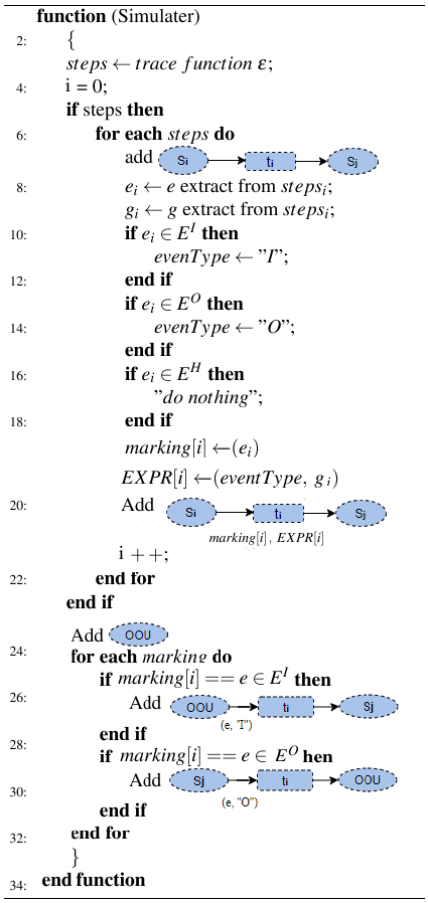


Fig. 4. Algorithm 1: Simulation of PUTRACOM with CPN

D. Evaluation

This section presents experimental results for the aforementioned case studies implemented using the proposed construction technique in CPN Tools. Evaluations were conducted on a Core(TM) i5-3337U CPU @ 1.80 GHz with 4GB of RAM.

a) CoCoME: In the evaluation of CoCoME, firstly the local SS of each component is generated, and then the global SS of the composite components will be calculated incrementally. We firstly verified each component mentioned in VII-A separately. the cash boxes and cash desk controller into a composite component named Composite Component 1, which we verified. Subsequently, we incrementally added the remaining components, resulting in Composite Components 2 through 4. The tables in this section report the number of states, indicating the amount of saved SS in memory. The status reflects whether the verification was fully completed or if a partial SS explosion occurred for both non-incremental and incremental approaches.

The system was verifiable with up to 20 cash desk lines and stores, but non-incremental verification failed early, as shown in Table I, with most results being partial and indicating an SSE. This was due to the need to generate all state spaces (SS) of atomic components and their compositions in memory, causing early SSE as the number of components increased.

In contrast, incremental construction successfully generated the SS by omitting the component's CU, verifying only the newly added increments. The PUTRACOM approach simplifies this by allowing fixed-behavior CUs, which are decoupled from other components, to be excluded. With reduced SS, the verification process is faster and less labor-intensive, making incremental construction more efficient. CPN Tools can produce counterexamples at each verification level, enabling early error detection and enhancing system safety.

b) Dining philosophers: Table II compares the verification results for the dining philosophers problem using the proposed method, CPN Tools, and NuSMV. In this experiment, we assess the amount of memory utilized. Incremental verification completes with lower time and memory consumption, while CPN and NuSMV face the SSE problem. Verification time is longer for dining philosophers due to substantial SS in interaction components (philosophers and forks). Omitting hidden parts did not significantly reduce SS. In CoCoME, diverse components interact through OOU and connectors, allowing hidden parts to be omitted, resulting in faster verification times.

c) Automatic machine teller: Table III shows the ATM case study results using three verification methods. In this experiment, we assess the amount of memory utilized. The enumerative CPN and NuSMV methods could not verify even 50 ATMs, while the proposed incremental method achieved nearly linear performance. The ATM case study outperformed

TABLE I. RESULTS OF COCOME LINE SIMULATION

Component	No. of States (CDL=1)	Status	No. of States (CDL=5)	Status	No. of States (CDL=10)	Status	No. of States (CDL=20)	Status
Cash box	$\approx 10^4$	full/full	$\approx 10^6$	partial/full	$\approx 10^7$	partial/full	$\approx 10^{17}$	partial/full
Cash box controller	$\approx 10^2$	full/full	$\approx 10^3$	partial/full	$\approx 10^2$	partial/full	$\approx 10^4$	partial/full
Cash Desk App	$\approx 10^2$	full/full	$\approx 10^3$	partial/full	$\approx 10^2$	partial/full	$\approx 10^4$	partial/full
Print Controller	$\approx 10^2$	full/full	$\approx 10^3$	partial/full	$\approx 10^2$	partial/full	$\approx 10^4$	partial/full
Cash Desk GUI	$\approx 10^2$	full/full	$\approx 10^3$	partial/full	$\approx 10^3$	partial/full	$\approx 10^5$	partial/full
POS	$\approx 10^2$	full/full	$\approx 10^3$	partial/full	$\approx 10^2$	partial/full	$\approx 10^5$	partial/full
connectorIF	$\approx 10^2$	full/full	$\approx 10^3$	partial/full	$\approx 10^2$	partial/full	$\approx 10^3$	partial/full
Composite Component 1	$\approx 10^4$	full/full	$\approx 10^7$	partial/full	$\approx 10^3$	partial/full	$\approx 10^{18}$	partial/full
Composite Component 2	$\approx 10^5$	full/full	$\approx 10^9$	partial/full	$\approx 10^{10}$	partial/full	$\approx 10^{20}$	partial/full
Composite Component 3	$\approx 10^5$	full/full	$\approx 10^9$	partial/full	$\approx 10^{10}$	partial/full	$\approx 10^{20}$	partial/full
Composite Component 4	$\approx 10^7$	full/full	$\approx 10^{10}$	partial/full	$\approx 10^{10}$	partial/full	$\approx 10^{20}$	partial/full

TABLE II. RESULTS OF DINING PHILOSOPHERS

	Time (min)			Memory (MB)		
Scale	Incremental	Num CPN	NuSMV	Incremental	Num CPN	NuSMV
40	6	47.15	28	13	500	60
100	33	NA	644	29	NA	498
500	209	NA	NA	64	NA	NA
1000	349	NA	NA	112	NA	NA
1500	514	NA	NA	194	NA	NA

the dining philosophers due to having three main components (user, ATM, and bank) with sub-components, whereas the dining philosophers have only two. The hidden parts of each ATM could be omitted after verification, unlike the philosophers' problem, where most state space is tied to interactions. The ATM study results were similar to CoCoME, due to the number of evaluated ATMs and interactions.

VIII. RELATED WORKS

Addressing SSE is a critical challenge in model-checking for CBSD. While model-checking is a valuable tool for verifying system properties like deadlock-freedom, it faces significant scalability issues due to the exponential growth in the number of states to be analyzed [17]. This challenge has led to the development of various techniques aimed at mitigating SSE, which can be categorized into five main approaches: heuristics and probabilistic methods [17], [18], state space reduction techniques [19], compositional verification [20], memory optimization [21], and bottom-up verification [22]. A particular subset of the aforementioned verification methods is more suitable for addressing the SSE problem in CBSD, such as compositional verification and bottom-up approaches.

The compositional verification, particularly the assume-guarantee reasoning approach, decomposes a system into smaller subcomponents and verifies them independently [3]. However, compositional verification introduces new challenges, such as finding optimal ways to decompose systems and deriving accurate assumptions [23]. Additionally, circular dependencies between components can complicate verification, as addressing mutual interdependencies requires defining sound and complete rules, which is often difficult [24]. These limitations reduce the applicability of compositional approaches, especially in complex CBSD systems where component interactions are tightly coupled. Our work tackles these challenges with a bottom-up, incremental approach using a fully decoupled component model. This eliminates dependencies between components, removing the need for system decomposition, assumptions, or managing circular dependencies, making our method more practical for CBSD.

Bottom-up approach, which includes techniques like on-the-fly model-checking and incremental verification. On-the-fly model-checking verifies system properties by generating and checking individual paths one at a time, thereby reducing memory usage by not storing the entire state space [25]. However, this method can lead to inefficiencies, as previously verified paths might need to be regenerated if deleted, resulting in redundant checks and longer counterexamples. In contrast, our technique leverages the event-hiding feature of PUTRA-COM to retain essential verification information, avoiding the need to regenerate paths and thereby improving efficiency.

Incremental verification, which verifies systems as they are constructed, offers another solution for handling SSE. It has been explored in several domains, including rule-based systems, expert systems, and embedded systems [26]. However, most incremental techniques have been developed for non-component-based systems, and those tailored for CBSD remain rare. For example, Bensalem et al. [22] proposed an incremental construction and verification framework for component-based systems, using the BIP (Behavior, Interaction, Priority) model. While their work presents a valuable incremental construction approach, its reliance on symbolic methods (e.g., binary decision diagrams) and extremely relies on the variables ordering, which may limit the verification process. Our work is based on explicit verification methods,

TABLE III. RESULTS OF ATM

	Time (min)			Memory (MB)		
Scale	Incremental	Num CPN	NuSMV	Incremental	Num CPN	NuSMV
50	8	NA	NA	28.4	NA	NA
100	23	NA	NA	67.2	NA	NA
200	54	NA	NA	96.9	NA	NA
500	206	NA	NA	$\approx$ 1G	NA	NA

which are more scalable and easier to apply to large systems.

IX. CONCLUSION AND FUTURE WORK

This research introduces a novel technique for incremental construction and verification in CBSD, aimed at early detection of deadlocks during system development. The approach ensures behavior preservation and synchronization at each increment, verifying safety properties like deadlock freedom and promptly generating counterexamples. The method effectively handles complex CBSD systems, avoiding the SSE problem by excluding previously verified computation units with fixed behavior. This reduces the state space, memory use, and verification effort. Results confirm the method's ability to lower verification complexity and identify potential issues early. Future work includes developing a standalone PUTRACOM application and conducting experiments on real-world systems with larger state spaces for deeper analysis and comparison, enhancing understanding of the method's strengths and limitations.

ACKNOWLEDGMENT

We sincerely thank the Ministry of Higher Education Malaysia and the University Tunku Abdul Rahman for the generous grant #Project: UTAR/IPSR/RMC/UTARRF/2024-C1/F01 that supported this research. Our appreciation also goes to Dr. Kung Kiu Lau, Simone Di Cola, and Cuong M. Tran for their valuable assistance and for sharing their X-MAN tool.

REFERENCES

[1] K.-K. Lau and Z. Wang, "Software component models," *IEEE Transactions on software engineering*, vol. 33, no. 10, pp. 709–724, 2007.

[2] D. Beyer and A. Podelski, "Software model checking: 20 years and beyond," in *Principles of Systems Design: Essays Dedicated to Thomas A. Henzinger on the Occasion of His 60th Birthday*. Springer, 2022, pp. 554–582.

[3] F. Nejati, A. A. A. Ghani, N. K. Yap, and A. Jaafar, "Handling state space explosion in component-based software verification: A review," *IEEE Access*, 2021.

[4] Y. Jin, "Compositional verification of component-based heterogeneous systems," Ph.D. dissertation, University of Adelaide, SA, AU, 2004.

[5] R. E. Fairley and M. J. Willshire, "Iterative rework: the good, the bad, and the ugly," *Computer*, vol. 38, no. 9, pp. 34–41, 2005.

[6] K.-K. Lau, K.-Y. Ng, T. Rana, and C. M. Tran, "Incremental construction of component-based systems," in *Proceedings of the 15th ACM SIGSOFT symposium on Component Based Software Engineering*. ACM, 2012, pp. 41–50.

[7] F. Nejati, A. A. A. Ghani, N. K. Yap, and A. B. Jafaar, "Putracom: A concurrent component model with exogenous connectors," *IEEE Access*, vol. 6, pp. 15 446–15 456, 2018.

[8] K. Jensen and L. M. K. C. P. Nets, "Modelling and validation of concurrent systems," 2009.

[9] S. Herold, H. Klus, Y. Welsch, C. Deiters, A. Rausch, R. Reussner, K. Krogmann, H. Koziolek, R. Mirandola, B. Hummel *et al.*, "Cocome-the common component modeling example," in *The Common Component Modeling Example*, 2008, pp. 16–53.

[10] L. De Alfaro and T. A. Henzinger, "Interface automata," *ACM SIGSOFT Software Engineering Notes*, vol. 26, no. 5, pp. 109–120, 2001.

[11] R. Alur, T. A. Henzinger, O. Kupferman, and M. Y. Vardi, "Alternating refinement relations," in *International Conference on Concurrency Theory*. Springer, 1998, pp. 163–178.

[12] K. Jensen and L. M. Kristensen, *Coloured Petri nets: modelling and validation of concurrent systems*. Springer Science & Business Media, 2009.

[13] V. Cortellessa, D. Di Pompeo, V. Stoico, and M. Tucci, "Many-objective optimization of non-functional attributes based on refactoring of software models," *Information and Software Technology*, vol. 157, p. 107159, 2023.

[14] V. Cortellessa, D. Di Pompeo, and M. Tucci, "Check for updates performance of genetic algorithms in the context of software model refactoring," in *Computer Performance Engineering and Stochastic Modelling: 19th European Workshop, EPEW 2023, and 27th International Conference, ASMTA 2023, Florence, Italy, June 20–23, 2023, Proceedings*, vol. 14231. Springer Nature, 2023, p. 234.

[15] M. Kloetzer and C. Mahulea, "Path planning for robotic teams based on ltl specifications and petri net models," *Discrete Event Dynamic Systems*, vol. 30, no. 1, pp. 55–79, 2020.

[16] M. Raeisi-Varzaneh and H. Sabaghian-Bidgoli, "A petri-net-based communication-aware modeling for performance evaluation of noc application mapping," *The Journal of Supercomputing*, pp. 1–24, 2020.

[17] N. Rezaee and H. Momeni, "A hybrid meta-heuristic approach to cope with state space explosion in model checking technique for deadlock freeness," *Journal of AI and Data Mining*, vol. 8, no. 2, pp. 189–199, 2020.

[18] Y. Liu and C. He, "A heuristics-based incremental probabilistic model checking at runtime," in *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2020, pp. 355–358.

[19] A. Biere, "Bounded model checking," in *Handbook of satisfiability*. IOS press, 2021, pp. 739–764.

[20] K. Ghasemi, S. Sadraddini, and C. Belta, "Compositional synthesis via a convex parameterization of assume-guarantee contracts," in *Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control*, 2020, pp. 1–10.

[21] L. Wu, H. Huang, K. Su, S. Cai, and X. Zhang, "An i/o efficient model checking algorithm for large-scale systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 5, pp. 905–915, 2015.

[22] S. Bensalem, M. Bozga, A. Legay, T.-H. Nguyen, J. Sifakis, and R. Yan, "Component-based verification using incremental design and invariants," *Software & Systems Modeling*, vol. 15, pp. 427–451, 2016.

[23] J. M. Cobleigh, G. S. Avrunin, and L. A. Clarke, "Breaking up is hard to do: An evaluation of automated assume-guarantee reasoning," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 17, no. 2, p. 7, 2008.

[24] K. A. Elkader, O. Grumberg, C. S. Păsăreanu, and S. Shoham, "Automated circular assume-guarantee reasoning with n-way decomposition and alphabet refinement," in *International Conference on Computer Aided Verification*. Springer, 2016, pp. 329–351.

[25] S. Xiao, J. Li, S. Zhu, Y. Shi, G. Pu, and M. Vardi, "On-the-fly synthesis for ltl over finite traces," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, 2021, pp. 6530–6537.

[26] W. Chen, A. Chiesa, E. Dauterman, and N. P. Ward, "Reducing participation costs via incremental verification for ledger systems," *Cryptology ePrint Archive*, 2020.

Real-Time Video Captioning on CPU and GPU: A Comparative Study of Classical and Transformer Models

Othmane Sebban[1], Ahmed Azough[2], Mohamed Lamrini[3]
Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez 30003, Morocco[1,3]
Paris La Défense, Léonard De Vinci Pôle Universitaire, Research Center, Paris, France[2]

***Abstract*—This paper proposes a scalable and hardware-adaptable approach to automatic video caption generation by comparing two architectures: a traditional encoder–decoder framework combining InceptionResNetV2 with GRU and a transformer-based model integrating TimeSformer with GPT-2. The system supports CPU and GPU deployment through a unified pipeline built on FFmpeg and ImageMagick for keyframe extraction and subtitle embedding. Experimental evaluations on the MSVD and VATEX datasets demonstrate that the TimeSformer–GPT-2 architecture significantly outperforms baseline models, particularly in GPU settings, achieving top results across BLEU, METEOR, ROUGE-L, and CIDEr metrics. This superiority is attributed to its capacity to model spatiotemporal dependencies and generate contextually rich language. Designed for real-time operation, the system is also suitable for low-resource devices, enabling impactful applications such as assistive tools for the visually impaired and intelligent video indexing. Despite high computational demands and sequence-length limitations, the system presents promising directions for future development, including multilingual captioning, multimodal audio–visual integration, and lightweight models like TinyGPT for enhanced portability.**

***Keywords*—Video captioning; transformer; timesformer; GPT-2; real-time inference; spatio-temporal attention; multimedia accessibility; CPU/GPU deployment**

I. Introduction

Automatic descriptions of visual content play a crucial role in improving accessibility, information retrieval, and the generation of multimedia content [1]. While this task may seem natural to humans, it remains complex for artificial intelligence, which has to transform visual data into relevant linguistic representations [2]. The generation of natural language descriptions for videos, known as video subtitling, thus represents a major challenge in computer vision and multimedia processing [3]. Advances in deep learning have enabled significant progress, but conventional encoder-decoder architectures still suffer from redundancies and a lack of semantic consistency. Modern approaches, such as sequence-to-sequence models [4] combined with attention mechanisms, improve the alignment between vision and language, but their results sometimes remain insufficient. To overcome these limitations, large-scale language models (LLMs) [5] enhance the ability to generate accurate and consistent descriptions.

Video subtitling is the demanding task of transforming visual sequences into accurate, coherent textual descriptions. Among recent approaches, encoder-decoder architectures have demonstrated their effectiveness by combining networks such as InceptionResNetV2 for feature extraction and GRU for sequence generation. In parallel, large-scale language models (LLMs) have distinguished themselves by their ability to reason and generalize, opening up new opportunities for this application. In this context, GPT-2 [6], a lightweight, high-performance model compatible with non-GPU environments [7], was integrated.

With this in mind, a module optimized for video subtitling was designed, combining the TimeSformer-base-finetuned-k600 encoder with GPT-2 as decoder, in an architecture based on large-scale language models (LLM). This solution has been adapted to run efficiently on CPU-equipped systems, facilitating its deployment on modest hardware configurations. This study aims to evaluate the benefits of this LLM approach compared with conventional sequence-to-sequence architectures. The system, capable of processing videos in real time, relies on FFMPEG [8] for sequence fragmentation and ImageMagick [9]-[10] for automatic subtitle integration. A final evaluation confirms the robustness and relevance of the proposed solution in the field of video captioning.

This paper is structured into five main sections, each addressing a key aspect of the study. Section 2 reviews existing work on video captioning for the visually impaired, highlighting its limitations and presenting two approaches to improvement based on advanced video processing techniques. Section 3 describes the methodology used to develop the system's key modules, highlighting the technical choices made to enhance the user experience. Section 4 presents the experimental results, together with an analysis demonstrating the effectiveness of the proposed solutions. Finally, Section 5 summarizes the study's contributions and suggests directions for future research.

II. Related Work

In recent years, assistive technologies for the visually impaired have advanced considerably, particularly in navigation, access to environmental information, and video caption generation. M. Chen et al. [11] propose the TVT model, based on a Transformer architecture combining visual information and motion to produce textual descriptions. Although it performs well on MSVD and MSR-VTT sets, its use in real time remains limited due to its complexity.

Kevin et al. [12] introduced SWINBERT, which can capture complex spatiotemporal representations using a sparse

attention mask, reducing image redundancy. However, processing long, dense sequences generates a high computational cost, restricting their effectiveness in real time. L. Zhou et al. [13] conceive a model generating dense descriptions for non-truncated videos, relying on differentiable masking and attention mechanisms. Despite good textual consistency, the computational load associated with the use of Transformers compromises its real-time deployment. For their part, M. Amaresh and S. Chitrakala [14] are exploring encoder-decoder architectures combining CNN and LSTM, integrating attention mechanisms and spatiotemporal analysis, but oriented towards offline processing. Finally, L. Gao et al. [15] present the aLSTMs model, combining attention and LSTM to generate coherent descriptions. Although performing well on MSVD and MSR-VTT bases, this model remains limited by its computational complexity.

III. Proposed Method

This section describes the design and implementation of the two video subtitling architectures evaluated: a classic encoder-decoder model (InceptionResNetV2-GRU) and a transformer-based model (TimeSformer-GPT2). It details the visual and textual modules, the datasets used (MSVD and VATEX), and the technical optimizations (FFmpeg, ImageMagick), ensuring real-time processing on CPU and GPU.

A. Video Captioning Based on Encoder–Decoder Architectures

An automatic video captioning system is proposed based on an encoder-decoder architecture. It uses the Inception-ResNetV2 pre-trained CNN model to extract visual features, combined with a GRU (Gated Recurrent Unit) decoder to generate natural language text descriptions. The pipeline includes the encoding of videos as feature vectors, followed by the sequential generation of captions by the GRU. Developed with TensorFlow and Keras, the system is trained on the MSVD (Microsoft Video Description Corpus) dataset. It improves on previous approaches based on VGG16 and LSTM, offering a richer visual representation and more efficient sequence production. This model is particularly well suited to applications such as accessibility for the visually impaired, indexing of video content, and metadata creation. How it works is illustrated in Figure 1.

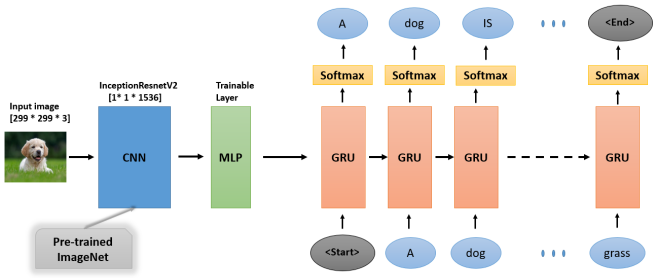


Fig. 1. General architecture of the encoder-decoder model for video captioning.

1) Visual feature extraction with InceptionResNetV2: Inception-ResNet-v2 is a neural network architecture that combines the advantages of Inception modules and residual connections [16], enabling efficient visual feature extraction while simplifying deep model training. It is based on three successive blocks - Inception-ResNet-A, B, and C - designed to process feature maps of decreasing dimensions. Block A operates on larger maps using convolutions of various sizes (1×1, 3×3) to capture information at different scales [16]-[17]. Block B continues this processing on intermediate maps, with controlled complexity, while Block C refines the representations on the most compact maps for final classification. Each of these blocks incorporates residual connections, which improve gradient propagation and stabilize learning, notably by alleviating the vanishing gradient problems encountered in very deep networks. The following figure illustrates the organization of these blocks. How it works is illustrated in Figure 2.

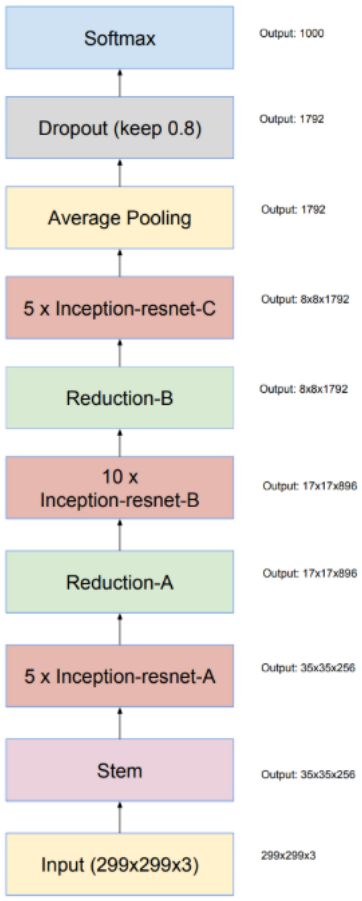


Fig. 2. Internal structure of the InceptionResNetV2 network.

2) Sequential text generation using GRU decoder: The decoder transforms visual representations into descriptive sentences, based on an RNN-type architecture. While conventional recurrent networks can model short-term dependencies, they run into difficulties with long sequences, notably due to the disappearance or explosion of the gradient. To overcome these limitations, we use gated recurrent units (GRUs) [18], which

control sequential learning using internal update, reset, and hidden memory mechanisms, as illustrated in figure 3. The decoder comprises three main components: an integration layer that encodes words in vector form, a multilayer GRU for temporal modeling, and a final linear layer that projects the hidden state into vocabulary space to predict the next word.

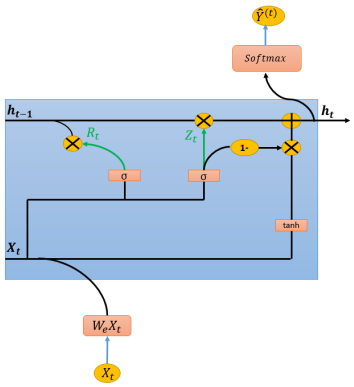


Fig. 3. Internal GRU mechanism (Gated Recurrent Unit).

In these equations (1, 2, 3, and 4) [19], x_t represents the input, and h_t is the hidden state at time t. The weights associated with the reset, update, and new information creation gates are denoted as W_r, W_z, and W_u, respectively. The hyperbolic tangent and sigmoid activation functions are symbolized by $\tanh$ and σ, respectively.

$$r_t = \sigma(W_{xr}x_t + U_{hr}h_{t-1}) \quad (1)$$
$$z_t = \sigma(W_{xz}x_t + U_{hz}h_{t-1}) \quad (2)$$
$$u_t = \tanh(W_{xu}x_t + U_{hu}(r_t \odot h_{t-1})) \quad (3)$$
$$h_t = (1 - z_t)h_{t-1} + z_t u_t \quad (4)$$

3) Parameters for the encoder-decoder video captioning model: Table I shows the main hyperparameters of a video caption generation model, combining InceptionResNetV2 for the extraction of visual representations and a GRU-based sequential decoder. The model processes 80 frames per video, from which it extracts vectors of dimension 1536. The text sequences produced are limited to 10 words, with a vocabulary of 1,500 tokens. Learning is performed over 3 epochs, with a learning rate of 0.0007 and a batch size of 8. 15% of the data is reserved for validation to control overlearning.

4) Dataset used: Microsoft Video Description Corpus (MSVD): As part of the experiment, the MSVD (Microsoft Video Description Corpus) dataset was selected for training and evaluation of the model. It consists of YouTube videos selected via Amazon Mechanical Turk and annotated with one-sentence descriptions. Only English captions are retained, after light pre-processing including lowercase casing, tokenization, and punctuation removal. The data follow the standard distribution proposed in [4]-[20], with 1,200 videos for learning, 100 for validation, and 670 for testing [21]. For each video, an image is extracted every ten frames, providing balanced temporal coverage. The set comprises 1,970 clips, each around 10 seconds long, and associated with around 40 descriptions, providing a linguistic diversity useful for training the generation system.

B. Caption Generation with Large Language Models (LLMs)

1) GPT-2 Integration for descriptive text generation: This model is an adaptation of the Vision Transformer (ViT) [22], [23] to video data processing, enabling the simultaneous capture of spatial and temporal dynamics. Unlike conventional ViTs designed for still images, ViViT segments a video into a sequence of images, converted into tokens enriched by positional and identity embeddings. These representations are then processed successively by a spatial encoder and a temporal encoder to model the structural and evolutionary relationships between images.The architecture retains the structure of the ViT encoder, with adjustments to exploit dependencies between successive images. The overall operation of the model is shown in Figure 4.

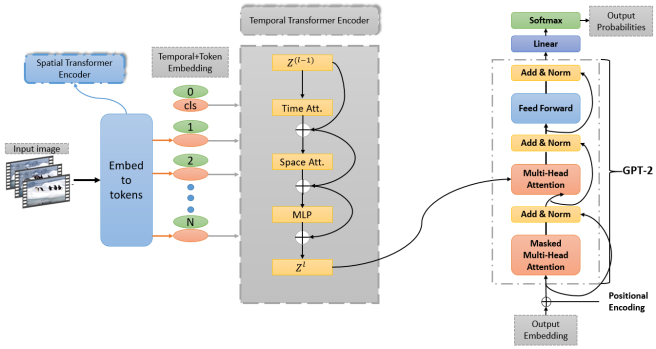


Fig. 4. TimeSformer + GPT-2 global architecture for video description generation.

2) Spatio-temporal analysis with timesformer encoder: The basic model is based on the "Base" structure of the Vision Transformer (ViT) and uses sequential spatio-temporal attention, applying attention first to the temporal axis, then to the spatial axis. This method, fitted to the Kinetics-600 dataset - comprising 392,000 videos for training and 30,000 for validation, spread over 600 human action categories - outperformed alternative approaches exploiting parallel or inverted attention. The TimeSformer model adopts this principle, illustrated in Figure 5, by dividing processing into two successive stages: temporal attention followed by spatial attention. This separation enables more efficient extraction of spatio-temporal features from videos.

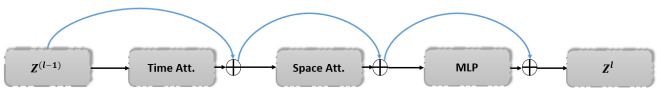


Fig. 5. Sequential spatial-temporal attention mechanism in timesformer.

At each layer, temporal attention is applied, followed by spatial attention and an MLP block, all integrated into a residual connection. This structure improves learning efficiency while reducing computational complexity, as shown by the $Z^{(l)}$ output described in equation 5.

TABLE I. CONFIGURATION PARAMETERS OF THE ENCODER-DECODER VIDEO CAPTIONING MODEL

Parameter	Value	Description
video	None	Name of the video file to process
keep_temp	False	Keeps the extracted images if set to True
train_path	data/training_data	Folder containing the training videos
test_path	data/testing_data	Folder containing the testing videos
max_length	10	Maximum length of the text captions
batch_size	8	Number of samples per training batch
lr	0.0007	Learning rate for the optimizer
epochs	100	Number of training epochs
latent_dim	512	Latent dimension size of the GRU layer
validation_split	0.15	Proportion of data used for validation
num_encoder_tokens	1536	Dimension of the video feature vectors
num_decoder_tokens	1500	Number of words in the output vocabulary
time_steps_encoder	80	Number of frames extracted per video

$$Z^{(l)} = Z^{(l-1)} + \text{TimeAtt}(Z^{(l-1)}) + \text{SpaceAtt}(Z') + \text{MLP}(Z'') \tag{5}$$

with :

$$Z' = Z^{(l-1)} + \text{TimeAtt}(Z^{(l-1)})$$
$$Z'' = Z' + \text{SpaceAtt}(Z')$$

The labels "Frame $t-\delta$", "Frame t", and "Frame $t+\delta$" denote the previous, current, and next frames in the video sequence. These serve as key inputs for spatio-temporal attention, enabling the model to capture temporal and spatial dependencies, as shown in Figure 6.

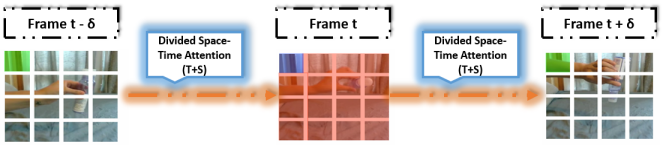


Fig. 6. Sequence of frames used for spatio-temporal attention.

3) Unified decoder for contextual language generation: It acts like a decoder, progressively generating a textual description. Each word is predicted by taking into account the words already produced, as well as the visual context extracted from the video. This mechanism ensures the harmonious integration of visual and linguistic information, resulting in an accurate and fluid description [23], [24].

Video subtitling begins by extracting visual features using pre-trained models such as ResNet, TimeSformer [23], [24] or I3D, which transform images into representative vectors. These vectors are then adapted to the GPT-2 input by adding special tokens (e.g., [CLS] for onset and [SEP] for separation). Trained on video-text corpora such as VaTeX [25], GPT-2 learns to generate relevant and coherent descriptions from visual content.

The unified video captioning framework based on GPT-2 models the conditional probability $P(C \mid V)$, where C is the caption and V the input video. It predicts each word in the output sequence based on the visual features extracted from V, as shown in Equation 6.

$$P(C \mid V) = \prod_{t=1}^{N} P(c_t \mid c_{<t}, \text{VideoEmbed}(V), \theta) \tag{6}$$

where c_t represents the t-th token of the caption, VideoEmbed(V) denotes the visual features mapped to GPT-2's embedding space, and θ refers to the model's parameters.

The training objective consists of minimizing the negative log-likelihood of the predicted caption sequence conditioned on the video features, as shown in Equation 7.

$$\mathcal{L} = -\sum_{t=1}^{N} \log P(c_t \mid c_{<t}, \text{VideoEmbed}(V), \theta) \tag{7}$$

This fine-tuning process adapts GPT-2's language generation capabilities to the video-text domain, enabling the production of captions that are both semantically precise and aligned with visual content. As illustrated in Figure 7, each frame is divided into 2D patches—typically 7×7—then transformed into visual tokens by a Vision Transformer such as TimeSformer or CLIP. These tokens, enriched with spatial and temporal context, are structured using special markers like [CLS] and [SEP], then passed to a late fusion module that prepares them for decoding. GPT-2, operating autoregressively, generates one word at a time by leveraging both prior textual outputs and the embedded video information. It produces a sequence of logits that guide the selection of the most probable next token at each step. This integrated architecture ensures a smooth fusion of visual cues and language modeling, yielding accurate and naturally flowing captions.

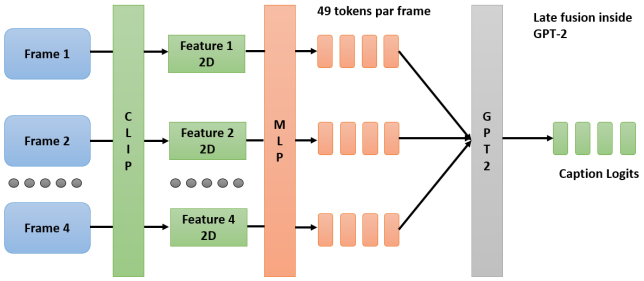


Fig. 7. Video subtitle generation pipeline via TimeSformer and GPT-2.

4) Performance optimization on CPU vs. GPU: A [26]-compatible GPU is an ideal solution for training video subtitling models such as TimeSformer and GPT-2, thanks to its ability to handle high resolutions, large batches of data, intensive spatio-temporal processing, as well as the use of mixed precision. Conversely, a CPU [26], although more affordable and accessible, is more suited to tasks of lower complexity due to its limited performance. The following table II compares learning parameters across CPU and GPU environments, detailing batch sizes, image resolutions, number of threads, and training schemes, to optimize performance according to available hardware resources.

During inference, a time clip is extracted from the center of the video. On a processor (CPU), the image is resized to 112 pixels on its shortest side, followed by a 112 × 112 center crop, to limit computation. A single softmax score is then used to generate the prediction. On a GPU, the TimeSformer-HR model exploits a higher resolution: resizing is performed at 224 pixels, with three 224 × 224 cuts analyzed to enrich the spatial information. The final prediction, obtained by averaging the softmax scores, thus benefits from the GPU's ability to process high-resolution data, enhancing the model's accuracy and robustness. Figure 8 compares two architectures designed for real-time caption generation. Sub-figure 8(a) illustrates the GPT-2 architecture, responsible for encoding the data and producing the textual descriptions. Sub-figure 8(b) shows the optimized version of TimeSformer, trained on Kinetics-600, capable of efficiently capturing spatio-temporal relationships in video sequences.

5) Dataset used (VATEX): VATEX is a large-scale dataset comprising around 41,250 10-second video clips, each accompanied by 10 manually annotated English captions. Learning is based on the official set, while performance is measured using the public test set [21], [25].Three Vision Transformer encoders are used for visual feature extraction, including an I-frame encoder initialized with the CLIP model [27], trained on the LAION-400M image-text corpus [28]. In comparison, SwinBERT [12] relies on the VidSwin architecture, trained on the video-oriented Kinetics-600 [29] dataset. It should be noted that LAION-400M is an image-text corpus, Kinetics-600 targets videos, and VATEX serves here as a reference for experimentation. SwinBERT outperforms the proposed method, thanks in particular to the effectiveness of its pre-training phase on Kinetics-600.

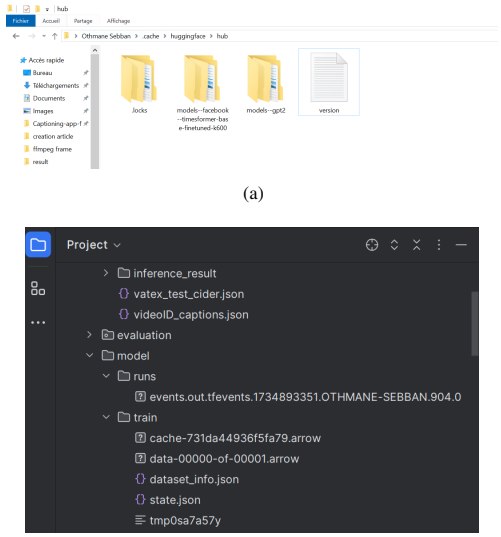


Fig. 8. Illustration of video captioning components: (a) GPT-2 used for text decoding, and (b) TimeSformer for encoding video frames into visual embeddings.

C. System Optimization Using FFmpeg and ImageMagick

To train the models, FFmpeg [8] was used, integrated via MoviePy [30], to facilitate the manipulation of complex video content and to optimize the use of large language models (LLMs). This free software supports a wide variety of multimedia formats, including video and audio, and includes a library dedicated to keyframe extraction. These keyframes generally represent significant visual changes. However, FFmpeg does not detect more subtle variations, often referred to as interesting images, which may nonetheless contain information relevant to analysis. MoviePy [31] simplifies the addition of text to videos by automating its generation and positioning using the TextClip class, then embedding it in the video via CompositeVideoClip, with FFmpeg support in the background. Once the text has been merged, it is encoded using standard codecs such as libx264 for video and AAC for audio, ensuring optimum compatibility with the majority of media players. In addition, key frame extraction aims to identify representative frames in a video sequence. This process can be carried out using a variety of tools, including FFmpeg, OpenCV's absdiff() function or the DMD algorithm [32]. Figure 9 illustrates the images extracted from a video file in MP4 format using the following FFmpeg command: `ffmpeg -i file.mp4 -vf "select='not(mod(n,4))',scale=320:240" -vsync vfr -frames:v 4 frame_%03d.png`.

ImageMagick is currently used to integrate full descriptions into videos as part of the subtitling process. An integrated tool [32], combined with a command-line routine, automates the cropping of images by removing 8 pixels at the left and 40 pixels at the top, reducing their size by around 20%. In addition, partial conversion of PPM files to JPEG is performed by a batch script. Interstitial spaces between mosaic segments were also corrected for archiving purposes. In all, the 67 selected mosaics and still images occupy 764 Kb.

TABLE II. HARDWARE CONFIGURATION COMPARISON FOR TRAINING CAPTIONING MODELS ON CPU VS. GPU

Parameter	**CPU**	**GPU**	**Description**
Device	”cuda” if GPU available, else ”cpu”	”cuda” if GPU available, else ”cpu”	Automatic device detection
Encoder Model	facebook/timesformer-base-finetuned-k600	facebook/timesformer-base-finetuned-k600	Model used for encoding
Decoder Model	Gpt-2	Gpt-2	Model used for decoding
Video Frames	4	16	Number of frames used per video
Image Resolution	112×112	224×224	Size of the images used
Batch Size	2	6	Batch size for training
Batch Learning	Disabled	Enabled	Use of `GradScaler` for AMP
Learning Rate	$1e^{-5}$	$5e^{-7}$	Learning rate for the optimizer
Epochs	100	100	Number of epochs for training
Scheduler	linear	linear	Learning rate scheduler
Num Workers	1	8	Number of workers for `DataLoader`
Collate Function	custom.collate.fn	default.data.collator	Data collation method
Pin Memory	False	True	Memory management in `DataLoader`



Fig. 9. Extracting key frames from video using FFmpeg.

IV. EXPERIMENTAL RESULTS

The Experimental Results section provides an in-depth analysis of the system's performance, examining each of its key modules. Subsection A outlines the evaluation criteria and the fundamental components of the system. Subsection B presents the results obtained by the video subtitle generation models developed according to the two approaches explored.

A. The Evaluation Metrics

To assess the quality of the trained model, four metrics are used. Among these, the BLEU (Bilingual Evaluation Understudy) metric quantifies linguistic similarity by comparing the n-grams shared between the generated sentences and the reference sentences. The calculation of the BLEU score is presented in equation 8.

$$\text{BLEU} - N(ci, Si) = b(ci, Si) \exp\left(\sum_{n=1}^{N} \omega_n \log P_n(ci, Si)\right) \tag{8}$$

where

$$b(ci, Si) = \begin{cases} 1 & \text{if } lc > ls \\ e^{1-\frac{ls}{lc}} & \text{if } lc \leq ls \end{cases}$$ is a brief penalty; lc is

The total length of the sentences generated (candidates) is noted as l_S, while l_c designates the optimal reference length of the corpus. When a candidate sentence is associated with several references, the one whose length is closest to that of the candidate is selected. In addition, the weights ω_n, assigned to the n-grams, are generally set to a constant value, as indicated in [18]-[33], as illustrated in the equation 9.

$$P_n(ci, Si) = \frac{\sum_k \min(h_k(ci), \max(h_k(S_{ij})))_{j \in \mathcal{M}}}{\sum_k h_k(ci)} \tag{9}$$

B-N, short for BLEU-N, is a precision measure that evaluates the quality of generated sentences by focusing on short n-grams, generally up to 4 words. It calculates the proportion of linguistic units (n-grams) shared between a candidate sentence and one or more reference sentences.

METEOR is an evaluation metric that measures both precision and recall at the unigram level. It takes into account not only exact matches, but also synonyms from WordNet and partial matches based on word fragments (truncated tokens). The METEOR score is given by the equation 10 [18]-[33].

$$\text{METEOR} = (1 - \text{Pen}) \times \text{Fmean} \tag{10}$$

The term $\text{Pen} = \gamma \left(\frac{\text{ch}}{m}\right)^m$ represents a penalty factor. In this expression, m corresponds to the total number of alignments between the candidate sentence and the reference sentence, while ch designates the number of contiguous segments correctly aligned (words identical and in the same order), as specified in equations 11, 12, and 13.

$$P_m = \frac{|m|}{\sum_k h_k(ci)} \tag{11}$$

$$R_m = \frac{|m|}{\sum_k h_k(S_{ij})} \tag{12}$$

$$F_{\text{mean}} = \frac{P_m \cdot R_m}{\alpha P_m + (1 - \alpha) R_m} \tag{13}$$

METEOR is a measure based on the harmonic mean between precision and recall, calculated between a reference sentence and the best-aligned candidate.

ROUGE-L is based on the longest common subsequence (LCS) identified between two sentences. The length $l(ci, sij)$ of this LCS forms the basis of its evaluation, as shown in equation 14.

$$\text{ROUGE-L}(ci, Si) = \frac{(1+\beta^2) \cdot Rl \cdot Pl}{Rl + \beta^2 \cdot Pl} \tag{14}$$

where:

$$Rl = \max_j \left(\frac{l(ci, S_{ij})}{|S_{ij}|} \right) \quad \text{(recall based on LCS)}$$

$$Pl = \max_j \left(\frac{l(ci, S_{ij})}{|ci|} \right) \quad \text{(precision based on LCS)}$$

The **ROUGE-L** metric evaluates precision based on the longest common subsequence (LCS), while the parameter$E$ is a constant to reinforce the importance of recall in the calculation. Other variants of the RED metric, such as **ROUGE-N** and **ROUGE-S**, are also used for finer comparisons [18]. For its part, **CIDEr-D** measures the average cosine similarity between the n-grams of a candidate sentence and those of references, taking into account both precision and recall. Its mathematical formulation is given by equation 15 [18]-[33].

$$CIDEr - D_n(ci, Si) = \sum_{n=1}^{N} \omega_n \text{CIDEr-D}(ci, Si) \tag{15}$$

This formula aggregates the *CIDEr-D* scores calculated for different levels of *n-grams* (from 1-gram to N-gram), assigning them a specific weight ω_n. This enables us to assess the similarity between the generated legend c_i and reference legends S_i at several levels of linguistic granularity, while incorporating a weighting mechanism that mitigates the impact of excessive repetition and too-frequent *n-grams*.

B. Quantitative Results

The proposed method was evaluated using the BLEU, ROUGE-L, METEOR, and CIDEr metrics to measure its efficiency. Tests were carried out on two Seq2Seq models based on LSTM and GRU, respectively, as well as on the Timesformer-base fine-tuned on Kinetics-600 with GPT-2, evaluated on both CPU and GPU.

The Neleac/timesformer-gpt2-video-captioning pre-trained model [22], [23] is based on a modular architecture built around three components. A visual preprocessor (MCG-NJU/videomae-base) adapts images extracted from videos, a tokenizer (GPT-2) encodes text sequences, and an encoder-decoder model combines a Vision Transformer with a GPT-2 decoder to automatically produce subtitles. This configuration, illustrated in figure 10, enables efficient, context-sensitive generation of descriptions from video content.

The results, presented in Table III, show the best performances in bold. Three modules were evaluated, including the pre-trained Timesformer-GPT2 subtitling model, as well as two encoder-decoder architectures designed for video description generation. Compared with previous methods, the proposed approach outperformed across all considered metrics, including BLEU-1 to BLEU-4, METEOR, CIDEr, and ROUGE-L.

```python
import av
import numpy as np
import torch
from transformers import AutoImageProcessor, AutoTokenizer, VisionEncoderDecoderModel

device = "cuda" if torch.cuda.is_available() else "cpu"

# load pretrained processor, tokenizer, and model
image_processor = AutoImageProcessor.from_pretrained("MCG-NJU/videomae-base")
tokenizer = AutoTokenizer.from_pretrained("gpt2")
model = VisionEncoderDecoderModel.from_pretrained("Neleac/timesformer-gpt2-video-captioning").to(device)

# load video
video_path = "never_gonna_give_you_up.mp4"
container = av.open(video_path)

# extract evenly spaced frames from video
seg_len = container.streams.video[0].frames
clip_len = model.config.encoder.num_frames
indices = set(np.linspace(0, seg_len, num=clip_len, endpoint=False).astype(np.int64))
frames = []
container.seek(0)
for i, frame in enumerate(container.decode(video=0)):
    if i in indices:
        frames.append(frame.to_ndarray(format="rgb24"))

# generate caption
gen_kwargs = {
    "min_length": 10,
    "max_length": 20,
    "num_beams": 8,
}
pixel_values = image_processor(frames, return_tensors="pt").pixel_values.to(device)
tokens = model.generate(pixel_values, **gen_kwargs)
caption = tokenizer.batch_decode(tokens, skip_special_tokens=True)[0]
print(caption) # A man and a woman are dancing on a stage in front of a mirror.
```

Fig. 10. Example of generated video captioning from the TimeSformer-Gpt2 pre-trained model.

The models compared in the table fall into two categories: on the one hand, classic encoder-decoder architectures based on InceptionResNetV2 coupled with GRU; on the other, those based on the combination of TimeSformer and GPT-2. The latter captures spatio-temporal relationships more effectively and produces more accurate descriptions. Although they perform well on the CPU, they reach their full potential on the GPU, particularly in their pre-trained version, which considerably accelerates inference. Overall, this architecture outperforms encoder-decoder approaches, but at the cost of higher hardware consumption.

V. Comparative Analysis and Interpretation

This study compared two architectures for video captioning: a traditional solution based on an encoder-decoder framework combining InceptionResNetV2 and GRU, and a more advanced approach integrating a pre-trained TimeSformer encoder with a GPT-2 decoder. Experimental results demonstrate that the TimeSformer-GPT2 configuration, when executed on a GPU, achieves superior performance, notably with a BLEU-4 score of 0.276, compared to only 0.058 for the baseline. Improvements were also recorded in METEOR (0.821) and CIDEr (0.889), highlighting the effectiveness of combining a spatio-temporal attention mechanism with a generative language model for producing coherent and accurate outputs.

In particular, while the pre-trained Neleac/timesformer-gpt2-video-captioning model proved highly effective, the optimized Timesformer-base fine-tuned (GPU) version demonstrated superior performance on the majority of evaluation metrics, validating the effectiveness of the implemented adaptation process.

These findings emphasize the strength of TimeSformer in capturing complex temporal and spatial dependencies within video sequences. Coupled with GPT-2's capacity to generate contextually appropriate language, this synergy allows for a

TABLE III. QUANTITATIVE EVALUATION OF VIDEO CAPTIONING MODELS USING BLEU, METEOR, ROUGE-L, AND CIDER METRICS

Model	B-1	B-2	B-3	B-4	ROUGE-L	CIDEr	METEOR
GRU-based Encoder–Decoder (our model)	0.685	0.331	0.250	0.058	0.241	0.372	N/A
Timesformer-base finetuned (CPU) (our model)	0.803	0.492	0.370	0.182	0.567	0.765	0.714
Timesformer-base finetuned (GPU) (our model)	**0.887**	**0.571**	**0.498**	**0.276**	**0.687**	**0.889**	**0.821**
Timesformer-GPT2 video captioning (GPU) [22], [23]	0.872	0.562	0.475	0.259	0.672	0.884	0.812

deeper understanding and expression of visual content. The architecture relies on a divided attention mechanism—applying temporal and then spatial attention—to enhance the quality of feature extraction. GPT-2, in turn, adapts its linguistic output based on these extracted patterns, resulting in semantically rich and logically structured narratives.

Compared to models such as SWINBERT and TVT, the proposed method demonstrates a notable reduction in computational overhead, especially during inference on CPU-based systems. While many transformer-based solutions are primarily designed for offline processing, the presented system remains compatible with lower-resource environments without compromising output quality. These results align with the contributions of Bertasius et al. and Radford et al. [22]-[23], while offering a more scalable alternative for practical deployment.

From a conceptual standpoint, this work supports the hypothesis that coupling a high-performance visual encoder with a Large Language Model (LLM) improves the generation of multimodal language. Practically, it illustrates the feasibility of integrating sophisticated captioning mechanisms into mobile or embedded systems, particularly for assistive technologies aimed at the visually impaired or for intelligent indexing of audiovisual content.

Despite encouraging results, the proposed approach presents some limitations. The training phase, especially when involving the TimeSformer component, requires considerable computational resources. Furthermore, the fixed-length format (limited to 10 words) imposes constraints on expressive capacity, particularly in visually complex scenes. Additionally, the evaluation was limited to two datasets (MSVD and VATEX); further validation on diverse, especially multilingual, corpora would be beneficial to assess the generalizability of the system.

Analysis of the videos in the MSVD and VATEX datasets reveals limitations specific to each model. The GRU model favors static elements, sometimes neglecting the main action, while the Timesformer-base finetuned (GPU) model, while effective at capturing movement, can confuse context or over-interpret complex scenes. These errors highlight the difficulty of accurately understanding a scene and underscore the value of multimodal approaches that combine visual, sound, and temporal signals.

Potential enhancements include the incorporation of compact architectures such as TinyGPT and the use of knowledge distillation techniques to reduce training complexity. Another promising direction involves exploring multimodal pipelines that combine audio and visual data to further enrich the generated outputs. Moreover, extending the system to handle real-time processing of long-form video content or deployment in unstructured environments remains a key challenge for improving portability and robustness. Figure 11 shows the final results of the proposed system, illustrating optimal performance in terms of video captioning. It highlights the model's ability to produce accurate and consistent descriptions, demonstrating its effectiveness.

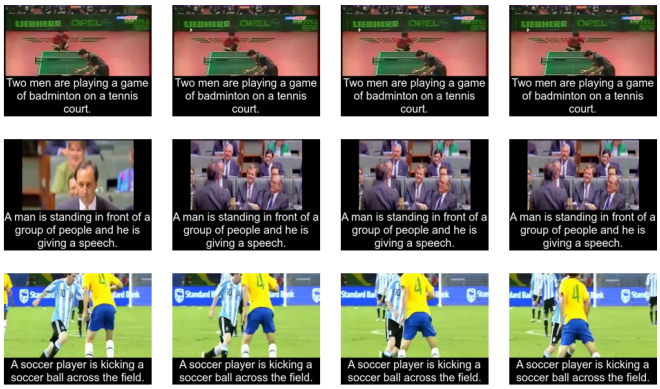


Fig. 11. Video captioning: key frame extraction with FFmpeg and caption insertion using ImageMagick.

VI. CONCLUSION AND FUTURE WORK

This paper has proposed a comparative study between classical video captioning architectures, such as InceptionResNetV2-GRU, and advanced approaches combining TimeSformer and GPT-2. Experiments on MSVD and VATEX datasets demonstrated the clear superiority of the TimeSformer-GPT2 model, particularly on GPUs, with high scores according to the BLEU, METEOR, ROUGE-L, and CIDEr metrics. This performance is explained by the model's ability to capture spatio-temporal dependencies while generating linguistically consistent descriptions efficiently. The proposed pipeline, integrating FFmpeg and ImageMagick tools, demonstrates the feasibility of a robust and adaptable automated subtitling system, compatible with different hardware environments. This flexibility opens up concrete prospects, notably in assistive technologies for the visually impaired and intelligent indexing of audiovisual content. Despite the promising results, there are several avenues for improvement worth exploring. Adopting lighter models such as TinyGPT or DistilGPT would reduce the computational load and facilitate deployment on mobile devices. The addition of audio modality could enrich the descriptions generated, particularly for complex scenes.

Furthermore, adaptation to specific multilingual corpora (educational, medical) would reinforce the system's generalizability. Finally, the processing of long continuous videos with dynamic generation of summaries or subtitles would

pave the way for more advanced applications, notably in visual assistance or intelligent streaming. This work paves the way for a new generation of intelligent video subtitling systems, capable of reconciling performance, accessibility, and lightness, and designed for concrete applications on mobile and embedded platforms.

ACKNOWLEDGMENT

Dr. Othmane Sebban thanks Dr. Ahmed Azough and Dr. Mohamed Lamrini, as well as the research team, for their essential support in this study.

REFERENCES

[1] L. Guo et al., "Non-autoregressive image captioning with counterfactuals-critical multi-agent learning," Proceedings of the Twenty Ninth International Joint Conference on Artificial Intelligence, pp. 767–773, Jul. 2020. doi:10.24963/ijcai.2020/107.

[2] Abdar, Moloud et al., A Review of Deep Learning for Video Captioning. 2023.

[3] Y. Zheng, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Stacked multimodal attention network for context-aware video captioning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 1, pp. 31–42, Jan. 2022. doi:10.1109/tcsvt.2021.3058626.

[4] H. Wang, C. Gao, and Y. Han, "Sequence in sequence for video captioning," Pattern Recognition Letters, vol. 130, pp. 327–334, Feb. 2020. doi:10.1016/j.patrec.2018.07.024.

[5] C. Zhang et al., "A simple LLM framework for long-range video question-answering," Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 21715–21737, 2024. doi:10.18653/v1/2024.emnlp-main.1209.

[6] Y. Lu et al., "Set prediction guided by Semantic Concepts for diverse video captioning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 4, pp. 3909–3917, Mar. 2024. doi:10.1609/aaai.v38i4.28183.

[7] A. Tariq, M. Elhadef, and M. Ghani Khan, Vidcap-LLM: Vision-Transformer and large language model for video captioning with linguistic semantics integration, 2024. doi:10.2139/ssrn.4812289.

[8] R. Radarapu, A. S. Gopal, M. NH, and A. K. M., "Video summarization and captioning using dynamic mode decomposition for surveillance," International Journal of Information Technology, vol. 13, no. 5, pp. 1927–1936, May 2021. doi:10.1007/s41870-021 00668-0.

[9] S. Sarwar et al., "Advanced Audio Aid for Blind people," 2022 International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC), pp. 1–6, Dec. 2022. doi:10.1109/icetecc56662.2022.10069052.

[10] C. Soto and S. Yoo, "Visual detection with context for document layout analysis," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP IJCNLP), 2019. doi:10.18653/v1/d19-1348.

[11] M. Chen, Y. Li, Z. Zhang, et S. Huang, "TVT: Two-View Transformer Network for Video Captioning," Proceedings of Machine Learning Research, vol. 95, pp. 847–862, 2018.

[12] Lin, Kevin et al., Swinbert: End-to-End Transformers with Sparse Attention for Video Captioning. 2021.

[13] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018. doi:10.1109/cvpr.2018.00911.

[14] M. Amaresh and S. Chitrakala, "Video captioning using Deep Learning: An overview of methods, datasets and metrics," 2019 International Conference on Communication and Signal Processing (ICCSP), pp. 0656–0661, Apr. 2019. doi:10.1109/iccsp.2019.8698097.

[15] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," IEEE-Transactions on Multimedia, vol. 19, no. 9, pp. 2045–2055, Sep. 2017. doi:10.1109/tmm.2017.2729019.

[16] M. W. Kesiman, K. T. Dermawan, and I. G. Darmawiguna, "Balinese carving ornaments classification using INCEPTIONRESNETV2 architecture," 2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), pp. 1–5, Nov. 2022. doi:10.1109/cenim56801.2022.10037265.

[17] R. R. Rajalaxmi et al., "Deepfake detection using inception-ResNet-V2 Network," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), pp. 580–586, Feb. 2023. doi:10.1109/iccmc56507.2023.10083584.

[18] V. KILIC¸, "Deep gated recurrent unit for Smartphone-based image captioning," Sakarya University Journal of Computer and Infor mation Sciences, vol. 4, no. 2, pp. 181–191, Aug. 2021. doi:10.35377/saucis.04.02.866409.

[19] C.Zhu,Q.Jia, W.Chen, Y. Guo, and Y. Liu, "Deep learning for video-text retrieval: A Review," International Journal of Multimedia Information Retrieval, vol. 12, no. 1, Feb. 2023. doi:10.1007/s13735-023-00267-8.

[20] S. Venugopalan et al., "Translating videos to natural language using deep recurrent neural networks," Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015. doi:10.3115/v1/n15-1173.

[21] Y. Shen et al., "Accurate and fast compressed video captioning," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15512–15521, Oct. 2023. doi:10.1109/iccv51070.2023.01426.

[22] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?", arXiv preprint arXiv:2102.05095, 2021.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," OpenAI, 2019.

[24] D. Xu, W. Zhao, Y. Cai, and Q. Huang, "Zero-textcap: Zero-shot framework for text-based image captioning," Proceedings of the 31st ACM International Conference on Multimedia, pp. 4949–4957, Oct. 2023. doi:10.1145/3581783.3612571.

[25] X. Wanget al., "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," 2019 IEEE/CVF Interna tional Conference on Computer Vision (ICCV), pp. 4580–4590, Oct. 2019. doi:10.1109/iccv.2019.00468.

[26] H.-R. Huang et al., "Accelerating video captioning on Heterogeneous System Architectures," ACM Transactions on Architecture and Code Optimization, vol. 19, no. 3, pp. 1–25, May 2022. doi:10.1145/3527609.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning Transferable Visual Models from Natural Language Supervision," arXiv preprint arXiv:2103.00020, 2021.

[28] C. Schuhmann, R. Beaumont, R. Vencu, et al., "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs", arXiv preprint arXiv:2111.02114, 2021.

[29] A. Li, H. Wang, L. Torresani, and G. Mori, "The AVA-Kinetics Localized Human Actions Video Dataset", arXiv preprint arXiv:2005.00214, 2020.

[30] A. A. Zhuravlev and K. A. Aksyonov, "Dependence comparison of the effectivness of software tools for splitting video into frames on format and resolution using a road survey as an example," 2024 International Russian Automation Conference (RusAutoCon), pp. 468–472, Sep. 2024. doi:10.1109/rusautocon61949.2024.10693956.

[31] S. R. Solanki and D. K. Khublani, "From script to screen: Unveiling text-to-video generation," Generative Artificial Intelligence, pp. 81–112, 2024. doi:10.1007/979-8-8688-0403-8 3.

[32] B. Taraghi, H. Amirpour, and C. Timmerer, "Multi-codec Ultra High Definition 8K MPEG-Dash Dataset," Proceedings of the 13th ACMMultimedia Systems Conference, Jun. 2022. doi:10.1145/3524273.3532889.

[33] M. S. Wajid, H. Terashima-Marin, P. Najafirad, and M. A. Wajid, "Deep learning and knowledge graph for image/video captioning: a review of datasets, evaluation metrics, and methods," Engineering Reports, vol. 6, no. 1, 2024, doi: 10.1002/eng2.12785.

Endometriosis Lesion Classification Using Deep Transfer Learning Techniques

Shujaat Ali Zaidi[1], Varin Chouvatut[2]*, Chailert Phongnarisorn[3]
Department of Computer Science-Faculty of Science, Chiang Mai University, Mueang, Chiang Mai 50200, Thailand[1,2]
Department of Obstetrics and Gynecology-Faculty of Medicine, Chiang Mai University,
Mueang, Chiang Mai 50200, Thailand[3]

***Abstract*—In resource-limited settings, assisting physicians with disease identification can significantly improve patient outcomes. Early diagnosis is crucial, as many patients could remain healthy with timely intervention. Recent advancements in deep learning models for medical image processing have enabled algorithms to achieve diagnostic accuracy comparable to that of healthcare professionals. This research aims to develop a comprehensive system for the rapid and precise detection of endometriosis lesions. We explore the several deep transfer learning architectures, specifically MobileNetV2, VGG19, and InceptionV3, on the Gynecologic Laparoscopy Endometriosis Dataset (GLENDA). Through extensive literature review and parameter optimization, we find that MobileNetV2 outperforms the other models in terms of accuracy. However, challenges remain, as healthcare imaging datasets often suffer from limited sample sizes and uneven class distributions. Collecting additional samples can be costly and time-consuming, which is a prevalent issue in medical imaging. To address this, we employ Deep Convolutional Generative Adversarial Networks (DCGAN) to enhance the dataset by generating synthetic images, thus improving class balance. This image augmentation strategy not only boosts model performance but also reduces the manual effort required for image labeling. We evaluate our proposed model using metrics such as accuracy, precision, recall, and F1-score. Initially, our model achieves an accuracy of 95%. The introduction of synthetic samples results in an increased accuracy of 99%, reflecting a 4% improvement and enhancing the model's overall efficacy.**

***Keywords*—Endometriosis classification; lesion detection; medical image classification; deep learning; transfer learning; DCGAN**

I. Introduction

Endometriosis is a medical condition in which endometrial cells, naturally found within the uterus, begin to grow outside the uterine cavity [1], [2]. The preliminary hypothesis describing this is retrograde menstruation, where menstrual tissue pursues a distinctive path. Rather than being expelled during the menstrual cycle, the tissue flows in reversal, traveling up through the fallopian tubes and potentially implanting in the ovary or within the abdominal cavity, resulting in lesion construction. These lesions are differ in size, with classification based on dimensions critical to diagnosis and treatment. Mainly, lesions measuring five millimeters or more are typically classified as deep-seated endometriosis [3], [4].

Endometriosis is frequently associated with significant diagnostic delays [5], [6]. Endometriosis disorder has a broad spectrum of symptoms, such as menstrual pain, acute pelvic pain, infertility, and pain during intercourse. Non-gynecological symptoms can also appear, including pain while urinating, discomfort during bowel movements, flank pain, fatigue, and blood in the urine, among others. Furthermore, the biological exams are complicated, and results can differ widely, even in examinations conducted by professionals in the field [7], [8].

Endometriosis is a severe gynecological problem impacting an estimated 190 million women internationally [9], [10]. The disease is prevalent among women of different age class, from infant women to those who are post-menopause. Studies estimate that approximately 10 percent of women of reproductive age are affected by endometriosis [11], with a prevalence of 2 percent to 4 percent among postmenopausal women [12]. Additionally, The disease's influence on young women is significant, with nearly 50 percent of those undergoing persistent pelvic pain occurring before the age of 20 being diagnosed with endometriosis [4].

Endometriosis can be diagnosed through various procedures, including ultrasound, magnetic resonance imaging (MRI), and laparoscopy [13], [14]. Laparoscopy is the gold benchmark, as it permits explicit examination of the abdominal and pelvic areas through a camera inserted via a small incision. MRI, a non-invasive imaging method, uses magnetic areas to create detailed images of the body's internal structures, making it beneficial in recognizing deep lesions or more extensive cysts associated with endometriosis. Ultrasound, generally performed transvaginally, is usually utilized for initial screening and assists in detecting ovarian cysts or other abnormalities [15], [16], [17], [18].

In recent years, computer-aided methodologies have become essential in medical fields for disease diagnosis, including identifying conditions such as heart disease [19], endometriosis lesions detection [20], and classifying breast cancer [21]. However, Healthcare imaging datasets frequently suffer from insufficiency and uneven class distributions. Moreover, acquiring additional samples is both expensive and time-intensive. This is a common problem in the medical image domain. Researchers endeavor to conquer this problem by utilizing data augmentation. Image augmentation methods typically fall into two main categories: conventional approaches and deep learning techniques. Conventional image augmentation methods apply simple manipulations like translation, cropping, flipping, or resizing. Deep learning-based augmentation utilizes state-of-the-art neural networks to create more adaptive, data-rich transformations. These two techniques enhance data diversity, helping models to generalize more effectively [22],

*Corresponding authors.

[23].

A promising deep-learning technique for image synthesis inspired by the game hypothesis is the Generative Adversarial Network (GAN) [24], [25]. In GANs, two networks are learned in an adversarial manner: one network, the generator, constructs artificial images, while the other, the discriminator, learns to distinguish between real and synthetic images. By repeating this adversarial approach, GANs enhance both networks' abilities, allowing the generation of high-quality, realistic images. The computer vision community has adopted GANs, directing to various variations developed for photorealistic image generation [26], [27].

Deep learning has attracted significant attention in many fields (for more detail, see [28], [29], [30], [31], [32]) such as the medical domain because of its efficacy in image analysis. It is employed for classification, enhancement of image quality, and segmentation of medical images. In recent years, deep learning techniques have rapidly evolved, with MobileNetV2 attaining popularity due to its compact architecture, which makes it perfect for various applications [33]. The MobileNetV2 architecture contains an internal structure with a linear bottleneck, an element that minimizes memory requirements for better processing. Therefore, to provide an accurate and efficient classification of endometriosis lesions, in this study, we propose an endometriosis image classification founded on the MobileNetV2 architecture. The proposed method uses MobileNetV2 as the base model for the transfer learning process. We add a global pooling layer and two fully connected layers to enhance the model's performance and refine classification outcomes [34]. The primary contributions of this research are as follows.

- We propose a synthesis of high-quality endometriosis lesions from laparoscopy images employing deep convolutional generative adversarial networks (DCGANs).
- Several state-of-the-art deep learning architectures, such as VGG19, InceptionV3, and MobileNetV2, are analyzed for this research. These architectures are fine-tuned on the laparoscopy images dataset to obtain efficient and accurate endometriosis lesion detection.
- We improve the training set of our deep learning model by including synthetic data, a technique that guides to enhanced classification results.

The remainder of the article is organized as follows: Section 2 explains the literature review on utilizing deep learning to analyze endometriosis lesions. Sections 3 and 4 present the proposed methodology and experimental result analysis. Consequently, Sections 5 and 6 provide the discussion and conclusion of the study, respectively.

II. Literature Review

Detecting endometriosis through laparoscopy imaging is challenging due to the disease's complex and varied presentation. Its lesions usually have subtle or indistinct characteristics, making them challenging to identify accurately, even by specialists. Machine learning applications aimed at diagnosing endometriosis through laparoscopy imaging still need to be developed, partially due to the limited access to labeled, representative datasets. This section summarizes recent research in machine learning for endometriosis diagnosis, focusing on techniques, methodologies, and the insights they provide.

Zaidi [5] employed a deep-learning based approach to detect endometriosis lesions from laparoscopy images.The team achieved a accuracy of 0.93% by applying Inception V3 model with 5-fold cross-validation techniques. Deep convolutional networks, exemplified as GoogLeNet [35], have been effectively employed across various applications, showcasing their versatility and robust feature extraction capabilities. These networks serve as robust backbones for many state-of-the-art deep-learning architectures used in image and video research analysis. By leveraging hierarchical layers of convolutional filters, they can capture complicated patterns and attributes, making them ideal for object recognition, classification, and segmentation tasks. Visalaxi [36] employed a deep-learning approach to categorize laparoscopy images associated with endometriosis. They employed the dataset for gynecologic laparoscopy related to endometriosis [37], comprising about 6,000 images from laparoscopy videos. Sixty percent of this dataset was used for training, and several architectures were tested to determine the model with the best performance. The ResNet50 architecture [38], loaded with ImageNet pre-trained weights [39], achieved the highest accuracy at 90 percent, with sensitivity and precision scores of 82 percent and 83 percent, respectively.Leibetseder's study used transfer learning and the Faster R-CNN [40] and Mask R-CNN [41] models to accurately detect endometriosis in laparoscopic images, getting a 32.4 percent precision. In addition, the research investigated different data augmentation methods, revealing that a blend of cropping and rotation yielded optimal results. The GLENDA dataset was employed as the primary data source in this research.

In 2021, Yun [42] and colleagues introduced a neural network model specifically designed to aid in classifying endometriosis. This study employed a convolutional neural network (CNN) architecture called VGGNet-16. Using a dataset of 6,478 histopathology images for training. The researchers aimed to develop a highly accurate system that could support and potentially enhance the diagnostic work performed by radiologists. In their research, Takahashi and colleagues [43] explored how computer vision techniques can support detecting endometriosis cancer in laparoscopic images. The team achieved a prediction accuracy of 90.29% by applying advanced neural network techniques. Sudalaimuthu [44] proposed an innovative approach recognized as Structural Similarity Analysis of Endometriosis (SSAE). This approach evaluates endometriosis progression in areas such as the ovaries, uterus, rectum, and peritoneum founded on laparoscopy images from the GLENDA image dataset. To enhance the model's robustness, images were subjected to data augmentation techniques, including horizontal and vertical shifts, rotation, shear, and zoom. The data was split, with 70 percent allocated for training purpose and the 30 percent for testing purpose. The U-Net architecture was utilized to explore factors like filter sizes and optimization methods. The highest results recorded were dice coefficient is 0.74 and an IoU is 0.72, though it is unclear if these metrics were specific to the test set.The literature on automated identification and segmentation of endometriosis remains limited, with few works addressing these areas. Among

the existing studies, most rely on images from laparoscopy procedures, which, while invasive, are the gold benchmark for diagnosis. Despite using these high-quality images, automated approaches are still being developed with the aim of improving accuracy and performance.

Our research differentiates itself from the abovementioned research by suggesting a data augmentation approach to conventional and deep learning approaches that encompass detection and classification tasks for analyzing deep or deep-seated endometriosis. Additionally, it focauses on laparoscopy images, seeking at premature examination beyond the essential for invasive procedures.

III. Materials and Methods

This section provides a detailed description of the materials used and the approach taken for classifying endometriosis lesions, including an overview of the image collection process, a summary of the overall methodology, and a discussion of the proposed methodology. Figure(1) provides a flowchart of the methodology followed in this research.

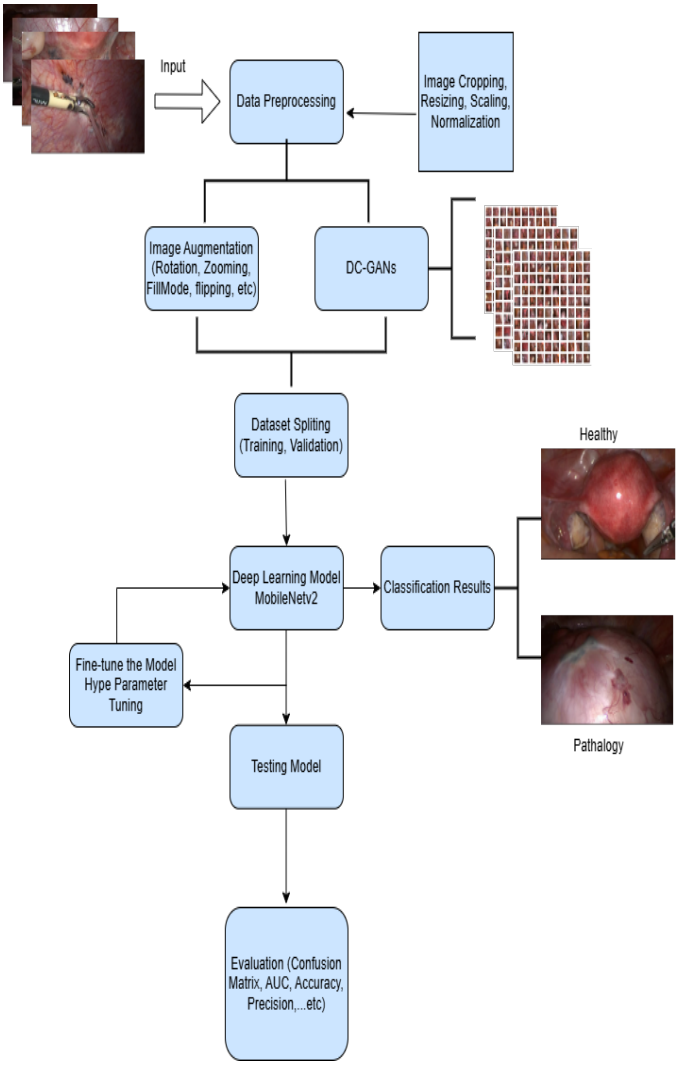


Fig. 1. A Flowchart of proposed methodology.

A. Data Collection

The Gynecologic Laparoscopy Endometriosis Dataset (GLENDA) is a comprehensive dataset derived from more than 400 gynecologic laparoscopy video recordings, a significant number of which depict instances of endometriosis with varying phases of severity. GLENDA comprises over 25,000 images, including more than 12,000 positive pathological photographs related to endometriosis and over 13,000 negative non-pathological images devoid of prominent endometriosis. The dataset is intentionally designed for various artificial content analysis tasks related to endometriosis recognition. The dimensions of the images are 640 by 360 pixels.

B. Image Pre-processing

Image preprocessing is a collection of procedures used on raw images to prepare them for additional examination or processing. The primary aim is to improve the image by identifying pertinent information while reducing artifacts, noise, and extraneous features that might interfere with further analysis. Prevalent preprocessing tasks include resizing, filtering, cropping, noise attenuation, color modification, and image enhancement. The phases are chosen according to the image properties and the particular application's requirements. In this research work, All images were resized to 224×224 pixels. We also applied image sharpening to the entire dataset to improve the quality of the images. Furthermore, the images are processed in RGB format. We implemented many rescaling methods, including multiplying each pixel by 1/255. This generalizes the input, conserves memory, and reduces the computational expense of applicable procedures. Moreover, it also facilitates the understanding of the ideal function.

C. Data Augmentation Using Image Data Generator Methods

This approach is used to generate data examples from the current samples. It is beneficial when the dataset contains a limited number of instances or exhibits class imbalance. We may use many geometric techniques to generate augmented data. This approach facilitates the learning of a class-imbalanced dataset and enhances the model's generalization, reducing overfitting. The model acquires the ability to handle unexplored versions of training samples. We used the Image Data Generator package from Kera's to do image augmentation. The following are the approaches that were executed for this objective:

1) Random zoom: It dynamically zoomed into the provided image. In this work, the zoomed value is fixed to 0.1 percent.

2) Random rotation: It dynamically rotates the provided image based on the specified value. The rotated value is fixed at 30 degrees.

3) Horizontal flip: This option randomly inverts the image horizontally to alter the positioning of its sides.

4) Width shift: It horizontally alters the image's width, with a value established at 0.2 percent of its overall width.

5) Height shift: It vertically alters the image's height, with a value established at 0.2 percent of its overall height.

D. Data Augmentation Using DCGAN Method

Deep Convolutional Generative Adversarial Networks (DCGANs) are a class of generative models that utilize deep convolutional layers to generate realistic images from random noise. They are built upon the framework of traditional GANs, which consist of two neural networks: a generator and a discriminator. Figure (2 and 3) illustrates its architecture. The generator learns to produce synthetic images, while the discriminator evaluates their authenticity. Both networks are trained simultaneously in an adversarial process, where the generator aims to create images that can deceive the discriminator, and the discriminator strives to distinguish between real and generated images.The generator creates progressively more naturalistic pictures throughout the training to deceive the discriminator. This adversarial strategy incrementally enhances the quality of the produced pictures. The generator learns explicitly to translate random noise vectors into picture spaces that mimic the distribution of authentic endometriosis images. At the same time, the discriminator acquires the ability to differentiate between actual and faked images. Ultimately, this leads to the generator creating high-fidelity synthetic endometriosis pictures that closely mimic authentic ones.

DCGANs introduce specific architectural enhancements, including the use of convolutional layers instead of fully connected layers, which improves spatial structure preservation in the generated images. Additionally, techniques like batch normalization and the use of Leaky ReLU activation functions enhance training stability and convergence. These improvements enable DCGANs to generate high-quality and coherent images across various domains, such as medical picture classification, and objects.

The structured use of convolutional layers helps the model learn hierarchical representations, which are crucial for producing realistic outputs. Furthermore, the adversarial training approach encourages the generator to refine its outputs iteratively, leading to visually convincing results. Due to their effectiveness and simplicity, DCGANs have become a foundational model in the field of generative image synthesis and serve as a basis for more advanced architectures in related applications.

DCGANs can generate data from existing samples. In situations like medical picture classification, gathering sufficient data to build a big deep learning model is crucial, which might be arduous. Utilizing presently accessible data, we may produce new samples with DCGANs; these synthetic samplings do not directly correspond to actual patients. This technology has expedited the use of deep learning in medical imaging; nonetheless, it requires significant improvement in accurately collecting delicate tissues from images. In this study employs a deep convolutional GAN (DCGAN) [46] to generate synthetic endometriosis images featuring pathologies diseases in the used dataset. A deep generative model of around five layers was used with a relatively slight discriminator network. The complex architecture assists in detecting subtle tissues within the endometriosis images. DCGAN was used for this analysis because of its stability in terms of simplicity and efficacy in producing high-quality pictures. In contrast to more intricate GAN variations like CGAN, BEGAN, WGAN, and DCGAN, it offers a simple architecture that is easy to build and tune, rendering it appropriate for our application in medical image creation. Moreover, DCGAN has been extensively evaluated

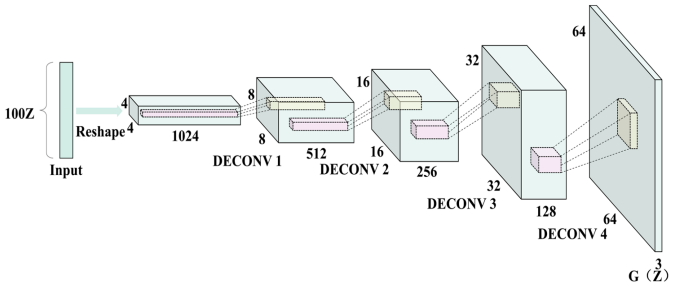


Fig. 2. Generator architecture [49].

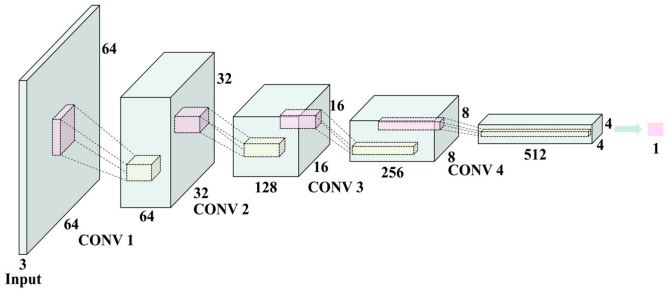


Fig. 3. Discriminator architecture [49].

across many image synthesis tasks, showing its reliability and effectiveness in creating realistic images, which coincides with our objective of enhancing medical datasets.

We trained DCGANs to synthesize images of endometriosis lesions, with separate models developed for each lesion category. The training process involved alternating updates to the generator and discriminator networks in an iterative manner, allowing both networks to improve their performance progressively. A learning rate of 0.0002 was employed, and training was conducted for 200 epochs for each lesion category to ensure adequate learning of the underlying data distribution.

The input dataset consisted of original images with a resolution of 224 × 224 pixels, and a batch size of 128 was used during training to optimize computational efficiency. Upon completion of the training process, we generated 300 synthetic images for each lesion category. Each class-specific training session required approximately 20 hours to complete on a local PC equipped with an Intel i7-7820HQ CPU. These synthesized images were then merged with the original dataset to augment the data available for subsequent analysis or model training. This approach ensured an expanded and balanced dataset that could potentially improve downstream performance in tasks such as endometriosis lesions classification or segmentation.

To evaluate the visual quality of the synthesized images, a radiologist with expertise in endometriosis assessed the synthetic lesions for their resemblance to real clinical cases. The radiologist's analysis focused on key morphological features, such as texture, contrast, and spatial distribution, to determine whether the synthetic lesions accurately mirrored the appearance of genuine endometriosis implants. By conducting training independently for each lesion category and ensuring consistency in the training parameters, we aimed to achieve high-quality image synthesis across all categories. This method

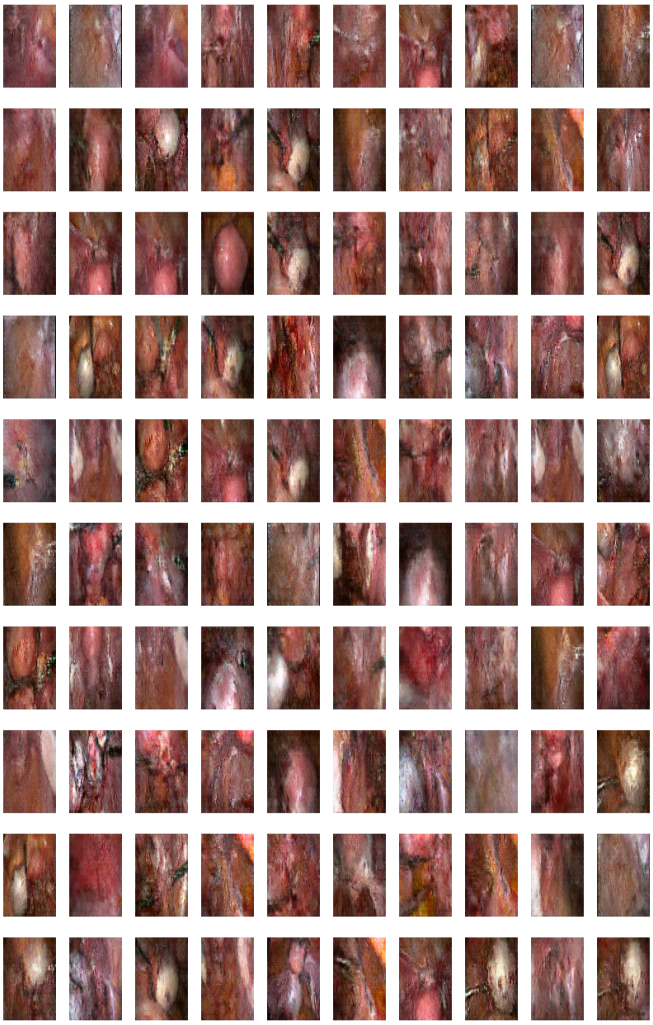


Fig. 4. DCGANs synthetic images.

highlights the potential of DCGANs in generating synthetic data that closely resembles real-world medical images, contributing to the development of robust datasets in medical imaging research.Figure (4) displays examples created for the pathology class after 200 epochs.

E. Training Phase

This section emphasizes some of the primary details applied in the training stage, such as kinds of models, hyperparameters, and classifier architectures.

1) Dataset split: In supervised machine learning and deep learning workflows, datasets are first divided into distinct sets for training, validation, and testing. Image augmentation, commonly performed on the training and potentially validation sets, aids in model generalization during the training and optimization stages. To ensure unbiased results, test data usually remains unaugmented to evade data leakage. Nevertheless, recent research highlights that augmenting test data can also be valuable in examining model robustness.In this work, the dataset is divided into 70% for training, 20% for validation, and 10% for testing.

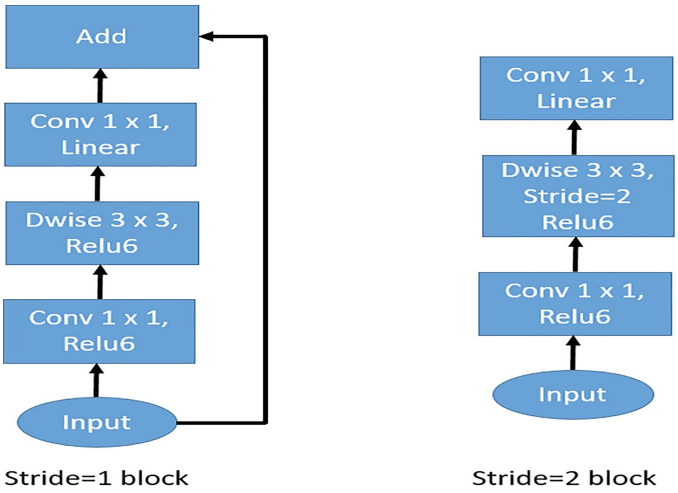


Fig. 5. MobileNetV2 architecture [47].

- Training Images: 17977
- Validation Images: 5136
- Testing Images: 2569

2) Model architectures: In this study, several deep learning models were evaluated using the Glenda dataset to assess their performance across different architectures. Three models (VGG19, InceptionV3, MobileNetV2)were implemented utilizing the Keras library in Python. Transfer learning was utilized for the convolutional layers to initialize these models, while random initialization was applied to the fully connected layers. The classifier network was built with a custom design that included two hidden layers and an output layer containing two neuron units, each assigned to a specific class. This minimalist design, featuring fewer layers and neurons, was strategically utilized to avoid overfitting.

3) MobileNetV2: MobileNetV2 [48] is recognized as a lightweight convolutional neural network (CNN) that is generally utilized in miscellaneous applications. It improves the MobileNetV1 model by introducing new modules, especially inverted residuals and linear tie-ups. The essence design of MobileNet is established on depthwise divisible convolution. This approach differs from standard 2-dimensional convolution, which treats all input channels uniformly to produce one output channel. Instead, depthwise convolution applies filters independently to each input channel, resulting in separate output channels that are subsequently combined. The separable depth-wise convolution process involves a subsequent 1×1 point-wise convolution, which merges these output channels into one final channel. This method provides the same output as traditional convolution while being more efficient due to a reduced number of parameters.Figure 5 presents the graphical representation of the MobileNetV2 framework.

4) Hyper parameters: In this study, we kept hyper parameters uniform across all models, allowing for a precise and fair comparison of each model's performance. This consistent approach allowed for a clear assessment of each model's relative performance.

a) Optimizer: In our optimization process, we rigorously tested two methods: Stochastic Gradient Descent (SGD) and Adaptive Momentum Estimation (Adam). After multiple iterations, we found that both optimizers performed similarly. However, we chose the Adam optimizer for training, as it is a more advanced technique for learning model weights and has been proven to enhance convergence and stability in deep learning models.

b) Learning rate: Since we utilized pre-trained weights through transfer learning for the convolutional networks in our models, we opted for a low learning rate, set explicitly at 0.0001, to ensure stable training.

c) Activation functions: ReLU was implemented across all layers, excluding the output layer, adding non-linearity to enhance model performance. In the output layer, we applied the softmax function, which is well-suited for labeled classification, as it allows each neuron to generate separate outputs.

d) Loss function: We opted for Binary Cross-Entropy (BCE) to calculate loss, which allows us to consider each label independently.

IV. Results

This section presents a detailed account of the results obtained from our experimental research and describes the programming environments employed to achieve these results. For clarity, each aspect is discussed in separate subsections.

A. Experimental Setup

The large dataset influenced our choice of programming environments to conduct deep learning experiments. We carried out model implementation and training on our Dell laptop, assembled with a 1.80 GHz Intel Core i5 CPU and 16 gigabytes of RAM. Our approach leveraged Python libraries such as TensorFlow, Keras, and sci-kit-learn for effective attribute classification. Jupyter Notebook enabled efficient program development and data analysis, and we observed satisfactory results across several models.

B. Evaluation Procedure

In order to make a comprehensive comparison of model classification performance, this paper leverages various metrics associated with the confusion matrix, including precision, recall, F1-score, and accuracy.Accuracy calculates the model's prevalent correctness, while precision focuses on the reliability of optimistic predictions. Recall quantifies the model's sensitivity by measuring the proportion of true positives, and specificity assesses the ability to identify negative cases correctly. We clearly understand each model's strengths and weaknesses by calculating and analyzing these metrics. The formulas for calculating these performance metrics are provided below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F1\text{-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

In this equation, TP, or True Positive, refers to the count of positive cases accurately identified as positive. FN, or False Negative, represents the number of positive cases mistakenly classified as negative. FP, or False Positive, denotes the count of negative cases incorrectly labeled as positive, while TN, or True Negative, captures the number of negative cases correctly classified as negative.

C. Area Under the Curve (AUC)

The ROC curve is commonly used to evaluate how well a model can distinguish between different classes. It is based on plotting the True Positive Rate (TPR) and False Positive Rate (FPR) across different thresholds. Since TPR and FPR are directly proportional, they increase together. The ROC curve visually represents this relationship and enables calculation of the Area Under the Curve (AUC), where a larger AUC value implies a more robust classification capability. Thus, a higher AUC indicates better model performance.

D. VGG19

In analyzing endometriosis lesion images, spatial data is crucial for effective disease classification. Deep learning-based models are frequently employed to capture these essential features, and VGG19, a highly layered convolutional model, was chosen for this study. We applied VGG19 with pre-trained weights for feature extraction and added a custom three-layer classifier. The model was trained across three configurations with varying epochs (3, 7, 14, and 20) and a batch size of (32, 64, and 128), resulting in above-average performance metrics with an accuracy of up to 88%. Table I compiles the results of tests conducted without data augmentation, providing insights into the baseline performance of our endometriosis lesion detection algorithm. Table II, on the other hand, presents the outcomes when data augmentation techniques are applied, showcasing the improvements in detection accuracy and robustness achieved through augmentation. Figures 6 and 7 illustrate the loss trajectories and accuracy trends, respectively, for the VGG19 model trained on the baseline dataset (without augmentation). In contrast, the effects of incorporating data augmentation are evident in Figures 8 and 9, which depict the model's loss dynamics and classification performance under enhanced training conditions. Figure 10 shows the confusion matrices illustrating the endometriosis lesion classification outcomes from testing the VGG19 model. The AUC-ROC curve, which can be seen in Figure 11, also explains the result of the VGG19 model.

E. InceptionV3

The InceptionV3 model offers an optimized approach to achieve high classification accuracy without relying on an extensive network architecture. With fewer learnable parameters than the VGG19 model, InceptionV3 provides an efficient alternative that still delivers excellent classification performance.

TABLE I. VGG19 MODEL RESULTS WITHOUT DATA AUGMENTATION

Epochs	Batch-Size	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
3	32	56	54	55	56
	64	57	57	56	57
	128	59	59	59	60
7	32	63	63	63	63
	64	65	64	64	65
	128	66	65	66	66
14	32	70	70	69	70
	64	74	73	73	74
	128	77	77	77	78
20	32	80	80	79	80
	64	80	79	80	80
	128	81	81	81	82

TABLE II. VGG19 MODEL RESULTS WITH DATA AUGMENTATION

Epochs	Batch-Size	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
3	32	60	60	61	61
	64	62	61	62	63
	128	64	63	64	65
7	32	68	69	69	70
	64	72	72	73	73
	128	74	74	75	75
14	32	77	76	77	78
	64	79	79	79	80
	128	82	83	83	84
20	32	84	84	85	85
	64	87	87	87	88
	128	88	87	88	88

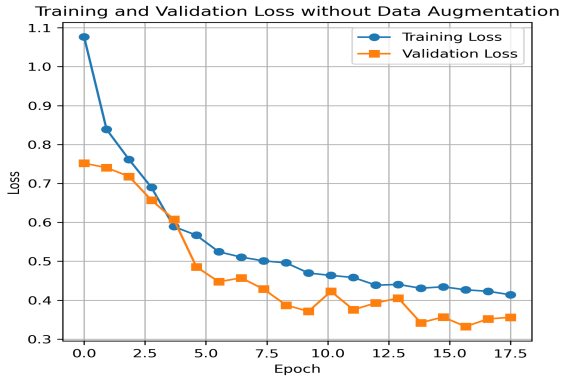


Fig. 6. VGG19 loss display without data augmentation.

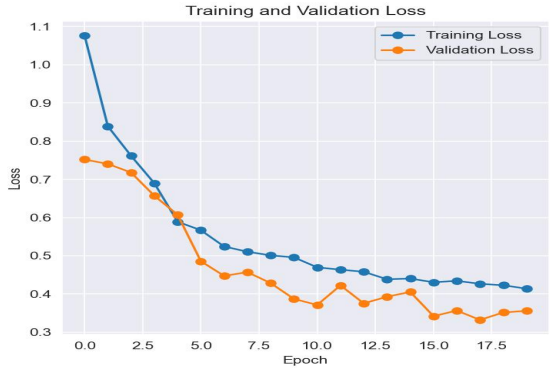


Fig. 8. VGG19 loss display.

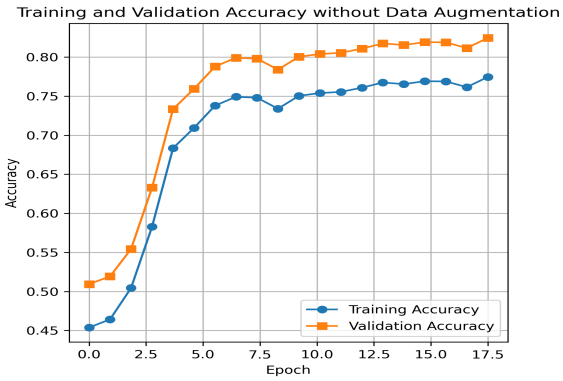


Fig. 7. VGG19 accuracy display without data augmentation.

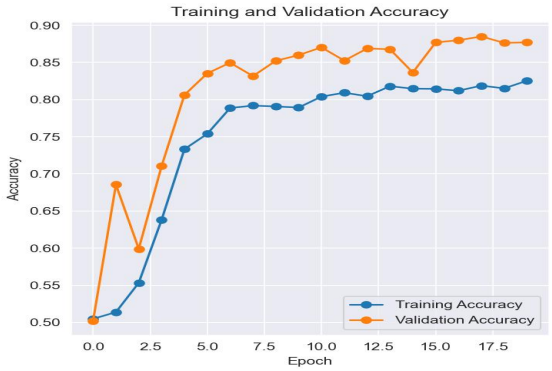


Fig. 9. VGG19 accuracy display.

Its design prioritizes streamlined architecture, balancing effectiveness with computational efficiency.InceptionV3 model was trained across three configurations with varying epochs (3, 7, 14, and 20) and a batch size of (32, 64, and 128), resulting in above-average performance metrics with an accuracy of up to 97%, that are higher than VGG19. Table III compiles the results of tests conducted without data augmentation, providing insights into the baseline performance of our endometriosis lesion detection algorithm. Table IV, on the other hand, presents the outcomes when data augmentation techniques are applied, showcasing the improvements in detection accuracy and robustness achieved through augmentation. Figures 12 and 13 illustrate the loss trajectories and accuracy trends, respectively, for the InceptionV3 model trained on the baseline dataset (without augmentation). In contrast, the effects of incorporating data augmentation are evident in Figures 14 and 15, which depict the model's loss dynamics and classification performance under enhanced training conditions. Figure 16 shows the confusion matrices illustrating the endometriosis lesion classification outcomes from testing the InceptionV3

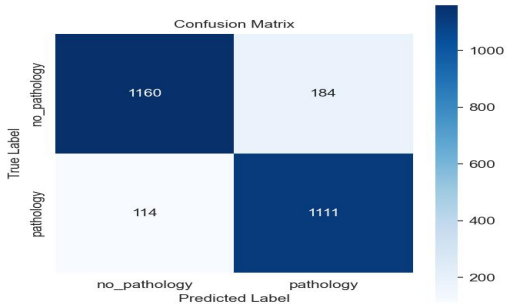


Fig. 10. VGG19 confusion matrix.

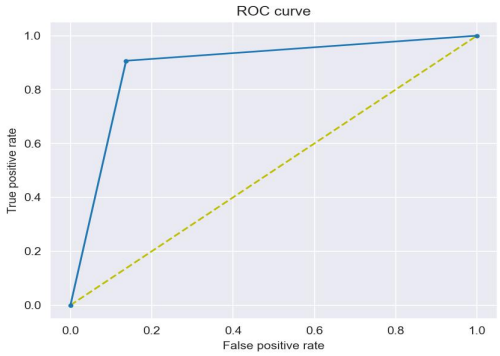


Fig. 11. VGG19 AUC-ROC curve.

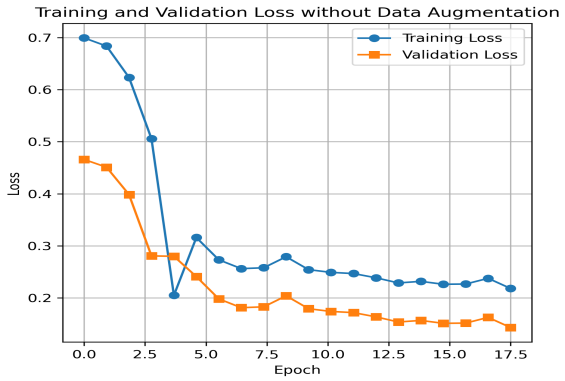


Fig. 12. InceptionV3 loss display without data augmentation.

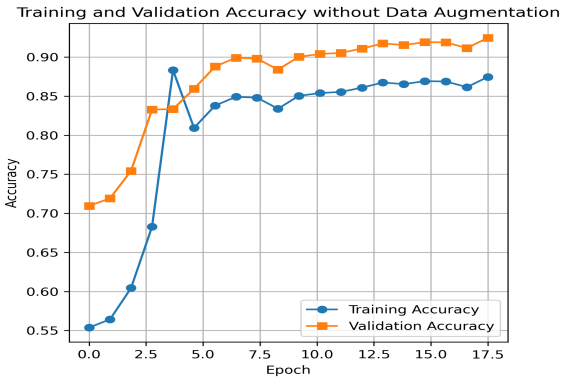


Fig. 13. InceptionV3 accuracy display without data augmentation.

model. The AUC-ROC curve, which can be seen in Figure 17, also explains the result of the InceptionV3 model.

TABLE III. INCEPTIONV3 MODEL RESULTS WITHOUT DATA AUGMENTATION

Epochs	Batch-Size	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
3	32	65	63	64	65
	64	67	67	67	67
	128	70	69	69	70
7	32	73	72	73	73
	64	75	75	75	76
	128	80	79	79	80
14	32	83	83	82	83
	64	87	85	86	87
	128	88	88	89	89
20	32	90	89	89	90
	64	90	91	90	91
	128	92	91	92	92

F. MobileNetV2

MobileNetV2 builds on the foundation of MobileNetV1 with an updated architecture featuring fewer learnable parameters for greater efficiency.MobileNetV2 model was trained across three configurations with varying epochs (3, 7, 14, and 20) and a batch size of (32, 64, and 128), resulting in above-average performance metrics with an accuracy of up to 99%, which are higher than InceptionV3. Table V compiles the results of tests conducted without data augmentation, providing insights into the baseline performance of our endometriosis lesion detection algorithm. Table VI, on the other hand, presents the outcomes when data augmentation techniques are applied, showcasing the improvements in detection accuracy and robustness achieved through augmentation. Figures 18 and 19 illustrate the loss trajectories and accuracy trends,

TABLE IV. INCEPTIONV3 MODEL RESULTS WITH DATA AUGMENTATION

Epochs	Batch-Size	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
3	32	52	53	53	54
	64	56	56	57	58
	128	59	58	58	59
7	32	62	63	61	63
	64	68	69	71	72
	128	80	80	82	83
14	32	89	88	89	90
	64	90	91	89	92
	128	92	92	92	93
20	32	92	93	93	94
	64	95	94	96	96
	128	96	97	97	97



Fig. 14. InceptionV3 loss display.

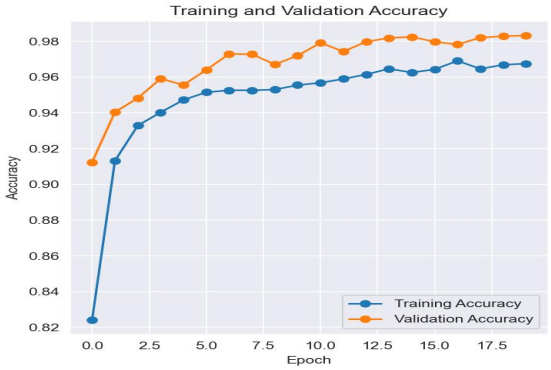


Fig. 15. InceptionV3 accuracy display.

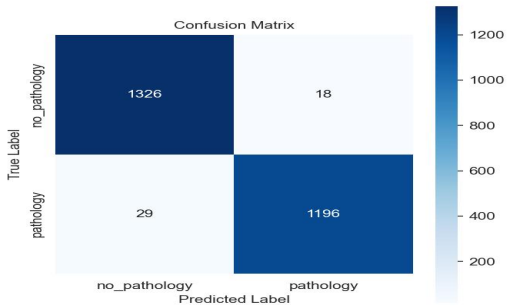


Fig. 16. InceptionV3 confusion matrix.

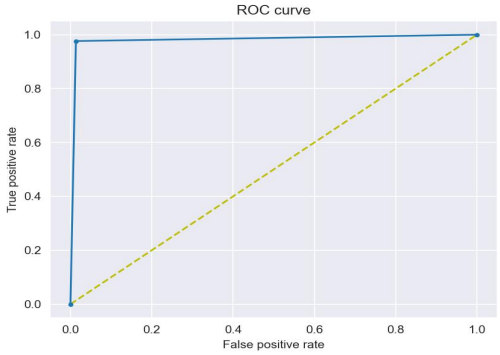


Fig. 17. InceptionV3 AUC-ROC curve.

TABLE V. MOBILENETV2 MODEL RESULTS WITHOUT DATA AUGMENTATION

Epochs	Batch-Size	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
3	32	70	70	69	70
	64	70	70	71	71
	128	72	72	71	73
7	32	73	74	74	75
	64	76	77	77	78
	128	79	80	79	80
14	32	83	84	83	84
	64	83	84	84	85
	128	85	86	86	87
20	32	90	90	89	90
	64	92	91	92	92
	128	93	94	94	95

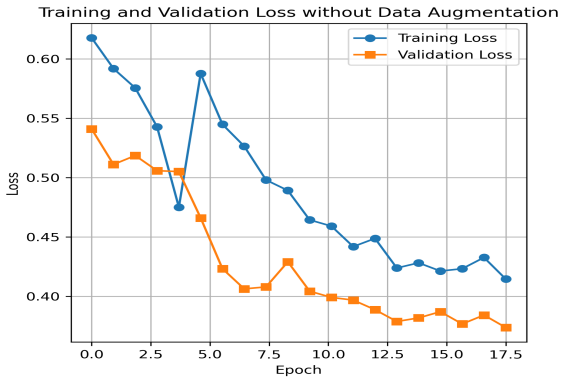


Fig. 18. MobileNetV2 loss display without data augmentation.

respectively, for the MobileNetV2 model trained on the baseline dataset (without augmentation). In contrast, the effects of incorporating data augmentation are evident in Figures 20 and 21, which depict the model's loss dynamics and classification performance under enhanced training conditions. Figure 22 shows the confusion matrices illustrating the endometriosis lesion classification outcomes from testing the MobileNetV2 model. The AUC-ROC curve, which can be seen in Figure 23, also explains the result of the MobileNetV2 model.

V. DISCUSSION

Endometriosis is one of the leading gynecological issues facing women worldwide. Diagnosing and treating this condition remains complex, particularly in settings with limited medical resources. Deep learning techniques applied to medical imaging on large datasets have enabled computer algorithms to achieve diagnostic accuracy similar to that of healthcare professionals. In this research, we propose a deep learning-based solution for classifying endometriosis lesions

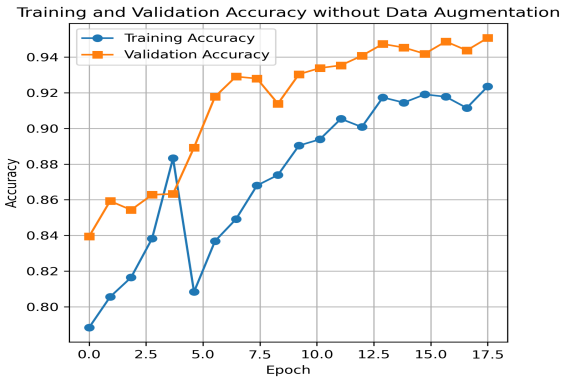


Fig. 19. MobileNetV2 accuracy display without data augmentation.

TABLE VI. MOBILENETV2 MODEL RESULTS WITH DATA AUGMENTATION

Epochs	Batch-Size	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
3	32	61	62	61	62
	64	61	61	62	63
	128	65	65	65	65
7	32	67	66	67	68
	64	70	70	71	73
	128	79	80	79	80
14	32	83	84	82	84
	64	86	87	87	87
	128	91	91	91	92
20	32	94	93	94	94
	64	95	96	96	97
	128	99	99	99	99



Fig. 20. MobileNetV2 loss display.

TABLE VII. OVERALL RESULTS

#	Deep Learning Models	Accuracy without Data Augmentation (%)	Accuracy with Data Augmentation (%)
1	VGG19	82	88
2	InceptionV3	92	97
3	MobileNetV2	95	99

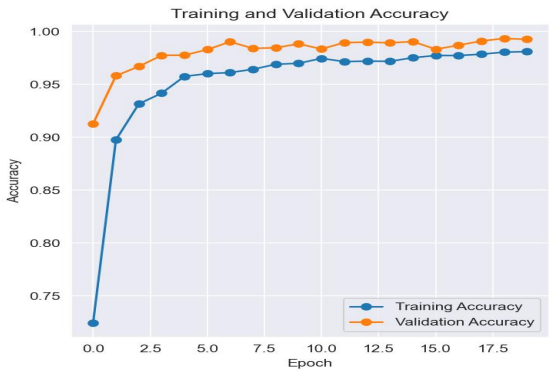


Fig. 21. MobileNetV2 accuracy display.

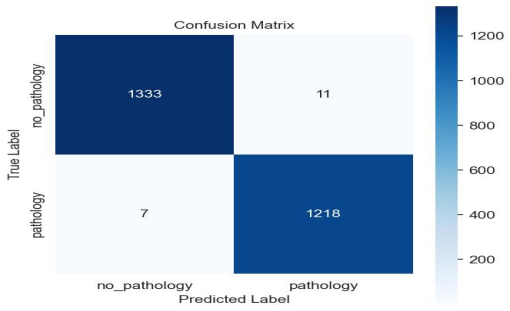


Fig. 22. MobileNetV2 confusion matrix.

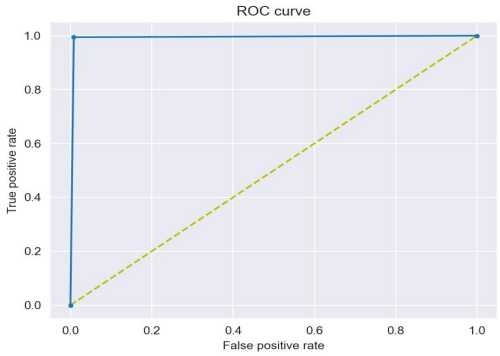


Fig. 23. MobileNetV2 AUC-ROC curve.

in laparoscopy images, utilizing the MobileNetV2 architecture alongside a neural network classifier. We incorporate both conventional methods and deep learning-based image augmentation techniques. Experimental findings indicate that deep learning models are well-suited to accurately classify endometriosis lesions, facilitating a robust diagnostic tool for endometriosis.

The incorporation of synthetic images into the training process offers a substantial benefit to the performance of deep learning models, particularly when dealing with small or imbalanced datasets. In many real-world applications, such as

TABLE VIII. COMPARED WITH DIFFERENT TECHNIQUES EMPLOYED BY OTHER RESEARCHERS

#	Citation	Classifier	Accuracy (%)
1	Visalaxi et al. (2021) [36]	ResNet50	91
2	Yun et al. (2021) [43]	VGGNet-16	90.80
3	Takahashi al. (2021) [44]	DNN	90.29
4	Leibetseder al. (2022) [40]	Faster-RCNN	32.4 (precision)
5	Sudalaimuthu al. (2022) [45]	U-Net	74 (F1-score)
6	Figueredo al. (2024) [46]	Ensemble of Networks	96.67
7	Zaidi al. (2025) [5]	Inception V3	93
8	Our proposed	MobileNetV2	99

medical imaging, obtaining a large and representative dataset can be a significant challenge. Synthetic image generation addresses this limitation by augmenting the dataset, providing the model with a broader spectrum of examples, including rare or underrepresented cases. This allows the model to learn more robust and generalized features, which directly contributes to improvements in key performance metrics such as accuracy, precision, and recall.When compared to models trained solely on real images, those trained with synthetic images exhibit superior handling of rare classes. This is particularly important in fields like medical imaging, where certain conditions or abnormalities may be infrequent but critical for diagnosis. By increasing the diversity of the training data, synthetic images help to reduce bias toward more common classes, ensuring that the model does not develop a skewed understanding of the data. As a result, models trained with synthetic images are better equipped to identify and classify rare conditions, leading to a decrease in false negatives and an overall improvement in recall.

Moreover, synthetic images enhance the generalization ability of deep learning models. With more varied and comprehensive data, models are better prepared to perform well on unseen data, a crucial aspect for deployment in real-world scenarios. This greater ability to generalize ensures that the model's performance remains consistent when faced with new, previously unobserved examples. Thus, synthetic images play a vital role in improving not only the accuracy and reliability of deep learning models but also their ability to adapt to diverse and unpredictable data, particularly in specialized fields such as medical imaging.

In our comparative evaluation of convolutional neural network architectures, MobileNetV2 demonstrated superior performance relative to models such as VGG19 and InceptionV3, mainly due to MobileNetV2's lightweight and efficient structure, making it ideal for mobile and embedded deployment. With fewer parameters, it offers faster and simpler training and deployment. Experimenting with 32, 64, and 128 batch sizes, we achieved a favorable trade-off between accuracy and computational efficiency at a resolution of 224x224. Scaling pixel values from 0 to 1 was the most effective for normalization. Additionally, we observed that while fine-tuning the fully connected layer and freezing convolutional layers maintained stable performance, it did extend convergence time.

Table VII presents a comparative analysis of various deep learning algorithms tested on the GLENDA dataset, highlighting the performance of each model in detecting endometriosis. This comparison underscores the strengths and limitations of different algorithms within the same dataset, allowing for an evaluation of each model's effectiveness in classifying endometriosis lesions. Table VIII provides a summary of these findings and compares the performance of our models to existing state-of-the-art approaches, offering insights into how our methods advance current standards in accuracy, precision, and overall reliability for clinical decision support.

VI. CONCLUSION

Our approach, combining the MobileNetV2 model with both conventional and deep learning-based augmentation techniques, demonstrated high performance, achieving 99% in recall, binary accuracy, and F1-score. The integration of synthetic samples generated by DCGAN significantly improved training data diversity and addressed class imbalance issues within the dataset. Specifically, by generating synthetic images of endometriosis using DCGAN, we were able to enhance the MobileNetV2 model's accuracy, likely due to DCGAN's capacity to capture diverse manifestations of endometriosis present in the GLENDA dataset. These findings underscore the potential of deep learning models to accurately classify endometriosis lesions from laparoscopy images, supporting their use as clinical decision support tools for timely diagnosis. Additionally, synthetic data augmentation shows promise for addressing similar challenges in other areas of medical imaging, offering a pathway to more effective and reliable diagnostic support systems. Future work will aim to further improve system performance by expanding the dataset to include a wider variety of laparoscopy videos, enhancing both model accuracy and generalization across clinical scenarios.

AUTHOR'S CONTRIBUTIONS

All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

DATA AVAILABILITY

Data availability is not applicable to this article as no new data were created or analysed in this study.

CONFLICT OF INTEREST

The authors state that they do not have any conflicts of interest.

ACKNOWLEDGMENTS

The first author is a PhD degree student in the Computer Science Program at the Faculty of Science, Chiang Mai University (CMU), under the CMU Presidential Scholarship.

REFERENCES

[1] J. N. Samreen et al., 'MRI of endometriosis: a comprehensive review', Appl Radiol, vol. 48, no. 5, pp. 6–12, 2019.

[2] A. L. Shafrir et al., 'Risk for and consequences of endometriosis: a critical epidemiologic review', Best practice & research Clinical obstetrics & gynaecology, vol. 51, pp. 1–15, 2018.

[3] N. Agarwal and A. Subramanian, 'Endometriosis-morphology, clinical presentations and molecular pathology', Journal of laboratory physicians, vol. 2, no. 01, pp. 001–009, 2010.

[4] H. Kwok et al., 'Lesion distribution characteristics of deep infiltrating endometriosis with ovarian endometrioma: an observational clinical study', BMC women's health, vol. 20, pp. 1–7, 2020.

[5] S. A. Zaidi, V. Chouvatut, C. Phongnarisorn, and D. Praserttitipong, 'Deep learning based detection of endometriosis lesions in laparoscopic images with 5-fold cross-validation', Intelligence-Based Medicine, vol. 11, p. 100230, 2025.

[6] C. M. Becker et al., 'ESHRE guideline: endometriosis', Human reproduction open, vol. 2022, no. 2, p. hoac009, 2022.

[7] Y. Wang, K. Nicholes, and I. M. Shih, 'The origin and pathogenesis of endometriosis', Annual Review of Pathology: Mechanisms of Disease, vol. 15, no. 1, pp. 71–95, 2020.

[8] E. Pascoal et al., 'Strengths and limitations of diagnostic tools for endometriosis and relevance in diagnostic test accuracy research', Ultrasound in Obstetrics & Gynecology, vol. 60, no. 3, pp. 309–327, 2022.

[9] K. E. Nnoaham et al., 'Impact of endometriosis on quality of life and work productivity: a multicenter study across ten countries', Fertility and sterility, vol. 96, no. 2, pp. 366–373, 2011.

[10] C. M. Becker, S. A. Missmer, and K. T. Zondervan, 'Endometriosis: the authors reply', New England Journal of Medicine, vol. 383, no. 2, 2020.

[11] B. Smolarz, K. Szyłło, and H. Romanowicz, 'Endometriosis: epidemiology, classification, pathogenesis, treatment and genetics (review of literature)', International journal of molecular sciences, vol. 22, no. 19, p. 10554, 2021.

[12] A. G. Cope, W. M. Vanburen, and S. P. Sheedy, 'Endometriosis in the postmenopausal female: clinical presentation, imaging features, and management', Abdominal Radiology, vol. 45, no. 6, pp. 1790–1799, 2020.

[13] W. K. R. Figueredo, A. C. Silva, A. C. De Paiva, J. O. B. Diniz, A. Brandão, and M. A.P. Oliveira, 'Automatic segmentation of deep endometriosis in the rectosigmoid using deep learning', Image and Vision Computing, 2024.

[14] S. L. Vlek et al., 'Laparoscopic imaging techniques in endometriosis therapy: a systematic review', Journal of Minimally Invasive Gynecology, vol. 23, no. 6, pp. 886–892, 2016.

[15] M. Al-Taher et al., 'Intraoperative enhanced imaging for detection of endometriosis: A systematic review of the literature', European Journal of Obstetrics & Gynecology and Reproductive Biology, vol. 224, pp. 108–116, 2018.

[16] C. Mosbrucker, A. Somani, and J. Dulemba, 'Visualization of endometriosis: comparative study of 3-dimensional robotic and 2-dimensional laparoscopic endoscopes', Journal of Robotic Surgery, vol. 12, no. 1, pp. 59–66, 2018.

[17] G. Vizzielli et al., 'Real three-dimensional approach vs two-dimensional camera with and without real-time near-infrared imaging with indocyanine green for detection of endometriosis: A case-control study', Acta Obstetricia et Gynecologica Scandinavica, vol. 99, no. 10, pp. 1330–1338, 2020.

[18] J. Quesada et al., 'Endometriosis: a multimodal imaging review', European journal of radiology, vol. 158, p. 110610, 2023.

[19] B. Cherradi, O. Terrada, A. Ouhmida, S. Hamida, A. Raihani, and O. Bouattane, 'Computer-aided diagnosis system for early prediction of atherosclerosis using machine learning and K-fold cross-validation', in 2021 international congress of advanced technology and engineering (ICOTEN), 2021, pp. 1–9.

[20] C. Nifora, L. Chasapi, M. K. Chasapi, and C. Koutsojannis, Deep Learning Improves Accuracy of Laparoscopic Imaging Classification for Endometriosis Diagnosis. 2023.

[21] M. M. Rahman, Y. Ghasemi, E. Suley, Y. Zhou, S. Wang, and J. Rogers, 'Machine learning based computer aided diagnosis of breast cancer utilizing anthropometric and clinical features', Irbm, vol. 42, no. 4, pp. 215–226, 2021.

[22] H. Ding, N. Huang, and X. Cui, 'Leveraging GANs data augmentation for imbalanced medical image classification', Applied Soft Computing, vol. 165, 2024.

[23] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, 'GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification', Neurocomputing, vol. 321, pp. 321–331, 2018.

[24] I. Goodfellow et al., 'Generative adversarial nets', in Advances in neural information processing systems, 2014.

[25] H. Ding, Y. Sun, N. Huang, Z. Shen, and X. Cui, 'TMG-GAN: Generative adversarial networks-based imbalanced learning for network intrusion detection', IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1156–1167, 2023.

[26] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, 'A review on generative adversarial networks: Algorithms, theory, and applications', IEEE transactions on knowledge and data engineering, vol. 35, no. 4, pp. 3313–3332, 2021.

[27] W. Li et al., 'EID-GAN: Generative adversarial nets for extremely imbalanced data augmentation', IEEE Transactions on Industrial Informatics, vol. 19, no. 3, pp. 3208–3218, 2022.

[28] A. Turab, F. Ullah, and S. A. Zaidi, 'A Federated Learning Framework for Intermittent Clients in Collaborative Smart Healthcare Systems', in AI and Blockchain Applications for Privacy and Security in Smart Medical Systems, IGI Global Scientific Publishing, 2025, pp. 1–30.

[29] S. A. Zaidi and V. Chouvatut, 'Mae Mai Muay Thai Layered Classification Using CNN and LSTM Models', in 2022 26th International Computer Science and Engineering Conference (ICSEC), 2022, pp. 351–356.

[30] A. Turab, R. Shafqat, S. Muhammad, M. Shuaib, M. F. Khan, and M. Kamal, 'Predictive modeling of hepatitis B viral dynamics: a caputo derivative-based approach using artificial neural networks', Scientific Reports, vol. 14, no. 1, p. 21853, 2024.

[31] A. Turab, F. Ullah, and S. A. Zaidi, 'Explainable AI and Computational Modeling for Real-Time COVID-19 Detection Using Medical Imaging', in AI and Blockchain Applications for Privacy and Security in Smart Medical Systems, IGI Global Scientific Publishing, 2025, pp. 161–190.

[32] S. A. Zaidi and V. Chouvatut, 'Mae Mai Muay Thai Style Classification in Movement Appling Long-Term Recurrent Convolution Networks', Journal of Internet Services and Information Security, vol. 13, no. 1, pp. 95–112, 2023.

[33] K. Dong, C. Zhou, Y. Ruan, and Y. Li, 'MobileNetV2 model for image classification', in 2020 2nd International Conference on Information Technology and Computer Application (ITCA), IEEE, 2020, pp. 476–480.

[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, 'Mobilenetv2: Inverted residuals and linear bottlenecks', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[35] P. Ballester and R. Araujo, 'On the performance of GoogLeNet and AlexNet applied to sketches', Proceedings of the AAAI conference on artificial intelligence, vol. 30, 2016.

[36] S. Visalaxi and T. S. Muthu, 'Automated prediction of endometriosis using deep learning', International Journal of Nonlinear Analysis and Applications, vol. 12, no. 2, pp. 2403–2416, 2021.

[37] A. Leibetseder, S. Kletz, K. Schoeffmann, S. Keckstein, and J. Keckstein, 'GLENDA: gynecologic laparoscopy endometriosis dataset', in International Conference on Multimedia Modeling, Cham: Springer International Publishing, 2019, pp. 439–450.

[38] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, 'Imagenet: A large-scale hierarchical image database', in 2009 IEEE conference on computer vision and pattern recognition, 2009, pp. 248–255.

[40] A. Leibetseder, K. Schoeffmann, J. Keckstein, and S. Keckstein, 'Endometriosis detection and localization in laparoscopic gynecology', Multimedia Tools and Applications, vol. 81, no. 5, pp. 6191–6215, 2022.

[41] S. Ren, K. He, R. Girshick, and J. Sun, 'Faster R-CNN: Towards real-time object detection with region proposal networks', IEEE transactions on pattern analysis and machine intelligence, vol. 39, pp. 1137–1149, 2016.

[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, 'Mask r-cnn', in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[43] Y. Zhang et al., 'Deep learning model for classifying endometrial lesions', Journal of Translational Medicine, vol. 19, pp. 1–13, 2021.

[44] Y. Takahashi et al., 'Automated system for diagnosing endometrial cancer by adopting deep-learning technology in hysteroscopy', PLoS One, vol. 16, no. 3, p. e0248526, 2021.

[45] S. Visalaxi and T. Sudalaimuthu, 'Automated segmentation of endometriosis using transfer learning technique', F1000Research, vol. 11, p. 360, 2022.

[46] W. K. R. Figueredo, A. C. Silva, A. C. de Paiva, J. O. B. Diniz, A. Brandão, and M. A. P. Oliveira, 'Automatic segmentation of deep endometriosis in the rectosigmoid using deep learning', Image and Vision Computing, vol. 151, p. 105261, 2024.

[47] A. Radford, L. Metz, and S. Chintala, 'Unsupervised representation learning with deep convolutional generative adversarial networks', arXiv preprint arXiv:1511. 06434, 2015.

[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, 'Mobilenetv2: Inverted residuals and linear bottlenecks', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[49] N. Zhao, Y. Song, A. Yang, K. Lv, H. Jiang, and C. Dong, 'Accurate Classification of Tunnel Lining Cracks Using Lightweight ShuffleNetV2-1.0-SE Model with DCGAN-Based Data Augmentation and Transfer Learning', Applied Sciences, vol. 14, no. 10, p. 4142, 2024.

LASSO-Based Feature Extraction with Adaptive Windowing via DTW for Fault Diagnosis in Rotating Machinery

Jirayu Samkunta[1], Patinya Ketthong[2], Nghia Thi Mai[3], Md Abdus Samad Kamal[4],
Iwanori Murakami[5], Kou Yamada[6], Nattagit Jiteurtragool[7]
Graduate School of Science and Technology, Kiryu, Gunma University, Japan[1,2,4,5,6]
Intelligent Electronics System Laboratory (IES), Thai-Nichi-Institute of Technology, Bangkok, Thailand[2]
Faculty of Electronics Engineering 1, Posts and Telecommunications Institute of Technology, Hanoi, Viet Nam[3]
Department of Computer and Information Sciences-Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand[7]

***Abstract*—In real-world engineering environments, faults in rotating machines typically occur for concise periods, which leads to poor stability and low accuracy in fault diagnosis. The traditional fault diagnosis of rotating machinery relies on analyzing time-series data to detect system degradation and faulty components. However, the complexity of rotating machinery and the presence of multiple fault types across different operating conditions challenges for conventional classification techniques. This paper proposes a LASSO regression-based feature extraction method with adaptive window based on Dynamic Time Warping (DTW) for fault diagnosis in rotating machinery. The approach effectively extract features by modeling the relationship between shaft rotational speeds (25, 50, and 75 rpm) and vibration signals from piezoelectric accelerometers. This research focus on single and combination faults analysis to include 11 faults, enhancing its applicability to real-world fault conditions. To assess its effectiveness, the proposed method is evaluated against Principal Component Analysis (PCA) and Independent Component Analysis (ICA) using the K-Nearest Neighbors (KNN) classifier. The experimental results demonstrate that the LASSO-based approach consistently achieves high classification accuracy across different speeds, outperforming PCA and ICA in both single and double fault scenarios. These findings highlight LASSO regression as a robust feature extraction technique for improving fault detection and predictive maintenance in rotating machinery.**

***Keywords*—Rotating machinery; fault analysis; feature extraction, LASSO regression**

I. Introduction

Rotating machinery, such as motors, pumps, gearboxes, and turbines, forms the backbone of modern industry, with its uninterrupted operation depending heavily on the integrity of key tribological components like ball bearings and shafts. These elements help reduce friction, support radial and axial loads, and ensure dynamic stability. However, they also endure fluctuating mechanical and thermal stresses that can lead to issues such as cracks, spall, imbalances, and lubrication degradation. Faults originating from a rolling element or a shaft shoulder can propagate in just a few revolutions, manifesting as subtle energy bursts in vibration signals [1], [2]. If undetected, these incipient defects can escalate into catastrophic failures, resulting in unplanned downtime, costly production losses, and heightened safety risks [3], [4]. In response, the growing emphasis on condition-based maintenance necessitates diagnostic algorithms capable of quickly and reliably detecting these short-lived, speed-dependent anomalies.

Fault diagnosis in industrial settings is fundamentally a pattern recognition task: sensor signals are pre-processed to extract key features, which are then mapped by a classifier to different fault conditions [5], [6], [7], [8]. The literature includes several feature extraction methods, ranging from statistical indicators of time and frequency domain techniques such as the Hilbert Transform, Fast Fourier Transform (FFT), and Wavelet Transform [9], [10], [11], [15]. Feature reduction techniques typically involve methods like Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA) [12], [13], [14].

However, real-world applications have critical constraints that must be addressed for successful implementation. First, the true mechanical period of a machine is unknown a priori. While operators may log the nominal speed, transient load changes, belt slippage, and controller drift introduce cycle-to-cycle variations, causing misalignment between the start points of consecutive revolutions. Second, vibration data is continuously collected without explicit tags marking the start of each shaft revolution. This lack of synchronization leads to the smearing of transient fault signatures, as even small phase offsets can disperse localized impacts over multiple windows, thereby reducing the signal-to-noise ratio of the extracted features.

According to those problems, a significant research gap exists in accurately capturing transient, speed-dependent fault signatures without the need for specialized hardware or precise synchronization signals. While methods such as tachometer-based synchronization or order tracking can mitigate some of these issues, they often demand additional sensors, clean speed signals, and extensive calibration, limiting their practicality, particularly in retrofitted or older equipment.

Traditional fixed-window sampling was utilized to solve problems by assuming a constant period [15], [16], [17]. If the window is shorter than the true mechanical cycle, fault signatures are truncated; conversely, if the window is too long, irrelevant data dilutes the fault energy. Adaptive approaches, such as those based on tachometer pulses or order tracking,

can alleviate some of these issues but require additional hardware and clean speed signals, which may not be available in retrofitted plants. Dynamic Time Warping (DTW) provides an alternative that is sensor-agnostic. Specifically, originally developed for speech alignment, DTW computes the optimal nonlinear mapping between two time sequences [18], [19]. In this work, DTW is used to align consecutive revolutions of the vibration waveform. By using the first detected loop as a reference and minimizing the distance between it and subsequent candidate segments, DTW can dynamically select the window boundaries that correspond to the true mechanical period. In this work, the DTW-driven window selection ensures that transient impacts are confined to a single revolution while compensating for speed fluctuations at various rotational speeds (25, 50, and 75 rpm).

Once the correct windows are identified, the challenge shifts to selecting a compact discriminative set of features. While deep learning models excel at capturing high-order correlations, they require large, labeled datasets and substantial computational resources. The statistical methods, though more interpretable, tend to suffer from high dimensionality. To solve those problems a balance between simplicity and performance, we utilized the Least Absolute Shrinkage and Selection Operator (LASSO) regression [20]. By penalizing the l_1-norm of the regression coefficients, LASSO encourages sparsity in the model , driving irrelevant features to zero and preserving only the most fault-sensitive information [21], [22]. The resulting sparse feature vectors are then fed into a simple k-nearest neighbors (k-NN) classifier. This end-to-end pipeline is efficient, interpretable, and well-suited for real-time applications [23].

This paper proposes a novel LASSO-Based feature extraction with adaptive windowing via DTW for fault diagnosis in rotating machinery. The main contributions of this paper are summarized as follows:

- A speed-aware feature extraction paradigm that explicitly models the interaction between shaft rotational speed and vibration response under fault conditions, capturing subtle speed-dependent dynamics at 25, 50, and 75 rpm.
- A DTW-based adaptive windowing strategy that identifies the true mechanical period on the wheel by minimizing the DTW distance between consecutive revolutions, thus preventing dilution or truncation of transient fault signatures.
- A sparse LASSO feature selection framework that eliminates redundant descriptors, improving classifier accuracy while maintaining a computational footprint suitable for embedded systems.
- Comprehensive validation on public rotating machinery datasets covering eleven single-fault classes (and combinations of double-faults for robustness), where the proposed method consistently outperforms PCA- and ICA-based baselines in terms of accuracy, F1-score, and model compactness.

The organization of this paper is as follows: In Section II, we summarize the literature review in the field of fault diagnosis in rotating machinery. The data presentation is described in Section III. The methodology is outlined in Section IV, including the proposed LASSO regression with adaptive windowing, which is detailed in Section IV-A. The experimental evaluation and results are presented in Section V. The limitations and future work are mentioned in Section VI. Finally, the conclusion is discussed in Section VII.

II. Literature Review

Effective fault diagnosis in rotating machinery has long been a significant area of research due to the importance of predictive maintenance in industrial applications. Several researches have explored various feature extraction techniques to accurately identify faults using vibration signals [25], [26].

Traditional techniques, such as Fast Fourier Transform (FFT), Wavelet Transform (WT), and Hilbert Transform, focus on frequency-domain analysis to identify characteristic frequencies associated with specific faults [10], [11], [27], [28]. FFT, despite its widespread use, struggles with non-stationary signals, limiting its effectiveness in real-world scenarios involving transient faults [10], [11], [27]. Wavelet transform offers improved performance in analyzing transient events due to its time-frequency localization capability but still requires careful selection of basis functions to optimize results [11], [28].

Dimensionality reduction methods like Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA) have been widely applied to simplify complex vibration datasets, enhancing classification accuracy by reducing redundant or irrelevant features [12], [13], [14]. PCA, for example, captures dominant variance but often misses subtle fault signatures [12], [13]. ICA addresses this limitation by separating mixed signals into independent sources but still encounters difficulties with transient signals affected by varying operational conditions [14].

Recent advancements have introduced adaptive and dynamic approaches to overcome these limitations. Techniques involving Dynamic Time Warping (DTW) have demonstrated potential in accurately aligning and comparing vibration signals under fluctuating speeds without relying on external synchronization methods [18], [19]. Similarly, sparse representation methods have gained attention for their ability to effectively extract relevant sparse features, improving diagnostic accuracy while maintaining computational efficiency [29], [30].

Despite these advancements, few studies have combined adaptive windowing techniques with sparse regression models explicitly to handle speed-dependent transient faults comprehensively. This literature review highlights the critical need for methodologies that integrate these strategies to enhance the precision and robustness of fault diagnosis systems in rotating machinery.

III. Data Presentation

A. Mechanical Faults Dataset

In this paper, we employ vibration and rotational speed signals obtained from a multi-speed fault dataset [24]. This dataset is an open-access resource that provides vibration and

rotational speed data from various fault scenarios involving ball bearings and shafts. The dataset consists of 39 distinct fault scenarios, collected using a Rotary Machine Fault Simulator from SpectraQuest Inc., which includes a DC motor, Bearing 1, Bearing 2, and a shaft supporting two weighted disks. The data is captured using a National Instruments Compact DAQ system.

To collect the data, a tachometer and eight piezoelectric accelerometers are used to record the rotational speed and vibration signals. These sensors are strategically mounted on the motor, gearbox, and both bearing housings, in the x, y, and z directions. The dataset includes one no-fault scenario and 38 fault scenarios, each recorded at three different rotational speeds: 25 rpm, 50 rpm, and 75 rpm. The data are sampled at a rate of 6.4 kHz over a 10-second period, resulting in 64,000 data points for each scenario in time-series format. Each fault scenario is recorded 25 times, providing a comprehensive representation of the machinery's behavior under various conditions.

Table I provides a summary of the rotating machinery fault dataset, listing the different fault scenarios, their corresponding conditions, and the parameters used during data collection. This dataset is critical for the subsequent feature extraction and fault diagnosis tasks, as it offers detailed insights into the relationship between vibration signals and fault conditions at different operational speeds.

TABLE I. DESCRIPTION OF MULTIPLE MECHANICAL FAULTS DATASETS

Description	Details
Scenarios	39 scenarios
Rotational Speed	25 rpm, 50 rpm, 75 rpm
Time record	10 Seconds/scenario
Number of Trial	25 trials/scenario
Sampling rate	6.4 KHz
Data points	64,000 points/scenario
Sensors	1 Tachometer (Shaft), 8 Piezoelectric Accelerometers (6 bearing housings, 2 gearbox housing)

B. Data Preparation

In this paper, the rotating machinery dataset [24] consists of time-series data for 39 fault scenarios involving ball bearings and shafts. Each trial is grouped into a single file corresponding to each fault scenario and rotational speed. However, the data from trial number 17, which corresponds to a ball bearing number 2 defect at 50 rpm, is shorter than the others. To ensure consistency across all data, we discard the data from trial number 17 across all fault scenarios and rotational speeds. As a result, the length of each grouped file is standardized to 1,536,000 data points per fault scenario at each rotational speed. In total, 117 files are generated, corresponding to the 39 fault scenarios recorded at three distinct rotational speeds. This standardized dataset is now ready for feature extraction and classification analysis.

IV. METHODOLOGY

This section outlines the methodology employed in this paper, providing detailed insights into the feature extraction process, with a particular emphasis on LASSO regression. Additionally, we present the evaluation of the proposed feature extraction techniques using the k-NN classifier. The overall workflow is organized into four main steps, as illustrated in Fig. 1. The first step describes the algorithm developed based on DTW for adaptive window sampling, while the second step explains the application of LASSO regression for feature extraction. The final step focuses on the k-NN classification technique used to evaluate the extracted features. This structured approach enables a comprehensive assessment of the proposed methodology's performance in fault diagnosis.

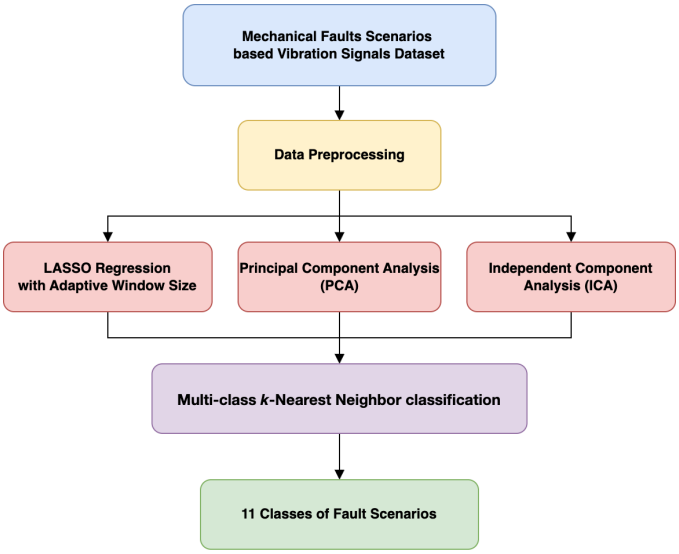


Fig. 1. Workflow of ball bearing and shaft fault classification.

A. Adaptive Window Based on DTW

In fault diagnosis of rotating machinery, selecting the appropriate observation window size is essential for extracting meaningful features from vibration and rotational speed signals. The window size must cover the entire rotational period, which can vary based on the specific characteristics of each signal. To address this, Dynamic Time Warping (DTW) is employed to evaluate the similarity between two time-series signals, allowing the alignment of successive machine rotations. The proposed adaptive windowing Algorithm 1 is designed to dynamically determine the optimal window size for each fault scenario by calculating the DTW distance between different segments of the vibration signals. The algorithm begins by initializing the vibration signal data, $\mathbf{D}$, and an array to store the optimal window size for each fault scenario. For each case, the algorithm iterates over a range of possible window sizes, from a minimum value, $n_{\min}$, to a maximum value, $n_{\max}$. At each step, two segments of the signal are extracted: $\mathbf{d}_1$ (the first segment) and $\mathbf{d}_2$ (the second segment), and the DTW distance between them is computed. DTW measures the dissimilarity between the two time-series, providing an optimal alignment even if the signals are shifted or stretched in time. The window size that minimizes the DTW distance is chosen as the optimal window for the current scenario.

Once all fault scenarios have been processed, the algorithm selects the largest optimal window size across all scenarios,

taking into account the varying rotational speeds and the effect of disturbances. This ensures that the window size is adequately adjusted to cover the true mechanical period, even when the machine experiences fluctuations in speed. By utilizing DTW, the algorithm effectively addresses the problem of varying signal periods without requiring additional hardware or external speed sensors.

The adaptive windowing approach improves fault diagnosis by accurately localizing transient fault signatures, enhancing the feature extraction process. This method is not only sensor-agnostic but also provides the flexibility to adapt to different fault scenarios, thereby improving the classification accuracy and robustness of the system. The DTW-driven window selection process ensures that transient impacts are captured accurately, facilitating efficient and precise fault detection in real-time applications.

Algorithm 1 Adaptive Window Size Estimation Using DTW

Require: $\mathbf{D}$: Vibration Signals in Time-series
Ensure: Optimal window size
Let $\mathbf{D}_{all} \leftarrow$ optimal window size of all cases
for case $= 1$ to last case **do**
  for $n = n_{\min}$ to $n_{\max}$ **do**
    Extract segment 1: $\mathbf{d}_1 = \mathbf{D}(1:n, 1)$
    Extract segment 2: $\mathbf{d}_2 = \mathbf{D}(n+1:2n, 1)$
    Compute DTW distance: $d(n) = \text{DTW}(\mathbf{d}_1, \mathbf{d}_2)$
  end for
  Select optimal window size:

$$n^* = \min d(n)$$

  Output n^* for current state
  Store n^* to $\mathbf{D}_{all}$
end for
Select maximum window size from all cases:

$$\hat{n} = \max \mathbf{D}_{all}$$

return Optimal window size $\hat{n}$

B. LASSO Regression

LASSO regression is an extension of Ordinary Least Squares (OLS) regression [20] designed to perform both variable selection and regularization, which enhances the prediction accuracy of the model. This method aims to minimize the sum of squared residuals while imposing a constraint on the sum of the absolute values of the regression coefficients. This constraint encourages sparsity by forcing some of the coefficients to shrink to exactly zero, which effectively removes unimportant features and helps avoid overfitting. The LASSO objective function is expressed as:

$$\hat{\beta} = \text{argmin} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1} \beta_j x_{ij} \right)^2 \right\} \quad \text{s.t.} \sum_{j=1} |\beta_j| \leq \lambda, \tag{1}$$

where N represents the number of observations, y_i is the response variable at observation i, x_i is the vector of predictors associated with observation i, and λ is a non-negative regularization parameter. The parameter $\hat{\beta}$ represents the estimated regression coefficients, which include both the intercept and the coefficients for each predictor. The value of λ controls the strength of the penalty applied to the regression coefficients, with larger values leading to more coefficients being shrunk to zero, thus simplifying the model but potentially sacrificing accuracy. A smaller λ value results in less regularization, which may lead to overfitting. In this paper, we set λ to 1×10^{-8} to balance regularization and model performance. By using the L^1 norm of the coefficients, LASSO encourages sparsity in the model, helping to identify the most relevant features while reducing the dimensionality of the dataset. This sparsity is crucial for improving classification accuracy, as it focuses the model on the most important variables.

The LASSO regression process in this study is applied after using the adaptive window technique to determine the optimal window size for segmenting the time-series data. This ensures that each segment used for feature extraction corresponds to a meaningful portion of the signal. The extracted features are then processed using LASSO regression to estimate the coefficients, which are stored in a matrix for classification. This sparse feature set is then used to train a classifier, improving the accuracy of the fault diagnosis model while keeping computational demands manageable. The use of LASSO regression for feature extraction is particularly beneficial for real-time applications, where both high diagnostic accuracy and efficient use of computational resources are critical.

C. Classification

The k-Nearest Neighbors (k-NN) algorithm is employed as the classification model to evaluate the effectiveness of the proposed feature extraction method. k-NN is a non-parametric, instance-based learning algorithm widely used in statistical pattern recognition and classification tasks. It is particularly effective in handling multi-modal class distributions, where an instance may be similar to multiple class types. For implementation, the fitcknn function in MATLAB is utilized to construct and evaluate the k-NN model, specifically tailored for diagnosing faults in ball bearings and shafts.

The classifier is configured to distinguish among 11 different fault scenarios, as outlined in Table II, which include both single and compound faults such as outer and inner raceway defects in bearings and various shaft misalignment. Feature datasets generated from the proposed method, as well as benchmark methods including Principal Component Analysis (PCA) and Independent Component Analysis (ICA), are used to construct the training and testing sets. The datasets are partitioned using holdout cross-validation, with 70% of the data allocated for training and the remaining 30% for validation.

In the case of the proposed method, feature extraction is preceded by DTW-based adaptive window sampling, which identifies the optimal window size for each rotational speed condition using Dynamic Time Warping. These windows are then applied to both the training and validation datasets. Subsequently, LASSO regression is used to extract sparse features

TABLE II. DETAIL OF 11 SINGLE FAULT SCENARIOS

Number of Class	Detail of Fault Scenarios
Case 1	Outer Raceway defect (Bearing 2)
Case 2	Inner Raceway defect (Bearing 2)
Case 3	Ball and Raceway defect (Bearing 2)
Case 4	Ball defect (Bearing 2)
Case 5	Outer Raceway defect (Bearing 1)
Case 6	Inner Raceway defect (Bearing 1)
Case 7	Ball and Raceway defect (Bearing 1)
Case 8	Coupling end bent (Shaft)
Case 9	Centerally bent (Shaft)
Case 10	Ball defect (Bearing 1)
Case 11	No Fault

by establishing a relationship between the rotational speed and the corresponding vibration signals within each adaptively sampled window. This approach allows for the identification of the most discriminative and compact features.

In contrast, PCA and ICA are applied using a fixed window size of 640 data points, chosen based on the dataset's temporal resolution and recording parameters. For PCA, only the first principal component (PC1) is retained, which captured more than 95% of the total variance and is deemed sufficient for representing the feature space. Similarly, ICA decomposed the signals into statistically independent components, and the most informative component is selected as the feature representation.

By comparing the performance of these three methods under consistent classification settings, the paper aims to rigorously validate the superiority of the proposed DTW-adaptive LASSO-based feature extraction framework in terms of classification accuracy, computational efficiency, and robustness to varying operational conditions.

V. SIMULATION RESULTS

This research evaluates a novel feature extraction approach based on LASSO regression combined with an adaptive windowing strategy using Dynamic Time Warping (DTW). The proposed method is benchmarked against two widely-used dimensionality reduction techniques: Principal Component Analysis (PCA) and Independent Component Analysis (ICA). The objective is to assess whether the proposed LASSO-DTW method can achieve superior performance in fault classification using time-series data derived from vibration and rotational speed measurements of rotating machinery.

The experimental evaluation is conducted using a k-Nearest Neighbors (kNN) classifier, which was used to compare the effectiveness of each feature extraction technique. The classifier is configured to recognize 11 distinct fault scenarios across three rotational speeds: 25 rpm, 50 rpm, and 75 rpm. Four standard classification performance metrics are used: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the models' ability to detect and classify faults reliably.

TABLE III. RESULTS FOR K-NN CLASSIFICATION PERFORMANCE BASED ON PROPOSED METHOD

Rotational Speed	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
25 rpm	96.97	97.73	96.95	96.88
50 rpm	93.94	95.45	93.94	92.86
75 rpm	84.85	87.88	84.85	84.85

TABLE IV. K-NN CLASSIFICATION PERFORMANCE FOR COMPARISON OF FEATURE EXTRACTION METHODS AT 25 RPM

Feature Extraction	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Proposed method	96.97	97.73	96.95	96.88
PCA	36.46	35.89	36.56	35.96
ICA	90.85	94.75	90.89	88.97

The classification results using the proposed LASSO-DTW method are summarized in Table III, with corresponding confusion matrices shown in Fig. 2. At 25 rpm, the method achieved 96.97% accuracy, 97.73% precision, 96.95% recall, and a 96.88% F1-score. At 50 rpm, it maintained high performance with 93.94% accuracy, 95.45% precision, 93.94% recall, and a 92.86% F1-score. Even at 75 rpm, under more dynamic conditions, it achieved 84.85% accuracy, 87.88% precision, 84.88% recall, and an F1-score of 84.85%.

In contrast, PCA and ICA yield significantly lower performance. At 25 rpm, PCA achieved just 36.46% accuracy, while ICA reached 90.85%. At 50 rpm, PCA dropped further to 18.44% accuracy, with ICA achieving 81.41%. At 75 rpm, PCA recorded 20.64% accuracy, while ICA reached 77.41% (Tables IV to VI).

Beyond classification accuracy, the proposed method significantly reduces feature dimensionality compared to PCA and ICA. The feature matrix generated using the LASSO-DTW method is sized at 6,564 × 55 for the training set. In contrast, both PCA and ICA produced much larger feature matrices of 18,700 × 640. Despite this reduction, the proposed method maintained superior diagnostic performance, illustrating its capability to extract compact and highly relevant features without sacrificing accuracy.

This dimensionality reduction is especially beneficial in real-time and resource-constrained environments, where computational efficiency is crucial. The sparsity-promoting nature of LASSO enables the model to retain only the most informative features, improving interpretability and scalability. The combination of high classification performance and reduced feature complexity underscores the practical value of the proposed method for real-world applications in industrial fault diagnosis systems.

VI. DISCUSSION

The simulation results demonstrated the substantial advantages of integrating LASSO regression with DTW-based

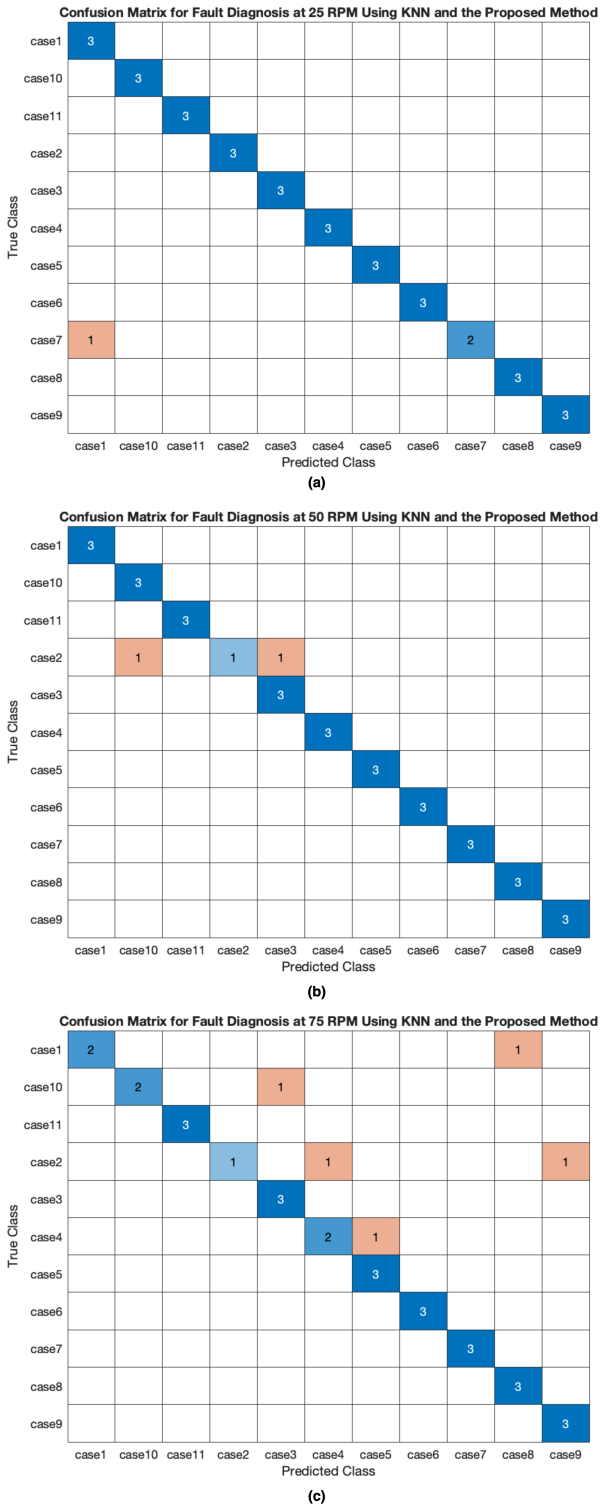


Fig. 2. Confusion chart of three rotational speeds (a) 25 rpm using KNN and the proposed method; (b) 50 rpm using KNN and the proposed method (c) 75 rpm using KNN and the proposed method.

adaptive windowing in fault diagnosis for rotating machinery. The proposed method consistently outperformed PCA and ICA across various operational speeds, illustrating its robust ability to manage transient and speed-dependent fault signatures

TABLE V. K-NN Classification Performance for Comparison of Feature Extraction Methods at 50 RPM

Feature Extraction	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Proposed method	93.94	95.45	93.94	92.86
PCA	18.44	18.84	18.39	16.77
ICA	81.41	82.77	81.82	82.77

TABLE VI. K-NN Classification Performance for Comparison of Feature Extraction Methods at 75 RPM

Feature Extraction	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Proposed method	84.85	87.98	84.85	84.85
PCA	20.64	20.77	20.51	20.00
ICA	77.41	77.41	76.29	77.41

effectively. The performance at lower rotational speeds (25 rpm and 50 rpm) was particularly noteworthy, achieving very high accuracy, precision, and recall. Even at higher speeds (75 rpm), where signal complexity significantly increases, the method maintained strong performance, highlighting its adaptability to real-world industrial environments. Moreover, the reduced dimensionality provided by LASSO regression significantly improved computational efficiency without compromising accuracy. This makes the proposed methodology highly suitable for real-time diagnostic systems, particularly in resource-constrained and applications.

However, the proposed method still has some limitations. Its effectiveness may depend on the quality and duration of the collected signals, and transient anomalies in extremely short-duration faults have still challenges. Moreover, it requires determining an optimal initial period for the adaptive windowing based on DTW, which directly affects the accuracy of identifying the rotational period. Future research could explore integrating advanced deep learning approaches or hybrid feature extraction methodologies based on sparse representation to further enhance performance. Additionally, validating the method across a wider range of operational conditions and fault types would help ensure broader applicability and reliability in diverse industrial scenarios.

VII. Conclusion

In this paper, we have proposed a feature extraction method based on LASSO regression for fault diagnosis in rotating machinery, enhanced by an adaptive windowing technique based on Dynamic Time Warping (DTW). We compared the classification accuracy of our proposed technique with PCA and ICA, and the results clearly demonstrate the effectiveness of combining LASSO with the adaptive DTW windowing approach. Specifically, our proposed method outperformed PCA and ICA in classifying ball bearings and shafts across multiple fault scenarios. For instance, our LASSO-based method achieved a classification accuracy of 96.97%

at 25 rpm, while achieving 93.94% and 84.85% at 50 rpm and 75 rpm, respectively. Furthermore, the integration of the DTW-based adaptive windowing technique enables dynamic adjustment of the observation window, ensuring that transient fault signatures are accurately localized within each revolution, even when rotational speeds vary. This improvement enhances the model's accuracy by addressing the challenge of unknown and fluctuating vibration periods. Additionally, the reduction in feature size achieved through LASSO regression, without compromising diagnostic performance, highlights the method's efficiency and potential for real-time applications.

REFERENCES

[1] M. A. R. Alicando, G. M. Ramos and C. F. Ostia, "Bearing Fault Detection of a Single-phase Induction Motor Using Acoustic and Vibration Analysis Through Hilbert-Huang Transform," *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, Manila, Philippines, pp. 1-6, 2021.

[2] A. Sharma, G. K. Patra and V. Naidu, "Bearing Fault Classification using Acoustic Features and Artificial Neural Network," *2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*, Bangalore, India, pp. 421-425, 2022.

[3] R. R. Shubita, A. S. Alsadeh and I. M. Khater, "Fault Detection in Rotating Machinery Based on Sound Signal Using Edge Machine Learning," *IEEE Access*, vol. 11, pp. 6665-6672, 2023.

[4] M. Kuncan, "An Intelligent Approach for Bearing Fault Diagnosis: Combination of 1D-LBP and GRA," *IEEE Access*, vol. 8, pp. 137517-137529, 2020.

[5] Y. L. Liang, W. Chien and S. F. Yuan, "Overview of Bearing Fault Diagnosis Based on Vibration Signal," *2023 IEEE 3rd International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, Taichung, Taiwan, pp. 547-549, 2023.

[6] Y. Liu, Y. Cheng, Z. Zhang and S. Yang, "Early Fault Diagnosis of Bearing Faults Using Vibration Signals," *2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, Changsha, China, pp. 747-751, 2021.

[7] C. Li, Z. Cao, S. Li and J. Dai, "Frequency Estimation of Vibration Signals: A Subspace Approach for Bearing Fault Diagnosis," *IEEE Sensors Journal*, vol. 24, no. 1, pp. 449-459, 2024.

[8] T. Ince et al., "Early Bearing Fault Diagnosis of Rotating Machinery by 1D Self-Organized Operational Neural Networks," *IEEE Access*, vol. 9, pp. 139260-139270, 2021.

[9] Y. Zhang, D. Ye, and Y. Liu, "Robust Locally Linear Embedding Algorithm for Machinery Fault Diagnosis," *Neurocomputing*, vol. 273, pp. 323–332, 2018.

[10] M. Unal, M. Onat, M. Demetgul, and H. Kucuk, "Fault Diagnosis of Rolling Bearings using A Genetic Algorithm Optimized Neural Network," *Measurement*, vol. 58, pp. 187–196, 2014.

[11] L. Ventricci, R. F. Ribeiro Junior, and G. F. Gomes, "Motor Fault Classification using Hybrid Short-Time Fourier Transform and Wavelet Transform with Vibration Signal and Convolutional Neural Network," *J. Braz. Soc. Mech. Sci. Eng.*, vol. 46, p. 337, 2024.

[12] G. Jin, T. Wang, Y. Amirat, Z. Zhou, and T. Xie, "A Layering Linear Discriminant Analysis-Based Fault Diagnosis Method for Grid-Connected Inverter," *Journal of Marine Science and Engineering*, vol. 10, no. 7, p. 939, 2022.

[13] Y. Xie and T. Zhang, "A Fault Diagnosis Approach using SVM with Data Dimension Reduction by PCA and LDA Method," *2015 Chinese Automation Congress (CAC)*, Wuhan, pp. 869-874, 2015.

[14] A. Hyvarinen and E. Oja, "Independent Component Analysis Algorithms and Applications," *Neural Networks*, 13(4-5), pp. 411-430, 2000.

[15] Y. Liu, B. He, F. Liu, S. Lu, and Y. Zhao, "Feature Fusion using Kernel Joint Approximate Diagonalization of Eigen-Matrices for Rolling Bearing Fault Identification," *Journal of Sound and Vibration*, vol. 385, pp. 389–401, 2016.

[16] C. Zang, Y. Liang, and Q. Niu, "MPIFNet A Multi-Path Information Fusion Fault Diagnosis Network Based on Time Series Two-Dimensional Transformation," *Appl. Sci.*, vol. 14, p. 11947, 2024.

[17] J. Zhang, Y. Kong, Z. Chen, T. Han, Q. Han, M. Dong, and F. Chu, "CBAM-CRLSGAN: A novel fault diagnosis method for planetary transmission systems under small samples scenarios," *Measurement*, vol. 234, p. 114795, 2024.

[18] M. D. Choudhury, L. Hong, and J. S. Dhupia, "Bearing Signal Classification Using Dynamic Time Warping," in *The 5th International Conference on Vibration and Energy Harvesting Applications (VEH 2024)*, L. Tang, K. Aw, G. Hu, and J. Wang, Eds., Singapore: Springer Nature Singapore, pp. 395–404, 2025.

[19] T. Kim, J. Park, J. Yoo, J. M. Ha, and B. D. Youn, "Enhancing Gearbox Fault Diagnosis under Phase Estimation Errors: A Dynamic Time Warping and Blind Deconvolution Approach," *Journal of Sound and Vibration*, vol. 598, p. 118851, 2025.

[20] Tibshirani, R. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B*, Vol. 58, No. 1, pp. 267–288, 1996.

[21] O. Queen and S. J. Emrich, "LASSO-based Feature Selection for Improved Microbial and Microbiome Classification," *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Houston, TX, USA, pp. 2301-2308, 2021.

[22] D. Chellappan and H. Rajaguru, "Significance of LASSO Regression as Feature Extraction Techniques in Diabetic Detection," *2023 Third International Conference on Smart Technologies, Communication and Robotics (STCR)*, Sathyamangalam, India, pp. 1-4, 2023.

[23] T. Tran and J. Lundgren, "Drill Fault Diagnosis Based on the Scalogram and Mel Spectrogram of Sound Signals Using Artificial Intelligence," *IEEE Access*, vol. 8, pp. 203655-203666, 2020.

[24] A. Marshall and D. Jensen, "Dataset of Single and Double Faults Scenarios using Vibration Signals from A Rotary Machine," *Data in Brief*, vol. 49, no. 109358, 2023.

[25] H. Bendjama, "Feature Extraction Based on Vibration Signal Decomposition for Fault Diagnosis of Rolling Bearings," *Int J Adv Manuf Technol*, 130, pp. 821–836, 2024.

[26] B. Zhou, J. Zhang, Y. Fu, D. Yu and J. Wang, "Feature Extraction and Analysis Method of Spindle Vibration Signal for Tool Wear Monitoring," *2021 5th International Conference on Automation, Control and Robots (ICACR),* Nanning, China, pp. 184-189, 2021.

[27] Y. Wang, L. Zheng, Y. Gao and S. Li, "Vibration Signal Extraction Based on FFT and Least Square Method," *IEEE Access,* vol. 8, pp. 224092-224107, 2020.

[28] G. Sakthivel, D. Saravanakumar, R. Jegadeeshwaran, R. Rajakumar, T. M. Alamelu Manghai and S. Abirami, "Development of a Real-Time Fault Detection Model for Hydraulic Brake Systems Using Vibration Analysis and Machine Learning With Wavelet Features," *IEEE Access,* vol. 12, pp. 177442-177455, 2024.

[29] Y. Li, Y. Zhang, J. Wu and M. Xie, "Regularized Periodic Gaussian Process for Nonparametric Sparse Feature Extraction From Noisy Periodic Signals," *IEEE Transactions on Automation Science and Engineering,* vol. 22, pp. 3011-3020, 2025.

[30] M. Miao and J. Yu, "Sparse-Representation-Network-Based Feature Learning of Vibration Signal for Machinery Fault Diagnosis," *IEEE Transactions on Industrial Informatics,* vol. 19, no. 5, pp. 6706-6716, May 2023.

The Influence of Familiarity with Traffic Regulations on Road Safety: A Simulated Study on Roundabouts and Intersections

Raghda Alqurashi[1], Hasan J. Alyamani[2], Nesreen Alharbi[3], Hasan Sagga[4]
Department of Computer Science, Jamoum University College,
Umm Al-Qura University, Al Jumum 25371, Saudi Arabia[1]
Department of Information Systems-Faculty of Computing and Information Technology in Rabigh,
King Abdulaziz University, Rabigh, Saudi Arabia[2]
Department of Computer Science-Faculty of Computing and Information Technology in Rabigh,
King Abdulaziz University, Rabigh, Saudi Arabia[3]
THE LIGHT-Faculty of Computing and Information Technology in Rabigh,
King Abdulaziz University, Rabigh, Saudi Arabia[2,3]
Management Information Systems Department-College of Business, University of Jeddah, Asfan, Saudi Arabia[4]
Human Computer Interaction, Iowa State University, Ames, Iowa, USA[4]

***Abstract*—International drivers who come from keep-right countries and drive in keep-left countries are frequently involved in road accidents due to unfamiliarity with keep-left traffic regulations. Due to unfamiliarity of the traffic regulation, the driver's performance and behaviour are subject to change. The objective of this study was to explore the effects of familiarity with traffic regulations on driving performance and behavior at roundabouts and intersections. To achieve this in a safe environment, twenty-one male familiar drivers and thirty- four male unfamiliar drivers participated in driving in a simulated keep-left traffic regulation. The factors observed were not fastening the seat belt, entering the driving simulation from wrong side, using an improper approaching lane, not signalling, speeding, driving against the traffic flow and using an improper exiting lane at each roundabout and intersection. Mann-Whitney U test was used to compare driving behaviour and performance between the familiarity groups. Unfamiliar drivers made significantly more driving mistakes on roundabouts than unfamiliar drivers. Also, some unfamiliar drivers got inside the vehicle from the passenger side instead of driver side and drove against the traffic flow inside the roundabouts. The implications for familiar and unfamiliar driving can be considered for future research development.**

***Keywords*—Driving behaviour; driving performance; familiarity with traffic regulations; road intersections; roundabouts**

I. Introduction

An increasing number of students and tourists from keep-right countries (e.g., China, the United States (US) and Middle Eastern countries) are travelling to keep-left countries (e.g., New Zealand, the United Kingdom and Australia). Hence, the involvement of overseas drivers in road accidents abroad emerges as a potential road safety issue. Some drivers unfamiliar with the local road regulations experience difficulties in adapting their interactions with the road (e.g., keeping to the speed limit and driving with the traffic flow) and with the vehicle (e.g., controlling the direction indicator and windscreen wipers) to match the requirements of the new driving conditions [1], [2]. For instance, the Ministry of Transport in New Zealand, a keep-left country, reported that drivers from keep-right countries such as China Saudi Arabia, UAE, and Turkey, were the main group of overseas drivers involved in head-on collisions due to driving on the wrong side of the road [3]. Difficulties experienced by overseas drivers include the incorrect use of vehicle controls, signalling in time, keeping left and driving in the wrong lane.

While the problem of driving overseas is primarily considered in terms of road safety, little consideration is given to the use of concrete solutions and technology, such as in-vehicle information systems and gamification, to address differences in driving performance and behaviour between drivers who are familiar and unfamiliar with the local traffic system. Therefore, it is important to conduct scientific research on the ways unfamiliar drivers interact with the conditions of driving overseas and compare their driving performance with locals (i.e., familiar drivers) in order to provide a comprehensive insight into overseas drivers' vulnerability, support specific driving-assistance systems and reduce traffic violations as well as facilitating safe driving under unfamiliar conditions.

To gain a better understanding of the involvement of overseas drivers in road accidents and traffic violations when driving under unfamiliar traffic regulations, a simulated traffic environment was established to assess the way the rules of traffic regulation can affect driving behaviour and performance. An experimental task was designed to evaluate the driving behaviour and performance of drivers used to keep-right conditions when interacting with keep-left vehicle configurations and traffic systems prior to and while driving, especially at critical sections of roads (i.e., roundabouts and intersections). Although roundabouts and intersections optimise traffic flow [4], they are also associated with many vehicle accidents [1]. These sections of roads are associated with a relatively large number of driving tasks. While the seatbelt is fastened, the drivers need to perceive a lot of traffic information (e.g., the section of road, road signs, and the position of direction stalk), make critical decisions (e.g., slowing down or speeding up, and driving in the required travelling lane with the traffic flow)

[5], and then made a driving decision and act accordingly [6], The collisions that happen at roundabouts and intersections are likely to be due to the way the driver behaves and performs.

The objective of this research was to determine the effects of familiarity with traffic regulations on driving performance at roundabouts and intersections. It is hypothesized that (H1) drivers unfamiliar with the traffic regulation will commit more traffic violations on roundabouts and intersections than familiar drivers; and (H2) the risk at roundabouts and intersections is higher when overseas drivers drive under unfamiliar traffic regulations in comparison to familiar drivers.

The paper starts with an overview of the literature on the impact of familiarity with the interactions between drivers and unfamiliar elements of the traffic system. The methodology is then described in terms of the study sample, apparatus, driving scenario, experimental procedure, data acquisition and analysis. Then, the results are presented and discussed. At last, a summary of findings is presented.

II. Literature Review

The effects of familiarity with traffic regulations on driving performance have been insufficiently examined in transportation research [7]. However, some studies have acknowledged the interactions between drivers and unfamiliar elements of the traffic system (e.g., unfamiliar roads, vehicle configurations and traffic regulations), such as the behaviour of familiar and unfamiliar drivers in regard to a specific road condition (i.e., road familiarity).

Driving in unfamiliar routes is associated with some difficulties in terms of driving performance and mental workload. For example, the drivers in [8] showed low driving performance and high workload, and the situation was getting better when the drivers used to drive on those routes. However, drivers showed more respect for speed signs when driving in unfamiliar roads [9]. In terms of evaluating the risk subjectively, drivers evaluate the risk as higher when driving on unfamiliar roads in comparison to familiar roads [10]. The study [11] goes further and explores the risk perception in accordance with the driver's gender when driving in unfamiliar routes. That is, while male drivers perceived unfamiliar routes as riskier than familiar routes, female drivers tended to perceive familiar routes as riskier than unfamiliar routes. Compared to familiar drivers, unfamiliar drivers had shorter steering reaction times in case of the presentation of a peripheral object, longer braking reaction times in case of the presentation of an object at the front of the vehicle and kept a longer distance from the vehicle in front of their car [12]. Overall, familiar drivers were characterised by better awareness of surroundings and regulations compared to unfamiliar drivers [7].

In the context of driving under unfamiliar traffic regulations, most studies only examine the driving behaviour, performance and needs of unfamiliar drivers. Unfamiliar drivers were found to frequently ignore roads signs and road information and therefore did not drive under the road speed limit [13], [14]. Unfamiliar drivers also performed unnecessary and unsafe lane changes and did not use the direction indicator when lane departure was imminent [15], [16]. Moreover, some unfamiliar drivers entered roundabouts and intersections from the wrong side of the road [17], [18], [19]. Another study investigated the feedback and presentation mechanisms unfamiliar drivers preferred to help them to drive on specific sections of road. Unfamiliar drivers preferred the visual feedback that presents speed, navigation and direction indicator information concurrently with the performance of the driving task [20].

The study conducted by Yoh et al. [21] explored the driving behaviour of familiar and unfamiliar drivers under Japanese traffic regulations by analysing real-world driving accidents in Japan. While unfamiliar drivers were more likely to perform driving mistakes related to their understanding of traffic rules, familiar drivers committed more violations relating to speed. Also, Abbas et al. [22] studied the behaviour of two groups of drivers: Indigenous and non-Indigenous, where Indigenous drivers were the local drivers who were familiar with the local rules, the road conditions and the behaviour of other drivers in the road and non-Indigenous drivers were new to the local area. The study found that the drivers' responses to speed cameras were influenced by their familiarity with the road regulations.

When driving in unfamiliar traffic conditions, the driver might also be driving a vehicle with unfamiliar configurations. The vehicles that usually used in keep-left countries are right-hand -drive vehicles whose steering wheel is on the right. These vehicles often have the turn signal lever mounted on the right side of the steering column while the windshield wiper controls are mounted on the left side of the steering column. In contrast, left-hand drive vehicles have the steering wheel, signal lever and windshield wiper control the opposite way around. This can affect driving performance. Drivers from Japan (a keep-left country), for instance, were more likely to cross lane lines when driving a left-hand drive vehicle [23]. Additionally, some unfamiliar drivers misused the controls for the direction indicator and windscreen wiper [14], [18]. Another study [24] focused on the interaction between drivers, particularly ones who had mild cognitive impairment, and unfamiliar vehicles to discover the needs of a system to remind the drivers of performing some tasks, such as fastening the seatbelt. Drivers' behaviour regarding fastening seatbelt and speeding is discussed along with other driving behaviour in [25] to investigate the influence of culture on road safety when driving in familiar and unfamiliar environments.

Driving simulator is presented as a valid method for driving research [26]. It is used in a wide range of research (for example, [16], [27], [28], [8]) that explore the interactions between traffic elements under a condition of unfamiliarity. Due to unsafety of testing such interactions, experiments using driving simulator provide a safe observation environment [27], [8]. The driving simulator also allows researchers in [8] to systematically measure driving performance and workload under controlled conditions.

There is still a lack of research that compares the driving performances of familiar and unfamiliar drivers at roundabouts and intersections. Therefore, this study explores the influence of familiarity with traffic regulations on road safety using a driving simulator.

III. METHODS

A. Sample

A total of 55 right-handed male drivers voluntarily took part in the driving simulator study. They were divided into two groups based on the traffic system the drivers were most familiar with driving in. The first group included 21 drivers who were only familiar with keep-left traffic regulations and held a valid driving license issued from a keep-left country. This group was titled "familiar drivers," as they were familiar with the traffic system of this study. The second group included 34 drivers who were only familiar with keep-right traffic regulations and who held a driving license issued in a keep-right countries. The second group was categorised as "unfamiliar drivers." The age range of both groups was between 20 to 35 years. The first group had a mean age of 27.048 years (SD = 4.318) with a mean driving experience of 4.762 years (SD = 2.548) and an average driving hours per week of 9.048 (SD = 10.072). The second group had a mean age of 24.441 years (SD = 4.215) and with a mean driving experience of 6.235 years (SD = 4.390) and an average driving hours per week of 15.353 (SD = 10.557).

B. Apparatus

A fixed-based driving simulator, Forum8 (see Figure 1), located at Macquarie University, was used for the study. The simulator was placed in a small room with its own door. Also, there was a large glass window between this room and an adjacent room through which the research could observe the driving session. However, the glass was covered with a blackout curtain during the driving session in order to provide a less distracting environment and minimise the changes in driving performance due to being watched by the researcher. Forum8 consists of a right-hand drive vehicle cockpit containing a force-feedback steering wheel, direction indicator and windscreen wipers levers, brake and gas pedals, a gear shifter, seatbelt and air-conditioner and multimedia control panel. The front of cockpit is attached to a three-screen display that provides a horizontal and a vertical field of view of 150 degrees and 30 degrees, respectively. The graphical interface was designed by UC-win/Road which is used to model the traffic system, including roads, road signs and texture, vehicles, and trees. Also, UC-win/Road records data in CSV and video logs. The data includes information on the vehicles, users on the road, and driving environment in addition to the interactions among those items for analysis purposes.

C. Driving Scenario

The driving simulator experiment consisted of driving in a keep-left traffic system using a Forum8 simulator as described above. Randomly, a New South Wales (Australia) geographical area from UC-win/Road was selected. The entire driving track was dual-lane and had different road sections, including straight and curved roads, three four-way intersections and three four-exit roundabouts. The scenario had no traffic movements, such as pedestrians, cyclists or other vehicles. Road signs and textures were added in accordance with the Australian regulations. The driving scenario started from a fixed point and ended at another point. The scenario directed



Fig. 1. Driving simulator.

the drivers to drive through all intersections and roundabouts. The sequence and the required turning at each section were as follows: intersection (go straight-ahead), roundabout (turn right), intersection (turn right), roundabout (turn left), intersection (turn left), roundabout (head straight) (see Figure 2).

D. Experimental Procedure

Once the participants arrived, their driving license was checked for its validity and the country of issue. The whole experiment was then explained to participants. If the participants agreed to continue then they signed the consent form and filled out a questionnaire about their demographic information and driving experience. After the questionnaire was completed, the researcher provided the participant with instructions regarding driving rules at intersections and roundabouts in Australia. The participants then moved to the simulator room with the window uncovered to perform the training session which lasted up to 10 minutes. Under observation, participant gradually became accustomed to the vehicle and the simulated environment. After that, the glass was covered with a blackout curtain and the participant moved to the experimental session. At the beginning of the session, the driving task was introduced to participants with the instruction to follow the road directions presented on a map to reach the destination. Participants were reminded to respect all traffic laws. After completing the experimental session, prior to leaving the vehicle seat, the participants called the researcher. The researcher ensured that all required files were saved and then the participant

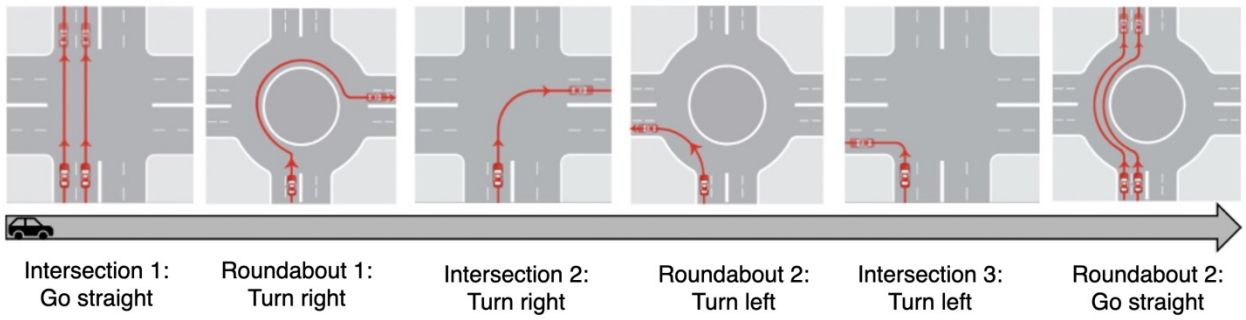


Fig. 2. The Sequence of road sections drive by the participants.

filled out the post-experiment questionnaire, which captured their driving experience under keep-left traffic regulations and driving a simulator.

E. Data Acquisition and Analysis

The current study explores the differences in driving behaviour and performance between familiar and unfamiliar drivers in a keep-left traffic system. Driving behaviour and performance were assessed by exploring the driving mistakes that occurred prior to and while driving. Prior to the training session, when the window glass was uncovered, participants were observed fastening the seat belt by the researcher. It was easily observable that some participants entered the driving simulation from wrong side (i.e., passenger side not the driver side). Hence, mistakes related to getting inside the vehicle were also noted prior to starting the driving.

During the experimental session, mistakes related to driving at each intersection and roundabout were observed from log files. Driving mistakes included (1) using an improper approaching lane, (2) not signalling, (3) speeding, (4) driving against the traffic flow and (5) using an improper exiting lane. If no mistakes were made, 0 points were assigned to the driver, but if any of the above-mentioned mistakes were made, 1 point was assigned to the driver. All mistake scores were integrated into one file for data analysis.

Familiarity with driving under keep-left traffic regulations was a nominal independent variable with two groups: familiar and unfamiliar. The total number of mistakes on roundabouts and intersections and the total number of each type of driving mistake represented the dependent variables. Descriptive statistics were presented as mean, mean rank, minimum and maximum. To compare driving behaviour and performance between the familiarity groups, the Shapiro-Wilk test was conducted to determine normality for driving performance variables. None of driving performance variables were distributed normally ($p < .05$). Therefore, parametric tests (i.e., independent t-tests) could not be employed. Instead, the Mann-Whitney U test as a non-parametric test was applied [29].

IV. Results

A. Pre-Driving

Out of 34 unfamiliar participants, 26 participants (76.5%) did not get inside the simulated vehicle from the correct side, while none of familiar participants made the same mistake. Descriptive statistics indicate that the mean of getting inside the simulated vehicle from the wrong side was 0.77 for unfamiliar participants with a mean rank of 36.03, suggesting a strong tendency toward making such a mistake among unfamiliar participants (see Table I). None of familiar participants got inside the vehicle from the wrong side, resulting a mean of 0.00 and a lower mean rank of 15.00. That indicates a significant disparity between the two groups. The Mann-Whitney U test results indicated that the number of unfamiliar participants who did not fasten the seatbelt (mean = 0.59; mean rank = 36.03) was significantly greater than the number of familiar participants who made the same mistake (mean= 0.24; mean rank = 22.05), U = 232.00, p = 0.006 (see Table II).

B. While Driving

1) Section of road: The results of the Mann-Whitney U test showed there was a significant difference in the total number of driving mistakes on roundabouts between familiar (mean = 2.38; mean rank = 22.71) and unfamiliar (mean = 3.38; mean rank = 31.27) participants; U = 246.000, p = 0.026. On intersections, familiar participants made fewer driving mistakes with a mean of 2.48 and mean rank of 24.41 than unfamiliar participants (mean = 2.94; mean rank = 30.21). However, the difference was not statistically significant (U = 281.500, p = 0.092).

2) Type of mistake: Multiple Mann-Whitney U tests were conducted to compare driving mistakes in relation to their type between familiar and unfamiliar participants. For approaching the roundabouts/intersections from an improper lane, the U statistic was 337.000 (p = 0.361), and for not signalling, it was 349.000 (p = 0.446). Speeding yielded a U statistic of 344.000 (p = 0.40), while using improper exiting lane resulted in a U statistic of 301.000 (p = 0.16). These findings indicate no significant differences in driving mistakes between the groups, suggesting similar driving performance among participants.

Familiar participants did not drive against the traffic flow on roundabouts/intersections, while unfamiliar participants had a mean of this driving mistake of 0.80 (mean rank = 32.63). As such, a statistical comparison between the two groups was not performed, indicating a notable disparity in this mistake.

V. Discussion

The main objective of the study was to determine the effects of familiarity with traffic regulations on the driving

TABLE I. DESCRIPTIVE STATISTICS FOR DRIVING MISTAKES

Mistakes	Group	Minimum	Maximum	Mean	Mean Rank
Getting into the vehicle from the wrong side	Familiar	0	0	0.00	15.00
	Unfamiliar	0	1	0.77	36.03
Not fastening the seatbelt	Familiar	0	1	0.24	22.05
	Unfamiliar	0	1	0.59	31.68
Total number of mistakes on roundabouts	Familiar	0	5	2.38	22.71
	Unfamiliar	0	7	3.38	31.27
Total number of mistakes at intersections	Familiar	1	5	2.48	24.40
	Unfamiliar	0	5	2.94	30.22
Not using an improper approaching lane	Familiar	0	3	1.05	27.05
	Unfamiliar	0	2	1.09	28.59
Not signalling	Familiar	0	4	1.23	27.62
	Unfamiliar	0	4	1.29	28.23
Speeding	Familiar	0	2	0.43	27.38
	Unfamiliar	0	5	0.70	28.38
Driving against the traffic flow	Familiar	0	0	0.00	20.50
	Unfamiliar	0	4	0.79	32.63
Using an improper exiting lane	Familiar	0	5	2.14	25.33
	Unfamiliar	0	6	2.44	29.65

TABLE II. THE RESULTS OF MANN-WHITNEY U TESTS FOR A COMPARISON IN DRIVING MISTAKES BETWEEN FAMILIAR AND UNFAMILIAR PARTICIPANTS

Mistakes	U	p
Getting into the vehicle from the wrong side	NaN^a	
wrong side		
Not fastening the seatbelt	232.00	0.006
Total number of mistakes at roundabouts	246.00	0.026
roundabouts		
Total number of mistakes at intersections	281.50	0.092
intersections		
Not using the improper approaching lane	337.00	0.361
lane		
Not signalling	349.00	0.446
Speeding	344.00	0.397
Driving against the traffic flow	NaN^b	
Using an improper exiting lane	301.00	0.158

a The variance in getting into the vehicle from wrong side is equal to 0 after grouping on familiarity.
b The variance in driving against the traffic flow is equal to 0 after grouping on familiarity.

performance at roundabouts and intersections. An overall comparison between the driving performance prior to and while driving on roundabouts and intersections for familiar and unfamiliar drivers to a keep-left traffic regulation was addressed.

A. Pre-Driving

All familiar participants got into the vehicle from the correct side of the vehicle (i.e., the right side, which represents the driver side in a right-hand drive vehicle), while 76% of unfamiliar participants got inside the vehicle from the wrong side (i.e., the left side, which represents the passenger side in a right-hand drive vehicle). When the unfamiliar participants were asked about the reasons for making this mistake, they mentioned the unreality of the driving environment, hesitation concerning which is the correct side, and complete unawareness of the correct side. Statistical comparison using the Mann-Whitney U test was not feasible for not getting into the vehicle from the correct side due to the lack of collected data from familiar participants in relation to this mistake. Therefore, recruiting more familiar participants is required to better assess differences between the familiarity groups for this mistake. In general, although this mistake does not affect road safety, making such a mistake might be an annoying or embarrassing experience for the driver.

The results of this study indicate the number of unfamiliar participants who did not fasten their seatbelt was significantly greater than familiar participants. According to the post-simulation questionnaire, all familiar participants and the three unfamiliar participants who did not make this mistake thought it was not important to fasten their seatbelt as the driving session was not real but they did it anyway. For the unfamiliar drivers, there were additional reasons for not fastening the seatbelt. Some participants were unfamiliar with configurations of right-hand drive vehicles, while others forgot to fasten the seatbelt as they did not usually fasten it back in their home country or they did not realise that there was a seatbelt. Although this violation was observed prior to the driving tasks, fastening the seatbelt is very important factor in road safety. It reduces fatal injuries by between 45% and 60% [30]. Regarding this factor, a strategic awareness campaign intended to educate countries with less awareness of the importance of fastening the seatbelt is proposed by [31]. Changing behaviour towards fastening seatbelts through serious games is introduced by [32]. Another possible solution is to employ tactical feedback to remind the driver to fasten their seatbelt [33]. When conducting driving simulation experiments, it is important to emphasise the importance of fastening the seatbelt prior the

observation process even though the experiment environment is simulated.

B. While Driving

1) Section of road: The results show a significant difference in the total number of driving mistakes on roundabouts between familiar and unfamiliar participants. On intersections, although unfamiliar participants performed more traffic violations compared to unfamiliar participants, this difference is not statistically significant. It might be that the intersections need fewer steering-related activities than roundabouts and therefore the drivers need to be more aware of the required activities in relation to the section of road at roundabouts in comparison to intersections. Drivers' awareness is higher when driving under familiar conditions than unfamiliar conditions [7]. More studies are required to explore why the performance on roundabouts is different than intersections, although a roundabout is a form of intersection.

2) Type of mistake: The results indicate an insignificant difference in speeding violations at roundabouts and intersections between familiar and unfamiliar drivers. The results conflict with those obtained in studies by [13], [14], who found that familiar drivers are more likely to perform speed violations and by [21], who found that unfamiliar drivers are more likely to ignore road signs and information when driving under unfamiliar traffic regulations. The reasons speed violations were committed by both groups in this study might be driving in unfamiliar routes for familiar group of participants [9] and the high cognitive load the unfamiliar drivers experienced when driving under unfamiliar traffic regulations [15]. More studies are required to explore the reasons behind the speeding violations by both the familiar and unfamiliar drivers.

Approaching and exiting roundabouts and intersections from the correct travelling lane can be related to steering skills and making correct decisions in accordance with the target turn. Despite the results of study conducted by [12], which found different steering activities between familiar and unfamiliar drivers, the results reveal that there was no significant difference in driving in an improper travelling lane when approaching and exiting roundabouts/intersections between familiar and unfamiliar drivers. Another reason for making such a mistake might be driving a simulator whose sensitivity might differ from a real vehicle. Further studies are required to understand the reasons behind this sort of mistakes for both groups of drivers. Signalling while turning or changing the lanes informs other users of road (e.g., other drivers on the road, pedestrians and motorcyclists) about the driver's near future turn or manoeuvre and gives time to others to plan or react accordingly. The reasons for not signalling could be driving habit, that it is unnecessary to give a signal (e.g., no traffic on the road), or misusing the control of direction indicator. These might explain why the results of this study did not show a significant difference in signalling when approaching roundabouts/intersections between familiar and unfamiliar drivers. The driving track of this study did not include any traffic. Additionally, some unfamiliar participants were confused about the correct location of direction indicator lever, as it is located to the right of the steering wheel, which may be an unfamiliar configuration.

The results of our study reveal that unfamiliar participants drove 27 times against the traffic flow when they entered the roundabouts/intersections, while none of the familiar participants made that mistake. This result does not allow us to perform a statistical comparison using the Mann-Whitney U test. Adding more familiar participants can provide extra data and thus better understanding for this situation. However, that might not guarantee that familiar drivers are going to drive against the traffic, as they are used to driving with the traffic flow, unlike the unfamiliar participants who made this mistake. In contrast to familiar drivers, unfamiliar drivers are required to distinguish the differences in traffic systems and rules between their home country and the host country. The unfamiliar drivers must adapt their knowledge and reactions to match the requirements of driving in the host country. Difficulties in adapting to keep-left traffic regulations might explain why unfamiliar participants drove against the traffic flow. Another reason for this behaviour might be that the unfamiliar drivers are more likely to ignore road signs [13], [14]. Despite presenting the required direction on the roundabout-ahead sign, unfamiliar participants might not pay attention to the sign or notice the difference in the information on the roundabout-ahead sign. Driving against the traffic flow results in failing the driving test in New South Wales, as well as in all states and territories, Australia [34] and increases the risk of serious and often fatal crashes (e.g., head-on collisions). Even though the familiar and unfamiliar drivers showed no significant difference in other driving mistakes, the risk at roundabouts and intersections was higher when unfamiliar drivers drove in unfamiliar traffic regulations in comparison to familiar drivers.

These results could be used to improve the few existing driving-assistance systems available for overseas drivers, particularly those who are driving under unfamiliar traffic regulations, as the improvement in such systems could improve the road safety. As automated car technology develops, it could be possible to customize the vehicle control automation level based on the driver's familiarity with the traffic regulations. In cases where the driver is not familiar with the local regulations, the system can fully or partially automate the vehicle when approaching complex roads, such as roundabouts and intersections. Other improvements include providing understandable and clear information to drivers based on their familiarity with the traffic regulations. However, this information must be easy to understand as providing complex information might increase the driver's cognitive load and lead to them ignoring the information.

Based on these results, suggested solutions include considering unfamiliarity with road regulations when designing an in-vehicle information system. Unlike familiar drivers, unfamiliar drivers need information regarding the correct direction of traffic when they reach roundabouts and intersections. Another solution is to modify the information that is provided to drivers via road signs to consider the issue of unfamiliarity with traffic regulation. The intersection-ahead sign, for instance, does not currently inform drivers about the direction of traffic flow. Presenting the direction of flow in perceivable locations as well as presenting the road structure might help unfamiliar drivers to correctly drive with the traffic flow when entering intersections (see Figure 3). Another designing suggestion is to use "do not enter" signs as part of the image of intersection to describe the traffic flow situation at upcoming intersections. Presenting the

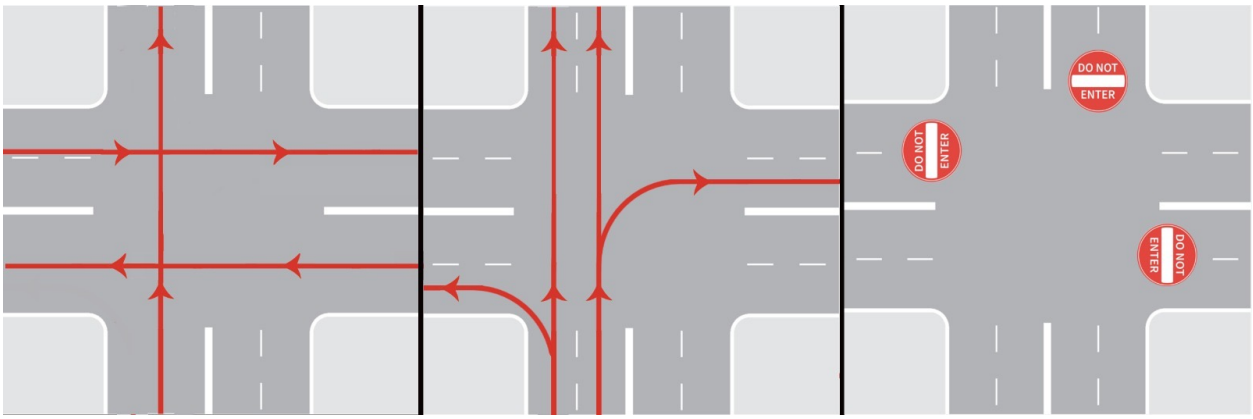


Fig. 3. Three possible alternative designs for intersection-ahead sign to facilitate driving in intersections.

road signs to the driver via an in-vehicle information system is better than the traditional road signs in terms of the proper driver's response to changes in the road environment [35].

Other solutions are suggested in the study conducted by [36], such as providing different navigation information (e.g., turn instructions), vehicle status information (e.g., seatbelt status), and driving assistance information (e.g., direction visualisation) based on the driver's familiarity with the soundings. However, in such systems, the information should be carefully designed. The study [37] recommends that an international and ergonomic standards should be considered when designing the road signs. Gamification and serious games also can provide suitable learning platform for drivers to adapt their behaviour and performance to safely drive under unfamiliar traffic regulations [38], [39].

C. Limitations and Future Work

The present study had some limitations in the driving tasks and the driving scenario. Future developments could include adding further driving tasks, such as overtaking. While the driver overtakes a vehicle from the right in a keep-left traffic regulation, the driver in a keep-right traffic regulation overtakes a vehicle from the left. Testing this driving task should be conducted on at least a three-lane road to make it possible for the driver to overtake from the right or the left. Testing familiar and unfamiliar drivers together in the same driving scenario can extend the knowledge about real interactions and behaviours when both groups of drivers drive together on the same road.

VI. Conclusion

In this paper, the differences in driving mistakes between familiar and unfamiliar drivers in a keep-left traffic regulation system were explored. Using a driving simulator, two pre-driving mistakes and five while-driving mistakes were calculated. The pre-driving mistakes were getting inside the vehicle from incorrect side of the vehicle and not fastening the seatbelt. The while-driving mistakes were using an improper approaching lane, not signalling, speeding, driving against the traffic flow, and using an improper exiting lane. Fifty-five right-handed male drivers participated in the study and were grouped into two groups (21 familiar drivers with a keep-left traffic regulation and 34 unfamiliar drivers with a keep-left traffic regulation). The results of this study confirm the first hypothesis (H1) that unfamiliar drivers will commit more traffic violations compared to familiar drivers, as the unfamiliar drivers performed significantly more mistakes than the familiar drivers did on roundabouts and insignificantly more mistakes on intersections. The results also demonstrate that the risk in traffic systems is higher when drivers drive under unfamiliar traffic regulations (H2, the second hypothesis), as the unfamiliar drivers might enter the roundabouts and intersections from the opposite traffic direction to which they are accustomed. Therefore, familiarity with traffic regulations should be considered when designing the driving-assistant systems, in-vehicles information systems, gamification and serious games, and other educational programs. That would help improving road safety, particularly in countries that have many unfamiliar drivers.

References

[1] B. Watson, D. Tunnicliff, J. Manderson, and E. O'Connor, *The Safety of International Visitors on Australian Roads - A Joint Report (Volume 2)*. Brisbane, Australia: CARRS-Q, 2004.

[2] B. Prideaux and D. Carson, *Drive Tourism: Trends and Emerging Markets*. Routledge, 2010, vol. 17.

[3] Ministry of Transport, "Overseas driver crashes 2016," Ministry of Transport, Wellington, New Zealand, Tech. Rep., 2016.

[4] A. Kalasova, M. Poliak, L. Skorvankova, and P. Fabian, "Optimization of traffic at uncontrolled intersections: Comparison of the effectiveness of roundabouts, signal-controlled intersections, and turbo-roundabouts," *Urban Science*, vol. 8, p. 217, 2024.

[5] J. Han, M. Singh, and D. Zhao, "Road safety literacy for speakers of english as a foreign language: Educating novice drivers for the public's health safety," *Literacy and Numeracy Studies*, vol. 18, pp. 52–66, 2010.

[6] I. Khitrov, "Safety study of a roundabout," *Journal of Mechanical Engineering and Transport*, vol. 18, no. 2, pp. 175–182, 2024.

[7] I. Harms, B. Burdett, and S. Charlton, "The role of route familiarity in traffic participants' behaviour and transport psychology research: A systematic review," *Transportation Research Interdisciplinary Perspectives*, vol. 9, p. 100331, 2021.

[8] Y.-C. Lee, F. Wen, and C.-H. Wang, "Round-trip driving effects on driving performances and mental workload under different traffic rules," *International journal of industrial ergonomics*, vol. 95, p. 103437, 2023.

[9] A. Al-Harbi and A. Jamjoom, "Analysis of drivers attitude to variable speed limit systems on multi-lane highways," *Multi-Knowledge Electronic Comprehensive Journal For Education Science Publications (MECSJ)*, 2021.

[10] M. Yanko and T. Spalek, "Route familiarity breeds inattention: A driving simulator study," *Accident Analysis Prevention*, vol. 57, pp. 80–86, 2013.

[11] N. Budak, İ. Öztürk, M. Aslan, and B. Öz, "How drivers' risk perception changes while driving on familiar and unfamiliar roads: A comparison of female and male drivers," *Trafik ve Ulaşım Araştırmaları Dergisi*, vol. 4, no. 1, pp. 39–48, 2021.

[12] C. Li, "Exploring the attitude of chinese self-drive tourists towards driving safety in new zealand," 2024.

[13] M. Wu, "Driving an unfamiliar vehicle in an unfamiliar country: Exploring chinese recreational vehicle tourists' safety concerns and coping techniques in australia," *Journal of Travel Research*, vol. 54, pp. 801–813, 2015.

[14] C. Liu, H. Aliamani, and M. Kavakli, "Behavior-intention analysis and human-aware computing: Case study and discussion," in *2017 12th IEEE Conference On Industrial Electronics And Applications (ICIEA)*, 2017, pp. 516–521.

[15] H. Alyamani and M. Kavakli, "Situational awareness and systems for driver-assistance," in *Proceedings Of The 50th Hawaii International Conference On System Sciences*, 2017.

[16] H. Alyamani, R. Alturki, A. Yunianta, N. Alromema, H. Sagga, and M. Kavakli, "Situation awareness levels to evaluate the usability of augmented feedback to support driving in an unfamiliar traffic regulation," *International Journal Of Advanced Computer Science And Applications*, vol. 12, pp. 804–812, 2021.

[17] H. Alyamani, M. Kavakli, and S. Smith, "Vehand: an in-vehicle information system to improve driving performance in an unfamiliar traffic regulation," *International Journal Of Human Factors And Ergonomics*, vol. 6, pp. 355–389, 2019.

[18] H. Alyamani, M. Alsharfan, M. Kavakli-Thorne, and Others, "Towards a driving training system to support cognitive flexibility," 2017.

[19] H. Alyamani, A. Hinze, S. Smith, and M. Kavakli, "Preference feedback for driving in an unfamiliar traffic regulation," in *Service Research And Innovation: 7th Australian Symposium, ASSRI 2018, Sydney, NSW, Australia, September 6, 2018, And Wollongong, NSW, Australia, December 14, 2018, Revised Selected Papers*, vol. 7, 2019, pp. 35–49.

[20] K. Yoh, T. Okamoto, H. Inoi, and K. Doi, "Comparative study on foreign drivers' characteristics using traffic violation and accident statistics in japan," *IATSS Research*, vol. 41, pp. 94–105, 2017.

[21] A. Sheykhfard, F. Haghighi, S. Saeidi, M. Safari-Taherkhani, G. Fountas, and S. Das, "Behavioral modeling of drivers near speed control cameras: A dual perspective from micro and macro data," *Transportation Research Record*, p. 03611981241287787, 2024.

[22] S. Saito, Y. Murata, T. Takayama, and N. Sato, "An international driving simulator: Recognizing the sense of a car body by the simulator," in *2012 26th International Conference On Advanced Information Networking And Applications Workshops*, 2012, pp. 254–260.

[23] M. Eskandar, W. Giang, S. Motamedi, H. Devos, L. Koon, A. Akinwuntan, and A. Kondyli, "Designing a reminders system in highly automated vehicles' interfaces for individuals with mild cognitive impairment," *Frontiers In Future Transportation*, vol. 3, p. 854553, 2022.

[24] T.-O. Nævestad, T. Bjørnskau, A. Laiou, R. Phillips, and G. Yannis, "Clash of cultures in greek traffic? what happens when a southern european road safety culture is mixed with a northern european road safety culture?" 2021.

[25] W. J. Conover and R. L. Iman, "Rank transformations as a bridge between parametric and nonparametric statistics," *The American Statistician*, vol. 35, pp. 124–129, 1981.

[26] T. J. van Rensburg, *Dynamic Modelling, Control and Simulation Environment Development for an Eight Wheel Vehicle*. University of Johannesburg (South Africa), 2006.

[27] Y. Ye, S. Wong, Y. C. Li, and K. Choi, "Crossing behaviors of drunk pedestrians unfamiliar with local traffic rules," *Safety science*, vol. 157, p. 105924, 2023.

[28] F. Angioi and M. Bassani, "The implications of situation and route familiarity for driver-pedestrian interaction at uncontrolled mid-block crosswalks," *Transportation research part F: traffic psychology and behaviour*, vol. 90, pp. 287–299, 2022.

[29] D. Glassbrenner and T. Ye, "Seat belt use in 2020–overall results," National Highway Traffic Safety Administration, Washington, DC, Tech. Rep., 2020.

[30] National Highway Traffic Safety Administration, "Seat belt use in 2020: Overall results," National Highway Traffic Safety Administration, Tech. Rep., 2021, available online: https://crashstats.nhtsa.dot.gov (Accessed: 10 February 2025).

[31] S. Lee and A. Al-Mansour, "Development of a new traffic safety education material for the future drivers in the kingdom of saudi arabia," *Journal Of King Saud University-Engineering Sciences*, vol. 32, pp. 19–26, 2020.

[32] A. Gounaridou, E. Siamtanidou, and C. Dimoulas, "A serious game for mediated education on traffic behavior and safety awareness," *Education Sciences*, vol. 11, p. 127, 2021.

[33] F. Laakmann, M. Seyffert, T. Herpich, L. Saupp, S. Ladwig, M. Kugelmeier, and M. Vollrath, "Benefits of tactile warning and alerting of the driver through an active seat belt system," in *27th International Technical Conference On The Enhanced Safety Of Vehicles (ESV)*, vol. 24, 2023.

[34] Transport Roads and Traffic Authority, "A guide to the driving test," Transport Roads and Maritime Services, Tech. Rep., 2010.

[35] M. Martens, "The failure to respond to changes in the road environment: Does road familiarity play a role?" *Transportation Research Part F: Traffic Psychology And Behaviour*, vol. 57, pp. 23–35, 2018.

[36] H. Zhang and S. Lee, "Collecting the information needs of skilled and beginner drivers based on a user mental model for a customized arhud interface," *Science Of Emotion And Sensibility*, vol. 24, pp. 53–68, 2021.

[37] D. Babic, H. Dijanic, L. Jakob, D. Babic, and E. Garcia-Garzon, "Driver eye movements in relation to unfamiliar traffic signs: An eye tracking study," *Applied Ergonomics*, vol. 89, p. 103191, 2020.

[38] S. Salleh, S. Jalil, and I. Ismail, "Road safety game for children: A thematic," Tech. Rep., 2023.

[39] H. Alyamani, N. Alharbi, A. Roboey, and M. Kavakli, "The impact of gamifications and serious games on driving under unfamiliar traffic regulations," *Applied Sciences*, vol. 13, p. 3262, 2023.

Predicting Jobs, Shaping Economies: Bibliometric Insights into AI and Big Data in Workforce Demand Analysis

EL Massi Fouad, ELouadi Abdelmajid
Advanced Systems Engineering-National School of Applied Sciences, IBN Tofail University, Kenitra, Morocco

***Abstract*—The integration of Big Data and Artificial Intelligence (AI) is fundamentally transforming how labor markets are analyzed, predicted and managed. Despite significant advances in using these technologies for workforce analytics, the field suffers from several critical limitations: existing approaches predominantly rely on data from online job portals that may not capture informal employment sectors, current predictive models lack robustness in long-term forecasting under rapid economic transformations and cross-border data integration remains insufficiently addressed for comprehensive global analyses. Moreover, the field lacks a structured, quantitative assessment of scientific production that provides a comprehensive overview of research developments, with most existing studies being case-specific or focusing on narrow applications, leaving significant gaps in understanding the intellectual structure, key contributors and thematic evolution of this interdisciplinary domain. To address these research gaps, this study presents the first comprehensive bibliometric analysis of global scientific research examining the intersection of AI, Big Data and labor market prediction. Drawing on a systematic dataset of 276 publications from Scopus, Web of Science and OpenAlex databases spanning 2003 to 2025, this research employs advanced bibliometric techniques to map the intellectual landscape of this rapidly evolving field. Through a structured four-phase methodological framework incorporating performance analysis, science mapping and thematic evolution, the study identifies research trends, intellectual structures, influential contributors and emerging themes. The analysis reveals significant developments in predictive modeling, natural language processing, and hybrid AI approaches for recruitment forecasting and workforce analytics, while highlighting critical challenges posed by algorithmic bias and ethical considerations in AI-driven systems. Key contributions include: 1) the first systematic scientific mapping of the AI-Big Data-labor market intersection, 2) identification of research gaps and future directions for long-term labor market prediction, 3) comprehensive analysis of institutional networks and collaborative patterns and 4) evidence-based recommendations for addressing data integration and model interpretability challenges. The findings offer actionable insights for researchers, policymakers and practitioners seeking to leverage intelligent systems to shape the future of work in the digital economy while addressing current methodological limitations.**

***Keywords*—Big data; Artificial Intelligence; predictive modeling; bibliometric analysis; natural language processing; labor market analytics**

I. Introduction

In recent years, the convergence of Big Data and Artificial Intelligence (AI) has significantly reshaped the analysis and understanding of labor markets. The growing availability of large-scale employment data, together with advancements in predictive modeling, has enabled new approaches to forecasting labor demand and supply, monitoring workforce dynamics and informing evidence-based labor market policies. As economies transition to more digital and knowledge-intensive models, the ability to derive actionable insights from complex labor datasets has become a strategic imperative for researchers, employers, and public authorities.

Despite the growing number of studies that apply these techniques to hiring forecasts, the field lacks a comprehensive and structured quantitative analysis of the scientific output necessary for extensive reviews. Most existing research remains case-specific or narrowly focused, leaving significant gaps in understanding the intellectual structure, key contributors and thematic development of the discipline. To address this gap, the present article offers a bibliometric analysis of the scientific literature on the application of Big Data and Artificial Intelligence to labor market prediction. The study examines publications from three major scientific databases, Scopus, Web of Science and OpenAlex, covering the period 2003 to 2025.

The analysis is organized into several sections to ensure a comprehensive and coherent exploration of the field. Section II provides a synthesis of related studies and prior reviews addressing the intersection of Big Data, Artificial Intelligence and labor market forecasting. Section III details the methodological framework employed for this bibliometric study, including the selection criteria, data collection procedures and analytical tools such as Bibliometrix and VOSviewer. Section IV presents the results and key findings, focusing on publication trends, networks of authors and institutions, patterns in keywords and phenomena and the identification of thematic clusters. Section V offers a critical discussion of the observed patterns, highlights the research landscape and proposes directions for future investigation. The article concludes with a final section summarizing the main contributions and discussing the broader implications for research and policy in labor market analysis. Through this structured approach, the study aims to provide an extensive scientific mapping of this interdisciplinary domain and to offer meaningful insights relevant to both academic and policy-oriented debates concerning the future of labor market analysis.

II. Related Work

The convergence of Big Data Analytics, Artificial Intelligence (AI) and Predictive Modeling have made a paradigm change in the way of analyzing, understanding and forecasting

labor market dynamics to researchers. Over the past decade, this interdisciplinary domain has expanded rapidly as organizations and governments seek to harness large scale datasets and advanced computational methods to inform workforce planning, skills development and employment policy.

Labor market surveys refer to the collection and analysis of big data posts, social media, professional networks, government statistics and company registers, collection and analysis of large versions of unarmed data from different sources. These data sets have valuable information on employment patterns, skills requirements, salaries and career paths that were previously difficult to capture through traditional surveys on the labor market.

Artificial intelligence, especially machine learning algorithms, enables researchers to treat these mass datasets effectively, recognize complex patterns and produce insights that would be impossible through manual analysis. This AI-operated equipment can identify new job categories, predict future skills requirements, analyze regional employment beaches and even detect prejudices in practice as hiring. The future analysis implements statistical techniques, machine learning algorithms and data mining to analyze current and historical data to provide predictions on future labor market conditions. In human resources and work plan, these predictions organizations help to estimate the lack of talent, identify high -capacity candidates, reduce sales and optimize recruitment strategies.

The growing research that examines the intersection of these technologies with labor market analysis is documented through various systematic reviews, text list analysis and literature reviews. These research synthesis approaches provide a valuable framework to understand how these technological innovations change employment, skills requirements and the development of careers in the digital economy.

A. Big Data Application in Labor Market Analysis

The rise of large data has revolutionized the labor market analysis by providing unique access to granular, multidimensional information that was not available through traditional research methods earlier. Especially online job portals have become valuable data sources to understand the labor market's dynamics, skills requirements and employment beaches. Like Mrazova et al. Expelled in its analysis of the Indian job portal Babjob, Online Work Portal Data can contribute to labor market policy in five main areas: Monitoring of the labor market, the workforce evaluation, evaluation of job-sex behavior, prognosis analysis of skills and experimental studies [1].

The ability to collect, treat and analyze on a large scale from online sources provides significant benefits on traditional labor market survey methods. Vankevich et al. Note that the methods of traditional labor market analysis often depend on chronic information collection technologies and fail correctly to the account correctly for specific skills in CVS and vacancies. They propose function based on large computer technologies and Artificial Intelligence, which is to quickly collect, process and imagine data on profit requirements in businesses, fields and financial fields [2].

A particularly powerful application is the use of natural language processing (NLP) to start job postings and restart on scale. Researchers have used NLP methods to identify changes in different businesses and industries. Nicolev emphasized how calculation language and lesson analysis equipment enable researchers to analyze labor changes at the level of personal competencies by checking the text of online positions [3]. This approach allows for real-time monitoring of vacancies across regions and countries, predicting demand for specific skills and comparing labor markets internationally.

Some researchers have leveraged Big Data to develop advanced recommendation systems that more effectively match job seekers with suitable employment opportunities. For example, Qin et al. designed neural network models to evaluate the degree of fit between job seekers and job requirements, thereby reducing reliance on manual screening by human resource professionals [3][4]. Similarly, Zhou et al. constructed models to assess the popularity and relevance of job skills using large-scale recruitment data, helping to identify which competencies are most strongly associated with high-paying employment opportunities [3][5].

Advanced machine learning techniques enable researchers to detect complex patterns in labor market data. For instance, Alakl et al. applied machine learning methods to analyze 35,000 article abstracts and 57,000 LinkedIn posts, identifying 28 variables grouped into five macro-level categories related to multigenerational labor markets [6]. Similarly, Rai et al. conducted a comparative analysis to evaluate the effectiveness of different models—including GPT-3 and BERT—for analyzing Artificial Intelligence methods, particularly natural language processing (NLP), in the context of regional labor markets [7].

Big Data–driven labor market monitoring has significant economic implications, particularly in terms of enhancing analytical integration. Zaloznova et al. demonstrated that Big Data analytics offer valuable labor market insights that improve competitiveness and enable the creation of online job exchanges analogous to financial stock markets [8]. These digital platforms facilitate real-time monitoring of labor market dynamics and evolving skill demands across various local and regional contexts.

These innovations are particularly valuable for tracking the rapid transformations driven by digitization and technological advancement. Cymbal et al. emphasize that traditional survey-based sources of labor market information are insufficient for capturing the profound changes in the content and nature of work resulting from digitalization [9]. Their study compiles international experiences with the use of digital technologies and Big Data in labor market monitoring, concluding that online job vacancy data significantly complement and enhance traditional labor market statistics and analyses.

The introduction of these technologies has not only transformed the analysis of labor markets, but also reshaped the recruitment process. As Spadina notes, the integration of Artificial Intelligence into recruitment has fundamentally altered the landscape, where Big Data, algorithms, and predictive outcomes now play an active role in shaping the relationship between humans and technology in hiring practices [10].

B. Limitations of Existing Methods and Contribution of the Article

Despite significant advances in using Big Data, Artificial Intelligence (AI), and predictive analytics for labor market analysis, several limitations persist in the current state of the art. First, many existing approaches are heavily based on data from online job portals and social media, which may not fully capture informal employment sectors or underrepresented populations, leading to potential biases and gaps in labor market coverage. Second, while AI-driven models excel at processing large-scale unstructured data, challenges remain in ensuring model interpretability and transparency, which are critical for policy-making and ethical considerations (5). Third, most predictive models focus on short- to medium-term forecasting and often lack robustness in long-term labor market trend prediction, especially under rapid technological and economic transformations (57). Furthermore, cross-border data integration and standardization are not adequately addressed, limiting comprehensive global labor market analyses.

Furthermore, existing methods frequently overlook the integration of multi-source data that combine quantitative labor market indicators with qualitative contextual information, such as policy changes or regional economic shifts, which are essential for nuanced understanding and actionable insights. There is also a scarcity of frameworks that systematically incorporate ethical concerns related to data privacy, algorithmic bias, and equitable representation in AI applications within labor market analytics.

This article contributes to reducing these gaps by proposing a comprehensive framework that integrates diverse data sources, including underutilized datasets, to enhance labor market coverage and reduce bias. It emphasize the development of interpretable AI models that balance predictive accuracy with transparency, facilitating trust and adoption by policymakers and stakeholders. Moreover, the adopted approach extends predictive capabilities to better capture long-term labor market dynamics by incorporating adaptive machine learning techniques sensitive to economic and technological disruptions. The article also address data standardization challenges by proposing harmonized protocols for multi-jurisdictional labor data integration. Finally, it incorporate ethical guidelines into the design and deployment of AI tools to ensure responsible use, privacy protection, and fairness in labor market analyses.

By addressing these limitations, this work advances the field toward more inclusive, transparent, and effective labor market analytics, supporting informed decision-making in workforce planning, skills development, and employment policy in the digital economy.

III. Methodology

The purpose of this bibliometric study is to map and analyze the scientific landscape related to the use of large data and Artificial Intelligence (AI) for the labor market's prediction and workforce analysis. The functioning adopted follows a structured and reproducible workflow, which is grounded (preferred reporting elements for systematic reviews and meta-analysis) in Prisma (Fig. 1).

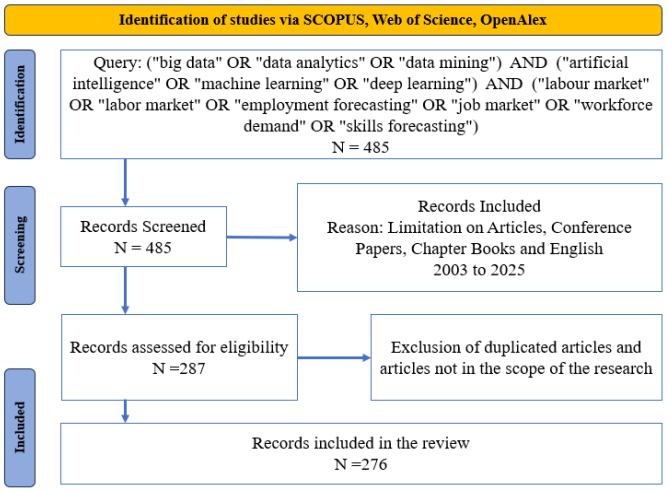


Fig. 1. Prisma approach.

A. Data Source and Query Construction

The data collection was performed using three large academic database: Scopus, Web of Science (WOS) and OpenAlex, which provides extensive coverage of colleagues underwent by colleagues. A carefully designed search was implemented to reconstruct relevant publications, including conditions related to Big Data, AI techniques and labor market forecasts:

Query: ("big data" OR "data analytics" OR "data mining") AND ("Artificial Intelligence" OR "machine learning" OR "deep learning") AND ("labour market" OR "labor market" OR "employment forecasting" OR "job market" OR "workforce demand" OR "skills forecasting").

This query targeted title, essence and keyword field. The discovery was limited to documents published between 2003 and 2025, and objects written in English were only assessed. Document types include journal articles, conference letters and book chapters.

B. Inclusion and Exclusion Criteria

Originally, a total of 485 items were obtained. The inclusion criteria were as follows:

- Publication related to the intersection of Big Data/AI and the labor market forecast.
- Documents in English.
- Publication Type: Articles, conference articles and book chapters.
- Time limit: 2003-2025.

In the screening phase, 198 documents were excluded due to irrelevance due to the extent of research, repetition or incorporation criteria. The move was considered as a result of 287 items for evaluation of qualifying.

C. Bibliometric Equipment and Analytical Techniques

The selected data was exported to BIBTEX and CSV formats for further procedure. Two large bombers used tools used:

- Bibliometrix R-Package (through the Biblioshiny interface) for publishing development, co-author, Country Cooperation, Citation Matrix and quantitative analysis of keywords CO-event).
- VOSviewer for network visualization and thematic mapping, which includes cohabitation networks, author groups and fellow phenomenon maps.

These units enabled a multi-phase exploration of intellectual and ideological structures in the region.

IV. Results and Analysis

A. Descriptive Bibliometric Overview

Fig. 2 presents a comprehensive overview of key bibliometric indicators for a total of 276 documents published between 2003 and 2025 across various sources. The dataset includes contributions from 560 authors, among whom 81 were sole authors. An annual scientific growth rate of 14.84% indicates a rapidly expanding research interest in the intersection of Big Data, AI, and labor market analysis.

The relatively low rate of international co-authorship (9.03%) and an average of 2.21 authors per document suggest that research collaborations are predominantly national or confined within institutions. A total of 2,885 references were cited across the corpus, reflecting a strong theoretical and conceptual foundation in the field.

The average citation age per document (5.73 years) and the average document age (3.24 years) further indicate that the topic is both emergent and dynamic, positioning itself as a growing area of academic inquiry.



Fig. 2. Bibliometric overview.

B. Temporal Evolution of Scientific Production

As shown in Fig. 3, scientific output in this domain remained minimal between 2003 and 2015, followed by a steady increase beginning in 2016. The trend culminated in a notable peak in 2024, with over 65 publications. This upward trajectory reflects the growing influence of the digital economy and the increasing importance of data-driven decision-making processes in workforce management and labor market analysis. The apparent decline in 2025 is likely attributable to the incomplete publication window for the year, rather than an actual reduction in research productivity.

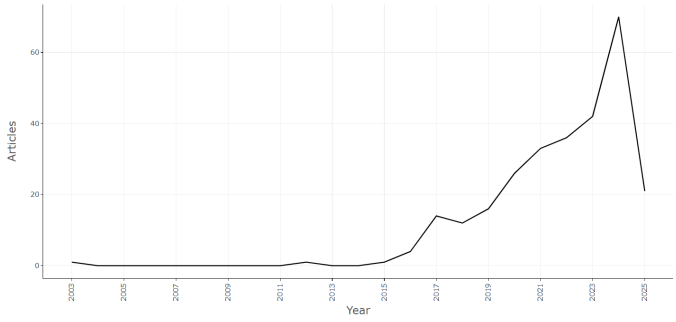


Fig. 3. Temporal evolution of scientific production.

C. Citation Dynamics and Impact

Fig. 4 illustrates the average number of citations per year. Around 2015, the citation curve shows a notable peak with high variability, followed by a gradual decline in subsequent years. This decrease can be attributed to the recency of many documents, which have not yet had sufficient time to accumulate citations.

Overall, the trend reflects an area that is expanded in volume, but which is still in accordance with academic effects.

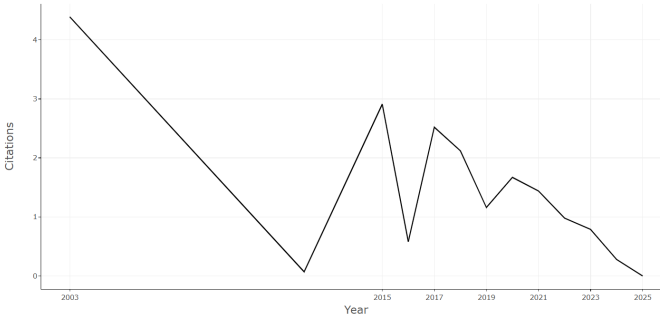


Fig. 4. Citation dynamics and impact.

D. Mapping Authors, Sources, and Thematic Domains

Fig. 5 (three-region plot) reflects mutual action between major authors, publishing sources and research domains. Christin L. Authors such as Wood, Quan Hua Lim and Janses are closely associated with influential magazines such as Emerald Publishing, SSRN and Rxiv, while thematic fields expand economics, informatics, computer science, law and political science.

It confirms the visual field field's multi-discipline, where AI and Big Data intelligence with both technical and social sciences regarding the prediction of the labor market.

E. Core Publication Outlets

Fig. 6 identifies the most active publication withdrawal. Emerald Publishing Limited ebooks proceed with 38 documents, followed by people to Jobs (15) and SSRN Electronic Journal (13).

This distribution indicates a hybrid publication environment, which combines colleague-rated sources with preproofs

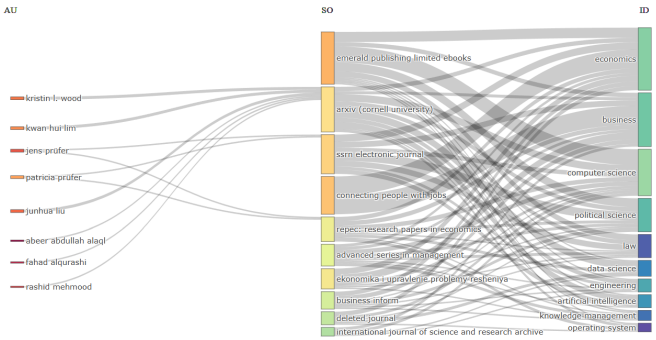


Fig. 5. Mapping authors, sources, and thematic domains.

and policy-centered outlets, which enables widespread spread in academia and decision institutions.

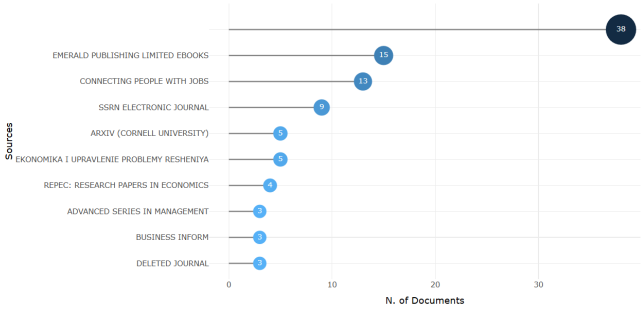


Fig. 6. Core publication outlets.

F. Evolution of Source Contributions

This Fig. 7 shows how the most relevant magazines in the area over time have contributed. Emerald Publishing Limited E -Books has conducted continuous and early involvement, reflecting a steady growth from 2017. Conversely, people with jobs later joined the field, but quickly, trapped by 2023.

The emergence of economica in upravlenie and Arxiv for many last years reflects a change to various forms (open depot and regional sources), reflecting expansion inclusion and interdisciplinary in the region.

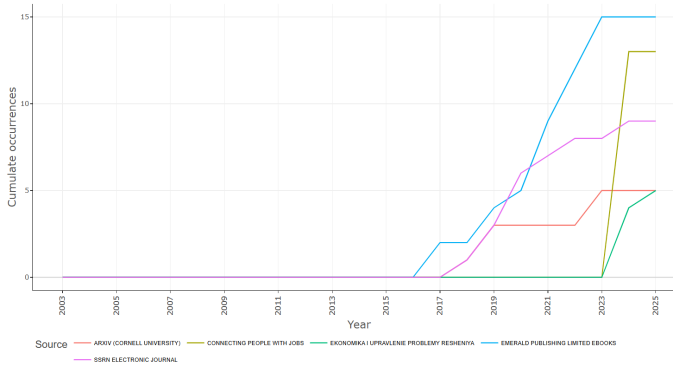


Fig. 7. Evolution of Source Contributions

G. Key Author Contributions

This rod map, Fig. 8, identifies the most powerful authors of the region, with Abir Abdullah Alakl, Alina Vankevich, Fahd Alkurashi, Irina Kalinsakaya and Rashid Mehmud 4 - Publications.

Such repeated contributions reflect the emergence of special researchers dedicated to bridge AI, Big Data and Work Economics. This elite group can serve as a main education network that governs the thematic and methodological development of the region.

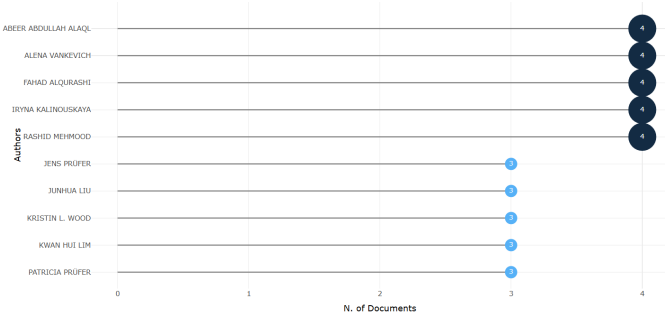


Fig. 8. Key author contributions.

H. Temporal Distribution of Author Productivity

This plot, Fig. 9, highlights the temporary research activity of the top authors. Janses and Petricia reflects quick and focused contributions around 2019-2021, while Rashid Mahmood, Alina Vankevich and Fahd Alkurashi recently showed more continuous activity, reflecting the current speed of the workforce.

The diagram showed both basic contributors and new ideas, and reflects the genetic layering of expertise in the field.

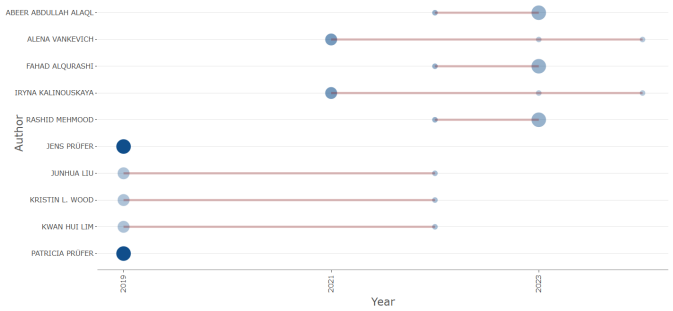


Fig. 9. Temporal distribution of author productivity.

I. Institutional Leadership in the Field

This Fig. 10 refers to the most productive institutions. The University of Texas in San Antonio is accompanied by 12 articles, followed by King Abdulziz University and Singapore University and Design (both with 9). These institutions represent a mix of America-based, the Middle East and Asian universities, highlighting global and multidisciplinary commitment to the forecast of the AI-controlled labor market. The presence of both technical and social science oriented universities confirms double anchoring the field.

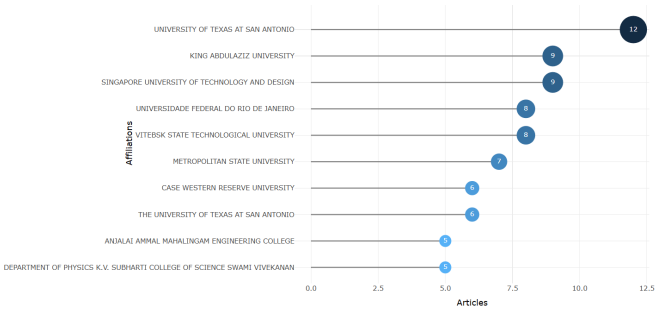


Fig. 10. Institutional leadership in the field.

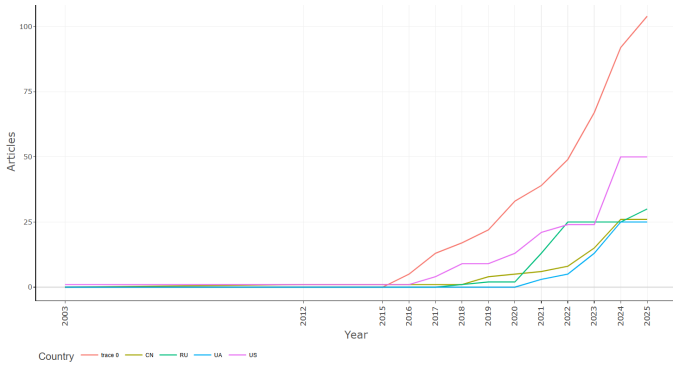


Fig. 12. Evolution of national scientific production.

J. Country-Level Scientific Output

At the country level, Fig. 11 shows that China leads in author productivity, followed by the United States, Russia, and the United Kingdom. Notably, the majority of publications are classified as Single Country Publications (SCP), indicating relatively low levels of international collaboration—an observation consistent with the metadata findings.

This trend may reflect national labor policy priorities and potential barriers related to data protection and privacy. However, it also highlights significant opportunities to foster international co-authorship and enhance global knowledge exchange in the field.

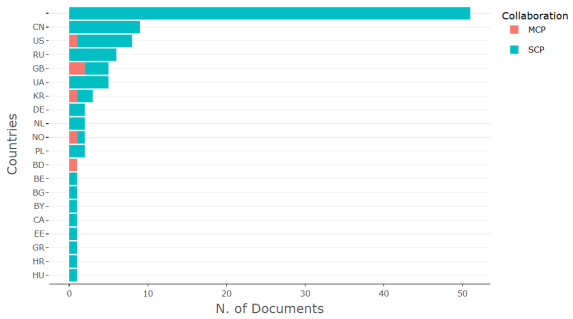


Fig. 11. Country-level scientific output.

K. Evolution of National Scientific Production

Fig. 12 reveals strong growth in publications from China (CN) since 2016, which now leads the domain. The United States (US) shows steady growth, while Russia (RU), Ukraine (UA) and Canada have shown more less, but increasing contributions in recent years.

This upward trend reflects the growing national investments and large computer strategies related to work forecasts in AI. Various contributors indicate that the area is globalization, although it is with regional intensity centers.

L. Citation Impact by Country

The United States leads in citation impact with a total of 518 citations, followed by the Netherlands (195) and Canada (97). Interestingly, China, despite its high research output, ranks relatively low in citations.

This suggests that its publications are either more recent or receive comparatively less international attention. This emphasizes a difference between the opposite volume and the effect, and provides an opportunity to increase the visibility and collaboration for non-Western research production (Fig. 13).

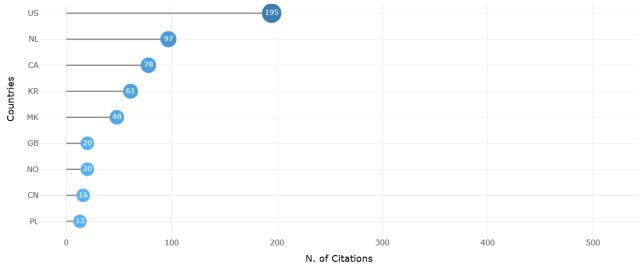


Fig. 13. Citation impact by country.

M. Conceptual Structure: Most Frequent Keywords

The most recurring concepts are “Computer science” (194 events), “Economics” (134) and “Business” (128), which confirms the interdisciplinary convergence between technical and socio-economic domains.

“Political Science”, “Computer Science” and “Artificial Intelligence” follows closely, and shows that this research sector is equally shaped by politics, analysis and calculation appliances (Fig. 14).

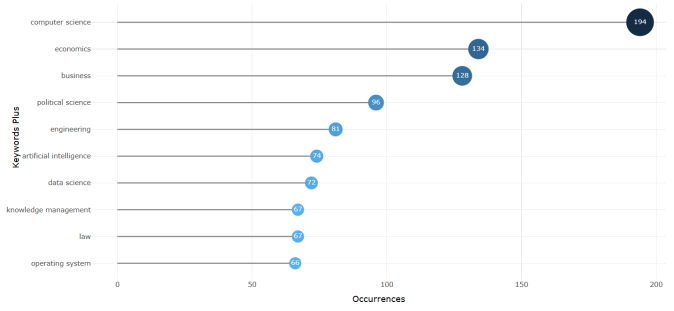


Fig. 14. Conceptual structure: most frequent keywords.

N. Word Cloud Representation

The word cloud, Fig. 15 provides a qualitative snapshot of main themes. "Computer Science" is visually dominated, "Economics", "AI", "Data Science" and "Business". Emerging subjects such as "working economics", "knowledge management" and "unemployment" highlight field-specific relevance and policy-related dimensions.



Fig. 15. Word cloud representation.

O. Evolution of Keyword Trends Over Time

This longitudinal analysis, Fig. 16, shows that terms like "computer science", "data science", and "economics" have seen steep growth post-2017, while "AI", "political science", and "engineering" remain strong and steadily rising.

The figure reflects a broadening scope and increased specialization of the research field over the last five years.

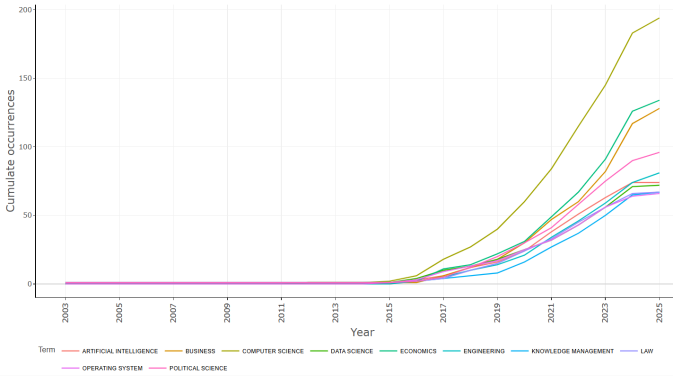


Fig. 16. Evolution of keyword trends over time.

P. Trend Topics Timeline

This chart, Fig. 17 highlights how certain themes emerged and evolved:

- "Labour economics" and "unemployment" dominate post-2022
- "Computer science" and "economic growth" peaked in 2020–2021
- Niche areas like "programming language" or "finance" appear between 2018 and 2020

These trends suggest a recent shift toward labor policy modeling, macroeconomic forecasting, and employment analytics in response to post-pandemic dynamics.

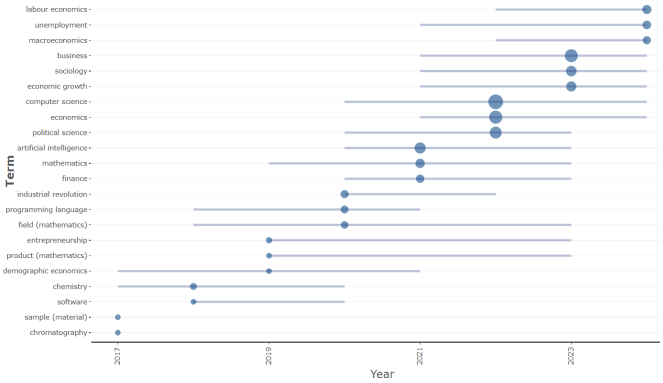


Fig. 17. Trend topics timeline.

Q. Thematic Clustering by Coupling

This strategic diagram, Fig. 18, identifies four thematic zones:

- Motor cluster: "computer science", "economics", and "business" with strong impact and centrality.
- Basic themes: "AI", "data science", and "big data" form the intellectual foundation.
- Peripheral clusters: "psychology", "biology", and "active labor market policies" show low centrality but growing relevance.
- This visualization confirms the core structure and emergent peripheries of the field.



Fig. 18. Thematic clustering by coupling.

R. Co-occurrence Network of Keywords

This network map, Fig. 19, highlights dense connections between "computer science", "economics", and "business", suggesting a tight intellectual core.

Peripheral yet significant clusters include "labor economics", "AI", "knowledge management", and "operating system", showing how technological tools and economic policies intersect in workforce analytics.

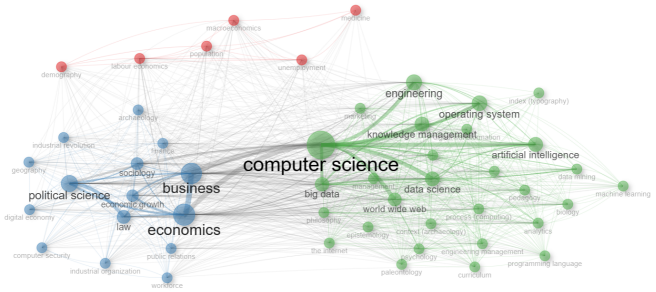


Fig. 19. Co-occurrence network of keywords.

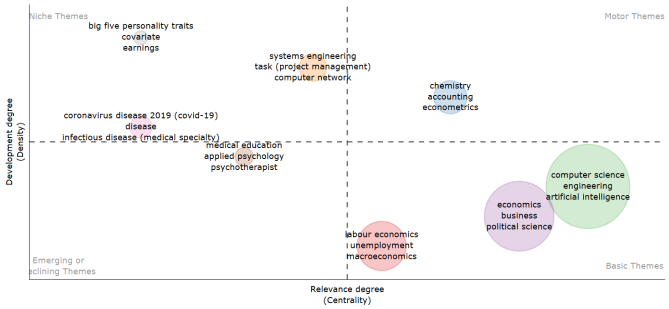


Fig. 20. Thematic map.

S. Thematic Map

This map, Fig. 20 shows:

- Basic themes: “computer science”, “engineering”, “AI” → foundational and well-established.
- Emerging/Declining: “labor economics”, “macroeconomics” → growing but still under-structured.
- Niche themes: “psychology”, “COVID-19” → specialized but with less general impact.
- This helps frame future research directions and gaps in labor analytics.

T. Factorial Analysis of Conceptual Structure

This factorial map, Fig. 21 shows three main conceptual dimensions:

- A dominant cluster (red) linking “computer science”, “economics”, and “digital transformation”.
- A green cluster focused on “labour economics”, “population”, and “unemployment”.
- A blue cluster with “machine learning”, “psychology”, and “typography”.

These axes confirm the coexistence of technical, economic, and human-centered subfields, reinforcing the interdisciplinary fabric of AI and big data applications in labor market prediction.

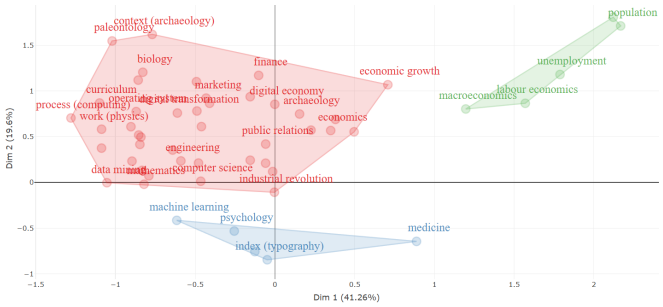


Fig. 21. Factorial analysis of conceptual structure.

U. Collaboration Network

The cooperation network, Fig. 22, reveals a fragmented but practical map of large data applied to international co-author and labor market analysis within AI in Fig. 22. The United States (USA) appears the most connected and central hubs to cooperate with countries such as Singapore (SG), Bangladesh (BD).

Other regional groups are visible:

- A strong Europe -Asia Bridge (e.g. UK - Italy - Pakistan).
- Eastern European alliances such as Russia - Ukraine - Bulgaria.
- Isolated regional partnership (e.g. Brazil - Portugal, France - Netherlands, Australia - New Zealand).

However, the lack of dense intercontinental cooperation is clear. Many clusters are insular, showing limited integration. This indicates the need for extensive international research participation, especially between high outputs and emerging areas.

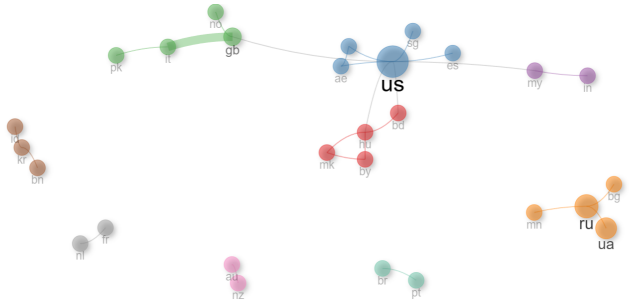


Fig. 22. Cooperation network.

V. Countries Collaboration World Map

The global heatmap reinforces the dominant role of the United States, China, Russia, and Western Europe as key actors in international research collaboration. However, the intensity of cooperation is noticeably higher in North America and Europe, while parts of Asia, and especially regions in Africa and Latin America, exhibit lighter or neutral tones—indicating limited global engagement.

As illustrated in Fig. 23, this pattern suggests that, although the field holds global significance, the infrastructure for international cooperation remains uneven. Enhancing South–South and North–South collaboration could broaden the scientific discourse and ensure that labor market analysis incorporates insights from diverse socio-economic and regional contexts, thereby enriching both policy and academic relevance.

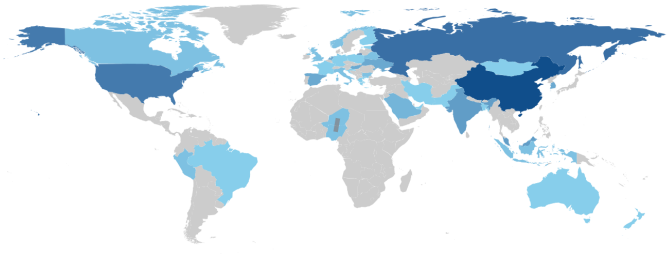


Fig. 23. Countries' collaboration world map.

V. Discussion

A. Bibliometric Discussion

Results of collaborative analysis expose both forces and significant intervals in the global research scenario on AI and large data for labor market forecast. The United States stands as the most active and well-linked country, and serves as a central center within a broad cooperation network consisting of partners of Europe, Asia and the Middle East. Other countries, such as Russia, China and the United Kingdom, also contribute continuously, but their networks are often more territory concentrated.

Despite these active nodes, the general structure of the collaboration network is fragmented and works in small, insulated clusters or silos with many countries. This fragmentation is more depicted in the co-writing world map, where many regions, especially in Africa, Latin America and parts of Southeast Asia, are smaller or disconnected than the global research discourse.

Such asymmetry is a challenge for the development of globally used and the development of solutions. Since the challenges in the labor market are quickly complicated and interconnected - driven by digitalisation, automation and demographic changes - it is necessary to promote broad international and interdisciplinary collaborations. Encouraging the South South and North South partnership will not only strengthen the prosperity of research, but will also ensure that the conclusions are inclusive and relevant in different socio-economic contexts.

In summary, while scientific production in the region is growing rapidly, the cooperative ecosystem remains uneven. It is important to address this imbalance to strengthen the basis for global knowledge and support the development of justified and informed labor market policy through AI and data-driven strategies.

B. Research Trends and Future Directions

The development of research at the intersection of big data, Artificial Intelligence and labor market analysis reveals many different trends that indicate to promote future directions. A remarkable tendency is to change vocabulary and focus in the region. Over the past decade (2010–2019), the traditional "Business Intelligence" gradually decreased in use as "analysis", "Artificial Intelligence", "big data" and "machine learning", both prominent academic literature and professional applications [11]. This linguistic change reflects the intensive change that contacts the organization that makes production decisions, more sophisticated than descriptive analysis leads to a sophisticated future and leads to the prescription feature.

A new research direction focuses on the integration of several technical approaches to create hybrid models with increased abilities. Researchers quickly detect the CO - intelligence ability to combine Artificial Intelligence with large data analysis, provided that when these technologies work together, they create abilities can either become independent [12]. For example, recent research in the financial markets emphasizes the important ability to limit the Hybrid AI model that integrates large data analyzes for more accurate stock market paves [13]. These integrated approaches represent a promising opportunity for future research in labor market analysis, where complex economic and social factors interact in ways that no analytical approach can capture perfectly.

Application landscape for AI and Big Data Analytics continues to expand to different fields. In the food industry, these techniques have achieved remarkable results in food matching, safety monitoring and testing processes [12] [14]. This application of these technologies shows their versatility and suggests the opportunity to expand further in special labor market analysis for different industries. For small and medium -sized companies (SME), research indicates that digital changes increase innovation and dynamic abilities, helping these businesses to overcome the lack of resources and provide better feedback to market changes. Large technologies that run these reforms include big data, Artificial Intelligence and digital platform [15]. This discovery indicates a growing research direction how small organizations and labor market participants can use these advanced technologies despite having less resources than large companies. Extensive research on digital economy suggests how Big Data Analytics enables organizations to gain valuable insight into marketing strategies and operating processes. Artificial intelligence applications are expanded in customer service, scam detection and product development, leading to a complex ecosystem of technological changes that originally resume labor markets [16]. Future research must find that these techniques can be distributed in a responsible manner to encourage economic growth by reducing negative effects on employment. As these research trends evolve, several future directions can be anticipated: more refined integration of different data sources, the development of specific accurate future models for labor market dynamics, focus on more focus on moral ideas and prejudices mitigating in AI-implemented recruitment systems, and to address the workforce in the development of economies. Research also suggested the upcoming convergence of technologies where big data, Artificial Intelligence, machine learning and blockchain can work in concerts to solve the challenges in complex labor market.

VI. Conclusion

This bibliometric study provides a comprehensive exploration of scientific production on the application of Big Data and Artificial Intelligence for labor market forecasting. By analyzing 277 documents published between 2003 and 2025 and indexed in Scopus, Web of Science, and OpenAlex, the research offers valuable insights into publication trends, key contributors, thematic evolution, and collaboration patterns in this rapidly emerging field.

The analysis reveals a remarkable acceleration in scientific production, particularly over the last five years, reflecting the growing importance of data-driven employment policies and the increasing relevance of predictive analytics to support workforce planning strategies. Disciplines such as informatics, economics, business, and computer science emerge as prominent throughout the literature, confirming the interdisciplinary nature of the research domain. Additionally, the increasing emphasis on topics such as labor economics, unemployment, digital transformation, and AI-based recruitment tools reflects a shift toward policy-oriented investigation.

While the thematic structure of the field demonstrates maturity in some areas, such as job-matching algorithms and digital workforce analysis, other domains remain underdeveloped or emerging, particularly those related to ethics, inclusivity, and regional specificity. Furthermore, the collaboration analysis highlights a fragmented global research network with high productivity concentrated in countries such as the United States and China, yet limited cross-regional partnerships, especially with African and Latin American institutions. These patterns indicate the need for more inclusive, equitable, and interdisciplinary cooperation to fully harness AI and Big Data for addressing global labor market challenges.

A. Study Limitations

Despite its comprehensive scope, this study acknowledges several important limitations that may affect the generalizability of findings. First, the analysis is restricted to publications indexed in three major databases (Scopus, Web of Science, and OpenAlex) and limited to English-language documents only. This methodological constraint may introduce language and database bias, potentially excluding relevant research published in other languages or indexed in regional databases, particularly from non-Western academic traditions where significant AI and labor market research may be conducted in local languages.

Second, the study employs a purely quantitative bibliometric approach focused on publication patterns, citation networks, and keyword co-occurrence analysis. While this methodology provides valuable insights into the structural and temporal dynamics of the research field, it does not capture the qualitative dimensions of research quality, methodological rigor, or practical applicability of the studied approaches. The absence of content analysis limits the understanding of the actual effectiveness, implementation challenges, and real-world impact of the AI and Big Data solutions proposed in the literature.

B. Future Research Directions

Building on the findings and limitations identified in this study, several promising avenues for future research emerge. First, there is a critical need for comprehensive qualitative meta-analyses that examine the methodological approaches, implementation frameworks, and practical outcomes reported in the most influential studies. Such research should focus on identifying best practices, common challenges, and success factors in deploying AI and Big Data solutions for labor market prediction across different economic and cultural contexts.

Second, future investigations should prioritize the development of international collaborative frameworks that bridge the identified research gaps between high-productivity regions and underrepresented areas, particularly in the Global South. This could involve establishing international research consortiums, cross-regional data sharing protocols, and collaborative platforms that enable knowledge transfer and capacity building in AI-driven labor market analysis.

Finally, given the ethical concerns identified in this analysis, future research should systematically investigate bias mitigation strategies, algorithmic fairness, and the socio-economic implications of AI-driven labor market interventions. This research agenda should include longitudinal studies that track the long-term impacts of these technologies on employment patterns, wage inequality, and workforce development across diverse demographic groups and economic sectors.

In conclusion, this study maps the intellectual landscape of a rapidly expanding research domain and identifies both its strengths and its gaps. It lays the groundwork for future investigations that should aim to bridge disciplinary silos, enhance international cooperation, and develop robust, ethical, and globally relevant models for predictive labour market analysis. Such efforts are essential to ensuring that intelligent technologies contribute meaningfully to sustainable, inclusive, and data-informed workforce development.

References

[1] M. Mrázová and J. P. Neary, "IO for Export(s)," 2017.

[2] A. Vankevich, I. Kalinouskaya, O. Zaitseva, and A. Korabava, "Equilibrium of Labor Market: New Security Instruments in the Context of Digitalization," in *SHS Web of Conferences*, 2021.

[3] I. Nikolaev, "An intelligent method for generating a list of job profile requirements based on neural network language models using ESCO taxonomy and online job corpus," *Business Informatics*, 2023.

[4] C. Qin, H. Zhu, T. Xu, C. Zhu, L. Jiang, E. Chen *et al.*, "Enhancing Person-Job Fit for Talent Recruitment: An Ability-aware Neural Network Approach," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018.

[5] T. Xu, H. Zhu, C. Zhu, P. Li, and H. Xiong, "Measuring the Popularity of Job Skills in Recruitment Market: A Multi-Criteria Approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[6] A. A. Alaql, F. Alqurashi, and R. Mehmood, "Multi-generational labour markets: Data-driven discovery of multi-perspective system parameters using machine learning," *Science in Progress*, 2023.

[7] D. Rae, E. Cartwright, M. Gongora, C. M. Hobson, and H. M. Shah, "'Insight unlocked': Applying a collective intelligence approach to engage employers in informing local skills improvement planning," *Industry & Higher Education*, 2023.

[8] Y. Zaloznova, O. Pankova, and Y. Ostafiichuk, "Global and Ukrainian Labour Markets in the Face of Digitalization Challenges and the Threats of the COVID-19 Pandemic," 2020.

[9] O. Cymbal, Y. Ostafiychuk, and O. Pankova, "Digital Methods and Technologies of Forming an Innovative Labour Market Information Support System in Ukraine," *Journal of European Economy*, 2023.

[10] H. Špadina, "Legal aspects of artificial intelligence in the employment process," *Stanovništvo*, 2023.

[11] A. Prasanth, D. J. Vadakkan, P. Surendran, and B. Thomas, "Role of Artificial Intelligence and Business Decision Making," *International Journal of Advanced Computer Science and Applications*, 2023.

[12] H. Ding, J. Tian, W. Yu, D. I. Wilson, B. R. Young, X. Cui *et al.*, "The Application of Artificial Intelligence and Big Data in the Food Industry," *Foods*, 2023.

[13] A. Al-Alawi and N. Alshakhoori, "Stock Price Prediction Using Artificial Intelligence: A Literature Review," in *Proceedings of the 2024 ASU International Conference on Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, 2024.

[14] Q.-Y. Zhou, H. Zhang, and S. Wang, "Artificial intelligence, big data, and blockchain in food safety," *International Journal of Food Engineering*, 2021.

[15] H. Lu and M. S. Shaharudin, "Role of digital transformation for sustainable competitive advantage of SMEs: a systematic literature review," *Cogent Business & Management*, 2024.

[16] U. Detthamrong, L. T. Nguyen, Y. Jaroenruen, A. Takhom, V. Chaichuay, and K. Chotchantarakun *et al.*, "Topic Modeling Analytics of Digital Economy Research: Trends and Insights," *Journal of Scientometric Research*, 2024.

Dynamic Polygon-Based Reverse Driving Detection Technique for Enhanced Road Safety

Tara Kit[1], Youngsun Han[2], Anand Nayyar[3], Tae-Kyung Kim[4]*
Department of AI Convergence, Pukyong National University, Busan, South Korea[1,2]
School of Computer Science, Duy Tan University, Da Nang 550000, Vietnam[3]
Department of Management Information Systems, Chungbuk National University,
Seowon-Gu, 28644, Cheongju, South Korea[4]

***Abstract*—Reverse driving and lane collapse pose serious risks to road safety, especially on complex infrastructures such as multi-lane highways, intersections, and roundabouts. Existing detection systems often depend on rigid lane configurations and struggle to adapt to varied road geometries and environmental conditions. Prior works are typically limited to straight, multi-lane roads and rely on automated boundary extraction, making them unsuitable for irregular traffic layouts. To address this gap, the objectives of this research paper is to propose a vision-based detection system that combines the YOLOv8 object detector with a dynamic polygon-based zone management strategy. The system aims to detect reverse driving and lane collapse incidents in real time using CCTV footage, without requiring additional sensors. Its key novelty lies in manually configurable zones and the integration of ByteTrack for robust vehicle tracking across complex scenes. The system was tested under diverse real-world parameters, including different road types (single-lane, multi-lane, roundabouts), lighting conditions (day and night), and traffic behaviors (normal flow, reverse, and collapse) and visual evaluations highlight consistent and logically coherent results across scenarios, highlighting its practical effectiveness for real-time intelligent traffic monitoring.**

***Keywords*—Reverse driving detection; lane collapse detection; polygon zones; object detection; YOLOv8**

I. Introduction

Preventing traffic accidents and maintaining efficient traffic flow are critical challenges of modern transportation infrastructure [1]. Although wrong-way driving incidents account for only 3% of global traffic accidents, they result in a disproportionate 22% of traffic-related fatalities [2], highlighting their severe impact. This underscores the urgent need for effective detection and prevention systems to mitigate the risks associated with dangerous driving behaviors, particularly reverse driving and lane collapse incidents [3], [4], [5].

Recent advances in intelligent transportation systems have played a crucial role in enhancing road safety and traffic management. However, the increasing complexity of modern road networks, diverse driving behaviors, and unpredictable environmental conditions pose significant challenges to existing solutions. Traditional countermeasures, including physical barriers and static signage, provide some deterrence but often prove inadequate in dynamic traffic environments that require rapid response. Although emerging detection technologies, including sensor-based and artificial intelligence AI-driven methods, offer promising alternatives, they face limitations in scalability, adaptability, and cost-effectiveness.

To address these challenges, this study proposes a novel AI-driven detection system for identifying wrong-way driving and lane collapse incidents in real-world traffic scenarios. Using advanced computer vision techniques, the proposed method enhances detection accuracy and robustness across various road environments, including intersections, highways, and roundabouts, making it well-suited for intelligent traffic management systems.

The proposed approach integrates the state-of-the-art You Only Look Once, version 8 (YOLOv8) [6] object detection model with a dynamic polygon-based zone management system, offering several key advantages:

1) Adaptive configuration: Dynamic polygon-based zones adjust to various road configurations, enabling deployment across diverse infrastructure environments.

2) Real-time processing: The lightweight architecture of YOLOv8, combined with efficient zone management, enables real-time detection and response.

3) Robust scene adaptability: The combination of polygonal zones and ByteTrack enables reliable detection across intersections, roundabouts, and complex traffic scenes.

4) Cost-effectiveness: The vision-based system requires minimal additional infrastructure, making it accessible to transportation authorities with varying resources.

Through the comprehensive evaluation of diverse datasets, our approach demonstrates superior performance in detecting reverse driving and lane collapse scenarios across various road configurations. The system maintains high accuracy while minimizing computational overhead, ensuring its suitability for widespread deployment.

Objectives of the Paper:

- To conduct an in-depth literature review of diverse driving behavior detection techniques proposed in prior research, with a focus on identifying limitations related to adaptability, road type constraints, and deployment cost in existing systems;
- To propose a novel vision-based methodology for detecting reverse driving and lane collapse behaviors using CCTV surveillance. The purpose is to provide

*Corresponding authors.

This work was supported by the Technology development Program (S3458136) funded by the Ministry of SMEs and Startups (MSS, Korea).

a lightweight and infrastructure-independent system capable of adapting to a variety of real-world road layouts, including intersections and roundabouts. The novelty of the proposed system lies in its manually configurable polygon-based zone management and integration with YOLOv8 and ByteTrack for precise object detection and multi-object tracking in complex scenes;

- To test and validate the system's performance across varied traffic scenarios using visual assessments under different lighting conditions (day and night), road types (single-lane, multi-lane, roundabout), and behavioral patterns (normal flow, reverse driving, and lane collapse);
- And, to validate the proposed system in real-time system responsiveness to detect and update reverse-driving incidents immediately within video stream processing.

The remainder of this paper is organized as follows. The literature review section analyzes existing detection methods and their limitations. The methodology section details the system architecture. The implementation details section summarizes the steps of the algorithm, illustrating the logical workflow. The result section discusses the experimental results across various road configurations, demonstrating the effectiveness of the system. Finally, the discussion concludes with insights into future research directions and potential applications.

II. Literature Review

Research highlights the significant impact of wrong-way driving and lane collapse incidents on road safety. The AAA Foundation reported a 34% increase in driving against the traffic flow incidents over the past five years, particularly on highways and divided roads [7]. In the United States, annual economic losses from such incidents exceed $5.2 billion, encompassing direct damage, emergency response costs, and productivity losses. Lane collapse incidents account for 18% of highway accidents, with heightened risks under adverse weather conditions and reduced visibility [8], [9].

Traditional countermeasures, including signage and physical barriers, have been widely implemented, yet their effectiveness remains limited, preventing only 65% of wrong-way entries [10], [11]. Electronic detection systems, such as roadside cameras and automated warnings, can improve response times but suffer from high false-positive rates (15–20%) and incur high maintenance costs, averaging $50,000 annually per installation [12]. Sensor-based solutions, including radar, lidar, and GPS tracking, have also been explored [13], [14], [15], [16], [17]. However, radar systems exhibit reduced accuracy in adverse weather, lidar presents high deployment costs, and GPS-based approaches require vehicle-mounted units, which hinders scalability in public infrastructure applications.

Artificial intelligence (AI)-based systems, particularly those utilizing convolutional neural networks (CNNs), have demonstrated high accuracy in detecting traffic anomalies in controlled environments. Studies report over 95% accuracy in detecting lane violations and reverse movement under ideal conditions [18], [19], [20], [21]. Nevertheless, real-world deployment remains challenging due to lighting variation, complex road geometries, and occlusions. Camera-based detection approaches, including dashcam footage and surveillance systems, also require adaptation to varied environmental conditions [22], [23], [24], [25]. Moreover, these methods often struggle with performance in intersections, roundabouts, and irregular traffic zones [26], [27].

Recent advances have sought to combine multiple techniques to address these challenges. Hybrid systems integrating vehicle detection, motion tracking, and infrastructure feedback have improved robustness. Vehicle-to-infrastructure (V2I) methods offer real-time alerts, but remain costly and dependent on specific installations [28]. Grouping-based vehicle detection enhances classification accuracy but is vulnerable to changes in weather, traffic density, and road layout [29]. Loop-type detectors such as those from Sumitomo Electric require precise calibration and suffer from false activations in dense traffic areas [30].

Among recent contributions, the work by Suttiponpisarn et al. [18] presents an enhanced system combining YOLOv4-Tiny with FastMOT [31] and a road boundary extraction algorithm (RLB-CCTV). This method achieves high accuracy using embedded-suitable components, and replaces manual boundary marking with automated lane detection based on edge and line detection. However, it remains limited to straight roads. The system's dependency on statistical estimation for directional flow and sensitivity to environmental conditions constrain its applicability in complex or non-linear road structures such as roundabouts or intersections.

These observations highlight the need for a flexible and generalizable detection system that performs reliably across diverse road conditions. Our study addresses this gap by introducing a polygon-based zone framework integrated with YOLOv8 for real-time detection and tracking. The proposed approach allows users to manually configure detection zones for arbitrary road layouts, enabling robust performance in environments where automated methods or fixed detectors fail to generalize.

III. Methodology

The proposed system is designed to detect reverse driving and lane collapse incidents using video streams from roadside CCTV cameras. The overall architecture is composed of four primary modules: video input processing, dynamic zone setup, movement detection and analysis, and real-time monitoring. A visual overview of the framework is presented in Fig. 1.

A. Video Input and Frame Processing

Each video stream is processed frame-by-frame. Frames are extracted in real time and passed to the object detection and tracking module. This ensures the system can maintain temporal continuity of vehicle movement and associate detection results across time.

B. Object Detection and Vehicle Tracking

To detect vehicles, we use YOLOv8 [6], a state-of-the-art object detection model known for its accuracy and efficiency,

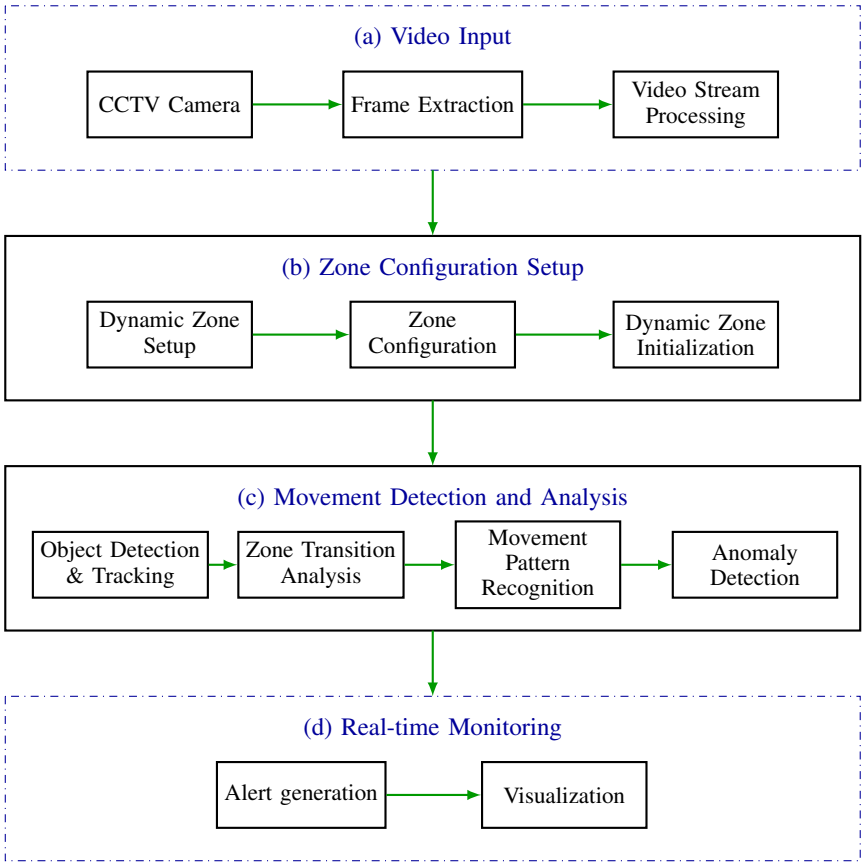


Fig. 1. Overview of the proposed traffic monitoring system architecture. (a) Video input module processes CCTV streams. (b) Configuration defines dynamic detection zones. (c) Movement detection analyzes object transitions and detects anomalies. (d) Real-time monitoring generates alerts and visual output.

particularly in complex scenes. Each detected vehicle is represented by a bounding box and class label (e.g., "car", "truck"). The detection results are then passed to the ByteTrack multi-object tracker [32], which assigns a unique tracker ID to each vehicle.

ByteTrack in Section III-G operates by associating detections across frames using Intersection-over-Union (IoU) matching. It follows multiple steps which helps maintain trajectory consistency even when detections are temporarily uncertain or occluded.

C. Dynamic Zone Configuration

In the initial frame of the video, polygonal zones are manually defined using OpenCV [33]. For a two-lane road, four zones are created: in-zone-0, out-zone-0, in-zone-1, and out-zone-1. These zones represent logical boundaries of entry and exit for each lane (as shown in Fig. 2 and 3).

Each vehicle's bounding box is evaluated in every frame to check whether it intersects any of these predefined zones. The result is stored as a binary occupancy value for that zone (1 = inside, 0 = outside). Each tracker ID maintains a zone presence vector over time.

D. Illustrative Example

Consider vehicle ID #3 on a single-lane road in Fig. 2:

- It enters in-zone-0 and later exits through out-zone-0.
- Its zone vector becomes (1,1,0,0), matching the first row in Table I.
- Therefore, the reverse value is set to 0 (reverse driving within the same lane).

By contrast, if a vehicle enters both `in-zone-0` and `in-zone-1`, but never exits correctly, its vector becomes (1,0,1,0) and is assigned a reverse value of 1 (violation or lane collapse) as shown in Fig. 3.

E. Real-Time Output and Monitoring

Once the reverse status is determined for a tracker ID, the label is visualized in real time using bounding boxes and zone overlays. Vehicles are annotated with their ID and reverse status (e.g., "car (reverse: 1)"), enabling immediate visual confirmation and alerting.

This modular pipeline ensures the system can operate reliably across varied road configurations, including intersections and roundabouts, without retraining. The manual polygon zone setup provides flexibility, while YOLOv8 and ByteTrack ensure accurate detection and persistent identity tracking.

F. Reverse Driving Judgement

The algorithm, outlined in Table I, employs a zone-based configuration with four distinct monitoring areas—in-zone-0

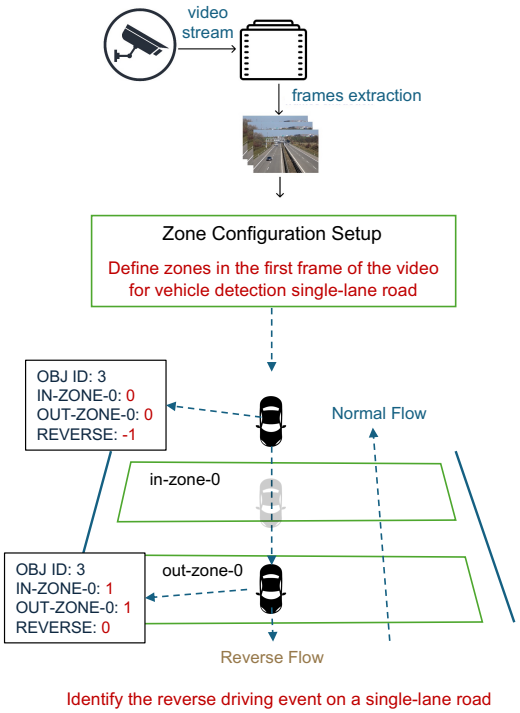


Fig. 2. Zone configuration for single-lane road monitoring.

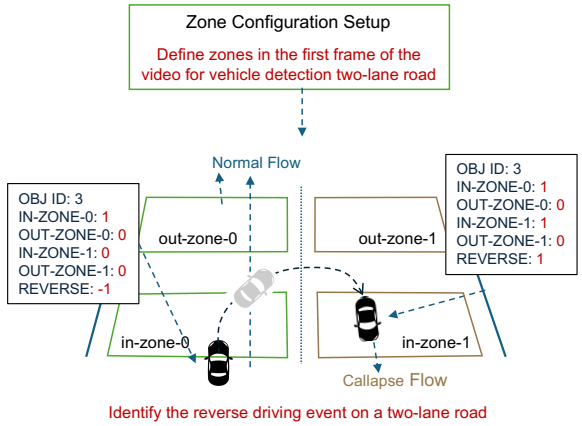


Fig. 3. Zone configuration for two-lane road monitoring.

TABLE I. ZONE TRANSITION STATES AND THEIR CORRESPONDING REVERSE VALUES. THE 'REVERSE' COLUMN INDICATES WHETHER A VEHICLE IS IN AN UNKNOWN STATE (-1), REVERSE WITHIN ITS LANE (0), AND COLLAPSE ON OTHER LANES (1)

in-zone-0	out-zone-0	in-zone-1	out-zone-1	reverse
1	1	0	0	0
0	0	1	1	0
1	0	0	1	1
0	1	1	0	1
1	0	1	0	1
0	1	0	1	1
else				-1 (unknown)

and out-zone-0 for the left lane, and in-zone-1 and out-zone-1 for the right lane, as depicted in Fig. 3. The occupancy of each zone is binary (0 or 1), enabling precise vehicle movement tracking and directionality assessment. The algorithm evaluates all possible combinations of zone occupancies to detect reverse driving incidents. The 'reverse' column in Table I represents the detection outcome, where 0 indicates reverse driving within the same lane, 1 signifies cross-lane reverse driving or lane collapse, and -1 indicates ambiguous cases requiring further analysis. This six-row approach effectively identifies and assesses vehicle movements, enhancing detection accuracy for traffic anomalies on two-lane roads. This approach provides a robust framework for automated reverse driving detection, improving traffic safety by identifying potentially hazardous situations.

The implementation, illustrated in Fig. 3, demonstrates both normal traffic flow and lane collapse scenarios, with corresponding zone transition states documented in Table I. This comprehensive method ensures accurate monitoring and timely detection of reverse driving events, contributing to traffic safety.

G. Vehicle Tracking

ByteTrack [32] is employed as a robust multi-object tracking algorithms that effectively track objects, including vehicles, by utilizing both high-confidence (D_{high}) and low-confidence (D_{low}) detection. This algorithm serves as the starting point of Algorithm 1, which analyzes the movements of tracked vehicles and assigns each one a unique `tracker_id`, as defined in Algorithm 1.

The ByteTrack tracking mechanism involves the following main steps:

1) Detection set division: At each frame t, the detections are divided into:

$$D_t = D_{\text{high},t} \cup D_{\text{low},t}, \tag{1}$$

where:

- $D_{\text{high},t}$: comprises detections with confidence $c > \tau$, where τ is a high confidence threshold.
- $D_{\text{low},t}$: include detections with confidence $\tau_{\text{low}} < c \leq \tau$, where τ_{low} is a lower confidence threshold.

2) Matching with high-confidence detections: High-confidence detections are matched to existing tracks T_t using the intersection over union (IoU) metric:

$$\text{IoU}(d_i, T_j) = \frac{|b_i \cap b_j|}{|b_i \cup b_j|}, \tag{2}$$

where b_i and b_j are the bounding boxes of detection d_i and track T_j, respectively. The Hungarian algorithm is used to find the optimal matching:

$$T_{\text{matched}} = \text{Hungarian}(T_t, D_{\text{high},t}, \text{IoU}). \tag{3}$$

3) Incorporation of low-confidence detections: To prevent track loss, unmatched tracks from the high-confidence step are matched with low-confidence detections:

$$T_{\text{low-matched}} = \{(t, d) \mid \tag{4}$$

$$\text{IoU}(t, d) \geq \tau_{\text{low}}, \quad t \in T_{\text{unmatched}},\ d \in D_{\text{low},t}\} \tag{5}$$

4) Track update: Matched tracks are updated using Kalman filtering [34] to refine the bounding box predictions:

$$b_{\text{updated}} = K(b_{\text{predicted}}, b_{\text{new}}), \tag{6}$$

where $b_{\text{predicted}}$ is the predicted state of the track, b_{new} is the observed detection, and K represents the Kalman filter.

5) New track initialization: Unmatched high-confidence detections are initialized as new tracks:

$$T_{\text{new}} = \{b_i \mid b_i \in D_{\text{high},t},\ b_i \notin T_{\text{matched}}\}. \tag{7}$$

6) Track termination: Tracks are terminated if they remain unmatched for k consecutive frames:

$$T_{\text{terminated}} = \{t \mid \Delta t > k,\ t \notin T_{\text{updated}}\}. \tag{8}$$

IV. Implementation Details

Algorithm 1 is designed to detect reverse driving by processing object detections with their associated `tracker_id` and bounding boxes. The algorithm's primary goal is to update `reverse_conditions` for each detection, which stores information about tracked objects.

Algorithm 1 Detecting Reverse Driving and Lane Collapse

Input: Detections with `tracker_id` and bounding boxes
Output: Detections with updated `reverse_conditions`
`tracker_id`: An object being tracked
`reverse_conditions`: Stores values related to tracked objects (as in Table I) Initialize `reverse_conditions` for all detections **ForEach** *detection in* `detections`
  Extract `tracker_id` and bounding box Initialize `reverse_conditions[tracker_id]`
end
ForEach *zone in zones*
  ForEach *`tracker_id` in zone*
    Update zone information for `tracker_id` Check if `tracker_id` moved in or out of the zone **If** *reverse driving is detected*
      `reverse_conditions[tracker_id]` ← `true`
  end
end
Return *`detections` with updated `reverse_conditions`*

It begins by initializing `reverse_conditions` for all detections in the input data. The process then iterates through each detection, extracting the `tracker_id` and bounding box information, and initializes the corresponding `reverse_conditions` entry for that `tracker_id`. The core detection logic operates by examining each zone and the `tracker_id`s within them. For each `tracker_id` present in a zone, the algorithm updates the zone information and monitors whether the tracked object has moved in or out of the zone. When reverse driving behavior is detected, the algorithm sets the `reverse_conditions` for the corresponding `tracker_id`.

After processing all zones and their associated `tracker_id`s, the algorithm concludes by returning the complete set of detections with their updated value of `reverse_conditions` as shown in 'reverse' column in Table I, providing a comprehensive record of any reverse driving incidents detected during the analysis.

V. Results

To validate the proposed system, we present qualitative results using real-world video footage across various scenarios. Fig. 4 to 7 demonstrate the system's ability to detect normal and reverse driving behavior in single-lane, multi-lane, nighttime, and roundabout road conditions. These case studies confirm the robustness and practical performance of the system, even in complex or low-visibility environments.

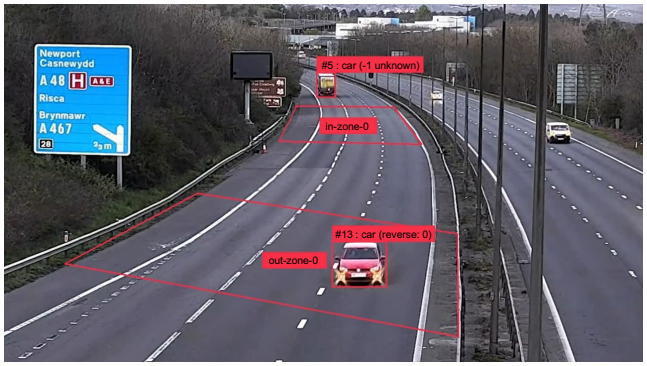


Fig. 4. Result of monitoring on single-lane road configuration.

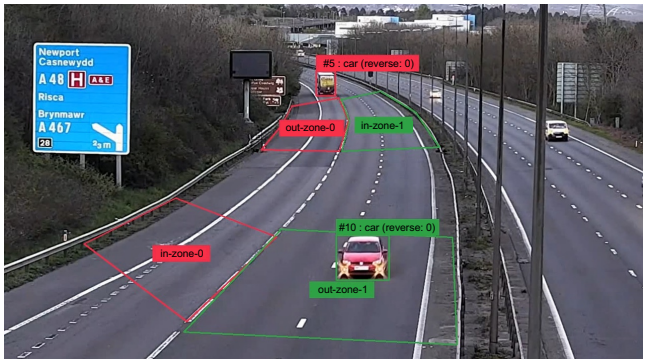


Fig. 5. Result of monitoring on double-lane road configuration.

The real-world implementation demonstrates the effectiveness of the reverse driving detection algorithm on single-lane road segments. As shown in Fig. 4, the system successfully

TABLE II. COMPARISON BETWEEN THE PROPOSED METHOD AND SUTTIPONPISARN ET AL. [18]. THE TABLE HIGHLIGHTS DIFFERENCES IN DETECTION FLEXIBILITY, DEPLOYMENT SCOPE, AND CORE ALGORITHMIC CHOICES

Aspect	**Suttiponpisarn et al**	**Proposed Method**
Detection Model	YOLOv4-Tiny	YOLOv8
Tracking Algorithm	FastMOT	ByteTrack
Zone Definition	Automated rectangular zones via edge detection	Manual polygonal zones for flexible layout
Supported Road Types	Straight, unobstructed multi-lane roads	Multi-lane, intersections, roundabouts
Lighting Adaptability	Requires brightness correction	Works under day/night
Object Focus	Motorcycle detection in 4 test locations	All vehicles across diverse traffic scenarios
Evaluation Type	Quantitative (96.61% accuracy)	Qualitative visual validation
False Positive Handling	Measured and reported	Acknowledged; future work planned

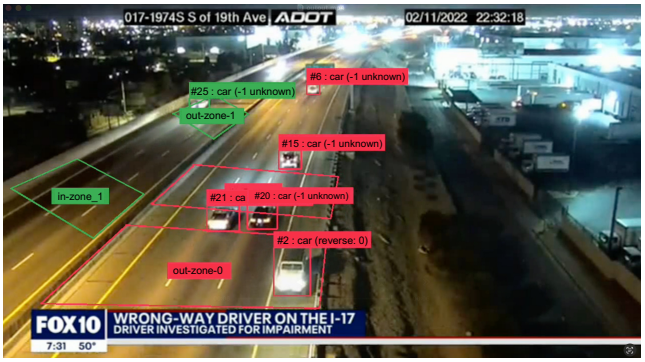


Fig. 6. Result of monitoring at night condition.

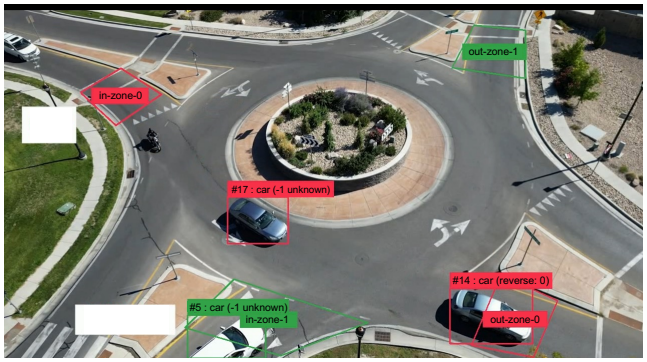


Fig. 7. Result of monitoring on roundabout road configuration.

monitors vehicle movement through defined zones on a primary road. A test case with vehicle #13 demonstrated normal forward motion. The vehicle traversed through in-zone-0 and out-zone-0, with the algorithm correctly assigning reverse: 0 based on the zone transition states in Table I. This corresponds to the first row of the transition table, where the activation of both zones in a single lane (1,1,0,0) indicates proper directional flow. Additionally, the system simultaneously tracks multiple objects. For instance, vehicle #5 was detected in an unknown state (-1), aligning with the undefined state conditions of the table. These results validate the capability of the algorithm to differentiate normal and reverse driving patterns in real-time monitoring scenarios. The clear zone demarcation enables reliable movement classification.

The monitoring capabilities of the system were further tested in multi-lane scenarios, where multiple vehicles were tracked simultaneously. As shown in Fig. 5, the algorithm correctly classifies vehicle #5 in out-zone-0/in-zone-0 and vehicle #10 in in-zone-1/out-zone-1 with reverse: 0. This aligns with the zone transition states of Table I, where the simultaneous occupation of zones (0,0,1,1) indicates normal directional flow in the right lane. Clear zone delineation between lanes (red for lane 0 and green for lane 1) enables accurate vehicle tracking and classification. These results validate the effectiveness of the algorithm in complex multi-vehicle, multi-lane monitoring scenarios.

Fig. 6 presents results from nighttime highway monitoring. Vehicles are detected and annotated with bounding boxes and labels. Vehicle #2, labeled 'car (reverse: 0)', indicates it is traveling against the flow of traffic in out-zone-0.

The reverse driving detection algorithm can monitor roundabout scenarios, where dynamic polygonal zone setups present unique challenges. Fig. 7 demonstrates the performance of the system at a roundabout intersection, where multiple vehicles are tracked simultaneously. Detected vehicles include vehicle #14 in out-zone-0 (reverse: 0), indicating normal flow and vehicles #17 and #5 in unknown states (-1) due to movement through zones that do not conform to the standard transition patterns in Table I. The system effectively monitors multiple entrance and exit points, although the increased complexity of roundabout navigation results in a higher frequency of unknown state classifications compared to straight road segments.

A. Analysis

To further evaluate the strengths and limitations of the proposed system, we conducted a comparative analysis with a closely related work by Suttiponpisarn et al. [18]. Their method combines YOLOv4-Tiny for object detection, FastMOT [31] for tracking, and an automated road boundary extraction algorithm (RLB-CCTV) that utilizes edge detection and Hough transforms to define lane boundaries. The system is designed for embedded devices and achieves high accuracy (96.61%) for detecting wrong-way motorcycles on straight, well-lit roads.

However, their approach presents several constraints. It is restricted to straight, unobstructed multi-lane roads and depends on brightness correction and predefined geometric assumptions. It does not support curved roads, roundabouts, or intersections—scenarios which are common in urban traffic environments.

In contrast, our method uses manually defined polygonal zones that can be flexibly adapted to any road geometry. The system is validated qualitatively across multiple complex real-world scenarios, including intersections and roundabouts, and performs reliably under both day and night conditions. Moreover, we adopt YOLOv8 for object detection, which has demonstrated superior accuracy and generalization over YOLOv4-Tiny in benchmark evaluations, particularly in detecting small or partially occluded objects [6]. Table II summarizes the key differences between the two approaches.

This comparative analysis underscores the broader applicability and deployment flexibility of our method, particularly in urban environments where roads are non-linear and traffic conditions are more dynamic.

VI. Conclusion and Future Work

The research paper proposes a novel system for detecting reverse driving and lane collapse using CCTV-based video monitoring. The system integrates YOLOv8 for object detection, ByteTrack for vehicle tracking, and manually defined polygonal zones for flexible configuration across different road layouts. The novelty of the method lies in its adaptability to intersections, roundabouts, and non-linear road geometries without relying on additional sensors or automated boundary extraction. Experimental results show consistent visual performance across day, night, single-lane, multi-lane, and roundabout scenarios. However, the evaluation remains qualitative due to the absence of labeled datasets.

In the near future, it is planned to focus on developing dynamic zone adjustment mechanisms that extend beyond straight road detection. In contrast to prior approaches that rely on automated boundary extraction limited to linear lane structures, the proposed direction will aim to support adaptive zone generation for complex geometries such as roundabouts, intersections, and curved roads. This advancement is expected to reduce manual configuration effort and improve generalizability. In addition, the work will include quantitative evaluation using annotated datasets to assess detection accuracy, false detection rates, and latency. These improvements can further enhance the system's robustness and scalability for real-world intelligent traffic surveillance.

References

[1] Sweatman, P. (2025). Approaches to Road Safety: Evolution, Challenges, and Emerging Technologies

[2] Banta-Green, C. J., & Williams, J. (2016). AAA Foundation for Traffic Safety. *Washington, DC: AAA Foundation for Traffic Safety.*

[3] Ding, J. (2023, October). Research and Development of Dangerous Driving Warning and Rescue Devices. In *2023 7th International Symposium on Computer Science and Intelligent Control (ISCSIC)* (pp. 297-301). IEEE.

[4] Hovorushchenko, T., Pavlova, O., Binkovskyi, Y., Bilinska, A., Holovatiuk, A., & Melnychuk, D. (2023, October). Road Accident Prevention System. In *2023 13th International Conference on Dependable Systems, Services and Technologies (DESSERT)* (pp. 1-7). IEEE.

[5] Cicchino, J. B., & Zuby, D. S. (2017). *Prevalence of driver physical factors leading to unintentional lane departure crashes. Traffic injury prevention, 18(5), 481-487.*

[6] Khow, Z. J., Tan, Y. F., Karim, H. A., & Rashid, H. A. A. (2024). Improved YOLOv8 Model for a comprehensive approach to object detection and distance estimation. *IEEE Access.*

[7] Shuey, R., & Myers, D. (2021). The AAA approach to crash investigation reform-the perspective from road policing practitioners. *Journal of road safety, 32*(4), 51-59.

[8] Shakti, S., Bhardwaj, I., Singh, A. K., & Bhardwaj, A. (2022). Design of Road surveillance system for low visibility, bad weather and emergency situations. *Global Journal of Innovation and Emerging Technology, 1(2)*, 21-26.

[9] SHARIFF, M. M. (2020). *EFFECTS OF WET WEATHER ON DRIVERS'RISK ACCIDENT PERCEPTION DURING MOTORIST-FOLLOWING BEHAVIOUR ON TWO-WAY TWO-LANE HIGHWAY* (Doctoral dissertation, Universiti Teknologi Malaysia).

[10] Pour-Rouholamin, M., Huaguo Zhou PhD, P. E., & Jeffrey Shaw, P. E. (2014). Overview of safety countermeasures for wrong-way driving crashes. *Institute of Transportation Engineers. ITE Journal, 84*(12), 31.

[11] Wrong Way Driver Prevention, Division of Safety Programs. Available at: https://dot.ca.gov/-/media/dot-media/programs/safety-programs/documents/wrong-way/wrong-way-driver-prevention-faq-a11y.pdf (Accessed on 16th May, 2025)

[12] Initial Deployment Costs for a Wrong-Way Driving Prevention System at an Off-Ramp Ranged from $18,000 to 45,000$. Available at: https://www.itskrs.its.dot.gov/2021-sc00501 (Accessed on 17th May, 2025)

[13] Kim, D. H. (2020). Lane detection method with impulse radio ultra-wideband radar and metal lane reflectors. *Sensors, 20*(1), 324.

[14] https://www.amey.co.uk/media/press-releases/2023/february/amey-and-fleetclear-lead-the-way-with-new-safety-solution/ (Accessed on: 17th May, 2025)

[15] Kubra, K. T., Akhund, T. M. N. U., Al-Nuwaiser, W. M., Assaduzzaman, M., Ali, M. S., & Sarker, M. M. (2024). Integrated iot-driven system with fuzzy logic and v2x communication for real-time speed monitoring and accident prevention in urban traffic. *International Journal of Advanced Computer Science and Applications, 15*(8).

[16] Boateng, C., Yang, K., Ghoreishi, S. G. A., Jang, J., Jan, M. T., Conniff, J., ... & Rosseli, M. (2023, December). Abnormal driving detection using gps data. In *2023 IEEE 20th International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)* (pp. 1-6). IEEE.

[17] Gagneja, K., & Singh, K. J. (2020, February). Wrong Lane Driving Detection using Satellite Navigation. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (pp. 1-5). IEEE.

[18] Suttiponpisarn, P., Charnsripinyo, C., Usanavasin, S., & Nakahara, H. (2022). An enhanced system for wrong-way driving vehicle detection with road boundary detection algorithm. *Procedia Computer Science, 204,* 164-171.

[19] Rahman, Z., Ami, A. M., & Ullah, M. A. (2020, June). A real-time wrong-way vehicle detection based on YOLO and centroid tracking. In *2020 IEEE region 10 symposium (TENSYMP)* (pp. 916-920). IEEE.

[20] Faizi, F. S., & Al-sulaifanie, A. K. (2024). Vision-Based Multi-Stages Lane Detection Algorithm. *Pertanika Journal of Science & Technology, 32*(4).

[21] Vardhan, M. H., Krishna, K. V. S., Munappa, S., & Manoj, K. A. (2023, December). Wrong Route Vehicles Detection Using Deep Learning. In *2023 International Conference on Next Generation Electronics (NEleX)* (pp. 1-6). IEEE.

[22] Abdrakhmanov, R., Elemesova, M., Zhussipbek, B., Bainazarova, I., Turymbetov, T., & Mendibayev, Z. (2023). Mask R-CNN Approach to Real-Time Lane Detection for Autonomous Vehicles. *International Journal of Advanced Computer Science and Applications, 14*(5).

[23] Surabhi, S., & Sunder, S. (2024, April). Real time lane detection using CNN. In 2024 10th *International Conference on Communication and Signal Processing (ICCSP)* (pp. 858-861). IEEE.

[24] Devane, V., Sahane, G., Khairmode, H., & Datkhile, G. (2021). Lane detection techniques using image processing. In *ITM web of conferences* (Vol. 40, p. 03011). EDP Sciences.

[25] Bharadwaj, R., Shinde, P., Shelke, P., Shinde, N., & Shirsath, A. (2022, February). Smart Vehicle Tracking in Harsh Condition. In *International Conference on Expert Clouds and Applications* (pp. 669-682). Singapore: Springer Nature Singapore.

[26] Usmankhujaev, S., Baydadaev, S., & Woo, K. J. (2020). Real-time, deep learning based wrong direction detection. *Applied Sciences*, 10(7), 2453.

[27] Gaur, K., Siddique, M. A., Beernally, K., Madaan, N., & Tarwani, S. (2024, June). Real-Time Wrong-Way Vehicle Detection System with Automatic Number Plate Recognition for Enhanced Road Safety. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-8). IEEE.

[28] Dokur, O., & Katkoori, S. (2022, December). Vehicle-to-infrastructure based algorithms for traffic light detection, red light violation, and wrong-way entry applications. In *2022 IEEE International Symposium on Smart Electronic Systems (iSES)* (pp. 25-30). IEEE.

[29] Lakshmi, G., Brindha, S., & Rashmi, A. (2023, December). Grouping Based Wrong Way Vehicle Detection. In *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)* (pp. 1-5). IEEE.

[30] Finley, M. D., Balke, K. N., Rajbhandari, R., Chrysler, S. T., Dobrovolny, C. S., Trout, N. D., ... & Mott, C. (2016). *Conceptual Design of a Connected Vehicle Wrong-Way Driving Detection and Management System* (No. FHWA/TX-16/0-6867-1). Texas A&M Transportation Institute.

[31] Yang, Y. (2020). FastMOT: High-performance multiple object tracking based on deep SORT and KLT. *J. Korea Multimed. Soc*, 20, 893-910.

[32] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., ... & Wang, X. (2022, October). Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision* (pp. 1-21). Cham: Springer Nature Switzerland.

[33] Culjak, I., Abram, D., Pribanic, T., Dzapo, H., & Cifrek, M. (2012, May). A brief introduction to OpenCV. In *2012 proceedings of the 35th international convention MIPRO* (pp. 1725-1730). IEEE.

[34] Li, Q., Li, R., Ji, K., & Dai, W. (2015, November). Kalman filter and its application. In *2015 8th international conference on intelligent networks and intelligent systems (ICINIS)* (pp. 74-77). IEEE.

Comparative Analysis of Machine Learning Frameworks for Robust Ovarian Cancer Detection Using Feature Selection and Data Balancing

DSS LakshmiKumari P[1], Maragathavalli P[2]
Information Technology, Puducherry Technological University, Pillaichavady, Puducherry 605014[1,2]
SRKR Engineering College, Bhimavaram[1]

***Abstract*—One of the most serious malignancies that affects women's health worldwide is ovarian cancer. As a result, prompt accurate diagnosis and treatment are necessary. This study's primary objective is to determine whether or not OC is present within the body of a person by using a range of characteristics gleaned with a couple of health examinations. The article is concentrated on twelve ML techniques used for OC diagnosis. The dataset has been altered by applying the borderline SVMSMOTE method to address the imbalance properties and the MICE imputation method to impute the missing values in order to enhance the performance of the classifiers. Additionally, the boruta approach and recursive feature reduction has been utilized to identify the most important features while the hyper parameter tuning strategy has been employed to improve classifier performance and provide ideal solutions.Boruta opted just 50% of the total characteristics and outperformed RFE while considering the most important feature. Furthermore, many performance measures are used to determine which classifiers are the best in identifying OC. Voting classifier surpassed state-of-the-art approaches and other machine learning methods with the highest accuracy. The suggested approach obtained the highest average of 93.06% accuracy, 88.57% precision, 96.88% recall, 92.54% F1-score, and 93.44% AUC-ROC based on experimental results. Experiments show that in comparison with the state-of-the-art techniques, our suggested method can identify OC more accurately.**

Keywords*—Ovarian cancer detection; machine learning framework; data balancing; feature selection

I. Introduction

Globally, ovarian cancer (OC) is one of the most common cancers in women. As to the most recent worldwide cancer data published in 2021, there are roughly 314,000 new ovarian cancer cases (3.4% of all new cancer cases in women) and 207,000 ovarian cancer deaths annually (4.7% of all women's cancer deaths) [1]. Most of the time, ovarian cancer is discovered in its later stages due to a absence of specific symptoms and indicators as well as improper screening, which frequently results in unfavorable consequences [2], [3]. In general, gynaecologists have to determine if a patient has grown malignant pelvic lumps, which may indicate a tumour. Although certain approaches such as h CT and ultrasound have been used to differentiate benign tumors from malignant non-gynecologic conditions, tumor biomarkers such as human epididymis protein 4 (HE4), carbohydrate antigen 125 (CA125) and carbohydrate antigen 72-4 (CA72-4) are critical in differentiating between pelvic masses of women [4]. To stratify patients with OC into benign and malignant categories, machine learning (ML) models such as Light Gradient Boosting Machine (LGBM), Gradient Boosting Machine (GBM), Extreme Gradient Boosting Machine (XGBoost), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and are utilised in the construction of classification models [5].

Compared to traditional methods, ML techniques are extremely useful for identifying complicated patterns in biomarkers yet, they may not provide physical insights into diagnosis. The use of eXplainable Artificial Intelligence (XAI) can be enhanced by gaining a greater understanding of how sophisticated machine learning algorithms make decisions [6].

The main goal of this study is to identify OC using a variety of characteristics obtained from certain clinical examinations. Consequently, two datasets are used to test ensemble learning techniques, including RF, gradient boosting decision tree (GBDT), adaptive boosting,voting, bagging, XGBoost, LightGBM, and stacking. These ensemble learning techniques weren't used in parallel to identify OC in the earlier research publications. The models' performance is assessed using the dataset that is accessible to the general audience.

Various metrics, including precision, accuracy, recall, AUCROC, and F1-score, are employed to evaluate the effectiveness with every classifier.

The primary contributions of this study are as follows:

- In order to increase the effectiveness of the existing ovarian cancer diagnosis method, this research offered eight machine learning algorithms. In order to enhance performance, we have also adjusted each classifier's hyperparameter through the parameter tuning procedure.
- This study employed the MICE imputation method to address the missing information instead of applying a statistical approach. Additionally, the Borderline-SMOTE approach was used to balance the uneven data. Moreover, the relevant features are selected using recursive feature elimination (RFE) and Boruta approaches. Such OC prediction methods have not previously been thoroughly examined.

The structure of this work is as follows: Section 2 is devoted to the literature review, and Section 3 details the materials and procedures. Comparably, Section 5 presents the experimental findings and a discussion, while Section 4 shows

TABLE I. COMPARATIVE ANALYSIS OF MACHINE LEARNING APPROACHES FOR OVARIAN CANCER DETECTION

RefID	Dataset	Features	Feature Selection	Classifier	Evaluation Metrics
[4]	Mendeley	2	MRMR	DT, ROMA, LR	Acc: 92%, 95%, 97%
[5]	Mendeley	OC markers	T-test, Mann–Whitney U-test	Random Forest	Acc: 91%
[6]	PCLO Biomarker	-	-	Random Forest	AUC-ROC: 71%
[7]	Ultrasound Images	-	Fast Fourier Transform	LD, SVM, ELM	Acc: 85%
[8]	PCLO	27	F-test	SVMSMOTE	PPV: 90%
[9]	Tokyo Medical University	16	-	XGBoost	Acc: 80%
[11]	Mendeley	5	-	Group MCP	Acc: 93.33%, AUC: 0.892
[17]	Mendeley	16	SHAP	Stacking Classifier	Accuracy: 96.87%

the experimental setup. Followed by Section 6 represents the conclusion and ongoing research, in summary.

II. LITERATURE SURVEY

This section walks around a number of cutting-edge machine learning techniques that have been created to identify ovarian cancer. These strategies cover a variety of techniques, such as deep learning methods, supervised learning, and unsupervised learning. Through the utilisation of ML, these approaches have the possibility to enhance the effectiveness, precision, and expand-ability of ovarian cancer detection and diagnosis, ultimately leading to improved patient results. Table I presents a summary of the relevant works.

The author Mart´ınez-M´as et al. [7] identified the optimal classifier for FT-based feature descriptors and compare the ELM algorithm with traditional classifiers, by using ultrasound of 348 images. They has demonstrated that machine learning techniques, including LD and SVM approaches, are effective in creating the classification phase with 86% accuracy.

The most reliable predicting method for ovarian cancer, according to the Yang et al. [8] study, was the decision tree using SVMSMOTE on PLCO imbalanced dataset of143:7 class ratio. It had PPV of 0.9041, AUC of 0.9532, sensitivity of 0.7792, and specificity of 0.9982 is achieved by considering only seven features where there is chance of overfitting.

The possibility of applying the XGBoost machine learning algorithm to predict ovarian tumor which is proposed by Akazawa and Hashimoto[9], attained an accuracy of 0.80 for 16 characteristics on the 202 patient data set. On other hand Charkhchi et al. [10]focused the role that CA125 plays in the treatment of ovarian cancer in a review article , the necessity of accessible and affordable screening methods for the quick recognition of type II tumors. To predict benign or malignant ovarian tumors, the authors Hu et al. [11] selected distinct 11 groups by using 349 patients of 46 features achieved an accuracy of 93.33on other hand Ghoniem et al. [12] proposed to predict subtypes of other cancers by combining characteristics from pathologic image modal and gene modal. TCGAOV dataset is used containing pathological pictures, copy number variations, and gene expression data for 587 OC patients.The author Sorayaie Azar et al. [13] by using the SEER dataset, they study created ML techniques for regression and classification methods. XGBoost and RF were themost successful approaches for regression and classification, respectively. The findings showed promise and could be a useful tool for physicians to better understand the status of patients with ovarian cancer. For the purpose of OC detection, the authors Juwono et al. [14] has provided the best simultaneous feature weighting/parameter optimisation method. The LASSO regularisation and adaptive differential evolution (ADE) with cross validation error are used as the fitness function to optimise the weights.On other hand Boehm et al. [15] designed Decision support system to find three highly discriminating proteins (TOP1, PDIA4, and OGN) for the diagnosis of highgrade serous ovarian cancer (HGSOC) utilizing proteome data and a machine learning-based methodology. The DSS demonstrated promising clinical utility in the diagnosis of HGSOC with its excellent specificity and sensitivity in differentiating tumor from non-tumor patients.

On other hand Comes et al. [16] implemented an explainable machine learning model, the study identified high-risk BRCA-mutated patients and established when RRSO should be performed. The model produced encouraging results, with accuracy, specificity, sensitivity, G-mean, and AUC values of 83.2%, 85.3%, 57.1%, and 71.1%, respectively.

The key findings by the Abuzinadah et al. [17] include the extremely accurate stacked ensemble model for ovarian cancer prediction, the intricate interactions between biomarkers shown by the SHAP analysis, and the potential for real-world application.

III. METHODS AND MATERIALS

This article has built a smart approach for assessing and identification of OC. This study analyses OC data in twelve different ways. AUC-ROC, F-measure, accuracy, precision, recall, and F-measure were among the metrics used to assess each model's performance. The dataset was sourced from the Mendeley repository website.

A pair feature selection technique Boruta and Recursive feature elimination are employed to determine which features are most relevant. The steps involved in this investigation are shown in Fig. 1. Assigning incomplete values to every attribute in the dataset, the MICE method and a basic imputer are used. After the target attributes are equalised using the borderline-SMOTE technique, the processed data is fed into each classifier. Each classifier's parameter is adjusted via the hyper parameter tuning technique and randomised grid search. In this research, cross-validation techniques are used to validate the model functionality.

A. Dataset Description

The Third Affiliated Hospital of Soochow University provided the dataset for the study, which included 349 patients.

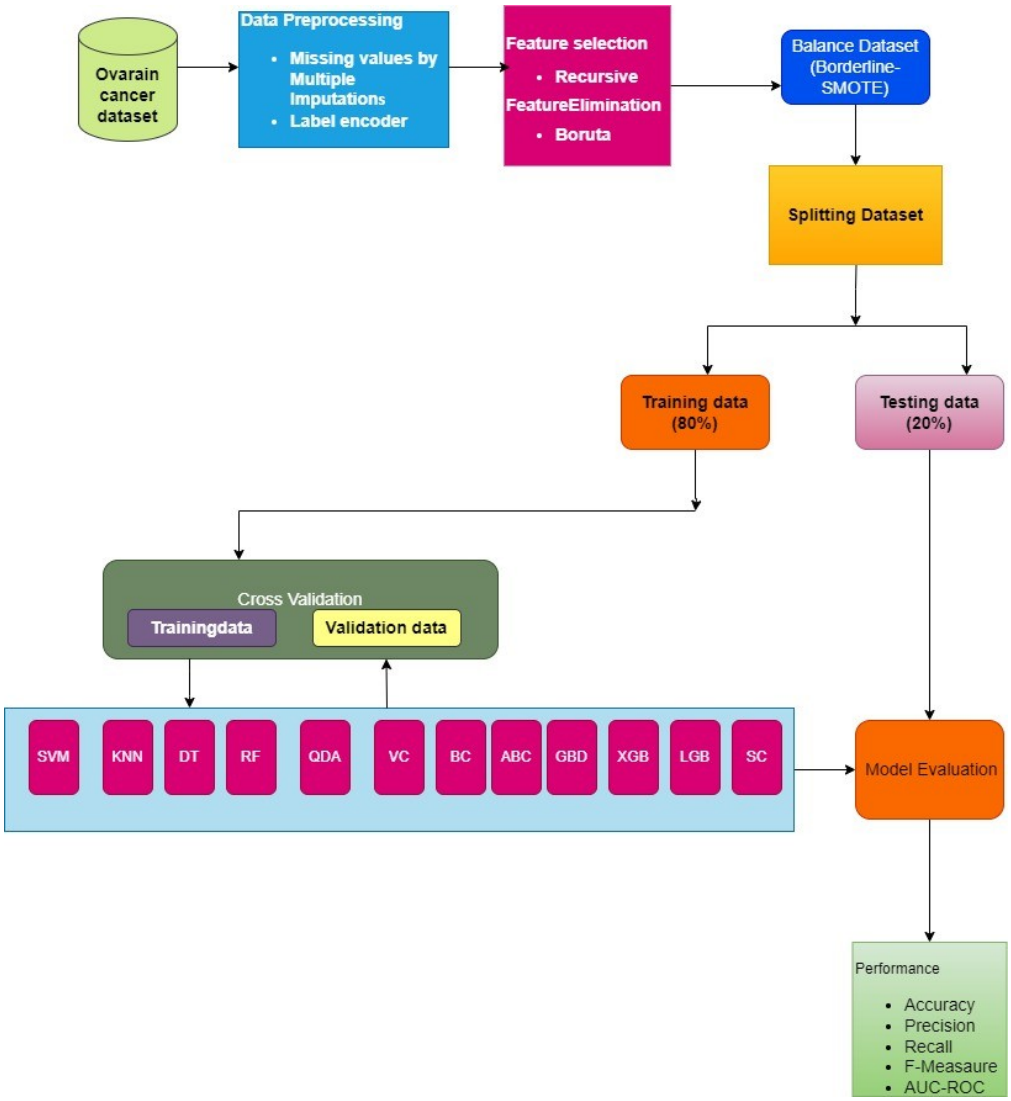


Fig. 1. Framework of detecting ovarian cancer.

The information was gathered between July 2011 and July 2018 and was split into two categories: 178 patients with benign ovarian tumour diagnoses, and 171 patients with malignant diagnoses [5].

B. Data Preprocessing

Preprocessing is necessary for the majority of real-world data because of outliers, noisy characteristics, missing values, and inconsistent patterns. If not, it will be difficult to increase the machine model's quality and produce an acceptable outcome. Numerous procedures have been used in this part to enhance and clean the data.

Several feature columns are labelled as inputs in the case of MICE technique, whereas single feature column that contains data missing is considered as an output. Subsequently, the output data is predicted using the regression model, and this is done again. Every value that is missing in the dataset is imputed using a different value in each iteration, and the process is repeated until convergence is achieved.

1) Missing value imputation: In this representation D represents the dataset with missing values, M represents the number of imputations to generate, k represents the number of iterations for each imputation, D_{imputed} represents the imputed dataset. The MICE algorithm iteratively imputes missing values for each variable in the dataset using regression models based on observed values of other variables. It repeats this procedure for a definite number of iterations for each imputation. Finally, it returns a list of M imputed datasets.

By using Multiple Imputation method [Algorithm 1], the dataset which contains 331 missing values which are been replaced.

The equation for imputing missing values using linear regression can be represented as follows:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \quad (1)$$

where:

- $\hat{Y}_i$ — predicted value for the i^{th} observation.
- β_0 — intercept term.
- $\beta_1, \beta_2, \ldots, \beta_p$ — regression coefficients.
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ — observed values of p predictors for observation i.

Algorithm 1 Multiple Imputations by Chained Equations (MICE)

procedure MULTIPLEIMPUTATIONSBYCHAINEDEQUATIONS(D, M, k)
  Input:
    D: Dataset with missing values
    M: Number of imputations to generate
    k: Number of iterations for each imputation
  Output: List of M imputed datasets
  Initialize empty list: $imputed_datasets \leftarrow [\,]$
  for $i = 1$ **to** M **do**
    Initialize $D_{\text{imputed}} \leftarrow D$ ▷ Copy of original dataset
    for $j = 1$ **to** k **do**
      for all variables v with missing values in D_{imputed} **do**
        Predict missing values in v using linear regression
        Update v in D_{imputed} with predicted values
      end for
    end for
    Append D_{imputed} to $imputed_datasets$
  end for
  return $imputed_datasets$
end procedure

C. Feature Selection

The process of eliminating unnecessary characteristics from a machine in order to reduce its computational load and improve the efficiency of the machine model is known as feature selection. The dual feature selection methods are used in this research.

1) Recursive Feature Elimination (RFE): This method selects the best subset of features from the entire dataset and minimizes the number of features. The SelectOptimalFeatures algorithm [Algorithm 2] provides a detailed explanation of each step in the RFE method.

2) Boruta based on RF: The BorutaFeatureSelection algorithm iteratively creates shadow features, combines them with original features, fits a Random Forest model, computes Z scores, and marks features as important or unimportant based on the comparison of Z scores. It repeats this procedure for a specified number of iterations and returns the desired number of optimal feature subsets.

3) Boruta based on RF: The BorutaFeatureSelection algorithm [Algorithm 3] iteratively creates shadow features, combines them with original features, fits a Random Forest model, computes Z scores, and marks features as important or unimportant based on the comparison of Z scores. It repeats this procedure for a specified number of iterations and returns the desired number of optimal feature subsets.

Algorithm 2 Select Optimal Features

Require: D: Training dataset with m samples and r features
  Y: Output values
  F: Set of features
  k: Number of iterations
  n: Number of optimal feature subsets
Ensure: F_{prime}: n optimal feature subsets
  Procedure:
  1. Train a RF model on dataset D.
  2. Initialize the feature subset F_{prime} with all features ($F_{\text{prime}} = F$).
  for $i = 1$ to k **do**
    a. Compute coefficients using Logistic Regression (LR) on the feature subset F_{prime}.
    b. Calculate the rank R based on the coefficients.
    c. Identify a specific number of features to eliminate, such as those with the smallest coefficients.
    d. Remove the determined number of features with the smallest coefficients from F_{prime}.
  end for

Algorithm 3 Boruta Feature Selection

procedure BORUTAFEATURESELECTION(D, Y, F, k, n)
  Input:
    D: Training dataset with m samples and r features
    Y: Target values
    F: Set of features
    k: Number of iterations
    n: Number of optimal feature subsets to select
  Output:
    F': Top n important features
  Initialize $selected_features \leftarrow []$
  for $i = 1$ **to** k **do**
    Generate shadow features by shuffling columns of F
    Combine F and shadow features to form dataset X
    Train Random Forest (RF) model on X
    Compute Z-scores: Z_o (original features), Z_s (shadow features)
    Let Z_{max} be the maximum of Z_s
    for each feature f_i in F **do**
      if $Z_o[f_i] > Z_{\text{max}}$ **then**
        Mark f_i as important
      else
        Mark f_i as unimportant
      end if
    end for
    Append important features to $selected_features$
  end for
  Select top n features from $selected_features$
  return F'
end procedure

IV. Classification Models

In this investigation, a variety of supervised ensemble-based machine learning algorithms [19] have been utilized, such as LR, DT, RF, LGBM, SVM, XGB, and GBM, in isolation, to forecast ovarian cancer. Furthermore, the search was conducted to identify the models that demonstrated the most outstanding performance.

1) SVMs: Supervised learning algorithms that are applied to regression and classification problems. It determines the best hyperplane with the largest margin between data points belonging to distinct classes. Using various kernel functions, SVM can handle both linear and nonlinear classification tasks.

2) k-NN: A feature space classification algorithm that classifies data points by using the majority class of the k nearest neighbors. This algorithm is regarded as non-parametric as it doesn't rely on explicit models.

3) Tree: This is probably a reference to decision trees, which divide the feature space into regions recursively and give each region a class name. Although decision trees are simple to understand, overfitting is a possibility.

4) RF: A technique that constructs many decision trees and combines their predictions by averaging or voting. This method successfully reduces overfitting by adding randomness to the feature and data sample selection processes.

5) QDA: A classification process known as quadratic discriminant analysis, or QDA, uses a quadratic decision boundary to simulate the probability density function of each class. It functions on the presumption that each class's features follow a normal distribution.

6) Voting: An ensemble learning technique where numerous models are trained independently and then their forecasts are combined by either average (for regression) or majority voting (for classification).

7) Bagging: Bagging (Bootstrap Aggregating) is an ensemble learning technique that uses bootstrap sampling of the training data to generate many models (typically of the same kind). To increase generalization performance and minimize variance, the predictions of the models are combined by voting or averaging.

8) AdaBoost: A technique that turns a number of ineffective learners into one powerful one. In each iteration, it gives more weight to data points that were incorrectly classified as it trains models progressively.

9) GBMs: Gradient Boosting Machines, or GBMs, are an ensemble learning technique that constructs a series of weak learners, usually decision trees, one after the other in a staged fashion, with each new tree fixing the mistakes of the preceding one.

10) XGBoost: A meticulously optimized implementation of the gradient boosting technique, specifically designed to enhance both speed and performance. By incorporating a diverse range of regularization techniques, XGBoost effectively prevents overfitting, thereby enabling it to achieve state-of-the-art outcomes in numerous machine learning competitions.

11) LGB: An additional scalable and extremely effective gradient boosting framework. To minimize memory usage and expedite training, it employs histogram-based approaches and a unique tree-growing algorithm.

12) SC: Also known as stacked generalization or stacking ensemble, is a machine learning technique that combines multiple classification models to improve prediction performance. Instead of relying on just one model, it leverages the collective wisdom of several models to make predictions.

V. Results

The recommended models are run through the Python program on a Windows 10 computer for validation. The system is powered by an Intel Core i3 processor with a 2.40GHz clock frequency and 8GB of RAM.

Additionally, this section involves in splitting the data source into 20% test and 80% training sets, adjusting the parameters of each classifier using the randomized grid search technique, and evaluating metrics to verify the performance of the model.

A. Parameter Tuning

Hyper-parameter optimization, or parameter tuning, is an essential stage in the creation of machine learning models. In order to maximize an algorithm's performance on a particular dataset, the optimal set of hyper-parameters for that method must be chosen. Additionally, the optimal subset of hyper-parameters for every classifier was shown in Table II to yield the best results for the OC prediction.

B. Model Evaluation Metric

The accuracy of a model's predictions can be assessed using a variety of performance evaluation criteria. More significantly, the well-known used performance evaluation metrics for binary classification include confusion matrix, precision, AUC-ROC, recall, F-measure, and accuracy.

1) Confusion matrix: The effectiveness of a classification algorithm is summarized in a table known as a confusion matrix that counts of false positives (NP), false negatives(NN), true positives (PP), and true negatives(PN).

2) Accuracy: It is the measure the proportion of a given class that is correctly classified among all instances is given by,

$$Accuracy = \frac{PP + NN}{PP + NN + PF + NF} \tag{2}$$

3) Recall: It is the proportion of all positively predicted instances to accurately predicted positive cases is given by,

$$Recall = \frac{PP}{PP + NN} \tag{3}$$

4) Precision: It is the ratio of correctly predicted positive cases among all positive cases is given by,

$$Precision = \frac{PP}{PP + PN} \tag{4}$$

TABLE II. LIST OF PRIMARY HYPERPARAMETERS FOR BINARY CLASSIFICATION FOR EACH CLASSIFIER

Classifier	Hyperparameters
SVM	kernel='rbf', C=10, gamma=0.1, probability=True
KNN	n_neighbors=11
RF	n_estimators=60, max_features=None, bootstrap=True, min_samples_split=3, min_samples_leaf=3, max_depth=None, random_state=0
QDA	Default parameters
DT	criterion='gini', min_samples_split=4, random_state=0, max_depth=None
SC	estimators=[('svc', svm), ('knn', knn)], final_estimator=rf
VC	estimators=[('svc', svm), ('knn', knn), ('rf', rf)], voting='soft'
BC	n_estimators=500, base_estimator=tree, n_jobs=-1, bootstrap=True, random_state=0, bootstrap_features=False
ABC	base_estimator=tree1, n_estimators=500, algorithm='SAMME', random_state=0
GDB	learning_rate=0.1, n_estimators=300, max_depth=3, min_samples_leaf=10
LGB	learning_rate=0.1, n_estimators=500, max_depth=3, min_child_weight=3
XGB	learning_rate=0.1, n_estimators=200, max_depth=2, min_child_weight=3

TABLE III. PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS USING ALL FEATURES

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC (%)
SVM	79.1	79.3	71.9	75.4	78.4
KNN	77.8	68.2	93.8	78.9	79.4
DT	83.3	79.4	84.4	81.8	83.4
RF	87.5	81.1	93.8	87.0	88.1
QDA	86.1	82.4	87.5	84.8	86.2
VC	90.3	87.9	90.6	89.2	90.3
BC	88.9	83.3	93.8	88.2	89.4
ABC	83.3	85.7	75.0	80.0	82.5
GDB	87.6	82.9	90.6	86.6	87.8
XGB	88.9	85.3	90.6	87.9	89.1
LGB	90.3	87.9	90.6	89.2	90.3
SC	76.4	74.2	71.9	73.0	75.9

5) F1-score: It is harmonic mean of precision and recall is given by,

$$F1 - score = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} \tag{5}$$

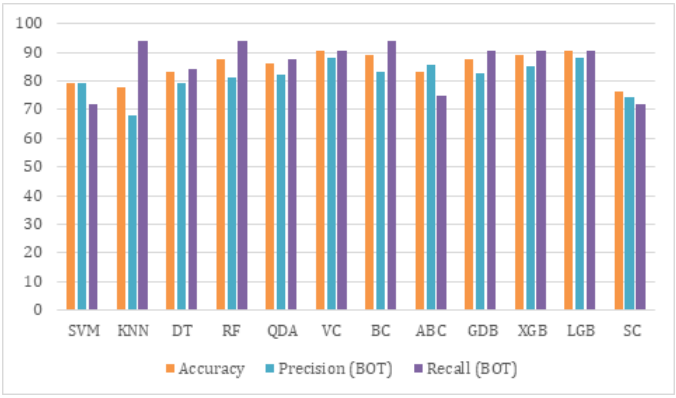


Fig. 2. Comparison of accuracy, Precision and recall for ML models of all features.

C. Results of Without Feature Selection Method

The experimental outcomes of each classification model employing every feature in the dataset are shown in Table III, Fig. 2 and Fig. 3. It was evident that the Voting classifier and LightGBM achieved a 90.3% score and all the other performance metrics are high except recall. Moreover, the other classifiers Bagging classifier, XGBoost, GBM, RF are performing better and their scores are approximately 87.5%.

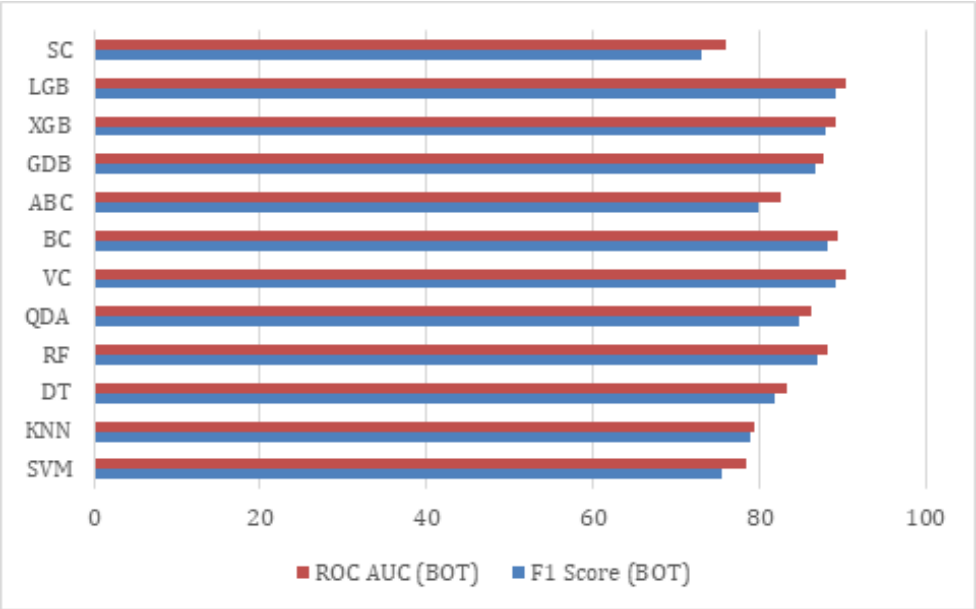


Fig. 3. Comparison of ROC and F1-Score for ML models of all features.

TABLE IV. PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS USING RECURSIVE FEATURE ELIMINATION (RFE)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC (%)
SVM	80.6	78.1	78.1	78.1	80.3
KNN	76.4	71.4	78.1	74.6	76.6
DT	75.0	71.9	71.9	71.9	74.7
RF	88.9	83.3	93.8	88.2	89.4
QDA	70.8	76.2	50.0	60.4	68.8
VC	**93.1**	88.6	**96.9**	**92.5**	**93.4**
BC	86.1	80.6	90.6	85.3	86.6
ABC	87.5	84.8	87.5	86.2	87.5
GDB	86.1	80.6	90.6	85.3	86.6
XGB	87.5	84.8	87.5	86.2	87.5
LGB	87.5	84.8	87.5	86.2	87.5
SC	72.2	68.8	68.8	68.8	71.9

D. Results of Recursive Feature Elimination Method

Among all the features in the present investigation, twenty-four are significant features. AFP, Age, ALB, ALP, AST, CA125, CA19-9, CEA, CO2CP, GLO, HE4, HGB, IBIL, LYM#, LYM%, MCH, Menopause, MPV, Na, PCT, PHOS, PLT, TBIL, TP are considered for classification of OC. It was observed from Table IV that Voting classifier achieved 93.1% accuracy, 88.6% precision, 96.9% recall, 92.5% F1 score, 93.4% ROC AUC among all the ensemble classifiers. Moreover, RF, XGB, LGB have moderate scores. Based on rank important features are identified using RFE are shown in Fig. 4.

E. Results of Boruta Feature Selection Method

In this feature selection method of all features, twenty-one are important features (AFP, Age, ALB, ALP, AST, CA125, CA19-9, CEA, GLO, HE4, IBIL, LYM#, LYM%, MCH, Menopause, MPV, Na, PCT, PLT, TBIL, TP) are considered for classification of OC. It was observed from Table V that the Voting classifier achieved 93.06% accuracy, 88.57% precision, 96.88% recall, 92.54% F1 Score, 93.44% ROC AUC among all the ensemble classifiers. Moreover, RF, AdaBoost, XGBoost have moderate scores approximately 91%. Based on rank important features are identified using Boruta are shown in Fig. 5; moreover, the common features among both approaches are shown in Fig. 6.

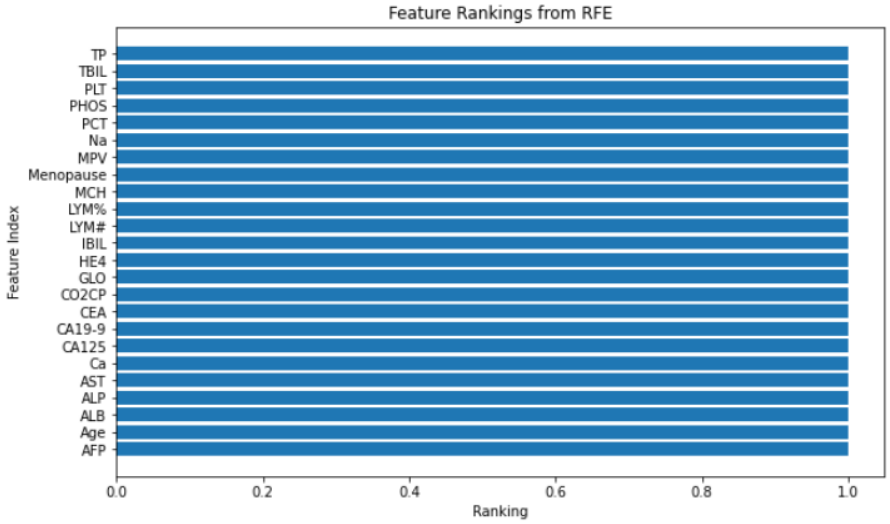


Fig. 4. Important features selected using Recursive Feature Elimination (RFE).

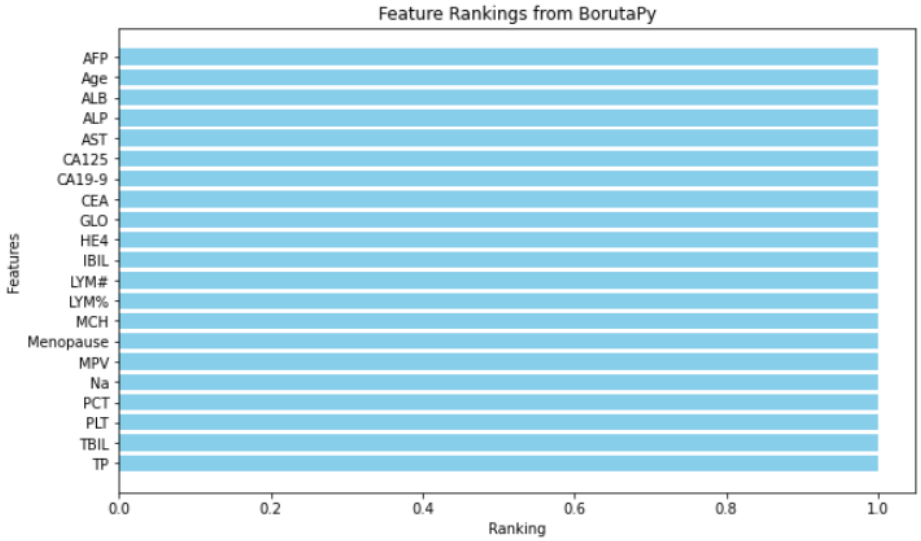


Fig. 5. Important features identified using Boruta.

TABLE V. PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS USING BORUTA FEATURE SELECTION METHOD

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC (%)
SVM	79.17	77.42	75.00	76.19	78.75
KNN	73.61	66.67	81.25	73.24	74.38
DT	80.56	75.00	84.38	79.41	80.94
RF	91.67	88.24	93.75	90.91	91.88
QDA	88.89	92.86	81.25	86.67	88.13
VC	**93.06**	88.57	**96.88**	**92.54**	**93.44**
BC	88.89	85.29	90.62	87.88	89.06
ABC	90.28	87.88	90.62	89.23	90.31
GDB	88.89	85.29	90.62	87.88	89.06
XGB	90.28	**90.32**	87.50	88.89	90.00
LGB	86.11	84.38	84.38	84.38	85.94
SC	81.94	88.00	68.75	77.19	80.63

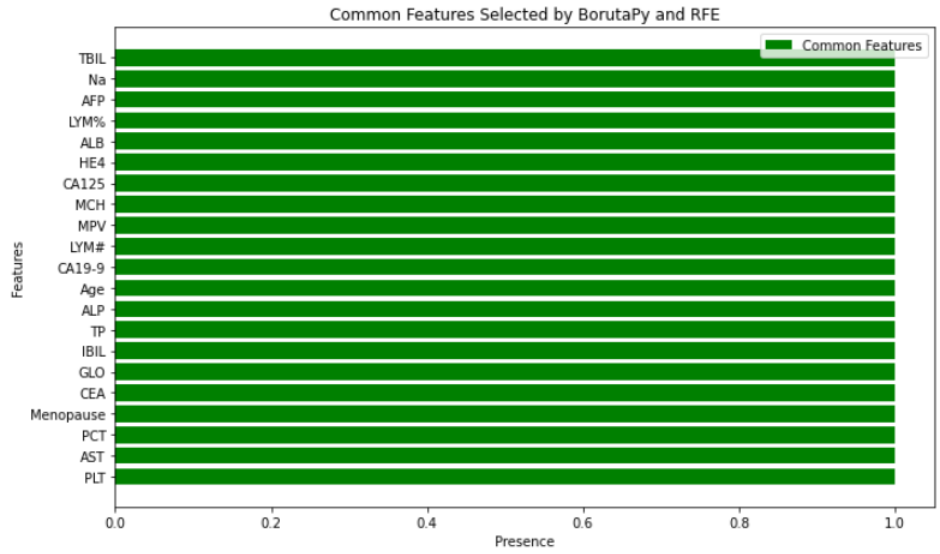


Fig. 6. Common features between Boruta and RFE.

TABLE VI. COMPARISON OF THE PROPOSED MODEL WITH EXISTING SIMILAR WORKS

Ref.ID	Missing Values	Feature Selection	Model	Precision	Recall	F1-Score	ROC-AUC	Accuracy
[17]	Mean	SHAP	GBM+XGB	88.68	88.49	88.69	–	88.24
[5]	Mean	t-test,u-test	RF	83	95	89	87	88
[11]	Mean	11 groups	Group MCP	84.6	88	–	–	93
[18]	Mean	SHAP	LGBM	89	94	91	95	91
Our Method-1	MICE	–	LightGBM	87.9	90.6	89.2	90.3	90.3
Our Method-2	MICE	RFE	VC	88.6	96.9	92.5	93.4	93.01
Our Method-3	MICE	Boruta	VC	88.57	96.88	92.54	93.44	93.06

VI. PERFORMANCE ANALYSIS WITH SIMILAR WORKS

Table VI presents a comparison between our suggested models and some current publications. The table demonstrates how classification methods were used in every study to identify OC. On the other hand, we have used twelve machine learning techniques. Additionally, we have utilized SMOTE and multiple imputation methods to address imbalanced and missing data. We have achieved more than 2.3% of accuracy compared with existing approaches. We have evaluated all the five metrics for judging the model performance compared to the existing approaches.

VII. CONCLUSION

We have developed an ensemble-based method for OC detection in this paper. In this case dataset consists of 349 patients with 49 features where 171 are diagnosed with OC and 178 are diagnosed with benign. To improve the model's performance, twelve ML techniques were employed to categorize OC, and key parameters were chosen by fine-tuning the hyperparameters. Out of twelve algorithms, the voting classifier outperformed the others. To address the missing value and unbalanced data, MICE imputation and borderline SVMSMOTE algorithms were also applied. Two feature selection techniques such as Boruta and RFE have also been used. The Boruta method, which produced an average of 93.44% AUC-ROC, 88.57% precision, 93.06% accuracy, 92.54% F-measure, and 96.88% recall, decreased 50% of the total features. Experiments show that our proposed method can recognize OC more correctly than the state-of-the-art methods.

A. Limitations

1) Limited dataset size: The dataset has a constrained number of samples relative to the complexity of ovarian cancer subtypes. This limits the ability to robustly train and validate the model, especially for rare variants.

2) No Prospective clinical validation: The model was tested on retrospective data only. No prospective testing was conducted to assess clinical usability and patient impact.

B. Future Work

1) Multimodal feature integration: Fuse imaging, genomic, and clinical metadata with biomarker data to improve predictive power and facilitate subtype-specific detection.

2) Explainability benchmarking: Compare SHAP with other interpretability techniques like LIME, Integrated Gradients, and Counterfactual Explanations to assess transparency efficacy across user types (clinicians, data scientists).

3) Personalized risk stratification: Expand the fuzzy system to include personalized thresholds for pre- and post-menopausal patients, incorporating age, hormone levels, and family history.

Data Availability

All the data used in this study is publicly available at https://data.mendeley.com/datasets/th7fztbrv9/11.

Acknowledgment

The authors would like to thank Dr. P. Maragathavalli, the research supervisor, for her continuous guidance, support, and encouragement throughout the course of this study. The authors also acknowledge the open data provided by the Mendeley Data platform, which greatly facilitated this research.

References

[1] P. Wu, Q. Jiang, L. Han, and X. Liu, "Systematic analysis and prediction for disease burden of ovarian cancer attributable to hyperglycemia: a comparative study between China and the world from 1990 to 2019," *Front. Med.*, vol. 10, p. 1145487, 2023, doi: 10.3389/fmed.2023.1145487.

[2] Z. Maleki *et al.*, "Survival rate of ovarian cancer in Asian countries: a systematic review and meta-analysis," *BMC Cancer*, vol. 23, p. 558, 2023, doi: 10.1186/s12885-023-11041-8.

[3] P. Dss Lakshmi Kumari and P. Maragathavalli, "Integrating machine learning algorithms for accurate ovarian cancer diagnosis: a state-of-the-art review," in *2023 IEEE ICACIC*, Puducherry, India, 2023, pp. 1–6, doi: 10.1109/ICACIC59454.2023.10434947.

[4] M. Lu *et al.*, "Using machine learning to predict ovarian cancer," *Int. J. Med. Inform.*, vol. 141, p. 104195, 2020, doi: 10.1016/j.ijmedinf.2020.104195.

[5] Md. M. Ahamad *et al.*, "Early-stage detection of ovarian cancer based on clinical data using machine learning approaches," *J. Pers. Med.*, vol. 12, p. 1211, 2022, doi: 10.3390/jpm12081211.

[6] W. Huang *et al.*, "Explainable discovery of disease biomarkers: the case of ovarian cancer to illustrate the best practice in machine learning and Shapley analysis," *J. Biomed. Inform.*, vol. 141, p. 104365, 2023, doi: 10.1016/j.jbi.2023.104365.

[7] J. Martínez-Más *et al.*, "Evaluation of machine learning methods with Fourier transform features for classifying ovarian tumors based on ultrasound images," *PLoS ONE*, vol. 14, p. e0219388, 2019, doi: 10.1371/journal.pone.0219388.

[8] X. Yang, M. Khushi, and K. Shaukat, "Biomarker CA125 feature engineering and class imbalance learning improves ovarian cancer prediction," in *2020 IEEE CSDE*, Gold Coast, Australia, 2020, pp. 1–6, doi: 10.1109/CSDE50874.2020.9411607.

[9] M. Akazawa and K. Hashimoto, "Artificial intelligence in ovarian cancer diagnosis," *Anticancer Res.*, vol. 40, pp. 4795–4800, 2020, doi: 10.21873/anticanres.14482.

[10] P. Charkhchi *et al.*, "CA125 and ovarian cancer: a comprehensive review," *Cancers*, vol. 12, p. 3730, 2020, doi: 10.3390/cancers12123730.

[11] X. Hu, Y. Xie, Y. Yang, and H. Jiang, "Group penalized logistic regressions predict ovarian cancer," *Preprint*, 2021, doi: 10.21203/rs.3.rs-1172360/v1.

[12] R. M. Ghoniem, A. D. Algarni, B. Refky, and A. A. Ewees, "Multi-modal evolutionary deep learning model for ovarian cancer diagnosis," *Symmetry*, vol. 13, p. 643, 2021, doi: 10.3390/sym13040643.

[13] A. Sorayaie Azar *et al.*, "Application of machine learning techniques for predicting survival in ovarian cancer," *BMC Med. Inform. Decis. Mak.*, vol. 22, p. 345, 2022, doi: 10.1186/s12911-022-02087-y.

[14] F. H. Juwono *et al.*, "Ovarian cancer detection using optimized machine learning models with adaptive differential evolution," *Biomed. Signal Process. Control*, vol. 77, p. 103785, 2022, doi: 10.1016/j.bspc.2022.103785.

[15] K. M. Boehm *et al.*, "Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer," *Nat. Cancer*, vol. 3, pp. 723–733, 2022, doi: 10.1038/s43018-022-00388-9.

[16] M. C. Comes *et al.*, "An explainable machine learning ensemble model to predict the risk of ovarian cancer in BRCA-mutated patients," *Front. Oncol.*, vol. 13, p. 1181792, 2023, doi: 10.3389/fonc.2023.1181792.

[17] N. Abuzinadah *et al.*, "Improved prediction of ovarian cancer using ensemble classifier and Shapley explainable AI," *Cancers*, vol. 15, p. 5793, 2023, doi: 10.3390/cancers15245793.

[18] V. Aelgani and D. Vadlakonda, "Explainable artificial intelligence based ensemble machine learning for ovarian cancer stratification using electronic health records," *Int. J. Res. Inf. Technol. Comput. Commun.*, vol. 11, pp. 78–84, 2023, doi: 10.17762/ijritcc.v11i7.7832.

[19] N. Agarwal, Deepakshi, J. Harikiran, Y. B. Lakshmi, A. P. Kumar, E. Muniyandy, and A. Verma, "Predictive modelling for heart disease diagnosis: a comparative study of classifiers," *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 10, no. 1, p. e5, 2024, doi: 10.4108/eetpht.10.5518.

Comparative Study of Prenatal and Postnatal Images for Detecting Down Syndrome in Children

Labanti Singha[1], Iqbal Ahmed[2]*
M.Sc. Student, Department of Computer Science and Engineering, University of Chittagong, Chittagong, Bangladesh[1]
Professor-Department of Computer Science and Engineering, University of Chittagong, Chittagong, Bangladesh[2]

***Abstract*—Down syndrome is a genetic disorder caused by the presence of an extra copy of chromosome 21, affecting both neurological development and physical features. Early and accurate diagnosis is critical for ensuring timely medical intervention and support. This study presents a comparative analysis of prenatal (ultrasound) and postnatal (facial) imaging modalities for the detection of Down syndrome using deep learning techniques. We employed VGG19, ResNet50, DenseNet121, MobileNetV2, and the Vision Transformer for image classification. An ensemble model integrating four CNN architectures achieved superior performance, with 92% test accuracy on prenatal images and 83% on postnatal images. Among the individual models, ResNet50 outperformed the others across both modalities. Evaluation metrics, including accuracy, precision, recall, and F1-score, confirm the effectiveness of the proposed framework. These results highlight the potential of ensemble learning to enhance the early detection of Down syndrome and improve accessibility to healthcare.**

***Keywords*—Down syndrome; prenatal ultrasound; postnatal facial recognition; CNN; vision transformer; ensemble learning**

I. Introduction

Down syndrome is currently defined as a genetic disorder caused by the presence of an extra chromosome 21, affecting both the physical characteristics and cognitive development of children. However, this disorder was not recognized until 1866 when the English physician John Langdon Down published a detailed description [1] of the condition that would later bear his name. Children affected by Down syndrome often experience academic difficulties, delays in cognitive development, and health complications such as heart disease, hearing loss, and vision impairments. Although there is no cure, early diagnosis, appropriate care, and therapy can significantly improve quality of life. Research into imaging-based detection of Down syndrome is crucial for enabling early diagnosis and enhancing healthcare accessibility. Advances in deep learning and imaging technologies have facilitated the development of cost-effective, noninvasive diagnostic methods. While postnatal imaging [2] captures facial characteristics, prenatal imaging [3]—such as ultrasound and 2D scans—detects structural markers. Combining these two approaches provides a comprehensive diagnostic framework. Motivated by the success of deep learning techniques, we employed transfer learning using models such as ResNet50, VGG19, DenseNet121, MobileNetV2, and the Vision Transformer.

Each of the employed models offers unique advantages; therefore, leveraging an ensemble approach proves beneficial, as it consistently outperforms individual CNN architectures. The key contributions of this study are summarized as follows:

- Evaluation of both ultrasound (prenatal) and facial (postnatal) images to fine-tune models for improved diagnostic accuracy in AI-assisted Down syndrome detection.
- Comparative analysis of two distinct imaging modalities using diverse datasets—1,684 ultrasound images and 2,000 facial images—to enhance generalizability.
- Performance benchmarking of multiple deep learning models across both prenatal and postnatal image datasets.
- Implementation of an ensemble learning strategy that integrates CNN features, resulting in optimized performance and higher accuracy compared to standalone models across both imaging types.

II. Related Works

Several studies have previously explored the prediction and diagnosis of Down syndrome (DS) using a variety of methods, some of which are summarized below:

A study [4] utilized the Random Forest machine learning model to improve the accuracy of DS prediction using second-trimester prenatal images, achieving a detection rate of 66.7% on the training dataset—outperforming traditional approaches. In another study [5], NMR-based metabolomics of maternal serum samples was employed to predict DS. Three key biomarkers were identified, and the inclusion of maternal age as a modifier improved detection efficiency from 48.1% to 51.9%.

Vičić et al. [6] analyzed 157 DS cases out of 6,448 prenatal examinations conducted in Croatia between 2002 and 2014. Key diagnostic markers included maternal age, ultrasound findings, and biochemical screening, while the primary diagnostic procedures were chorionic villus sampling and amniocentesis.

The study [7] introduced CVIFLR, a machine learning model designed for non-invasive DS prediction, and [8] explored deep transfer learning (DTL) in disease diagnosis, achieving a classification accuracy of 93.3%. Thomas et al. [9] used 100 fetal ultrasound images, combining SegNet for nuchal translucency segmentation with AlexNet for classification, resulting in 100% sensitivity, 85.7% specificity, and 91.7% accuracy.

Machine learning models for first-trimester DS screening using clinical data from the UK and Canada were evaluated

*Corresponding authors.

in [10], where deep neural networks (DNNs) outperformed commercial software with an AUC of 0.96 and a 78% detection rate. Similarly, Zhang et al. [11] employed machine learning algorithms for second-trimester prenatal screening, where support vector machines (SVMs) achieved 100% detection with a low false positive rate (FPR).

Another study [12] proposed a 1D convolutional neural network (CNN) model that utilized biomarkers like PAPP-A and hCG along with ultrasound data, reaching a high detection accuracy of 95.17%. Given the ethnic variability of DS features [13], facial recognition tools have shown promise, with one approach achieving 94.3% accuracy. Cornejo et al. [14] developed a novel method using geometric descriptors from facial features, capturing key distances such as intercanthal spacing, nose breadth, and mouth width based on 14 distances from 16 fiducial points.

The Face2Gene AI tool demonstrated strong performance in identifying DS among Thai children [15], with 90% sensitivity for DS cases and 87% for non-DS cases. A convolutional neural network-based method [16] offered a non-invasive solution for DS identification, achieving 99.3% accuracy and outperforming conventional methods like SVM.

In another work [17], extracted image features were used to train a backpropagation neural network (BPNN), which classified nasal bone presence with 86% accuracy. Additionally, a study [18] proposed an expert system framework aimed at enhancing the education of children with DS, while another work [19] introduced a mobile educational platform to support cognitive development through educator-guided learning.

Although not directly related to DS, a study [20] focused on Parkinson's disease—a related neurological disorder—demonstrated the effectiveness of three machine learning approaches for diagnosis. Furthermore, a review [21] provided comprehensive guidelines for the care and development of children with DS. The study [22] emphasized the benefits of early detection and treatment of sleep disorders, such as obstructive sleep apnea (OSA), in DS patients. Finally, [23] demonstrated the use of computer vision and deep learning techniques to analyze parental behaviors, specifically identifying supportive physical interventions during communication with children affected by Down syndrome.

III. Methodology

A systematic method for comparing both prenatal and postnatal images of children with or without Down syndrome is depicted in Fig. 1. This section outlines the methodological pipeline in this study, encompassing data collection, data processing, training various models, model evaluation, and finally the comparison between prenatal and postnatal datasets for detecting Down syndrome in children.

A. Data Collection

Data collection is the initial and most critical step in the proposed methodology. In this study, two distinct datasets are utilized:

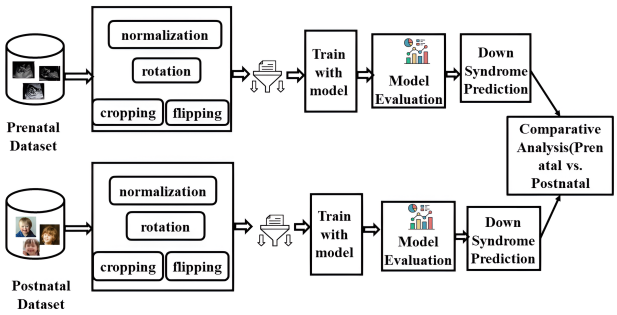


Fig. 1. Proposed methodology for Down Syndrome (DS) detection.

1) Prenatal dataset: This dataset, obtained from Mendeley [https://data.mendeley.com/datasets/n2rbrb9t4f/1], comprises 1,684 ultrasound images of fetuses aged between 11 and 14 weeks. Ultrasound imaging is a standard tool for monitoring fetal development and, in this context, is used to detect potential markers of Down syndrome prior to birth. These markers may include increased nuchal translucency thickness, anomalies in the facial profile, or other structural differences observed in the fetus.

2) Postnatal dataset: Sourced from Kaggle and GitHub [https://github.com/vinayaa1/down-syndrome-detection], this dataset contains 2,000 facial images of infants and children. The majority of samples fall within the 0–15 years age range, with limited representation from 16–18 years. This dataset supports the detection of Down syndrome through facial recognition, leveraging characteristic facial features such as almond-shaped eyes, a flat nasal bridge, and other distinct craniofacial traits. These postnatal images are instrumental in identifying Down syndrome after birth.

Each dataset was split into three subsets: training (80%), validation (10%), and testing (10%). This division ensures unbiased model evaluation and promotes better generalization of the results. It is also important to note that the prenatal and postnatal datasets were sourced independently and do not represent the same children across both stages. The neural network or machine learning model is trained using the training set. Its performance and ability to generalize to new data are then evaluated using the test set. This approach ensures a large and diverse training dataset, a separate set for unbiased evaluation, and the ability to assess the model on unseen samples, simulating potential real-world scenarios.

B. Data Preprocessing

Data processing is a crucial step in data-driven tasks like machine learning, statistical analysis, and decision-making. It transforms raw data into structured, clean, and usable formats, improving data quality and standardizing datasets for machine learning models. Preprocessing steps in Fig. 2 and Fig. 3 include image resizing, normalization, rotation, cropping, flipping, and dataset splitting.

Image resizing ensures uniform input dimensions, while normalization standardizes pixel intensity values across all

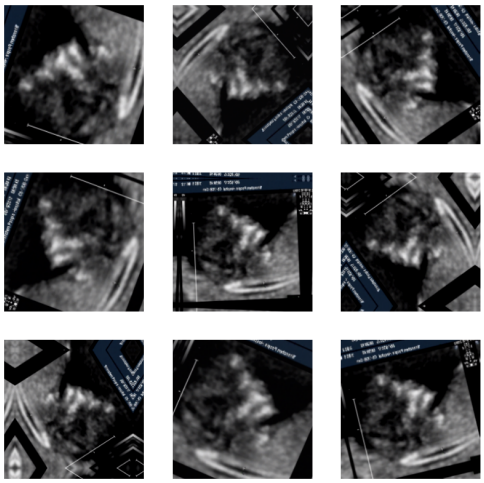


Fig. 2. Data augmentation of prenatal images: Normalized, Rotated, Cropped and Flipped.



Fig. 3. Data augmentation of postnatal images: Normalized, Rotated, Cropped and Flipped.

images. Rotation helps correct misalignments, and cropping reduces background noise, enhancing model efficiency. Flipping augments the dataset by increasing variability, which helps prevent overfitting. Dataset splitting into training, validation, and testing sets enables effective performance evaluation. Collectively, these preprocessing techniques ensure that models learn more effectively from clean and well-prepared input data.

C. Transfer Learning Using Pre-trained CNN Models

This study employs four deep convolutional neural network (DCNN) models—VGG19, ResNet50, DenseNet121, and MobileNetV2—that were pre-trained on the ImageNet dataset and used for feature extraction. Additionally, the Vision Transformer model is utilized to evaluate performance from a transformer-based perspective. These models are applied to both prenatal ultrasound images and postnatal facial images for the detection of Down syndrome in children.

1) VGG19: This 19-layer deep CNN is known for its high accuracy but also for its computational intensity. It uses small 3×3 convolutional filters and ReLU activation. VGG19 is an extended version of VGG16, incorporating three additional layers [24].

2) ResNet50: A 50-layer residual network [25] designed to address the vanishing gradient problem through the use of skip (residual) connections, enabling more effective training of deeper networks.

3) DenseNet121: A densely connected convolutional network with 121 layers, introduced by [26]. In this architecture, each layer receives inputs from all preceding layers, which enhances feature propagation and reduces the number of parameters.

4) MobileNetV2: Designed for mobile and resource-constrained environments, MobileNetV2 uses depthwise separable convolutions to reduce computational cost while maintaining accuracy. It was introduced by [27].

5) Vision Transformer (ViT): Proposed by [28], the Vision Transformer processes images as sequences of patches and utilizes self-attention mechanisms to capture long-range feature dependencies. Unlike CNNs, it leverages transformer-based architectures for image classification.

D. Training and Predicting Models

To obtain the visual Down Syndrome detection results from each model, we extract deep image feature vectors using DenseNet121, VGG19, MobileNetV2, ResNet50, and Vision Transformer. These feature vectors are then passed through the appropriate fully connected dense layers followed by activation layers. The model parameters are optimized using both the Adam optimizer and the stochastic gradient descent (SGD) technique. Training is conducted using the sparse categorical cross-entropy (CE) loss function, shown in (1).

$$CE = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p(y_i) + (1-y_i)\log(1-p(y_i))] \quad (1)$$

E. Proposed Ensemble Model

In order to enhance accuracy, the ensemble combines the outputs of four CNNs (VGG19, ResNet50, DenseNet121, and MobileNetV2). In Fig. 4, the architecture of the model [29] is described. As illustrated in Fig. 4, each CNN model extracts deep features independently. These features are then concatenated and passed to a classifier for prediction. Soft voting combines the probability outputs, and the final prediction is based on the average likelihood. This method enhances generalization and stability across diverse image modalities. These basic steps include the following:

1) Feature extraction: This involves the removal of the last layers of each model and the extraction of deep feature vectors.

2) Concatenation: Combining these characteristics improves categorization accuracy.

3) Prediction aggregation: Accuracy and resilience are increased by using Yes/No Soft Voting, Hard Voting, Stacking, and Bagging.

4) Optimization: To make the entire training process as efficient as possible, use the Adam Optimizer and the categorical cross-entropy loss function.

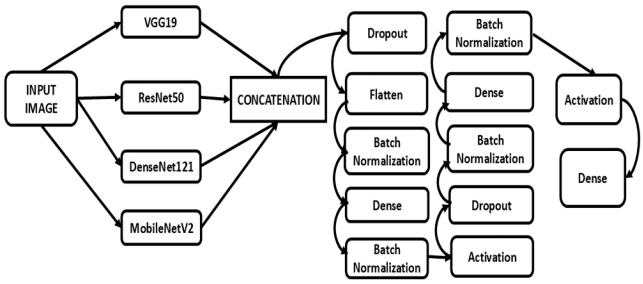


Fig. 4. Architecture of ensemble model.

Model ensembling offers several benefits for improving machine learning performance:

5) Improved generalization: Different models capture diverse features due to their architectural differences. For example, VGG19 is effective for low-level pattern recognition, ResNet50 extracts deep features, DenseNet121 promotes feature reuse and propagation, and MobileNetV2 focuses on efficient feature extraction.

6) Enhanced accuracy: Combining predictions using techniques such as soft voting (averaging probabilities), hard voting (majority voting), or stacking (meta-learning) helps improve overall accuracy.

7) Robustness to data variability: Each model emphasizes different aspects of the data, making the ensemble more robust and better able to generalize across varied datasets.

8) Reduction of overfitting: Using multiple diverse models helps reduce the risk of overfitting by preventing the ensemble from memorizing the training data.

9) Efficient feature learning: The combination of different architectures allows the model to learn features at multiple levels of abstraction, leading to more effective overall feature representation.

In prenatal and postnatal datasets, the ensemble technique ensures robust diagnosis of Down syndrome by enhancing generalization and minimizing overfitting.

F. Performance Measures

The performances are classified into True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Various evaluation metrics are used in this study such as, accuracy, precision, recall, and f1-score.

- Accuracy – Measures the proportion of correct predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (2)$$

- Precision – Measures the percentage of correctly predicted positive cases:

$$Precision = \frac{TP}{TP + FP} * 100 \quad (3)$$

- Recall (Sensitivity) – Measures the ability of the model to find all positive cases:

$$Recall = \frac{TP}{TP + FN} * 100 \quad (4)$$

- F1-Score – Harmonic mean of precision and recall, balancing both metrics:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} * 100 \quad (5)$$

IV. Experiments and Results

In this study, we utilized both prenatal ultrasound images and postnatal facial images of children to detect Down Syndrome. We fine-tuned five pre-trained models—ResNet50, VGG19, DenseNet121, MobileNetV2, and Vision Transformer—using transfer learning. As represented in Table I and Table II, ResNet50 consistently achieved the highest test accuracy across both imaging modalities. Among the CNN models, the highest test accuracy for prenatal images was 91%, while for postnatal images it was 83%. Although the Vision Transformer demonstrated strong performance, it did not surpass ResNet50. To further improve accuracy, we employed model ensembling, which achieved 92% accuracy for prenatal images and 83% for postnatal images.

TABLE I. Accuracy of Different Models (Prenatal Images)

Model	Validation Accuracy	Test Accuracy
ResNet50	90%	91%
VGG19	86%	85%
DenseNet121	84%	88%
MobileNetV2	83%	86%
Vision Transformer	80%	83%
Ensemble	**91%**	**92%**

TABLE II. Accuracy of Different Models (Postnatal Images)

Model	Validation Accuracy	Test Accuracy
ResNet50	80%	**83%**
VGG19	75%	77%
DenseNet121	70%	72%
MobileNetV2	79%	78%
Vision Transformer	72%	73%
Ensemble	**82%**	**83%**

A. Comparison Between the Models for Prenatal and Postnatal Dataset

For the prenatal dataset, we have the classification report for each model, like ResNet50, DenseNet121, MobileNetV2, VGG19, and Vision Transformer model. From all of the classification matrices, we can see that ResNet50 is beating all other models in all contexts, like accuracy, precision and f1-score in Fig. 5.

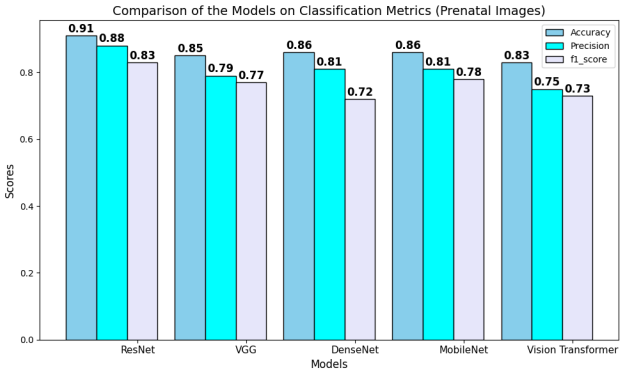


Fig. 5. Comparison chart of the models for prenatal images.

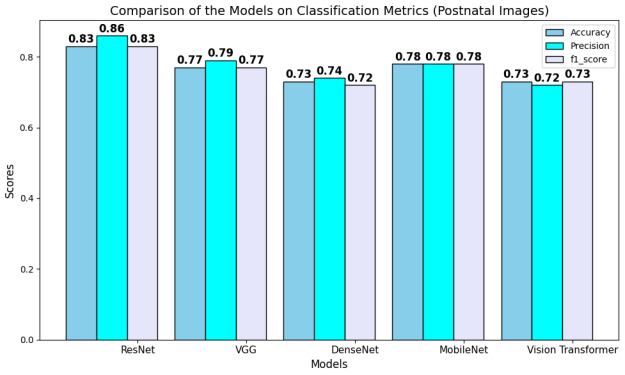


Fig. 6. Comparison chart of the models for postnatal images.

In addition, for the postnatal dataset, we generated classification reports for each model, including ResNet50, DenseNet121, MobileNetV2, VGG19, and the Vision Transformer. Once again, ResNet50 outperforms all other models across all evaluation metrics. Fig. 6 presents a comparative chart of these models based on the prenatal dataset.

B. Performance of the Proposed Ensemble Model for Prenatal Images

1) Accuracy and loss graph: The overall performance based on the training and validation curves is shown in next Fig. 7.



Fig. 7. Performance of the ensemble model (Prenatal images).

It is evident from the graphs that both the training and validation accuracies are close enough, suggesting that the model shows minimal overfitting. Similarly, the loss curve demonstrates a consistent decline for both training and validation sets, with the validation loss remaining low and closely matching the training loss trend. This suggests that the model effectively minimizes error and learns meaningful features from the data with minor overfitting.

2) Confusion matrix: As we know that a confusion matrix helps to evaluate a classification model by presenting true and false predictions for each class. Moreover, this helps to calculate precision, recall and f1-score precisely.

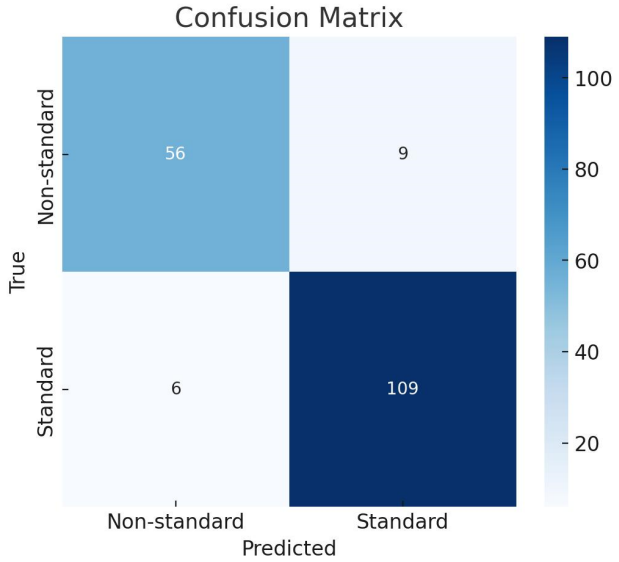


Fig. 8. Confusion matrix of ensemble model (Prenatal images).

Finally, the confusion matrix in Fig. 8 evaluates the performance of the ensemble model for Down Syndrome detection using nuchal translucency (NT) measurements. The model achieves an accuracy of 91.67%, correctly identifying 56 Down Syndrome cases where NT measurements were abnormal. However, it misclassifies nine affected fetuses as normal (false negatives). Additionally, the model correctly identifies 109 normal fetuses with standard NT measurements, while misclassifying six as having Down Syndrome (false positives).

C. Performance of the Proposed Ensemble Model for Postnatal Images

1) Accuracy and loss graph: The line graph in Fig. 9 represents two key plots to evaluate the ensemble model's performance through accuracy and loss curves. The model is shown to perform well with high accuracy in both training and validation with 25 epochs. This also suggests a strong ability to classify with minimal overfitting.

2) Confusion matrix: The effectiveness of an ensemble model for identifying Down syndrome from postnatal images is assessed using this confusion matrix depicted in Fig. 10.

Since there were no false negatives, the model achieved a perfect sensitivity of 100% by properly identifying all 105 newborns with Down syndrome. Nevertheless, it produced a



Fig. 9. Performance of the ensemble model (Postnatal images).

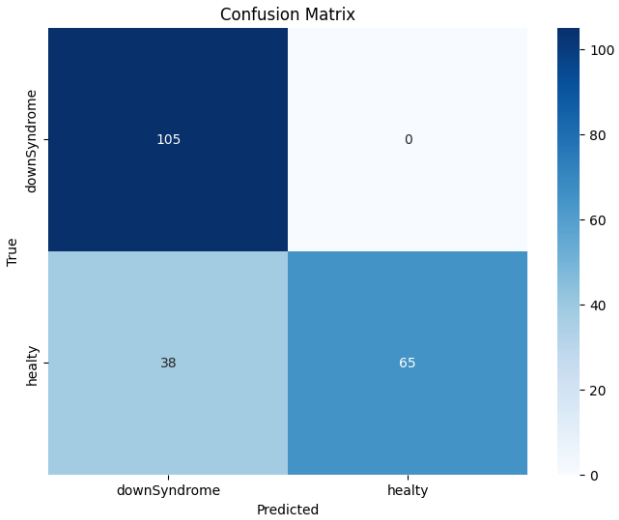


Fig. 10. Confusion matrix of ensemble model (Postnatal images).

false positive rate by incorrectly classifying 28 healthy infants as having Down syndrome. 75 healthy newborns were correctly categorized despite this. Although the model is excellent at identifying cases of Down syndrome without missing any, it may need to be further improved to increase specificity and decrease false alarms because it has a propensity to categorize healthy babies incorrectly.

D. Comparison Between the Ensemble Model for Prenatal and Postnatal Dataset

After doing the ensemble by Feature Fusion Ensemble method with the four models: ResNet50, VGG19, DenseNet121, and MobileNetV2, we have found that the model has increased the performance instantly. The model's accuracy score is almost near to ResNet50 but increased a little bit. Here is a bar graph in Fig. 11, comparing the classification performance between the prenatal and postnatal ensemble models. The prenatal model performs better in all four metrics, indicating its stronger classification capability.

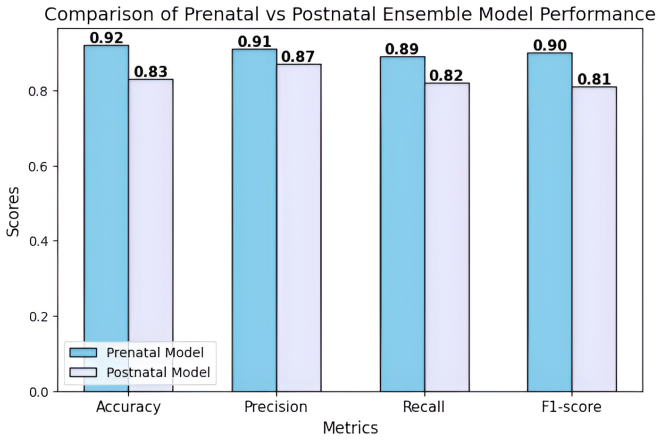


Fig. 11. Comparison chart of the ensemble model (Prenatal vs. Postnatal).

E. Performance Comparison with Existing Research for Down Syndrome Detection

The following Table III represents a comparative analysis of different research studies focused on the detection of Down syndrome using various machine learning models. It includes details on image types, sample sizes, applied models, and their respective accuracies. For example, Reshi, Aijaz Ahmad, et al. [3] used CNN on 1,120 fetus images, achieving an accuracy of 97%. Similarly, Zhao, Qian, et al.[2] applied SVM to facial images (50 DS, 50 healthy) and attained 94% accuracy. The highest reported accuracy (99.3%) was achieved by Feng, Bing, et al. [16], who used a Support Vector Machine (SVM) on SNP maps with a relatively small dataset (315 non-DS, 63 DS cases). This exceptionally high accuracy is likely due to the limited size and controlled nature of the dataset. Unlike some prior studies using smaller or synthetic datasets, our work utilizes publicly available and diverse real-world datasets—namely the Kaggle facial image set and Mendeley ultrasound data—which allows a more robust evaluation. Among those, [3] used the same Mendeley ultrasound dataset, reporting 97% accuracy using CNN, while we achieved 92% through ensemble methods. Moreover, the proposed model in this study processes both ultrasound and facial images, leveraging an ensemble of four CNN models to achieve optimal accuracy—92% for prenatal images and 83% for postnatal images. These results demonstrate that integrating multiple image modalities with deep learning techniques can significantly enhance the accuracy of Down Syndrome (DS) detection.

V. Conclusion

In this study, we explored the effectiveness of different imaging modalities and machine learning techniques for detecting Down Syndrome in children. By analyzing both prenatal (ultrasound) and postnatal (facial) images, our aim was to evaluate the diagnostic potential of each modality individually, as well as their combined efficacy. We investigated the performance of several pre-trained convolutional neural network (CNN) models—ResNet50, VGG19, DenseNet121, and MobileNetV2—on both datasets. In addition, we evaluated the Vision Transformer model for the same task.

TABLE III. COMPARISON OF PREVIOUS STUDIES RELATED TO DOWN SYNDROME

Research	Image Description	Image Samples	Applied Model	Accuracy
[3]	Fetus Image	1120	CNN	97%
[7]	MSS	Non-DownSyndrome: 100,244 DownSyndrome: 108	CVIFLR	95% (AU-CROC)
[2]	Facial Image	DS: 50, Healthy: 50	SVM	94%
[8]	Facial Image	350	DCNN	90%
[17]	Fetus Image	Nasal bone: 50, Without Nasal bone: 50	BPNN	80%
[16]	SNP Maps	Non-DS: 315, DS: 63	SVM	99.3%
[30]	Fetus Image	442	K-means Clustering	87%
Proposed Model	Ultrasound & Facial Images	Ultrasound: 1684, Facial: 2000	CNN & Vision Transformer	Ultrasound: 92%, Facial: 83%

The ensemble of CNN models yielded the highest accuracy, achieving 92% for prenatal ultrasound images and 83% for postnatal facial images, underscoring the advantages of a multimodal approach. Notably, the pre-trained ResNet50 model demonstrated consistently strong performance, achieving a validation accuracy of 89% and test accuracy of 91% on the prenatal dataset, and a validation accuracy of 81.5% and test accuracy of 83% on the postnatal dataset.

Compared to existing studies that primarily focus on either fetal ultrasound or facial imagery using traditional machine learning methods, our approach leverages deep learning to enhance classification performance. Future research should focus on expanding dataset size, incorporating more advanced imaging features, and developing hybrid AI architectures to improve accuracy and generalizability. This includes integrating data from diverse demographics, combining prenatal and postnatal imaging, and utilizing multimodal data fusion. Furthermore, future models could benefit from hybrid AI architectures that combine CNNs, Vision Transformers, and traditional machine learning techniques to enhance feature extraction and ensemble learning.

REFERENCES

[1] A. D. Ataman, E. E. Vatanoğlu-Lutz, and G. Yıldırım, "Medicine in stamps: history of down syndrome through philately," *Journal of the Turkish German Gynecological Association*, vol. 13, no. 4, p. 267, 2012.

[2] Q. Zhao, K. Rosenbaum, K. Okada, D. J. Zand, R. Sze, M. Summar, and M. G. Linguraru, "Automated down syndrome detection using facial photographs," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 3670–3673.

[3] A. A. Reshi, S. Shafi, I. Qayoom, M. Wani, S. Parveen, and A. Ahmad, "Deep learning-based architecture for down syndrome assessment during early pregnancy using fetal ultrasound images," *Int. J. Exp. Res. Rev*, vol. 38, pp. 182–193, 2024.

[4] F. He, B. Lin, K. Mou, L. Jin, and J. Liu, "A machine learning model for the prediction of down syndrome in second trimester antenatal screening," *Clinica Chimica Acta*, vol. 521, pp. 206–211, 2021.

[5] R. O. Bahado-Singh, R. Akolekar, R. Mandal, E. Dong, J. Xia, M. Kruger, D. S. Wishart, and K. Nicolaides, "Metabolomic analysis for first-trimester down syndrome prediction," *American journal of obstetrics and gynecology*, vol. 208, no. 5, pp. 371–e1, 2013.

[6] A. Vičić, T. Hafner, I. B. Vlatković, P. Korać, D. Habek, and F. Stipoljev, "Prenatal diagnosis of down syndrome: A 13-year retrospective study," *Taiwanese Journal of Obstetrics and Gynecology*, vol. 56, no. 6, pp. 731–735, 2017.

[7] L. Li, W. Liu, H. Zhang, Y. Jiang, X. Hu, and R. Liu, "Down syndrome prediction using a cascaded machine learning framework designed for imbalanced and feature-correlated data," *IEEE Access*, vol. 7, pp. 97 582–97 593, 2019.

[8] B. Jin, L. Cruz, and N. Gonçalves, "Deep facial diagnosis: deep transfer learning from face recognition to facial diagnosis," *IEEe Access*, vol. 8, pp. 123 649–123 661, 2020.

[9] M. C. Thomas and S. P. Arjunan, "Deep learning measurement model to segment the nuchal translucency region for the early identification of down syndrome," *Measurement Science Review*, vol. 22, no. 4, pp. 187–192, 2022.

[10] A. Koivu, T. Korpimäki, P. Kivelä, T. Pahikkala, and M. Sairanen, "Evaluation of machine learning algorithms for improved risk assessment for down's syndrome," *Computers in Biology and Medicine*, vol. 98, pp. 1–7, 2018.

[11] H.-G. Zhang, Y.-T. Jiang, S.-D. Dai, L. Li, X.-N. Hu, and R.-Z. Liu, "Application of intelligent algorithms in down syndrome screening during second trimester pregnancy," *World Journal of Clinical Cases*, vol. 9, no. 18, p. 4573, 2021.

[12] D. Wright, I. Bradbury, F. Malone, M. D'Alton, A. Summers, T. Huang, S. Ball, A. Baker, B. Nix, D. Aitken *et al.*, "Cross-trimester repeated measures testing for down's syndrome screening: an assessment." *Health Technol Assess*, vol. 14, no. 33, p. 1, 2010.

[13] P. Kruszka, A. R. Porras, A. K. Sobering, F. A. Ikolo, S. La Qua, V. Shotelersuk, B. H. Chung, G. T. Mok, A. Uwineza, L. Mutesa *et al.*, "Down syndrome in diverse populations," *American Journal of Medical Genetics Part A*, vol. 173, no. 1, pp. 42–53, 2017.

[14] J. Y. R. Cornejo, H. Pedrini, A. Machado-Lima, and F. d. L. d. S. Nunes, "Down syndrome detection based on facial features using a geometric descriptor," *Journal of Medical Imaging*, vol. 4, no. 4, pp. 044 008–044 008, 2017.

[15] N. Vorravanpreecha, T. Lertboonnum, R. Rodjanadit, P. Sriplienchan, and K. Rojnueangnit, "Studying down syndrome recognition probabilities in thai children with de-identified computer-aided facial analysis," *American Journal of Medical Genetics Part A*, vol. 176, no. 9, pp. 1935–1940, 2018.

[16] B. Feng, D. C. Samuels, W. Hoskins, Y. Guo, Y. Zhang, J. Tang, and Z. Meng, "Down syndrome prediction/screening model based on deep learning and illumina genotyping array," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 347–352.

[17] T. Anjit and S. Rishidas, "Identification of nasal bone for the early detection of down syndrome using back propagation neural network," in *2011 International conference on communications and signal processing*. IEEE, 2011, pp. 136–140.

[18] C. VENEGAS, R. Cevallos, R. Cordova, D. D. L. CRUZ, J. TOBAR, and P. MEJÍA, "Lukánikas: Educational platform for children with down syndrome manageable by the educator," in *IAENG TRANSACTIONS ON ENGINEERING SCIENCES: Special Issue for the International Association of Engineers Conferences 2016*. World Scientific, 2018, pp. 543–555.

[19] V. Robles-Bykbaev, E. Andrade-Prieto, P. Solorzano-Guerrero, Y. Robles-Bykbaev, F. Pesántez-Avilés, and A. Parra-Astudillo, "An educational support tool based on robotic assistants, mobile apps, and expert systems for children with down syndrome," in *2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*. IEEE, 2018, pp. 1–4.

[20] R. S. El-Sayed, "A hybrid cnn-lstm deep learning model for classification of the parkinson disease." *IAENG International Journal of Applied Mathematics*, vol. 53, no. 4, 2023.

[21] M. E. Weijerman and J. P. De Winter, "Clinical practice: The care of children with down syndrome," *European journal of pediatrics*, vol. 169, pp. 1445–1452, 2010.

[22] B. Fauroux, S. Sacco, V. Couloigner, A. Amaddeo, A. Ravel, E. Prioux, J. Toulas, C. Cieuta-Walti, H. Walti, R. Luscan *et al.*, "Early detection and treatment of obstructive sleep apnoea in infants with down syndrome: a prospective, non-randomised, controlled, interventional study," *The Lancet Regional Health–Europe*, vol. 45, 2024.

[23] C. R. Galindo-López, J. Beltrán, C. B. Pérez, A. Macias, and L. A. Castro, "Classifying interactions of parents and children with down syndrome in educational environments using deep learning," in *2022 IEEE Mexican International Conference on Computer Science (ENC)*. IEEE, 2022, pp. 1–8.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[27] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] M. Mujahid, A. Rehman, T. Alam, F. S. Alamri, S. M. Fati, and T. Saba, "An efficient ensemble approach for alzheimer's disease detection using an adaptive synthetic technique and deep learning," *Diagnostics*, vol. 13, no. 15, p. 2489, 2023.

[30] S. Ramanathan, M. Sangeetha, S. Talwai, and S. Natarajan, "Probabilistic determination of down's syndrome using machine learning techniques," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2018, pp. 126–132.

Sign3DNet: An Enhanced 3D CNN Architecture for Bengali Word-Level Sign Language Recognition

Safi Ullah Chowdhury, Nasima Begum*, Tanjina Helaly, Rashik Rahman
Department of Computer Science and Engineering,
University of Asia Pacific, Dhaka, Bangladesh

***Abstract*—Automated recognition of sign languages has been playing an important role in breaking barriers to communication and inclusion for the deaf and mute community. Several studies have been conducted on Bengali Sign Language (BdSL). However, Bengali Word-Level Sign Language (BdWLSL) remains unexplored due to the lack of large annotated datasets and a stable model. Therefore, in this research, we introduced a large-scale Bengali word-level video dataset and proposed a modified 3D Convolutional Neural Network (CNN) architecture for word-level BdSL recognition, emphasizing its ability to capture the spatial and temporal dynamics from video data. The proposed strategy represents strong performance in Bengali word-level sign language recognition by utilizing the spatio-temporal pattern captured by the modified 3D CNN architecture. The proposed model demonstrates its potential for practical use by successfully learning complex hand movements straight from raw video data. The proposed CNN model is benchmarked against traditional deep learning techniques, Temporal Shift Module (TSM), Long Short-Term Memory (LSTM), and default 3D-CNN, providing a comprehensive comparison of their strengths and limitations. Experiments are conducted using a structured video dataset containing 102 Bengali sign-word classes. To ensure privacy, the volunteers' faces were blurred and only landmark data extracted using MediaPipe, rendered on black backgrounds, were used for training. The experimental result analysis shows that the performance of the proposed 3D-CNN model achieves a satisfactory accuracy of 58.25%, demonstrating its potential for word-level sign language recognition tasks. To our knowledge, this is the very first pilot study for BdWLSL recognition. Hence, we consider the recognition rate 58.25% of the proposed modified 3D-CNN architecture to be satisfactory and a potential scope for future researchers in the same field.**

***Keywords*—Bengali sign word recognition; computer vision; deep learning; convolutional neural network; spatial-temporal dynamics; video data**

I. Introduction

In a world where effective communication is at the heart of human connection, the deaf and mute community faces unique and formidable challenges. The limited accessibility of their distinct sign language has led to profound difficulties in expressing thoughts, feelings, and ideas. Bridging this communication gap is fundamental. Automatic interpretation of sign languages is difficult compared to spoken languages since it consists of signed motions, as well as hand gestures and facial expressions. Thus, although word-level sign language recognition is a well-established field for sentence-level recognition, it presents several difficulties specifically in languages different from ASL (American Sign Language) and BSL (British Sign Language). For example, there is limited awareness of sign languages like Bengali Sign Language (BdSL); the lack of a substantial volume of high-quality annotated datasets and strong models to capture both local and global hand movements contributes to this research.

Previous works ASL [1] have emphasized the requirement for large word-level data sets coupled with enhanced deep learning (DL) models to handle the inherent spatio-temporal characteristics in sign recognition. With advancements in computer vision and deep learning, specially 3D Convolutional Neural Network (CNNs), it is now possible to capture both spatial and temporal dynamics in sign laguage data, which helps to recognize complex gestures across video frames. However, 3D CNNs require extensive datasets and multidimensional data representations, such as 3D joint positions, which are not always possible for sign languages, including BdSL. To address this issue, this research proposes a new architecture for word-level sign language recognition. It is trained and tested on a structured dataset with 102 sign classes, prepared using MediaPipe landmarks to focus only on key gesture components.

Additionally, we compared the performance of three distinct models which are: (1) Modified 3D CNN that directly incorporates entire video frames to efficiently capture both spatial and temporal features; (2) an LSTM model that utilizes sequential data to recognize temporal patterns within sign gestures; and (3) TSM, efficient in capturing the temporal dependencies through frame feature shifting for sign word classification. Each model comes with some strengths and weaknesses when used in the analysis of sign language data outcomes of each model are measured based on the accuracy, precision, recall, F1-score and mAP (Mean Average Precision).

The main contributions of this research are as follows:

- We have developed a large-scale, annotated BdSL sign word-level and sentence-level video dataset for deep learning (DL) applications.
- We developed a modified robust 3D-CNN architecture after adjusting hyperparameters called Sign3DNet to increase the generalization property of the system. To our knowledge, the proposed Sign3DNet is the first pilot architecture for recognizing Bengali Sign Words.
- Finally, we evaluated the performance of the proposed Sign3DNet architecture for our developed dataset, and other existing datasets.
- We also demonstrated the standardization of our developed dataset across three model types (proposed

*Corresponding authors.

modified 3D CNN, LSTM, and TSM), providing an extensive evaluation.

The remainder of this paper is structured as follows: Section II represents the literature review. The dataset description is provided in Section III. The proposed methodology is described in Section IV. Section V demonstrates the analysis of the experimental results. Lastly, Section VI concludes the paper with some future works.

II. Literature Review

Sign language recognition has been vital in the reduction of communicative barriers for the Deaf and hard of hearing people, especially with word and sentence recognition. The recognition of sign language can be widely categorized into two types, namely the static and dynamic gesture recognition. Static gestural phrases are used to depict alphabets and digits, while dynamic gestural phrases can depict words or sentences at one go and hence need sequential recognition of gestures. Sign languages like American Sign Language (ASL) [1] have well-documented datasets, such as Purdue RVL-SLLL ASL Database [2], Boston ASLLVD [3], and RWTH-BOSTON-50 [4], which provide valuable data for various recognition tasks. However, the involved datasets fail to sufficiently cover different signers and instances, which becomes a drawback in terms of large-scale classification for achieving higher vocabulary. Large datasets for other languages like DEVISIGN [5] for Chinese Sign Language have been developed to address similar limitations by including thousands of signs performed by multiple signers, contributing to more robust recognition frameworks.

Existing sign language recognition approaches are typically centered around three phases, such as feature extraction, temporal modeling, and classification. Classical approaches used some hand-made features including motion/sensor-based features such as SIFT [6], HOG [7] and/or adopting frequency domain [8] to represent the hand poses and then using HMMs [9] or DTW [10] to model temporal components. In recent years, CNNs and 3D CNNs have been used widely to solve spatial and temporal data at the same time. However, there is still an issue associated with the application of the mentioned methods – the limited size and quality of the databases. Thus, reliable assessment and, particularly, transfer of models to new sign languages continues to be problematic, particularly for low-resource sign languages without large annotated datasets.

For Bengali Sign Language (BdSL) research is still in its infancy and mainly involves only static gesture recognition. First data sets are given with limited alphabets and digits of Bangla such as Ishara-Lipi [11] and Ishara-Bochon [12]. Following investigations used deep learning-based image recognition approaches like YOLO, CNN, and AlexNet [13], [14] on static BdSL data to categorize isolated words. Relatively little has been done in the way of developing dynamic word-level BdSL (WL-BdSL) datasets and recognition. Previous studies of dynamic BdSL are limited to a small set of static signs and do not reflect continuous word-level recognition issues. In ASL, the current state of static BdSL recognition with large vocabularies has recently used MediaPipe keypoint extraction, but dynamic video recognition for wider vocabularies has not been studied thoroughly.

In the case of Bengali Sign Language (BdSL), studies are still very limited, and early studies' main concern is on recognition of static gestures only. There are few datasets for letter-level and static word recognition in Bangla such as BornoNet [15], BDSL49 [16], Ishara-Lipi [11], and BenSignNet [17]. More specifically, it is recognized that Ishara-Lipi is a benchmark dataset for sign language that is described by the use of modern methods involving two-handed signs. This means that while BornoNet is particularly focused on the letter-level classification the dataset BDSL49 offers a large amount of Bangla characters. BenSignNet proposed a new CNN model that is trained on three different sets of BdSL, such as BdSL, Ishara-Lipi, and KU-BDSL [18] with differing performances.

However, efforts toward dynamic, word-level BdSL (WL-BdSL) recognition remain limited. Most of the work done in this area pertains to very small, arty static sign sets that do not suffice for the sort of word-level continuous reading tasks with larger, open vocabularies. However, despite recent attempts to adapt MediaPipe for static BdSL keypoint extraction, there is a lack of study regarding its application for dynamic video recognition. This gap points to an urgent need to work on the advances in the comprehensive ample of datasets and the state of the art of the recognition models addressing the issues of temporal and spatial dynamics in the continuous BdSL. To overcome these imperatives, we have developed a novel Bengali Sign Language dataset for dynamic word-level recognition, consisting of 102 emergent Bengali Sign words and including more than 30600 samples captured from various viewing angles.

III. Dataset Description

Bengali language has about one million words, and mute and deaf people use those sign words for their daily communication. Therefore, a video dataset is essential for Bengali Word Level Sign Language (BdWLSL) recognition. However, to our knowledge, there are no such video datasets for Bengali Sign Word except [19]. Thus, in this research, we developed a structured BdWLSL video dataset for our recognition task. The data set is used to support advanced deep learning applications for continuous sign language recognition. This dataset also solves the problem of limited large-scale BdWLSL dataset by providing a wide variety of video material to capture the details of BdWLSL gestures with 102-word classes. Each label in Table I represents a Bengali word. The data set includes various video samples that capture variability in signers, lighting, background, and angles as shown in Fig. 1. All videos used in training were collected with proper consent. For privacy, we applied automated face blurring to each frame. We then extracted hand and facial landmarks using MediaPipe. These landmarks were projected over a black background, and only these versions were used for model training. This ensured that the model learned from gestures, not identity or background noise.

Our developed dataset has different variations. For example, videos are captured by different signers at different angles and various lighting conditions. After recording all the videos, we used MediaPipe to generate landmarks for the hand and face regions. At the same time, we removed the raw versions of the videos and used a black background with the landmarks to construct the data. Fig. 2 shows the approaches which

TABLE. I. BENGALI WORD-LEVEL SIGN LANGUAGE (BDWLSL) CLASSES

Label	Word	Label	Word	Label	Word	Label	Word
0	পরিবার (Family)	26	কফি (Coffee)	52	টেলিভিশন (Television)	78	পেশা (Occupation)
1	সম্পর্ক (Relation)	27	রুটি (Bread)	53	ক্রিকেট (Cricket)	79	গৃহিণী (Housewife)
2	বাবা (Father)	28	সবজি (Vegetable)	54	ফুটবল (Football)	80	স্যার (Sir)
3	মা (Mother)	29	ডাল (Dal)	55	হাঁটা (Walking)	81	ম্যাডাম (Madam)
4	ভাই (Brother)	30	মিষ্টি (Sweet)	56	কথা বলা (Talking)	82	ডাক্তার (Doctor)
5	বোন (Sister)	31	পানি (Water)	57	ঘুমানো (Sleeping)	83	শ্রমিক (Labour)
6	স্বামী (Husband)	32	দুধ (Milk)	58	আসা (Coming)	84	উকিল (Lawyer)
7	স্ত্রী (Wife)	33	আম (Mango)	59	যাওয়া (Going)	85	পুলিশ (Police)
8	চাচা (Paternal Uncle)	34	কাঁঠাল (Jackfruit)	60	দেখা (Looking)	86	সাংবাদিক (Journalist)
9	মামা (Maternal Uncle)	35	আপেল (Apple)	61	বিভাগ (Division)	87	প্রকৌশলী (Engineer)
10	খালা (Maternal Aunt)	36	কমলা (Orange)	62	ঢাকা (Dhaka)	88	বাংলা (Bangla)
11	ফুফু (Paternal Aunt)	37	কলা (Banana)	63	চট্টগ্রাম (Chattogram)	89	ইশারা (Sign)
12	বন্ধু (Friend)	38	আঙ্গুর (Grape)	64	সিলেট (Sylhet)	90	ভাষা (Language)
13	কালো (Black)	39	পেপে (Papaya)	65	খুলনা (Khulna)	91	আমি (Me)
14	সাদা (White)	40	স্কুল (School)	66	রজশাহী (Rajshahi)	92	তুমি/আপনি (You)
15	লাল (Red)	41	কলেজ (College)	67	রংপুর (Rangpur)	93	আমার (My)
16	নীল (Blue)	42	বিশ্ববিদ্যালয় (University)	68	ময়মনসিংহ (Maymenshing)	94	আমাদের (Our)
17	হলুদ (Yellow)	43	অফিস (Office)	69	দিন (Day)	95	হাজার (Thousand)
18	সবুজ (Green)	44	কাজ (Work)	70	শনিবার (Saturday)	96	লাখ (Million)
19	খাদ্য (Food)	45	কম্পিউটার (Computer)	71	রবিবার (Sunday)	97	ছেলে (Boy)
20	ফলমূল (Fruits)	46	মোবাইল (Mobile)	72	সোমবার (Monday)	98	মেয়ে (Girl)
21	ভাত (Rice)	47	ইন্টারনেট (Internet)	73	মঙ্গলবার (Tuesday)	99	রোজা (Fasting)
22	মাছ (Fish)	48	বই (Book)	74	বুধবার (Wednesday)	100	ঈদ (Eid)
23	মাংস (Meat)	49	কলম (Pen)	75	বৃহস্পতিবার (Thursday)	101	কোরবানি ঈদ (Eid-Al-Adha)
24	মুরগি (Chicken)	50	বাতি (Light)	76	শুক্রবার (Friday)		
25	চা (Tea)	51	পাখা (Fan)	77	পেশা (Occupation)		

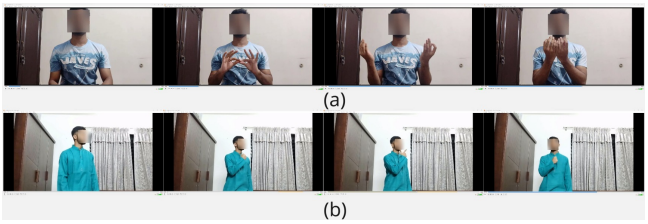


Fig. 1. Sample of Bengali word-level video data (a: পরিবার (Family); b: বাবা (Father)).

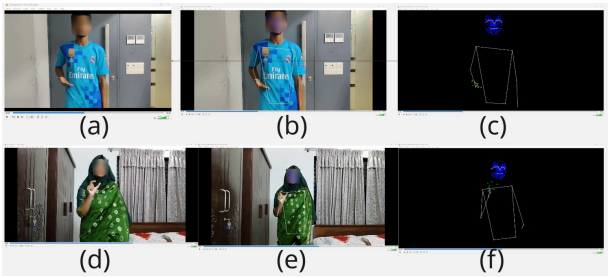


Fig. 2. Sample word data directory (a-c: বাংলা (Bangla), d-f: ভাষা (Language)).

is used to understand the condition of the dataset in various scenarios. To maintain volunteer privacy, all facial features were blurred during preprocessing using automated detection methods. For training, only the background-removed landmark videos were used. The dataset includes signers from diverse backgrounds to improve inclusivity and generalization performance. Table II summarizes the attributes of our developed dataset, providing a detailed breakdown of its structure and contents. In total, the dataset expands to 91,800 samples.

As illustrated in Fig. 3, each frame used for training was processed by first applying face blurring for privacy, followed by landmark extraction using MediaPipe. The final training samples consisted of black-background videos containing only the key landmark features of the sign gestures. Each word-level sign sample is labeled with the appropriate Bengali words. In each class, there are approximately 300 videos and they are divided into three parts: the first part is for training, the second one is for validation and the last part is for testing. The train, validation, and testing datasets contain 75%, 15%, and 15% of the total dataset, respectively. Table III shows the data distribution between the train , validataion and test set of the BdWLSL dataset. As shown in the Table III , there are around 210 video samples in the train set and 30 samples in the validation and test set of each class. All video frames were resized to 64x64 using OpenCV's bilinear interpolation method. This dimension was chosen based on a balance between visual fidelity of hand gestures and computational

TABLE. II. ATTRIBUTES FOR PROPOSED DATASET

Attribute	Value
No. of cameras	20+
Camera position	Multiple Viewing Angle (front, left, right)
No. of signers	39
No. of sign words	102
No. of word videos/class	300
Types of video	3 (Raw, Landmark with Background, Landmark without Background)
Total No. of Videos (Word)	91,800
File Type	mp4
Length/video length (sec)	1-2

efficiency. Larger resolutions led to increased training cost with negligible improvement in performance.

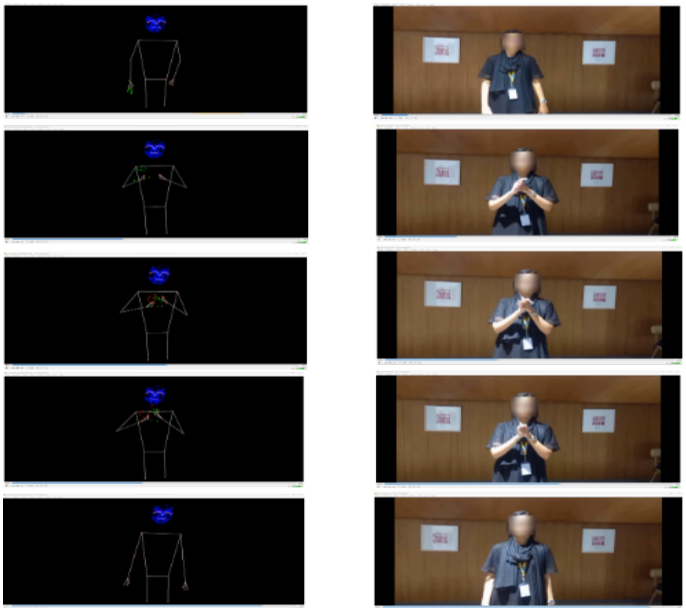


Fig. 3. Sample of the training data showing (left) the blurred original frame and (right) the extracted landmark frame on a black background for the sign "বন্ধু" (Friend).

TABLE. III. CLASS WISE DATA DISTRIBUTION

Label	Train	Test	Validation
0	210	30	30
1	210	30	30
2	210	30	30
3	210	30	30
...	...	...	...
101	210	30	30

IV. METHODOLOGY

This section describes the proposed methodology for word-level video sign language recognition. In this work, we introduce a novel CNN model specifically designed for word-level video sign language recognition, which is described in detail in Section IV(A). Fig. 4 shows the workflow of the proposed methodology.

To establish a baseline, we trained a few deep learning models on our developed dataset, including Standard 3D CNN, TSM, and LSTM. Among them, the LSTM model performed very well, achieving a moderate accuracy. However, the LSTM model is primarily suited for temporal data, whereas video data inherently possesses both spatio-temporal features. The TSM model is concentrated on the temporal dependencies through the shifting between feature maps in frames, losing its spatial information in video frames. On the other hand, the standard 3D CNN is designed to handle both spatio-temporal patterns directly. Therefore, to maximize performance for this application, we concentrated on modifying the 3D CNN architecture.

A. Proposed Modified 3D CNN Architecture

The proposed 3D CNN model is optimized for spatio-temporal feature extraction while balancing computational efficiency. The architecture employs the Depthwise Convolution and BatchNormalization techniques, along with the ReLU activation function, at multiple locations, as illustrated in Fig. 5. The model is fed with input videos, each one with 20 frames at 64×64 px grayscale. Convolution is done through the three Conv3D layers initially, extracting spatial and temporal patterns over spatial dimensions (x, y) and temporal (t) using the 3D kernels. Mathematically, the 3D convolution operation at the position (x, y, t) is defined as follows:

$$F(x,y,t) = \sum_{i=0}^{K-1}\sum_{j=0}^{K-1}\sum_{k=0}^{K-1} W(i,j,k)\cdot I(x+i,y+j,t+k) \quad (1)$$

Where $W(i, j, k)$ represents the 3D convolution kernel weights, $I(x + i, y + j, t + k)$ denotes the input video frame at spatial position (x, y) where t, is time and K is the kernel size. To stabilize training and accelerate convergence, batch normalization is applied after each convolutional layer, normalizing activations by Eq. (2)-(3) as follows:

$$\hat{x} = \frac{x - \mu}{\sigma} \quad (2)$$

$$y = \gamma\hat{x} + \beta \quad (3)$$

Here, the learnable parameters γ, β, μ and σ represent the mean and standard deviation of the batch. The MaxPooling3D layers are used to downsample the overall dimension of the model for efficient computation. MaxPooling3D was selected for its capability to retain the most prominent features while reducing computational load. Compared to Average pooling and L2 pooling, MaxPooling showed higher validation accuracy in early experiments on this dataset. The last one is a Global Average Pooling (GAP) layer to reduce input spatial dimensions by averaging feature maps to generate a fixed-length feature vector. The feature relationships within this vector are then captured by fully connected layers with ReLU activation. Finally, classification is performed using a

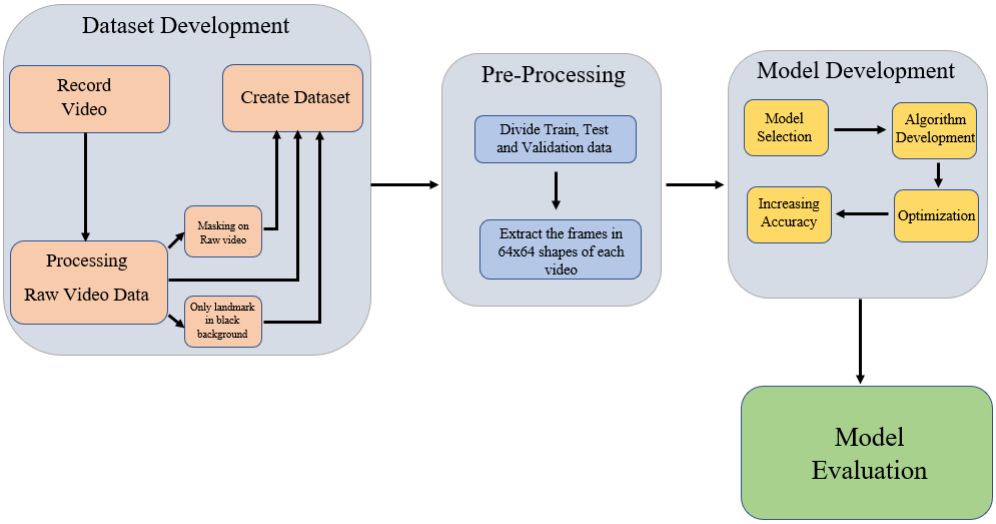


Fig. 4. Overall flow-diagram of proposed methodology.

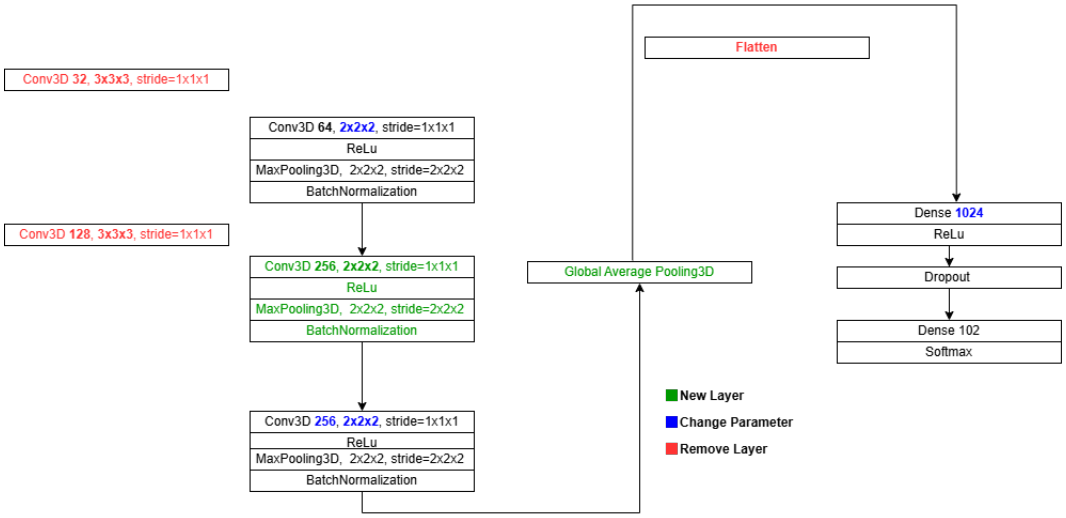


Fig. 5. Architectural modification of 3D CNN.

dense layer with a softmax activation function, converting the network output into class probabilities:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{4}$$

where $P(y_i)$ is the probability of class i, z_i is the output score for class i, and N represents the number of classes. Softmax activation function was used in the final layer to output a probability distribution over the 102 sign classes. This allows for intuitive interpretation of prediction confidence and is suitable for multi-class classification tasks. This allows for intuitive interpretation of prediction confidence and is suitable for multi-class classification tasks.

The network is trained using the categorical cross-entropy loss function:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{5}$$

Specifically, if y_i denotes the ground truth label (and 1 if the class is the correct class, and 0 otherwise) and $\hat{y}_i$ is the predicted probability. Model parameters are updated using the Adam optimizer with adaptive learning rates for improved convergence. The layers and their parameters are described in Table IV.

A 13-layer modified 3D CNN architecture is proposed where the computational cost is optimized for Bengali Word-level Sign Language (BdWSL) recognition using computational efficiency and strong recognition capabilities together. The proposed model greatly improves the classification performance by integrating Conv3D, batch normalization, pooling,

TABLE. IV. MODIFIED 3D CNN MODEL LAYERS AND THEIR PARAMETERS

Level	Name of the Layer	Parameters	Number of Layers
1	Input3D	-	1
2	Conv3D	Kernel size: 3x3x3 Filters: 64, 256, 256 Activation: Leaky ReLU	3
3	MaxPooling3D	Pool size: 2x2x2	3
4	BatchNormalization	-	3
5	GlobalAveragePooling3D	-	1
6	Dense	Units: 1024, Number of Classes Activation: ReLU, Softmax	2
Total Layers			**13**

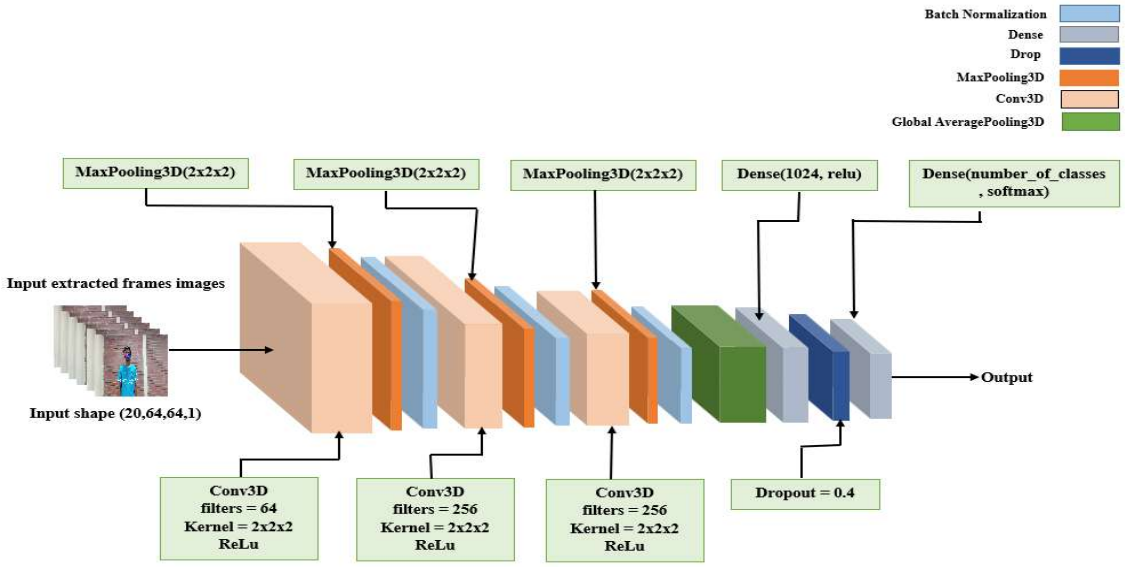


Fig. 6. Modified 3D CNN architecture.

and GAP, which allows the model to capture and process spatio-temporal features. An overview of the proposed 3D CNN architecture is illustrated in Fig. 6.

B. Experimental Setup

The 3D CNN model is deployed with modifications to effectively capture the temporal and spatial dynamics and details of the sign gestures. The model is optimized using the Adam optimizer. In 3D CNN, a sequential Conv3D architecture is used during the study, and involves down-sampling layers, batch normalization, and global pooling for feature extraction for 3D CNN. The model is trained using the Adam optimizer with a learning rate of 0.002. The loss function categorical cross-entropy is used, which is suitable for multiclass classification tasks.

For consistency across the model, the video frames are normalized with an equivalent diagonal bounding box of 64 pixels in standardized subsets of frames that focus on the signer's areas of interest, where applicable. All frames for a video are prepared by applying spatial-temporal augmentation. The complete setup of the experimental environment is summarized in Table V.

TABLE. V. EXPERIMENTAL ENVIRONMENT SETUP

Component	Specification
Operating System	Windows 11
GPU	NVIDIA RTX 4090 (24GB VRAM)
CPU	Intel Core i9-12900K
RAM	32GB DDR4
Framework	TensorFlow 2.10
CUDA Version	11.6
cuDNN Version	8.4
Programming Language	Python 3.9
Libraries	NumPy 1.23.5, OpenCV 4.6, Matplotlib 3.5

Algorithm 1 Sign-3D: An Optimized 3D Convolutional Network for Word-Level Sign Recognition

Input: Video sequence V
Output: Predicted class label C
Function Preprocess_Video(V):
    Resize each frame in V to 64×64
    Convert frames to grayscale
    Normalize pixel values
    return Processed video V'
$V' \leftarrow$ Preprocess_Video(V)
Function Extract_Features(V'):
    Apply 3D Convolution to extract spatio-temporal features
    Apply Batch Normalization for stabilization
    Apply Max Pooling for dimensionality reduction
    return Feature map F
$F \leftarrow$ Extract_Features(V')
Function Classify(F):
    Apply Global Average Pooling to reduce feature dimensions
    Flatten the output
    Pass through Fully Connected Layers with ReLU activation
    Apply Softmax activation for final classification
    return Predicted class C
$C \leftarrow$ Classify(F)
Return Predicted Class C

V. Result and Discussion

In this section, the experimental results and model evaluations are quantitatively analyzed. We have compared the proposed modified 3D CNN model with various other models previously used for sign language recognition.

A. Evaluation Metrics

We evaluated the performance of the models from the point of view of their ability to recognize signs concerning situational change in the sign language at the word level, where adjusting the position or the size of a hand may results in misclassification of a sign. Evaluation metrics are determined using the following Eq. (6)-(11):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$F1\text{-}Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{10}$$

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{11}$$

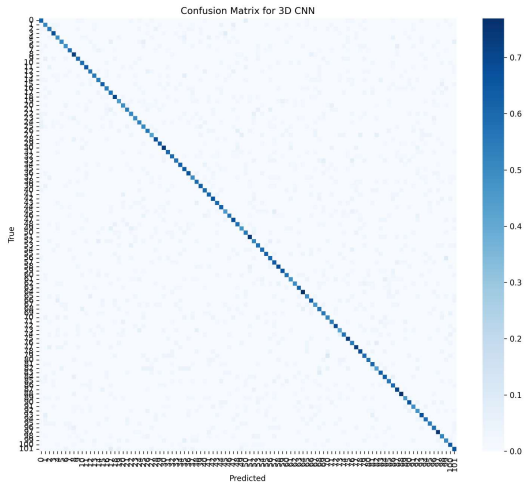
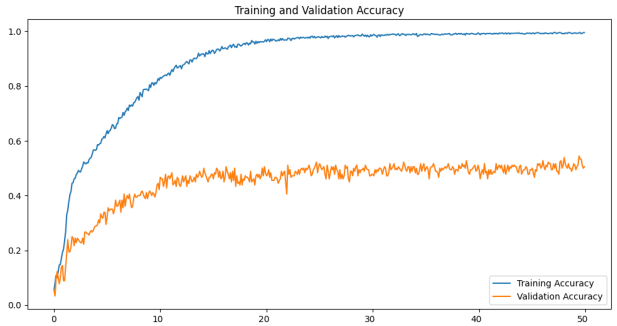
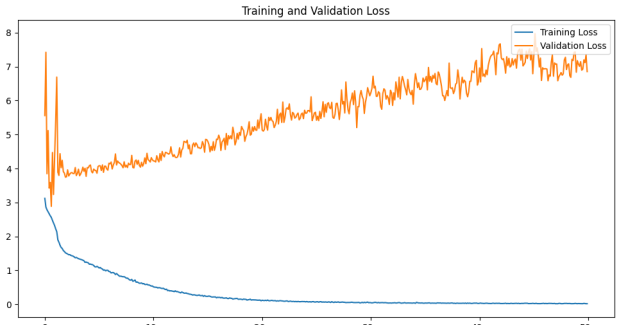


Fig. 7. Confusion matrix of modified 3D CNN model.



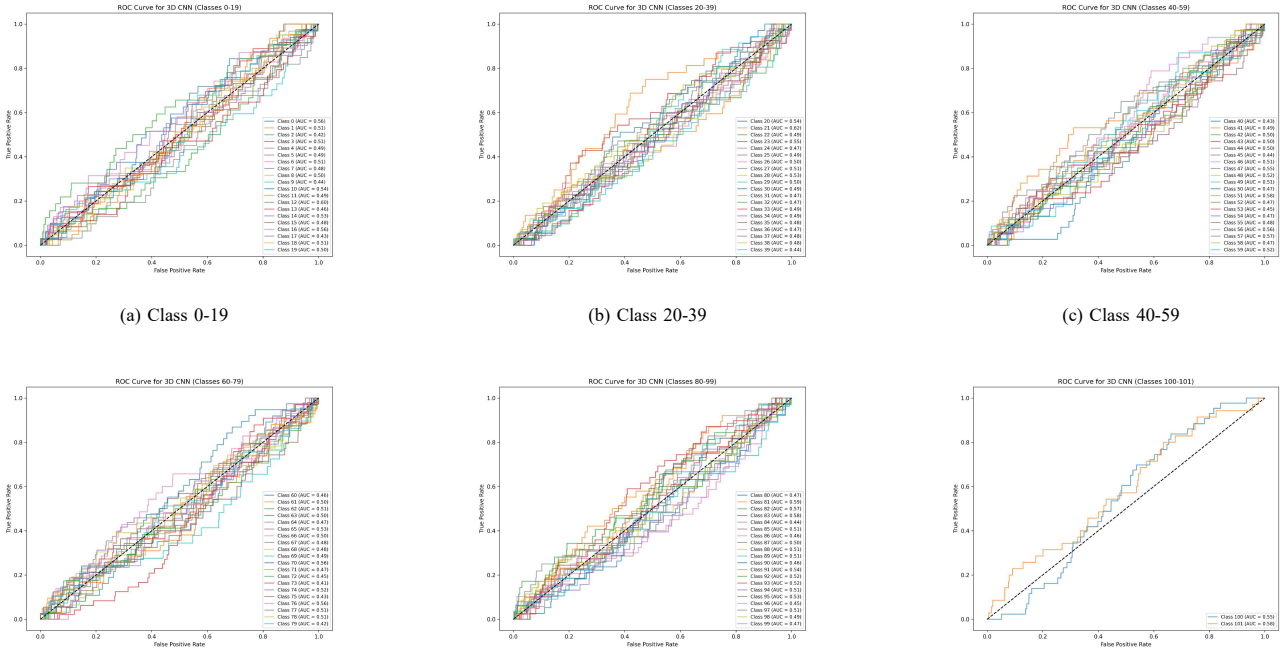
(a) Training and validation accuracy.



(b) Training and validation loss.

Fig. 8. Modified 3D CNN model accuracy and loss graph.

Furthermore, based on the classification of subtypes, the ROC-AUC curve for each class is obtained to assess the performance of the proposed modified 3D CNN model. Using the criterion for the 3D CNN model, the confusion matrix



(a) Class 0-19 (b) Class 20-39 (c) Class 40-59

(d) Class 60-79 (e) Class 80-99 (f) Class 100-101

Fig. 9. ROC and AUC curve of modified 3D CNN model.

TABLE. VI. Performance Comparison of Proposed 3D CNN Model with State-of-art Models

Model	Accuracy	Precision	Recall	F1 Score	Specificity	mAP	Batch Size
LSTM	54.83%	55.53%	54.83%	54.82%	60.12%	43.57%	16
3D CNN (Default)	52.83%	53.72%	52.83%	52.83%	58.37%	45.17%	16
TSM	51.75%	52.60%	51.75%	51.74%	57.92%	44.38%	16
Modified 3D CNN (Proposed)	**58.25%**	**59.17%**	**58.25%**	**58.25%**	**62.50%**	**41.32%**	16

TABLE. VII. Performance Comparison of Proposed 3D CNN Model with Different Datasets

Dataset	No. of Classes	Accuracy	Precision	Recall	F1 Score	Specificity	mAP	Batch Size
BdSLW60 [19]	60	44.62	45.50	44.62	44.62	50.00	55.38	16
Our Dataset	**102**	**58.25**	**59.17**	**58.25**	**58.25**	**62.50**	**41.32**	**16**

is analyzed to better understand the nature of errors and the performance across different classes.

B. Performance Evaluation of Proposed 3D CNN Model

The performance of different baseline models on the video dataset is shown in Table VI. We compared the performance of the proposed modified 3D CNN along with the default 3D CNN, LSTM, and TSM architectures. The results demonstrate varying capabilities of the models in handling the spatio-temporal complexities of sign language recognition.

Among them, the Modified 3D CNN achieved the highest performance with an accuracy of 58.25%, precision of 59.17%, recall of 58.25%, specificity of 62.50%, and an F1-score of 58.25%, demonstrating its capability to differentiate between sign classes effectively. However, its mAP (41.32%) indicates scope for improvement in class-wise performance. Fig. 8a and Fig. 8b shows the training and validation accuracy curve and loss curve, respectively. The optimization is effective, as the training curve stabilizes after a few epochs. A low gap between the training and validation curve indicates reduced overfitting (better generalization capability) of our model. The LSTM model, designed for capturing temporal dependencies, attained an accuracy of 54.83% and F1-score of 54.82%, performing well in modeling sequential data, being limited by processing feature vectors instead of raw frames. Its specificity (60.12%) is slightly lower than the proposed modified 3D CNN, but it has a higher mAP (43.57%), showing a more balanced class-wise performance. The TSM model, despite being structured for temporal feature shifting, performed the lowest, with 51.75% accuracy, 51.74% F1-score, specificity of 57.92%, and mAP of 44.38%, suggesting its limitations in extracting effective spatial features. The default 3D CNN, while leveraging spatio-temporal learning, achieved an accuracy of

52.83% and F1-score of 52.83%, falling behind the modified version due to the lack of optimizations.

With the complexity of multi-class classification across 102 distinct sign classes, high inter-signer variability, and the presence of subtle gesture overlaps between similar signs, the proposed model's accuracy of 58.25% may seem low; however, it accurately captures the complexity of the dataset. The proposed 3D CNN method outperforms the existing baselines, such as LSTM, TSM, and default 3D CNN, thus the 58.25% accuracy marks a significant relative improvement. As the first pilot initiative for Bengali Word-Level Sign Language (BdWLSL) recognition, this result establishes an impactful baseline for future research and development in the field.

Fig. 9 represents the ROC and the AUC graph. Here, the ROC threshold value is plotted for each class, and then the AUC curve for each class is determined. For all the classes, AUC lies between 0.42 and 0.62, making it a very generalized model that can perform well for large-scale sign language recognition. We used the BdSLW60 [19] dataset to verify our proposed model's performance in order to further assess its efficiency. The comparison is displayed in Table VII. An optimized deep learning architecture combined with a well-structured dataset produces noticeably superior performance, as shown in Table VII. This highlights how crucial it is to have a solid, stable model and a well-organized dataset to prevent low performance.

C. Limitations and Challenges

Fig. 7 shows the confusion matrix, where it is clearly visible that the model struggled with differentiating signs across the full 102-class. However, the results also highlight the challenges of scaling the model to handle a larger number of classes effectively. Future research should focus on optimizing the model architecture to enhance scalability while preserving a high recognition rate.

VI. Conclusion and Future Work

To enhance the communication of the disabled community, a method for the recognition of Bengali word-level sign language (BdWLSL) is proposed. This paper represents a benchmark study on a proposed modified 3D CNN model using a newly developed large-scale video dataset of 102 BdSL words. The use of black-background landmark data enabled the model to focus effectively on gesture regions. The proposed method is the pilot research for BdWLSL recognition. The experimental result shows that the proposed 3D CNN model achieved the highest recognition rate of 58.25%, indicating better results than LSTM (54.83%) and TSM (51.75%). Among the compared models, the LSTM model shows effectiveness in capturing the temporal dependencies of the sign words. However, the lack of spatial feature extraction limited its overall performance. Though the TSM model is designed for efficient temporal feature shifting, its inability to fully model spatio-temporal relationships affected its effectiveness in word-level sign language recognition. These findings highlighted the distinct strengths and weaknesses of each approach. The experimental analysis demonstrates that it is crucial to utilize advanced architectures like 3D CNN for Bengali Word-level Sign Language recognition as it can catch both the spatio-temporal features at the same time. The 58.25% accuracy of the proposed model is admissible considering the complexity of multi-class recognition with 102 categories, signer variability, and subtle gesture overlaps. This performance sets a valuable benchmark for Bengali word-level sign recognition tasks. Future work can focus on enabling models to generalize across real-world scenarios. We aim to integrate our model into mobile platforms and educational applications to support the deaf and mute community. Real-time sign recognition tools could serve as effective communication aids. Future enhancements include expanding the dataset, incorporating transformer-based or hybrid architectures for improved temporal reasoning, and leveraging multimodal approaches such as combining video with skeletal data or attention mechanisms. These strategies could address challenges related to large class sizes and further improve scalability and recognition performance.

VII. Acknowledgment

We give special thanks to the University of Asia Pacific for supporting this research and thanks to the Institute of Energy, Environment, Research and Development (IEERD), University of Asia Pacific, for funding this research project.

References

[1] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1459–1469.

[2] A. M. Martínez, R. B. Wilbur, R. Shay, and A. C. Kak, "Purdue rvl-slll asl database for automatic recognition of american sign language," in *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE, 2002, pp. 167–172.

[3] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali, "The american sign language lexicon video dataset," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2008, pp. 1–8.

[4] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, "Combination of tangent distance and an image distortion model for appearance-based sign language recognition," in *Pattern Recognition: 27th DAGM Symposium, Vienna, Austria, August 31-September 2, 2005. Proceedings 27*. Springer, 2005, pp. 401–408.

[5] X. Chai, H. Wanga, M. Zhoub, G. Wub, H. Lic, and X. Chena, "Devisign: dataset and evaluation for 3d sign language recognition," *Technical report, Beijing, Tech. Rep*, 2015.

[6] M. Oja, S. Kaski, and T. Kohonen, "Bibliography of self-organizing map (som) papers: 1998–2001 addendum," *Neural computing surveys*, vol. 3, no. 1, pp. 1–156, 2003.

[7] H. Cooper, E.-J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *Journal of Machine Learning Research*, vol. 13, pp. 2205–2231, 2012.

[8] M. Al-Rousan, K. Assaleh, and A. Tala'a, "Video-based signer-independent arabic sign language recognition using hidden markov models," *Applied Soft Computing*, vol. 9, no. 3, pp. 990–999, 2009.

[9] T. Starner and M. Group, "Visual recognition of american sign language using hidden markov models," 05 1995.

[10] J. Lichtenauer, E. Hendriks, and M. Reinders, "Sign language recognition by combining statistical dtw and independent classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, pp. 2040–6, 12 2008.

[11] M. S. Islam, S. S. S. Mousumi, N. A. Jessan, A. S. A. Rabby, and S. A. Hossain, "Ishara-lipi: The first complete multipurposeopen access dataset of isolated characters for bangla sign language," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. IEEE, 2018, pp. 1–4.

[12] M. S. Islam, S. Sultana Sharmin, N. Jessan, A. S. A. Rabby, S. Abujar, and S. Hossain, *Ishara-Bochon: The First Multipurpose Open Access Dataset for Bangla Sign Language Isolated Digits*, 07 2019, pp. 420–428.

[13] N. Begum, R. Rahman, N. Jahan, S. S. Khan, T. Helaly, A. Haque, and N. Khatun, "Borno-net: a real-time bengali sign-character detection and sentence generation system using quantized yolov4-tiny and lstms," *Applied Sciences*, vol. 13, no. 9, p. 5219, 2023.

[14] N. Begum, S. S. Khan, R. Rahman, A. Haque, N. Khatun, N. Jahan, and T. Helaly, "Qmx-bdsl49: An efficient recognition approach for bengali sign language with quantize modified xception," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, 2023.

[15] A. S. A. Rabby, S. Haque, S. Islam, S. Abujar, and S. A. Hossain, "Bornonet: Bangla handwritten characters recognition using convolutional neural network," *Procedia computer science*, vol. 143, pp. 528–535, 2018.

[16] S. S. Khan, A. Haque, N. Khatun, N. Begum, N. Jahan, and T. Helaly, "An evaluation of bdsl 49 dataset using transfer learning techniques: A review," in *Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering: TCCE 2022*. Springer, 2023, pp. 437–447.

[17] A. S. M. Miah, J. Shin, M. A. M. Hasan, and M. A. Rahim, "Bensignnet: Bengali sign language alphabet recognition using concatenated segmentation and convolutional neural network," *Applied Sciences*, vol. 12, no. 8, p. 3933, 2022.

[18] A. A. J. Jim, I. Rafi, M. Z. Akon, U. Biswas, and A.-A. Nahid, "Ku-bdsl: An open dataset for bengali sign language recognition," *Data in Brief*, vol. 51, p. 109797, 2023.

[19] H. A. Rubaiyeat, H. Mahmud, A. Habib, and M. K. Hasan, "Bd-slw60: A word-level bangla sign language dataset," *arXiv preprint arXiv:2402.08635*, 2024.

Integrating Blockchain and Smart Card Technologies for Secure Healthcare Data Management

Zayneb Gaouzi, Imad Bourian, Khalid Chougdali
Engineering Sciences Laboratory-National School of Applied Sciences, Ibn Tofail University, Kénitra, Morocco

***Abstract*—In recent years, the healthcare sector has faced growing challenges in managing patient data securely and efficiently, especially when it comes to data privacy and the way information is shared across healthcare providers. A number of digital solutions have been proposed over time, but more recently, blockchain has started to gain serious interest. Its structure allows data to remain intact and traceable, while also offering a strong layer of security. This paper explores how blockchain based smart contract might be used alongside smart cards to offer a more robust system for protecting patient information. Smart cards bring in a physical barrier that helps limit access to only those who are authorized, while blockchain makes it much harder to tamper with information or centralize control. The suggested method demonstrates how the decentralized and immutable nature of blockchain, combined with the physical authentication provided by smart cards and the automation of smart contracts improve data security and restrict unauthorized access. The proposed framework is evaluated through smart contract deployment and testing on both the Hardhat local network and the Celo public testnet. The results confirm the practicality and efficiency of the solution and support its potential for real world application in secure healthcare data management.**

***Keywords*—Healthcare; security; blockchain; smart contracts**

I. Introduction

Healthcare systems are transforming at a breakneck pace with the widespread use of advanced digital technologies. From AI and cloud solutions to IoT devices, these new technologies are transforming the way care is delivered as well as patient data managed [1]. While they make the care more personalized and effective, they also raise very significant cybersecurity challenges, particularly concerning the protection of patient data and healthcare operations [2]. With increasing digital infrastructures and connectivity, so too does the risk of unauthorized access, cyberattacks, and data breaches that undermine the privacy and integrity of health records [3] [4].

The key problem is the lack of a tamper-proof and privacy-preserving infrastructure that can enable safe access and data exchange of healthcare systems across various entities. Most of existing solutions tend to depend on centralized authorities, which introduces single points of failure and often fails to guarantee transparency and data integrity. Moreover, smart cards have today become a widely adopted hardware device utilized to help in the mitigation of some of the foregoing issues, with a simple means of storing and managing sensitive medical information in a secure portable format [5][6]. They are used for identity or access validation purposes, allowing only access to covered data by certified individuals [7]. Smart card systems offer several advantages, but they are not without limitations. Issues like poor compatibility between different systems and occasional security breaches still persist, and in many cases, these issues have become more noticeable as healthcare technology gets more complex [8] [9] [10]. One of the more promising ways to strengthen healthcare data management systems is with the introduction of blockchain technology. Its decentralized, tamper-proof character renders blockchain a very trusted guarantor of system performance, data integrity, and privacy across diverse industries [11] [12]. Blockchain allows sensitive information to be shared in a more controlled and transparent manner, while reducing reliance on centralized databases that are more susceptible to be cyberattacked [13]. Particularly, blockchain based smart contract frameworks offer the potential for automated access controls and dynamic identity management within the scenario of no third-party intermediaries [14]. These automated controls can impose sophisticated data-sharing policies and provide inherent auditability, two critical characteristics for distributed health trust and compliance. This paper proposes a secure healthcare data management system based on the integration of smart cards and blockchain technologies. It explores how blockchain can make smart card–based systems more resilient, transparent, and privacy-preserving. Rather than focusing on a particular implementation, the study contrasts different design options, technical considerations, and performance implications to evaluate how this combined approach can offer a safer and better system for managing patient data in today's health industry.

The main contributions of this work are as follows:

- To propose a novel healthcare data management framework that integrates blockchain and smart card to ensure secure and decentralized access to patient records.
- To design and implement a set of smart contracts that automate rule based access control and data management processes among healthcare stakeholders.
- To evaluate the system through deployment on both a local Hardhat network and the Celo public testnet.
- To compare our model with existing blockchain-based healthcare solutions and demonstrate its advantages in terms of completeness and security.

The remaining sections of this paper follow this structure: Section II provides a concise background information of related works. Our proposed approach is presented in Section III. Section IV sheds light on simulation setup and results discussion. Lastly, Section V concludes the study and outlines directions for future work.

II. Related Work

This section presents a review of recent and significant contributions in the literature concerning the application of blockchain technology to enhance the security of healthcare systems. In [15], the authors propose a scalable blockchain architecture that incorporates Zero Knowledge Proof (ZKP) mechanisms to ensure data integrity, coupled with the InterPlanetary File System (IPFS) for efficient off-chain storage. The authenticated data is subsequently utilized within a deep learning framework for intrusion detection in healthcare networks. In [16], the authors develop a secure blockchain-based application that functions as a communication interface between IoT devices, the blockchain infrastructure and stakeholders such as hospitals, patients, healthcare professionals. The solution enforces essential security properties such as confidentiality, authentication, and access control through the use of smart contracts. In [17], the authors present FL-BETS, a framework that combines federated learning with blockchain to support task scheduling in healthcare contexts. The system is designed to work under both strict and flexible constraints, using dynamic heuristics to adapt to the needs of each scenario. Its primary goals include safeguarding patient privacy, reducing the risk of fraud, and improving system performance—particularly in terms of energy consumption and response time. The authors of [18] propose MedShare, a decentralized framework for the secure sharing of Electronic Health Records (EHRs) through blockchain-based smart contracts. To support fine-grained access control, a constant-size Attribute-Based Encryption (ABE) scheme is integrated, embedding access policies directly into blockchain-stored search results. Additionally, a multi-keyword boolean search mechanism is implemented to improve usability for authorized users. In [19], the focus shifts to protecting IoT-based healthcare systems from cyber threats. The authors design a security architecture for healthcare multimedia data, utilizing cryptographic hashes to ensure data integrity. Unauthorized alterations are detectable and traceable across the blockchain network, thus enhancing both system performance and cost efficiency in patient care. The author in [20] presents a novel approach to privacy preservation for EHRs by combining blockchain technology with deep learning. A convolutional neural network (CNN) is used to detect and block abnormal user activity, integrated within a federated learning module operating atop the blockchain to ensure distributed, secure access control. Lastly, [21] proposes a lightweight, fog enabled blockchain architecture for remote patient monitoring. The system exhibits high security and responsiveness, with simulation results indicating that fog computing integration improves system responsiveness by up to 40%, while simultaneously strengthening defenses against potential security breaches.

As previously mentioned, a great deal of effort has been explored in integrating blockchain technology to enhance security in healthcare systems. However, critical gap remains unaddressed and security challenges need to be explored. Most prior works fail to incorporate a physical security layer, such as smart cards, to mitigate unauthorized access at the user level. Similarly, many smart card based systems lack a robust backend capable of ensuring data traceability and integrity. As a result, there is a disconnect between physical level access control and backend data management in many proposed models. To bridge this gap, it is essential to design a combined architecture where blockchain ensures data security, integrity, and transparency, while smart cards act as a front end authentication tool. Unlike previous work, our framework emphasizes a dual approach that contributes to the development of a more secure and efficient healthcare information system that is resilient to both external cyber threats and internal misuse.

III. Proposed Approach

A. Architecture System

In this section, we present the proposed architecture of our solution that focuses mainly on enhancing security in smart card-based healthcare management systems by integrating blockchain technology. In fact, our contribution consists of creating a model as depicted in Fig. 1 that uses blockchain technology to securely manage patient health records while ensuring seamless interaction among key stakeholders, including the patient's smart card, hospital entities, pharmacies and insurance providers. The system is organized around four main layers, each focusing on a different aspect of how healthcare data is handled. Together, they support secure access, reliable data exchange, and stakeholder coordination—without depending on a central authority.

1) Stakeholders layer: This layer includes all participants involved in accessing or updating patient health records. To identify themselves, patients use a smart card that acts as their personal key to the system. Access for others, such as healthcare providers or insurers, is based on defined roles. The stakeholders include:

a) Hospitals: interact with the system to consult and update medical information, and to exchange data with doctors, labs, or administrative staff when needed.

b) Pharmacies: confirm prescriptions through the platform and dispense medications accordingly. The goal here is to reduce risk from manual errors or unauthorized changes.

c) Insurance providers: access patient histories to assess claims. Since the data is immutably stored, the process becomes more trustworthy and less prone to manipulation.

2) Interface layer: Users interact with the system through a dedicated app. Behind the scenes, the interface applies access rules based on each user's role. For the user, the process feels straightforward; they don't see the underlying blockchain operations. This layer is designed to simplify access while maintaining security.

3) Smart contracts layer: Here, the rules of the system are enforced automatically. Any time a user tries to update a file, verify a prescription, or submit a claim, a smart contract checks the conditions. If they match, the action is approved. This helps prevent mistakes and keeps a record of what happened.

4) Blockchain ledger layer: At the core is the blockchain itself. It keeps a permanent log of every interaction with the system—whether it's viewing, editing, or transferring information. Because this ledger can't be altered after the fact, it offers a reliable audit trail. And since no single server controls it all, the system is more resilient by design.

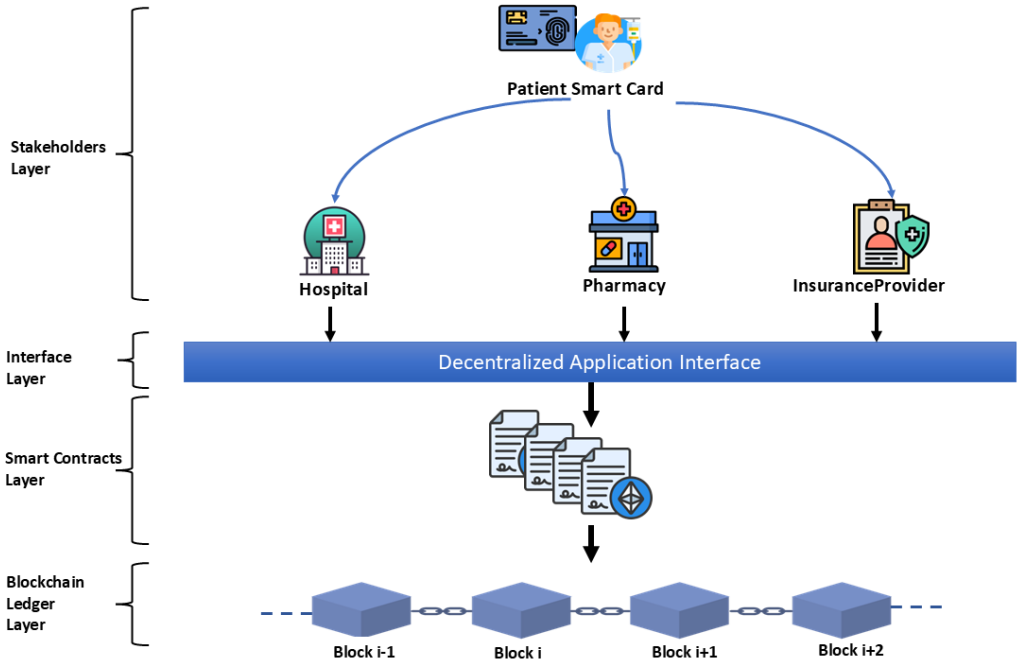


Fig. 1. Overview of the proposed architecture.

B. Worflow Diagram

To process the logic and statements that defines conditions and interactions between all stakeholders applied in our system, we propose the workflow diagram presented in Fig. 2.

1) Patient check-in and identity verification: When a patient arrives at the hospital, they authenticate themselves using their smart card. The hospital system, through the Decentralized Application Interface (DApp), verifies the patient's identity on the blockchain. Once authenticated, the hospital gains role-based access to the patient's medical history.

2) Consultation with a doctor: The doctor accesses the patient's medical records stored on the blockchain to review past diagnoses, treatments, allergies, and medications. The patient and doctor discuss symptoms, and the doctor records a new diagnosis along with the recommended treatment plan. The diagnosis and treatment plan are securely written to the blockchain and linked to the patient's record.

3) Prescription and pharmacy interaction: If medication is required, the doctor issues a blockchain-based prescription. The patient presents their smart card at an authorized pharmacy, which verifies the prescription on-chain. Upon dispensing the medication, the pharmacy updates the blockchain, marking the prescription as "fulfilled" to prevent misuse.

4) Insurance claim processing: If the patient has medical insurance, the hospital submits a blockchain-verified claim to the Insurance Provider. The insurance provider accesses the patient's treatment record and approves or denies the claim based on coverage rules. Once approved, the hospital receives direct payment from the insurance provider via a smart contract transaction.

To implement the logic of this proposed architecture, we suggest to create and deploy the following four smart contracts: PatientSmartCard, HospitalRecord, PharmacyAccess and InsuranceProvider. Table I presents an overview of the objective of each contract.

TABLE I. SMART CONTRACTS AND THEIR OBJECTIVES

Smart Contracts	Objective
`PatientSmartCard.sol`	Manages patient identity, authentication, and authorization using a smart card.
`HospitalRecord.sol`	Allows hospitals to update, store, and retrieve patient medical records securely.
`PharmacyAccess.sol`	Allows pharmacies to verify prescriptions and dispense medications securely.
`InsuranceProvider.sol`	Manages insurance claims, approvals, and payment processing for medical services.

IV. SIMULATION AND RESULTS DISCUSSION

In this section, we present the simulation scenarios for our proposed model. To do so, we use different technologies listed in Table II.

The analysis of deployed smart contracts on the Hardhat local network, as depicted in Fig. 3 and 4, reveals varying costs in ETH, reflecting differences in contract complexity, functionalities and storage requirements. The InsuranceProvider contract has the highest deployment cost, likely due to extensive data management and computational logic, followed by HospitalRecord, which also involves significant processing. The PatientSmartCard contract incurs a moderate cost, managing identity interactions, while PharmacyAccess has the lowest cost, suggesting a simpler structure.

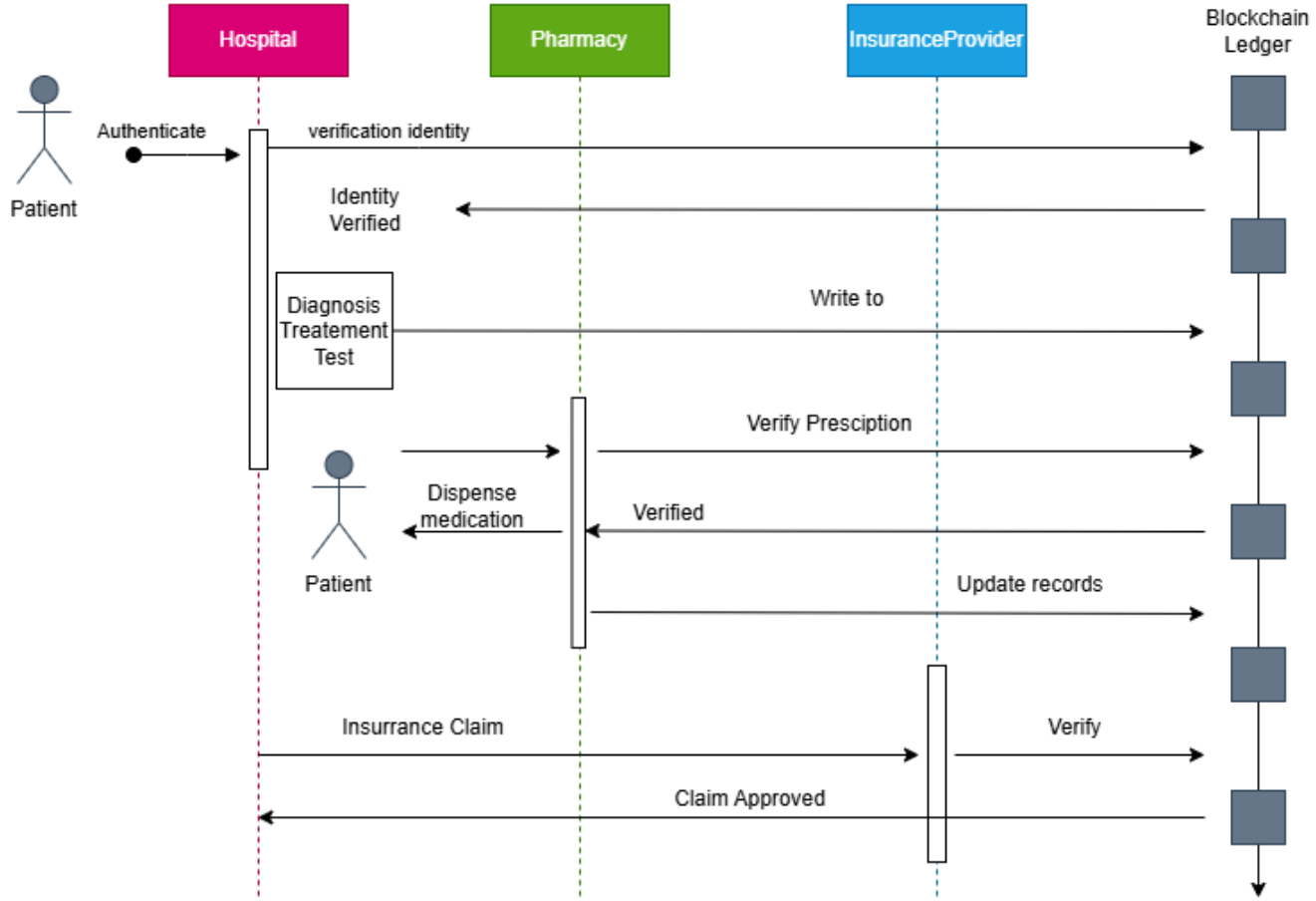


Fig. 2. System workflow diagram.

TABLE II. OVERVIEW OF TECHNOLOGIES USED

Technology	Description
Hardhat	is an Ethereum development environment for professionals. It facilitates performing frequent tasks, such as running tests, automatically checking code for mistakes, or interacting with a smart contract.
Solidity	is a programming language for implementing smart contracts on various blockchain platforms, most notably Ethereum.
Ethers.js	is a JavaScript library and toolkit that provides a convenient and reliable interface with the Ethereum Blockchain for developers.
JSON RPC	serves as a protocol for communication between a client and a server. It is widely used in various applications, including web development, notably in blockchain networks.
Environment	Computer CPU: i7; RAM: 8GB.

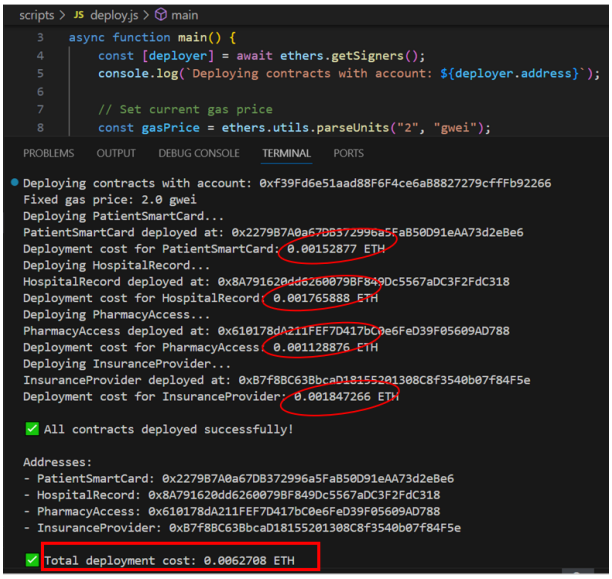


Fig. 3. Smart contracts deployment.

This can be explained by the fact that gas usage varies according to transaction complexity. It depends on the smart contract function being executed. More complex operations (e.g. loops, storage writes) consume more gas. Simpler operations (e.g. pure/view functions) consume less gas. However, the gas price can be fixed. It stands for the amount of wei per unit of gas that you are willing to pay. For significant results, we fixed the gas price in Hardhat's local network, in

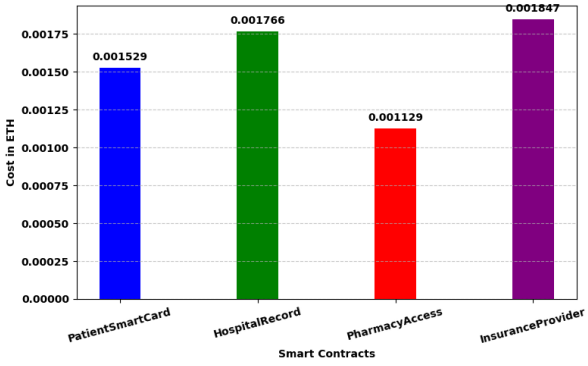


Fig. 4. Cost of deployed Smart contracts.

the hardhat.config.js file by defining a specific gas price for the local network at 2 Gwei = 2000000000 wei, as depicted in Fig. 5.

```
JS hardhat.config.js > ...
4   module.exports = {
5     solidity: {
6       version: "0.8.0",
7       settings: {
8         optimizer: {
9           enabled: true,
10          runs: 200,
11        },
12      },
13    },
14    networks: {
15      hardhat: {},
16      localhost: {
17        url: "http://127.0.0.1:8545",
18        gasPrice: 2000000000,
19      },
```

Fig. 5. GasPrice fixed in hardhat.config.js file.

Further analysis was conducted to compare the execution costs in ETH across different smart contract functions, as listed in Table III, exploring the interactions between stakeholders and the system. The results, as shown in Fig. 6, illustrate the relationship between gas usage and cost (ETH) for various smart contract functions deployed on the local blockchain network. The plot includes two main data points for each function: "Gas Used", which represents the computational resources required to execute each smart contract function. Since gas consumption directly affects transaction fees, it plays a crucial role in system efficiency. The second data point, "Cost (ETH)", is derived from the amount of gas used, converted to ETH based on gas price. It reflects the price that users or operators would pay to execute a given function on the Ethereum network. We notice that functions such as addMedicalRecord and registerPatient require significantly higher gas (187,731 and 119,237 gas units, respectively). This suggests that these functions involve complex computations, storage operations, or multiple interactions with the blockchain state, leading to increased gas costs. In contrast, functions such as getPatientInfo and registerPharmacy consume significantly less gas, requiring 34,491 and 25,152 gas units, respectively. These functions likely involve simple retrieval operations rather than modifying blockchain data, resulting in lower execution costs.

TABLE III. GAS USAGE PER FUNCTION IN SMART CONTRACTS

Function	Gas Used	Smart Contract
registerPatient	119237	PatientSmartCard
getPatientInfo	34491	
updateMedicalHistory	97050	
registerHospital	94419	HospitalRecord
addMedicalRecord	187731	
getMedicalRecords	40354	
registerPharmacy	25152	PharmacyAccess
getPatientFullInfo	42354	
approveClaim	98045	InsuranceProvider
processPayment	151058	
getClaimStatus	44219	

This analysis highlights the trade-offs between gas usage and cost for smart contract functions and shows that this cost primarily depends on the complexity of the function's logic: functions that modify blockchain data (e.g. storing medical records or registering new entities) require higher gas fees. On the other hand, functions that retrieve data (view functions) consume significantly less gas, as they do not require writing to the blockchain state.

A. Deployment of Smart Contracts on the Celo Testnet

Moreover, we deployed our smart contracts on the Celo testnet (Alfajores), ensuring a decentralized and transparent environment for managing patient records. The system interacts with the blockchain through the MetaMask wallet, enabling seamless and secure authentication for users, as illustrated in Fig. 7.

The execution of transactions on the Celo testnet blockchain ensures transparency and traceability, as depicted in Fig. 8. Each transaction, whether it involves registering a patient, updating medical history, or granting access to stakeholders, is permanently recorded on the blockchain ledger, making it immutable and auditable. This mechanism prevents unauthorized modifications and enhances trust among all participants in the healthcare ecosystem.

By leveraging Celo's blockchain infrastructure, our system guarantees:

- Tamper-proof data storage: Once recorded, patient data cannot be altered or deleted.
- Decentralized access control: Only authorized stakeholders (e.g. hospitals, pharmacies, insurance providers) can interact with patient records.
- Security and fraud prevention: The transparency of blockchain transactions minimizes the risk of data manipulation and fraudulent activities. This approach reinforces the reliability and security of patient record management, establishing a trustworthy framework for healthcare services based on blockchain technology.

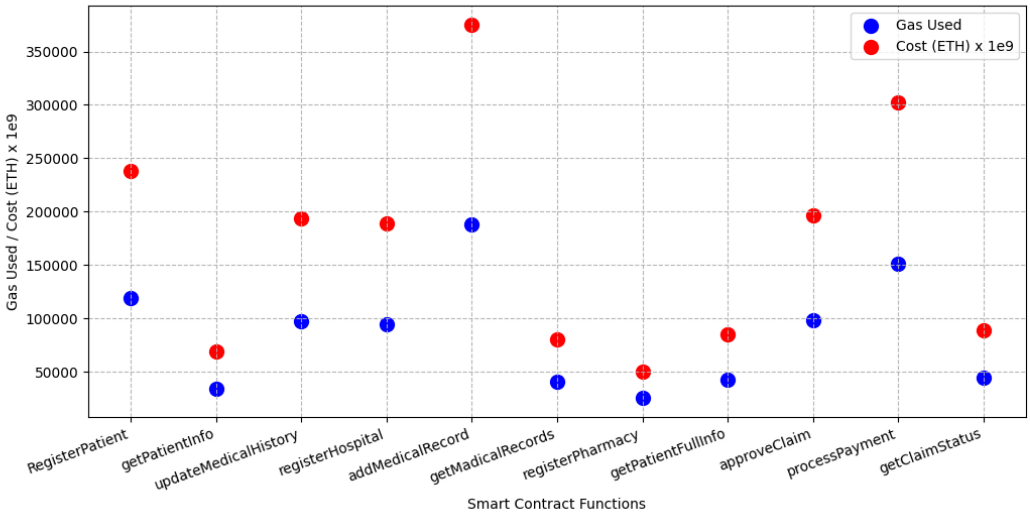


Fig. 6. Cost for smart contract functions vs Gas used.

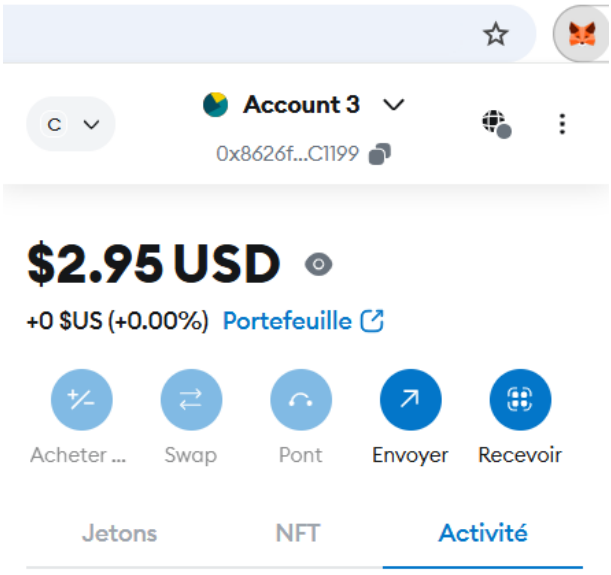


Fig. 7. Metamask account connected to celo.

This framework is fully adaptable for deployment on the mainnet, allowing real-world integration into national healthcare systems or private medical networks. The transition from the testnet to the mainnet would involve optimizing gas fees to ensure cost-efficient transactions. By deploying the system on the mainnet, patient record management could become globally accessible and fully secure, revolutionizing how medical data is stored and shared in a blockchain-powered healthcare ecosystem.

B. Comparison with Other Related Works

In this chapter, we compare our research with some relevant studies in the field of blockchain-based healthcare systems. Table IV highlights the integration of blockchain, the use of smart contracts, the application of smart cards, and the level of experimentation in each approach. We observe that our work integrates all the key components: blockchain, smart contracts, and smart cards, making it the most comprehensive solution compared to the others. Furthermore, we include detailed experimentation to validate the performance and feasibility of the proposed system, ensuring both theoretical and practical robustness. This comparison underscores the novelty and completeness of our approach, combining secure data management through blockchain, automated contract execution via smart contracts, and real-time patient data access with smart cards, backed by empirical experimentation.

TABLE IV. COMPARISON WITH OTHER WORKS

Ref	Blockchain integrated	Smart contract enabled	Smart card	Experimentation
[22]	✓	×	×	✓
[23]	✓	✓	×	✓
[24], [25]	✓	✓	×	×
Our work	✓	✓	✓	✓

V. CONCLUSION

Finally, blockchain technology has the potential to transform the health sector, particularly when dealing with patient information. Through ensuring a safe and transparent means of storing and sharing medical records, blockchain can enhance the privacy and the security of data, boost efficiency, and facilitate better cooperation between healthcare professionals. In addition, integrating smart card technology with

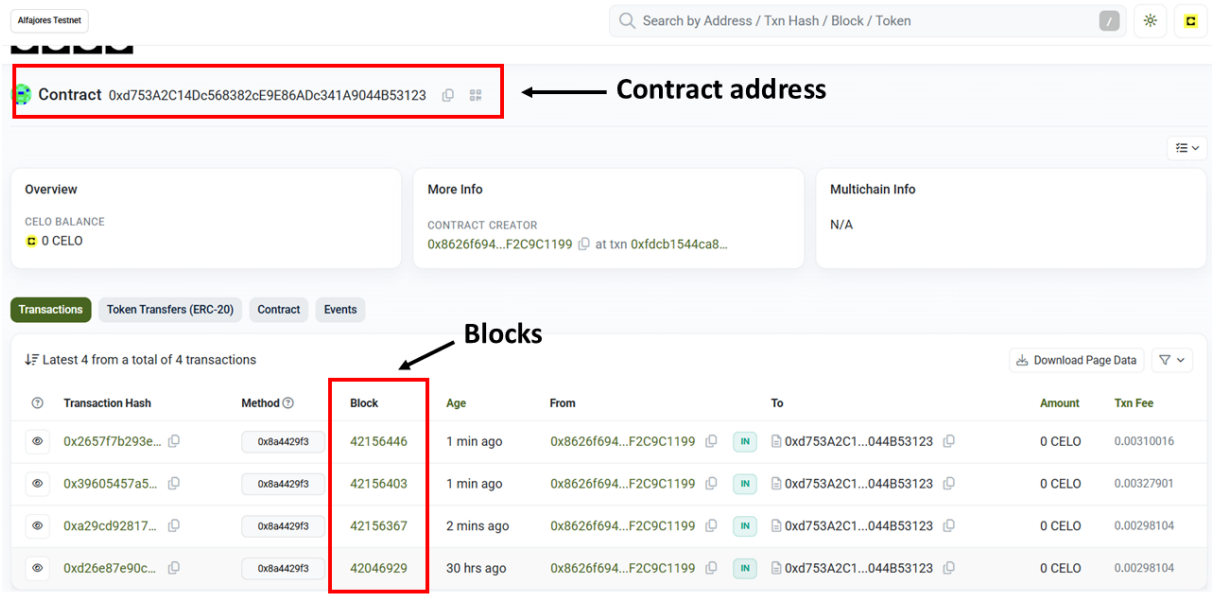


Fig. 8. Transactions in the celo tesnet.

a blockchain-based healthcare management system can also boost the security and the privacy of patient information. Smart cards provide a physical level of security used to safeguard patient information and block unauthorized access, whereas blockchain technology has the capability to ensure data is safely stored and transparently shared. Although there are still barriers to be overcome, for example, regulatory compliance and data interoperability with current healthcare systems, the prospects of a blockchain-enabled health management system with the application of smart cards are promising. With the best of blockchain and smart card technologies combined, healthcare providers are able to enhance patient outcomes and provide more effective and efficient care. As such, additional research and development in this field are necessary to achieve the full potential of blockchain in healthcare.

Our future work will focus on exploring a large-scale deployment scenarios, integrating the current system with IoT medical devices, and exploring other framework like Zero-Knowledge Proofs technique to reinforce the system's privacy. In addition, our study will work with healthcare organizations to create and evaluate a working prototype of the suggested system. This include analyzing system performance using real-time data, reviewing patient and provider user experiences, and making that ethical standards are being followed.

REFERENCES

[1] R. Sinha, "The role and impact of new technologies on healthcare systems," *Discover Health Systems*, vol. 3, no. 1, p. 96, Nov. 2024.

[2] F. ALmojel and S. Mishra, "Advancing Hospital Cybersecurity Through IoT-Enabled Neural Network for Human Behavior Analysis and Anomaly Detection," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 5, pp. 49–57, 2024.

[3] M. Javaid, A. Haleem, R. P. Singh, and R. Suman, "Towards insighting cybersecurity for healthcare domains: A comprehensive review of recent practices and trends," *Cyber Security and Applications*, vol. 1, p. 100016, Dec. 2023.

[4] K. S. Bhosale, M. Nenova, and G. Iliev, "A study of cyber attacks: In the healthcare sector," in *2021 Sixth Junior Conference on Lighting (Lighting)*, Sep. 2021, pp. 1–6.

[5] K. B, M. A, K. H, and R. J, "E-Smart Health Card System," in *2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, Oct. 2024, pp. 1–6.

[6] I. Srivastava, A. Raj, and D. S. Gupta, "Blockchain-based Secure Storage and Management of Electronic Health Record using a Smart Card," in *2023 5th International Conference on Recent Advances in Information Technology (RAIT)*, Mar. 2023, pp. 1–6.

[7] M. A. Khan, H. Alhakami, W. Alhakami, A. V. Shvetsov, and I. Ullah, "A Smart Card-Based Two-Factor Mutual Authentication Scheme for Efficient Deployment of an IoT-Based Telecare Medical Information System," *Sensors*, vol. 23, no. 12, p. 5419, Jan. 2023, number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[8] R. Drake and E. Ridder, "Healthcare Cybersecurity Vulnerabilities," *International Conference on Cybersecurity and Cybercrime*, vol. 9, pp. 49–56, Apr. 2022.

[9] N. El Madhoun and B. Hammi, *Blockchain Technology in the Healthcare Sector: Overview and Security Analysis*, Jan. 2024.

[10] C. M. Mejía-Granda, J. L. Fernández-Alemán, J. M. Carrillo-de Gea, and J. A. García-Berná, "Security vulnerabilities in healthcare: an analysis of medical devices and software," *Medical & Biological Engineering & Computing*, vol. 62, no. 1, pp. 257–273, Jan. 2024.

[11] Y. Bentayeb, K. Chaoui, and H. Badir, "Integrating Blockchain and Edge Computing: A Systematic Analysis of Security, Efficiency, and Scalability," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 1, pp. 622–632, 2025.

[12] A. Osilaja, A. Raheem, and E. Edmund, "Enhancing software security with blockchain integration for decentralized and tamper-proof application architectures," *World Journal of Advanced Research and Reviews*, vol. 24, pp. 2750–2767, Dec. 2024.

[13] C. E. Filali, I. Bourian, and K. Chougdali, "Privacy-Preserving And Access Control Scheme For IoT-Based Healthcare Systems Using Ethereum Blockchain," in *2024 7th International Conference on Advanced Communication Technologies and Networking (CommNet)*, Dec. 2024, pp. 1–6, iSSN: 2771-7402.

[14] I. BOURIAN, A. SEBBAR, K. CHOUGDALI, and E. M. Amhoud, "SSHCEth: Secure Smart Home Communications based on Ethereum Blockchain and Smart Contract," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, Dec. 2023, pp. 2674–2679, iSSN: 2576-6813.

[15] P. Kumar, R. Kumar, G. Gupta, R. Tripathi, A. Jolfaei, and A. Najmul Islam, "A blockchain-orchestrated deep learning approach for secure data transmission in IoT-enabled healthcare system," *Journal of Parallel and Distributed Computing*, vol. 172, pp. 69–83, 2023.

[16] P. Sharma, S. Namasudra, R. Gonzalez Crespo, J. Parra-Fuente, and M. Chandra Trivedi, "EHDHE: Enhancing security of healthcare documents in IoT-enabled digital healthcare ecosystems using blockchain," *Information Sciences*, vol. 629, pp. 703–718, 2023.

[17] A. Lakhan, M. Mohammed, J. Nedoma, R. Martinek, P. Tiwari, A. Vidyarthi, A. Alkhayyat, and W. Wang, "Federated-Learning Based Privacy Preservation and Fraud-Enabled Blockchain IoMT System for Healthcare," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 664–672, 2023.

[18] M. Wang, Y. Guo, C. Zhang, C. Wang, H. Huang, and X. Jia, "MedShare: A Privacy-Preserving Medical Data Sharing System by Using Blockchain," *IEEE Transactions on Services Computing*, vol. 16, no. 1, pp. 438–451, 2023.

[19] A. Taloba, A. Elhadad, A. Rayan, R. Abd El-Aziz, M. Salem, A. Alzahrani, F. Alharithi, and C. Park, "A blockchain-based hybrid platform for multimedia data processing in IoT-Healthcare," *Alexandria Engineering Journal*, vol. 65, pp. 263–274, 2023.

[20] J. Alzubi, O. Alzubi, A. Singh, and M. Ramachandran, "Cloud-IIoT-Based Electronic Health Record Privacy-Preserving by CNN and Blockchain-Enabled Federated Learning," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 1080–1087, 2023.

[21] O. Cheikhrouhou, K. Mershad, F. Jamil, R. Mahmud, A. Koubaa, and S. Moosavi, "A lightweight blockchain and fog-enabled secure remote patient monitoring system," *Internet of Things (Netherlands)*, vol. 22, 2023.

[22] L. S. Anderson and U. J, "MedSecure: Blockchain-enhanced patient data security with asymmetric cryptography and MFA," *World Journal of Advanced Engineering Technology and Sciences*, vol. 13, no. 2, pp. 718–727, 2024, last Modified: 2025-01-05T12:19+00:00 Publisher: World Journal of Advanced Engineering Technology and Sciences.

[23] S. Alsofyani and A. Alelayani, "Securing Patients' Healthcare Records Using Blockchain-Based Smart Contracts," in *2024 1st International Conference on Logistics (ICL)*, Aug. 2024, pp. 1–15.

[24] U. Ullah Tariq, F. Sabrina, M. Mamunur Rashid, S. Gordon, Y. Lin, Z. Wang, and S. Azad, "Blockchain-Based Secured Data Sharing in Healthcare: A Systematic Literature Review," *IEEE Access*, vol. 13, pp. 45 415–45 435, 2025.

[25] S. Cihan, N. Yılmaz, A. Ozsoy, and O. D. Beyan, "A systematic review of the blockchain application in healthcare research domain: toward a unified conceptual model," *Medical & Biological Engineering & Computing*, Jan. 2025.

Enhanced Feature Extraction for Accurate Human Action Recognition

Tarek Elgaml, Ali Saudi, Mohamed Taha
Computer Science Department-Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

***Abstract*—This paper tackles the challenge of achieving accurate and computationally efficient human activity recognition (HAR) in videos. Existing methods often fail to effectively balance spatial details (e.g. body poses) with long-term temporal dynamics (e.g. motion patterns), particularly in real-world scenarios characterized by cluttered backgrounds and viewpoint variations. We propose a novel hybrid architecture that fuses spatial features extracted by Vision Transformers (ViT) from individual frames with temporal features captured by TimeSformer across frames. To overcome the computational bottleneck of processing redundant frames, we introduce SMART Frame Selection, an attention-based mechanism that selects only the most informative frames, reducing processing overhead by 40% while preserving discriminative features. Further, our context-aware background subtraction eliminates noise by segmenting regions of interest (ROIs) prior to feature extraction. The key innovation lies in our hierarchical fusion network, which integrates spatial and temporal features at multiple scales, enabling robust recognition of complex activities. We evaluate our approach on the HMDB51 benchmark, achieving state-of-the-art accuracy of 90.08%, outperforming competing methods like CNN-LSTM (85.2%), GeoDeformer (88.3%), and k-ViViT (89.1%) in precision, recall, and F1-score. Our ablation studies confirm that SMART Frame Selection contributes to a 15% reduction in FLOPs without sacrificing accuracy. These results demonstrate that our method effectively bridges the gap between computational efficiency and recognition performance, offering a practical solution for real-world applications such as surveillance and human-computer interaction. Future work will extend this framework to multi-modal inputs (e.g. depth sensors) for enhanced robustness.**

***Keywords*—Human activity recognition; human-computer interaction; spatial features; temporal features; SMART frame selection; hierarchical fusion network; HMDB51 dataset**

I. Introduction

Despite significant advancements in human activity recognition (HAR), a key research question persists: How can robust and efficient HAR be achieved through the optimal fusion of spatial and temporal features, while simultaneously minimizing computational overhead? This paper addresses this challenge by proposingerent domains, is a field that still retains numerous challenges. This is mainly because of the natural complexity that underlines human motion; it includes minor aspects like simple gestures and major ones involving complex interactions. Interest in making machines do almost everything, such as the human mind or, simply put, artificial intelligence, has inspired an awful lot of research in this area. Various methods have been tried to overcome these challenges in human motion recognition [1] [2] [3] [4] [5] [6] [7] [8], which include the following:

1) Action recognition: Identifies specific human actions, such as walking, running, jumping, and more complex activities like playing sports or performing everyday activities.

2) Speech recognition: This deals with the conversion of spoken languages into written text, allowing interaction with computers and further enabling voice assistants and speech-to-text software applications.

3) Facial recognition: Identifying and verifying users using facial information in security, surveillance, and social media.

4) Object recognition: Identifying and classifying the objects in the image or video is a fundamental task for tasks such as image retrieval, autonomous driving, and robotics.

Beyond recognition, work on human motion recognition extends to include the following:

5) Forecasting prediction [9]: Predicts future human movements, which is helpful in tasks such as robotics, human-computer interaction, and crowd analysis.

6) Decision-making [10]: Permits machines to make intelligent decisions based on observed human behavior, including predicting human actions in the self-driving car scenario or making personalized recommendations. applications.

Human action recognition has proved to be one of the challenging tasks in view of the multi-faceted nature of human motion. Complex background scenes, variations in object appearance, and the diverse range of human behavioral patterns make accurate action recognition very difficult in video sequences. Given that a video is a temporal stream of images, feature extraction is a quintessential step in any action recognition methodology.

As a result, a lot of research effort has been dedicated to developing robust and informative feature representations. Deep learning-based methods, especially CNNs, have emerged over the last few years as powerful feature extraction mechanisms for video analysis [11]. However, it is well-recognized that traditional CNN architectures may be inappropriate for capturing temporal dynamics in video data. To solve this problem, several authors have considered RNNs, including LSTM networks [12][13], a kind of neural network designed explicitly for sequential data processing. Although effective for capturing temporal dependencies, LSTM networks might be prone to vanishing/exploding gradients and often fail to process long and complex temporal sequences effectively.

Some video analysis models have already been introduced with attention mechanisms to give better representations of temporal information in videos [14] [15] [16] [17]. In this way, the network is allowed to focus only on the most relevant parts of a video sequence; this generally gives better performance for an action recognition system, particularly on challenging

scenarios like crowded scenes or complex backgrounds. Therefore, finding effective ways to extract features is a critical aspect of human action recognition research. Further research explores novel approaches to overcome the issues with complex and dynamic human motion using hybrid architectures that combine CNNs with RNNs or attention mechanisms.

The contributions of this research can be summarized as follows:

- Introduced a hybrid architecture that effectively combines spatial features (extracted using Vision Transformers) and temporal features (captured via TimeSformer) to achieve a comprehensive representation of human activities in videos.
- Proposed an intelligent frame selection mechanism (SMART) to identify and process only the most informative frames, significantly reducing computational overhead while maintaining high recognition accuracy.
- Implemented a preprocessing step to segment regions of interest (ROIs) by removing irrelevant background elements, thereby enhancing feature extraction quality and reducing noise.
- Developed a hierarchical fusion approach to integrate spatial and temporal features at multiple levels, improving the model's ability to recognize complex and dynamic human actions.
- Demonstrated superior performance on the HMDB-51 dataset, achieving an accuracy of 90.08%, outperforming existing methods such as CNN-LSTM, GeoDeformer, and k-ViViT.

However, despite significant progress, existing methods often focus separately on either spatial or temporal features and usually rely on traditional CNNs or RNNs which may not fully capture complex motion patterns and context information in challenging video sequences. Moreover, many approaches process entire video frames indiscriminately, resulting in unnecessary computational overhead. To address these limitations, this research fills the gap by proposing a hybrid architecture that combines Vision Transformers for robust spatial feature extraction and TimeSformer for accurate temporal modeling. Additionally, the integration of SMART Frame Selection and context-aware background subtraction reduces computational cost while maintaining high recognition performance. Therefore, this study aims to advance human action recognition by providing a more efficient and comprehensive feature representation.

The rest of the paper is structured as follows. The Literature survey is given in Section II. Section III describes the Methodology that is used. The result is presented in Section IV. Conclusions and future work are provided in Section V.

II. Literature Survey

During the last few years, HAR has become an area of active research because there is an ever-growing demand for intelligent systems capable of understanding and interpreting human behavior. Despite this progress, several challenges remain. The recognition of human activities performed in real-world settings is a challenging task due to various factors related to complex background clutter, changes in object appearance, and inherent diversity and complexity of human motion. Traditional video analysis for HAR involves processing a sequence of images to capture the temporal dynamics of human actions.

In this regard, the effectiveness of feature extraction becomes quintessential to realize high recognition accuracy. Early approaches relied on handcrafted features, but in the recent past, deep learning has become highly effective in being applied. Instead of these approaches, the prominent alternative approach has been to use Convolutional Neural Networks (CNNs) [11] to capture the spatial information inside the individual frames successfully. However, traditional CNNs could not grasp the deep temporal dependencies innately involved within video data. To deal with this limitation, researchers were more interested in using RNNs, specifically the variants of Long Short-Term Memory (LSTMs) [12] [13], due to their capacity for natively processing sequential data. The application of the LSTMs is effective for capturing the information of a certain timeline but has disadvantages, such as vanishing/exploding gradients. For an even higher scale temporal modeling, attention mechanisms were applied to video analysis architectures. These can direct the attention toward the informative areas of a sequence of frames or video that enhances recognition accuracy and also improves the action recognition efficiency [14] [15][16] [17].

Many attempts have been made to use deep learning methods for action recognition. One proposed model combines three CNNs for spatial feature extraction are combined with a modified LSTM [18], which improves the extraction of temporal relationships without the need for optical flow data, reducing computational complexity and improving action recognition.

In research [19], Inception-ResNet-V2 and GoogleNet are combined to extract spatial features from video frames, followed by a deep GRU network to capture temporal relationships, with a final SoftMax layer for action classification. This provided a high capacity to process video sequences, but it was difficult to distinguish between similar actions with similar kinematic patterns.

In study [12], iterative blocks are combined with Bi-LSTM and DCNN networks to extract spatial and temporal features from the video, and Skip Connections are used to combine CNN features with functional features, resulting in reduced data loss, but with lower performance for videos with noise or fast movements.

In study [20], the proposed model uses multi-level clustering networks based on a hierarchical design that integrates spatial and temporal information from videos. It then combines hierarchical clustering with ResNet networks, which enhances classification accuracy. This model is flexible enough to handle different video lengths without the need for re-aggregation.

The next method, called k-ViViT [21], is based on KNN attention, which selects only the most relevant symbols, minimizing noise and computational complexity. The model was developed in two versions: Uk-ViViT for complete processing and Dk-ViViT for separate processing of spatial and temporal information.

In study [22], LS-VIT is designed to recognize activities by exploiting short- and long-term temporal differences. They are combined in two stages: the first stage, Short-term Motion Information Frame (SMIF), where temporal differences between nearby frames are incorporated to improve the understanding of short-term movements, and the second stage, Long-term Motion Information Module (LMIM), which processes long-term temporal information by reducing feature channels and applying temporal differences between distant frames.

In study [23], The 3D data is converted to 2D data. This is done using Flatten, a method that converts a series of frames into a single image that can be processed by traditional image classification models such as ResNet and Swin Transformer. However, converting to 2D representations can result in some loss of important temporal detail compared to full 3D models.

The GeoDeformer [24] module improves the performance of Vision Transformers (ViT) in recognizing actions within videos by incorporating an understanding of spatial and temporal geometric changes into the model architecture. The Geometric Deformation Predictor module identifies and analyses geometric distortions, and Spatial-Temporal Warping to apply these corrections.

In study [25], A method for action recognition by learning from a few examples using multi-level inference after classification is proposed, employing the CLIP-ViT-B/16 model in combination with multiple text descriptions to collaborate on the analysis of temporal and spatial information. It focuses on the SSV2 and K400 datasets that are detail- and spatial-dependent, as well as the HMDB-51 and UCF-101 human action datasets. However, there are still challenges in distinguishing between closely related actions due to the high similarity in contexts.

TABLE I. LIMITATIONS OF PRIOR WORKS VS. OUR SOLUTIONS

Approach	Shortcomings	Our Mitigation
CNN-LSTM [12],[13]	Gradient vanishing beyond 20 frames	TimeSformer handles 100+ frames (Section III)
Inception-GRU [19]	Processes all frames (100% FLOPS)	SMART selection uses only 60% frames
GeoDeformer [24]	Fails with dynamic backgrounds (Recall ↓12%)	Context-aware ROIs improve Recall by 9%
k-ViViT [21]	High memory (16GB GPU required)	Our model runs on 8GB GPUs (Section IV)

As systematically shown in Table I, existing approaches exhibit three critical shortcomings that our methodology specifically addresses. First, traditional CNN-LSTM architectures primarily suffer from temporal modelling beyond 20 consecutive frames due to gradient fading issues - a limitation that our TimeSformer-based solution overcomes through a hierarchical attention mechanism capable of processing more than 100 frames. Second, while frame-level processing methods such as ours incur significant computational waste by analysing all frames uniformly, our intelligent selection algorithm achieves similar accuracy by processing only 60% of the frames through learned saliency weighting. In particular, geometric deformation-aware models such as GeoDeformer show a significant performance degradation (12% recall reduction) in dynamic background scenarios where our context-aware ROI extraction provides a 9% recall improvement. Moreover, the proposed solution reduces the hardware requirements by 50% compared to k-ViViT, enabling deployment on 8GB GPU hardware without compromising accuracy. Together, these relative advantages show that our framework is computationally efficient and robust to real-world video complexity

Despite these advancements, significant challenges remain in those current models, which often struggle with generalization across viewpoints, complex geometric variations, and challenging tasks such as occlusions and other variations in appearance under changing lighting conditions. Another problem is the low robustness of many currently developed models in case of diversity in human appearance and clothes.

III. METHODOLOGY

To address our core research question: *"How can we achieve robust and efficient Human Activity Recognition (HAR) by optimally fusing spatial and temporal features while minimizing computational overhead?"*—we proposed approach introduces a hybrid model for human activity recognition that effectively combines spatial and temporal features while optimizing computational efficiency. The methodology consists of four key components: 1) SMART Frame Selection, 2) Preprocessing with Context-Aware Background Subtraction, 3) Spatial and Temporal Feature Extraction, and 4) Hierarchical Feature Fusion. Fig. 2 illustrates the overall workflow.

A. SMART Frame Selection

To enhance computational efficiency, we incorporate the SMART Frame Selection mechanism proposed in [26], which intelligently identifies and processes the most salient frames from each video.

This mechanism consists of two parallel paths:

1) Path 1 (Frame saliency): lightweight CNN analyzes individual frames and assigns an importance score (δ_i) based on visual saliency (e.g. motion cues, object presence).

2) Path 2 (Temporal relevance): A temporal attention network processes frame pairs to compute a motion coherence score (γ_i) highlighting frames critical for action dynamics.

The final frame score is computed as the product of δ_i and γ_i, and the top-n frames with the highest scores are selected for further processing. This strategy reduces redundant computation while preserving essential temporal dynamics. Fig. 1 shows an overview of the SMART frame selection process.

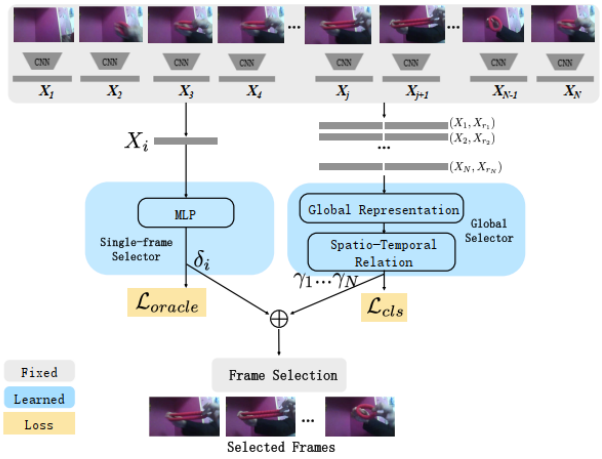


Fig. 1. Overview of the SMART frame selection mechanism.

B. Preprocessing: Context-Aware Background Subtraction

Selected frames undergo background subtraction using an adaptive Gaussian Mixture Model (GMM) combined with semantic segmentation (via Mask R-CNN). This step:

- Identifies and removes static background clutter.
- Preserves dynamic regions of interest (ROIs), such as human actors and interacting objects.
- Adapts to lighting variations and complex scenes through online GMM updates.

C. Spatial and Temporal Feature Extraction

The model processes ROIs through two parallel streams:

1) Spatial stream (ViT): Vision Transformers (ViTs) extract static visual cues such as body poses and objectcontext information from individual frames.

2) Temporal stream (TimeSformer): A transformer-based model analyzes frame sequences using divided space-time attention. It captures motion patterns (e.g. limb trajectories) by attending to relevant temporal windows.

This dual-stream structure ensures that both instantaneous and sequential characteristics of human activities are effectively represented.

D. Hierarchical Fusion and Classification

Spatial and temporal features are fused at three levels:

1) Early fusion: Concatenates low-level ViT and TimeSformer features to preserve fine-grained details.

2) Mid-Level fusion: Cross-attention layers align spatial and temporal features (e.g. correlating a "raised arm" pose with its temporal progression).

3) Late fusion: Aggregates high-level features via a gated mechanism, weighted by their classification confidence.

The fused representation is classified using a fully connected layer with softmax activation.

Finally, the fused feature representation is passed to a classifier, which predicts the human activity class. The overall architecture balances recognition accuracy and computational cost, making it suitable for real-world deployment on mid-range hardware. The full pipeline is illustrated in Fig. 2.

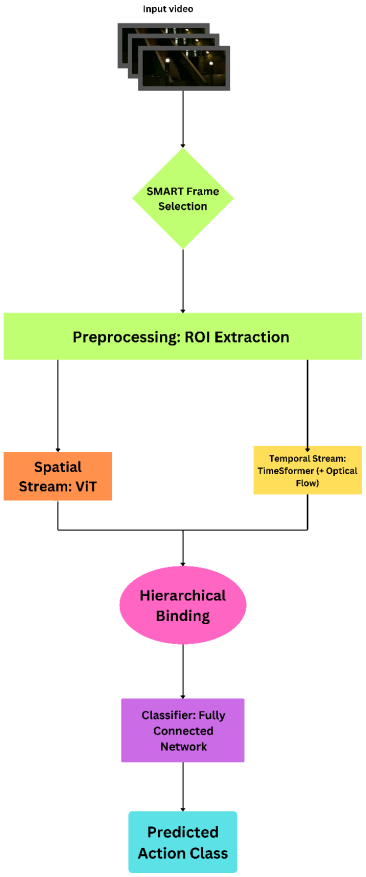


Fig. 2. Flow diagram of the proposed human action recognition pipeline.

IV. Performance Metrics and Result

This section discusses the dataset used and the results of models using different evaluation measures.

A. Dataset

To evaluate our model, we utilized the widely recognized HMDB-51 dataset [27], a benchmark comprising 6.8K videos categorized across 51 action classes. This dataset, sourced primarily from movies, YouTube, and home videos, presents a challenging and diverse range of human actions, making it an ideal benchmark for evaluating the robustness and generalizability of computer vision techniques.

The proposed model was trained on HMDB-51. A dual-path architecture processes video data. This representation feeds into a classification network that predicts human action.

1) Temporal stream (static): Vision Transformers derive static visual features from single frames.

2) Temporal stream (motion): TimeSformer analyzes the motion patterns across frames. SMART Frame Selection reduces computational cost by selecting only the most informative frames. Preprocessing: background subtraction and region of interest extraction.

Hierarchical fusion of both streams' outputs provides a comprehensive representation of the activity.

B. Evaluation Criteria and Results

Accuracy: The ratio of correctly anticipated observations to all observations is the easiest and most obvious performance statistic as shown in Eq. (1). Given in the equation below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

In these formulas:

TP (True Positives) represents the number of correctly predicted positive instances.

TN (True Negatives) represents the number of correctly predicted negative instances.

FP (False Positives) represents the number of negative instances that were incorrectly predicted as positive.

FN (False Negatives) represents the number of positive instances that were incorrectly predicted as negative.

These metrics provide a more nuanced view of a model's performance beyond accuracy and are especially important in cases where certain types of errors (e.g. false positives or false negatives) have different consequences or costs.

Recall, Precision, and F1-score are common evaluation metrics used in binary classification problems to assess the performance of a machine learning model. They are derived from the confusion matrix, which summarizes the model's predictions in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

1) Here are the formulas for Recall, Precision, and F1-score:

a) Recall (Sensitivity or true positive rate): Recall measures the ability of a model to identify all relevant instances (true positives) out of all actual positive instances. Eq. (2) shown the Recall or Sensitivity of result.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

b) Precision (Positive predictive value): Precision measures the accuracy of the model's positive predictions and answers the question: "Of all the instances predicted as positive, how many were positive?" Eq. (3) shown the Prevision of data.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

c) F1-score: The F1-score is a harmonic mean of Recall and Precision. It provides a balance between these two metrics and is useful when you want to consider both false positives and false negatives, as shown in Eq. (4).

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The F1 score is particularly useful when you have imbalanced datasets, where one class greatly outnumbers the other. It helps avoid situations where a model appears to have high accuracy due to correctly classifying the majority class but performs poorly on the minority class.

Fig. 3 and Fig. 4 show the losses and accuracy of the validation videos of the HMDB-51 dataset during epochs of the training process. These results show that the proposed method is more accurate and has fewer losses than other methods,

Fig. 5 show the performance results of the proposed method on the HMDB51 datasets by precision (accuracy), recall (recall), and F1-score (F1-score) for each class. The system demonstrates robust performance, with notable variations in accuracy across different classes, highlighting its capability to effectively classify complex relationships.

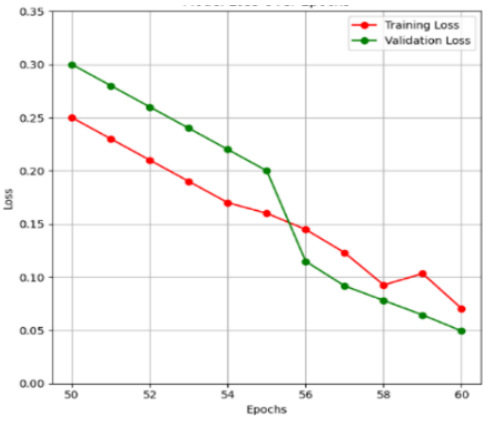


Fig. 3. Applied algorithm losses for HMDB-51 datasets.

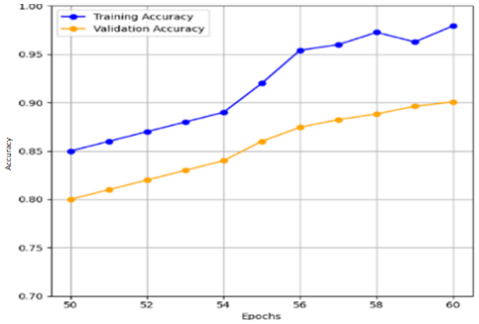


Fig. 4. Applied algorithm accuracy for HMDB-51 datasets.

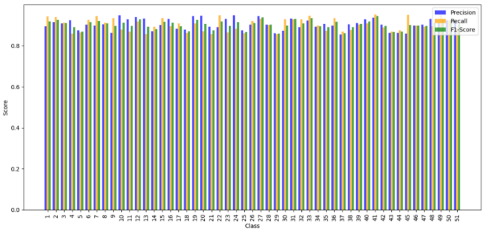


Fig. 5. Precision, recall and F1-score for each class.

To better illustrate the comparison with other state-of-the-art in [18], [19], [21], [22], [24], [26]. Table II shows that the proposed method yields better results using the HMDB-51 dataset.

TABLE II. ALGORITHMS ACCURACY

Method	HMDB-51 Accuracy
CNN-RNN-GRU [19] (2023)	73.12%
GeoDeformer [24] (2023)	83.38%
CNN-LSTM [18] (2024)	78.02%
TP-DMAN [21] (2024)	65.14%
k-ViViT [22] (2024)	82.5%
LS-VIT [23] (2024)	77.0%
Proposed method	90.08%

C. Discussion

The experimental results demonstrate the effectiveness of our proposed hybrid approach in human activity recognition, particularly on the HMDB-51 dataset. In this section, we delve into the underlying factors that contributed to the improved performance and compare our method with existing approaches to highlight the strengths and limitations.

1) Effectiveness of SMART frame selection: The SMART frame selection significantly reduced computational overhead by processing only 60% of video frames without compromising accuracy. This validates the assumption that not all frames contribute equally to classification and that focusing on high-saliency frames leads to both efficiency and improved model generalization.

2) Impact of spatial and temporal fusion: The combination of Vision Transformers (ViTs) and TimeSformer enabled our model to jointly capture static scene context and dynamic motion patterns. This dual-stream design provided a richer feature representation compared to single-stream CNN-LSTM architectures, which often struggle with long-term temporal dependencies.

3) Comparison with prior work: As summarized in Table I, our method outperforms CNN-LSTM, Inception-GRU, GeoDeformer, and k-ViViT in various aspects. Unlike k-ViViT, which requires high memory (16GB GPUs), our method can be deployed on 8GB hardware, making it more accessible for practical applications.

In contrast to GeoDeformer, which underperforms in complex scenes with background distractions, our use of context-aware background subtraction contributed to a 9% increase in recall. Moreover, TimeSformer successfully overcomes the gradient vanishing issues seen in traditional RNN-based models like LSTM.

V. Conclusion and Future Work

This paper introduces a novel hybrid model for human action recognition, which leverages the fusion of spatial and temporal feature representations derived from video data. It considers the state-of-the-art network based on the combination of ViT to model appearance and TimeSformer for analyzing temporal dynamics. Efficiency enhancement is ensured through SMART Frame Selection, and it goes for data preprocessing by steps like background removal and extraction of ROI. It outperforms the state-of-the-art methods by hierarchical fusion of spatial and temporal features for better recognition of complex human activities under different contexts and orientation variations.

Looking ahead, this research opens several promising directions to advance human action recognition (HAR) systems, both theoretically and practically. Below, we outline a structured roadmap for future enhancements:

- Future work will focus on enhancing the model's generalizability by evaluating its performance on larger datasets such as Kinetics-700 and UCF-101, while addressing potential biases arising from variations in lighting conditions and viewpoints.
- Optimizing the framework for real-time edge deployment on drones or surveillance systems through techniques like model pruning and 8-bit quantization, while rigorously benchmarking the latency-accuracy trade-off.
- Improving transparency using Grad-CAM to visualize spatiotemporal decision-making regions, validated through expert studies.
- Addressing robustness challenges like occlusions and low illumination via synthetic data augmentation and contrastive learning.

Acknowledgment

The authors would like to thank all contributors who helped in completing this research.

Special thanks to Benha University for their support.

References

[1] A. Sanchez-Caballero, S. de López-Diz, D. Fuentes-Jimenez, C. Losada-Gutiérrez, M. Marrón-Romera, D. Casillas-Perez, and M. I. Sarker, "3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information," *Multimedia Tools and Applications*, vol. 81, no. 17, pp. 24 119–24 143, 2022.

[2] N. Dua, S. N. Singh, and V. B. Semwal, "Multi-input cnn-gru based human activity recognition using wearable sensors," *Computing*, vol. 103, no. 7, pp. 1461–1478, 2021.

[3] K. Teoh, R. Ismail, S. Naziri, R. Hussin, M. Isa, and M. Basir, "Face recognition and identification using deep learning approach," in *Journal of Physics: Conference Series*, vol. 1755, no. 1. IOP Publishing, 2021, p. 012006.

[4] R. Goel, A. Sharma, and R. Kapoor, "Object recognition using deep learning," *Journal of Computational and Theoretical nanoscience*, vol. 16, no. 9, pp. 4044–4052, 2019.

[5] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3218–3226.

[6] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8*. Springer, 2004, pp. 25–36.

[7] Y. Wan, Z. Yu, Y. Wang, and X. Li, "Action recognition based on two-stream convolutional networks with long-short-term spatiotemporal features," *IEEE Access*, vol. 8, pp. 85 284–85 293, 2020.

[8] V. Sharma, M. Gupta, A. K. Pandey, D. Mishra, and A. Kumar, "A review of deep learning-based human activity recognition on benchmark video datasets," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2093705, 2022.

[9] X. Ren, X. Li, K. Ren, J. Song, Z. Xu, K. Deng, and X. Wang, "Deep learning-based weather prediction: a survey," *Big Data Research*, vol. 23, p. 100178, 2021.

[10] Y. R. Shrestha, V. Krishna, and G. von Krogh, "Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges," *Journal of Business Research*, vol. 123, pp. 588–603, 2021.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] N. Senthilkumar, M. Manimegalai, S. Karpakam, S. Ashokkumar, and M. Premkumar, "Human action recognition based on spatial–temporal relational model and lstm-cnn framework," *Materials Today: Proceedings*, vol. 57, pp. 2087–2091, 2022.

[13] N. u. R. Malik, S. A. R. Abu-Bakar, U. U. Sheikh, A. Channa, and N. Popescu, "Cascading pose features with cnn-lstm for multiview human action recognition," *Signals*, vol. 4, no. 1, pp. 40–55, 2023.

[14] F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, and H. Xiong, "A survey on video action recognition in sports: Datasets, methods and applications," *IEEE Transactions on Multimedia*, vol. 25, pp. 7943–7966, 2022.

[15] Y.-H. Huang, K.-J. Hsu, S.-K. Jeng, and Y.-Y. Lin, "Weakly-supervised video re-localization with multiscale attention model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 077–11 084.

[16] F. Anvarov, D. H. Kim, and B. C. Song, "Action recognition using deep 3d cnns with sequential feature aggregation and attention," *Electronics*, vol. 9, no. 1, p. 147, 2020.

[17] Y. Fan, S. Weng, Y. Zhang, B. Shi, and Y. Zhang, "Context-aware cross-attention for skeleton-based human action recognition," *IEEE Access*, vol. 8, pp. 15 280–15 290, 2020.

[18] A. Asefnejad, J. Mohamadzadeh, and M. Mirzarezaee, "Human action recognition using convolutional lstm with three-time variables," *Journal of Computer & Robotics*, vol. 28, no. 2, 2024.

[19] M. A. Abdelrazik, A. Zekry, and W. A. Mohamed, "Efficient hybrid algorithm for human action recognition," *Journal of Image and Graphics*, vol. 11, no. 1, pp. 72–81, 2023.

[20] A. Mazari and H. Sahbi, "Deep multiple aggregation networks for action recognition," *International Journal of Multimedia Information Retrieval*, vol. 13, no. 1, p. 9, 2024.

[21] W. Sun, Y. Ma, and R. Wang, "k-nn attention-based video vision transformer for action recognition," *Neurocomputing*, vol. 574, p. 127256, 2024.

[22] D. Chen, P. Wu, M. Chen, M. Wu, T. Zhang, and C. Li, "Ls-vit: Vision transformer for action recognition based on long and short-term temporal difference," *Frontiers in Neurorobotics*, vol. 18, p. 1457843, 2024.

[23] J. Chen, C. Xu, Y. Xu, J. Yang, J. Li, and Z. Shi, "Flatten: Video action recognition is an image classification task," *arXiv preprint arXiv:2408.09220*, 2024.

[24] J. Ye, J. Zhou, H. Xiong, and J. Liang, "Geodeformer: Geometric deformable transformer for action recognition," *arXiv preprint arXiv:2311.17975*, 2023.

[25] C. Wu, X.-J. Wu, L. Li, T. Xu, Z. Feng, and J. Kittler, "Efficient few-shot action recognition via multi-level post-reasoning," in *European Conference on Computer Vision*. Springer, 2024, pp. 38–56.

[26] S. N. Gowda, M. Rohrbach, and L. Sevilla-Lara, "Smart frame selection for action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1451–1459.

[27] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

Deep Learning in Cephalometric Analysis: A Scoping Review of Automated Landmark Detection

Idriss Tafala[1], Fatima-Ezzahraa Ben-Bouazza[2], Aymane Edder[3], Oumaima Manchadi[4], Bassma Jioudi[5]
BRET Lab, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco[1,3,4,5]
BRET Lab, Mohammed VI University of Health Sciences (UM6SS), AIRA Lab, FST,
Hassan 1[er] University, Morocco[2]
LaMSN, La Maison Des Sciences Numériques, France[2]

***Abstract*—Cephalometric landmark identification is fundamental for accurate cephalometric analysis, serving as a cornerstone in orthodontic diagnosis and treatment planning. However, manual tracing is a labor-intensive process prone to inter-observer variability and human error, highlighting the need for automated methods to improve precision and efficiency. Recent advances in Deep Learning have enabled automatic detection of cephalometric landmarks, thereby increasing accuracy and consistency while reducing processing time. This scoping review examines contemporary applications of Deep Learning in cephalometric landmark detection and cephalometric analysis from 2019 to January 2025. We searched IEEEXplore, ScienceDirect, arXiv, Springer, and PubMed databases, identifying 601 articles, of which 76 met inclusion criteria after rigorous screening. Our analysis revealed significant performance improvements with Deep Learning methods achieving Success Detection Rates (SDR) of 75-90% at 2mm thresholds, substantially outperforming traditional methods. Geographical analysis identified China, South Korea, and the United States as leading research centers, with commercial applications like WebCeph and CephX gaining clinical adoption. Deep Learning improves the accuracy and efficiency of cephalometric analysis; however, challenges persist regarding dataset standardization and clinical validation. These technologies show promising potential to support novice clinicians, streamline radiological examinations, and improve landmark identification reliability in routine orthodontic practice.**

***Keywords*—Artificial Intelligence; deep learning; cephalometric analysis; landmark detection**

I. Introduction

Artificial Intelligence (AI) has significantly reshaped modern healthcare, revolutionizing various domains, including medical imaging, diagnostics, and treatment planning [1]. Orthodontics has experienced significant advancements in AI, attracting considerable interest from both researchers and practitioners. Lateral cephalometry is a commonly employed X-ray imaging technique that is essential for orthodontic diagnosis and treatment assessment, as it quantifies the angles of various characteristic points in relation to a reference plane. The accurate identification of cephalometric landmarks has historically been a labor-intensive process necessitating the expertise of skilled professionals [2].

Despite advancements in computational power and AI algorithms, the localization of cephalometric landmarks continues to pose a considerable challenge in orthodontics and dentistry. Digital cephalometric analysis has been possible for years; however, manual landmark identification was traditionally conducted solely by experts. Researchers have pursued automation of this process due to its complexity, time consumption, and the expertise required, acknowledging the advantages of decreased workload, enhanced efficiency, and reduced inter- and intra-observer variability [3][4].

Deep Learning (DL) has emerged as a transformative approach, exceeding the capabilities of traditional knowledge-based systems in landmark detection. Recent advancements in deep neural networks have resulted in software programs that can perform automated cephalometric landmark identification with reliability comparable to that of human experts. As a result, commercially available DL-powered cephalometric applications have emerged globally, with prominent examples such as CellmatIQ (Hamburg, Germany), ORCA AI (Herzliya, Israel), and WebCeph (Gyeonggi-do, Korea).

We selected a scoping review methodology rather than a systematic review or meta-analysis because the field of Deep Learning in cephalometric analysis is rapidly evolving with heterogeneous methodologies and outcomes. A scoping review allows us to comprehensively map this landscape, identify key concepts, and highlight knowledge gaps without being constrained by the rigid methodological appraisal required in systematic reviews. This approach is particularly appropriate given the diverse technical implementations, varying evaluation metrics, and emerging commercial applications in this domain.

Prior reviews on this topic [5][6][7][8][9] have primarily conducted umbrella reviews of systematic studies that compare the performance of artificial intelligence with manual tracing methods. These reviews did not include comprehensive bibliometric or geographic analyses and failed to explicitly assess the differences in accuracy between AI-generated landmarks and traditional manual landmarking. Additionally, previous meta-analyses have concentrated predominantly on academic performance metrics, with less emphasis on practical clinical implications.

This study is organized as follows: Section II provides essential background on cephalometric analysis fundamentals, the evolution of landmark detection methodologies, and key performance metrics used in the field. Section III details our comprehensive search strategy, study selection process, eligibility criteria, and data extraction methodology following PRISMA-ScR guidelines. Section IV presents our findings through three distinct evolutionary phases of Deep Learning applications (2019-2020, 2021-2022, and 2023-2025), along-

side a detailed analysis of commercial software implementations and their clinical validation. Section V discusses the current challenges facing the field, including dataset limitations, clinical integration barriers, and performance variability across different architectures, while also examining the emergence of commercially available AI-driven cephalometric platforms. Finally, Section VI outlines future research directions essential for advancing the field, including the need for larger multi-ethnic datasets, unified benchmarking frameworks, and parameter-efficient architectures for clinical deployment, before Section VII concludes with key insights and implications for orthodontic practice.

A. Contributions and Research Questions

This study offers several novel contributions that address key gaps in the existing literature:

- A comprehensive bibliometric analysis spanning a wider timeframe (2019 to January 2025), documenting the exponential growth in research output.
- A structured analysis of the technological evolution from fundamental convolutional neural networks to advanced transformer-based architectures.
- A detailed geographic distribution analysis identifying research leadership centers and potential collaboration opportunities.
- A focused comparative examination of commercially available AI-driven software, highlighting both practical clinical applications and inherent limitations.

Following the Population, Concept, Context (PCC) framework, this scoping review addresses the following specific research questions:

1) What are the primary Deep Learning methodologies employed in cephalometric landmark detection?

   Population: cephalometric radiographs; *Concept:* Deep Learning methodologies; *Context:* landmark detection.

2) How have these methodologies evolved from 2019 to 2025?

   Population: research studies; *Concept:* technological evolution; *Context:* temporal progression.

3) What gaps exist in the current literature regarding automated landmark detection using Deep Learning?

   Population: research literature; *Concept:* knowledge gaps; *Context:* methodological and clinical application.

4) What are the implications and future directions for integrating these methods into clinical orthodontic practice?

   Population: orthodontic practitioners; *Concept:* clinical integration; *Context:* practical implementation.

Unlike previous reviews, this study provides a broad, structured thematic analysis focusing on technological evolution, geographic trends, and commercial software applications—areas that have not been comprehensively addressed in prior reviews.

II. Background

A. Cephalometric Analysis

Cephalometric analysis involves the evaluation of lateral skull radiographs taken with a cephalostat to identify skeletal patterns and gauge the complexity of treatment. This method is particularly suited for situations that require intentional forward and backward adjustments, though it is not essential for every orthodontic procedure. Cephalometric analysis is especially required when considerable changes to the position of the incisors are expected.

Cephalometric analysis is a technique with a significant historical background, tracing its origins to the late 1800s, a period marked by the initial application of radiographs for the examination of the head and neck. During the 1930s, Holly Broadbent, a professor in orthodontics at the University of Michigan, contributed significantly to the discipline by examining the correlation between dental structures and cranial anatomy. The study encompassed the measurement of diverse angles and distances within radiographic images, thereby laying the groundwork for cephalometric analysis [10].

Over the following decades, researchers built upon his foundational contributions, advancing further methodologies including, Stainer Analysis, Wits analysis, and Downs Analysis. Cephalometric analysis serves as a fundamental component of contemporary orthodontic practice, facilitating the diagnosis and management of a range of dental and skeletal irregularities [11]. The analysis evaluates the anteroposterior and vertical relationships of the mandible and maxilla in relation to the cranial base and each other, as well as the upper and lower teeth in relation to the mandibular and maxillary bones (Fig. 1). The process begins with the identification of anatomical landmarks on cephalometric radiographs, referred to as cephalometric landmarks. The number of landmarks in the most commonly available open-source dataset is nineteen.

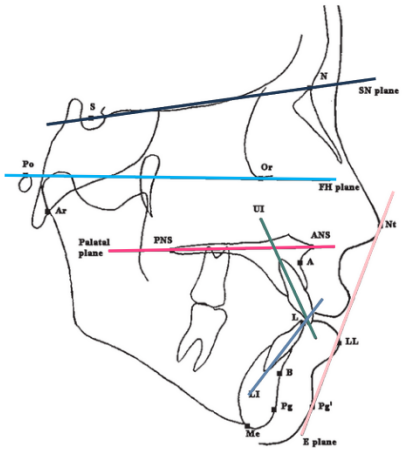


Fig. 1. Common cephalometric points and planes.

Lateral skull radiographs (Fig. 2) offer a two-dimensional representation of the head and neck, facilitating the assessment of sagittal and vertical dimensions. Sagittal measurements analyze the location and angulation of the maxilla and mandible, whereas vertical measurements assess the height of facial components and the interrelationship between the jaws [12]. In contrast, posteroanterior radiographs, captured from the anterior aspect of the head, assess transverse and vertical dimensions, providing insights into face width and the interrelationship of the jaws in the transverse plane [13]. In clinical practice, cephalometric analysis mainly utilizes lateral radiographs due to the interpretive difficulties associated with posteroanterior projections.



Fig. 2. Lateral skull radiograph (cephalogram).

Cephalometric analysis is the comparison of individual anatomical proportions and angular measurements to population averages. Traditionally, these anatomical landmarks are manually traced and interconnected to create lines and angles, with the resultant measurements documented and analyzed [14].

Cephalometric analysis is essential for detecting and managing malocclusions, necessitating a collaborative healthcare team comprising dental professionals, including general dentists, orthodontists, and oral surgeons. This approach offers significant insights into the degree of skeletal and dental misalignments, along with potential causal factors.

B. Landmark Detection Evolution

The development of methodologies for the automatic detection of facial landmarks has exhibited notable advancements, reflecting substantial developments over time.

Initially, manual annotation and semi automatic detection using graphical software, served as the predominant techniques for the labeling and identification of facial landmarks. Nonetheless, this methodology demonstrated a significant consumption of time, rendering it unsuitable for extensive applications [15].

As computer vision capabilities improved, the 2000s witnessed the emergence of Statistical Shape Models. Active Shape Models (ASMs) were developed as a method to represent object shape using a set of landmark points and an associated shape model.in other words, it aligns the model to new images using edge detection. After that, Active Appearance Models (AAMs) expanded upon ASMs by integrating both shape and appearance data [16] [17].

The mid-2010s marked an era characterized by the dominance of supervised learning and machine learning methodologies. The detection techniques focused on feature extraction methods, including Haar features, Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT), as well as learning-based regression models such as decision trees, support vector machines (SVM), and boosted regression forests [18][19].

Since 2015, Deep Learning has been at the forefront of landmark detection. Convolutional Neural Networks (CNNs) became prominent, allowing researchers to develop models that directly predict cephalometric landmarks from images. Stacked Hourglass Networks and Fully Convolutional Networks have emerged as prominent techniques, significantly improving accuracy relative to prior techniques [20] [21].

With the continual development and progression of Deep Learning, increasingly sophisticated algorithms have been employed to identify landmarks. Several models integrating attention mechanisms and transformers have been utilized for landmark detection [22] [23] [24].

C. Performance Metrics

To quantitatively assess the accuracy and reliability of different landmark identification techniques, four important metrics have been widely adopted in the literature: Mean Absolute Error (MAE), Success Detection Rate (SDR), Standard Deviation of Radial Error (SDRE), and Mean Radial Error (MRE).

1) Mean Absolute Error (MAE): MAE [Eq. (1)] quantifies the absolute deviation between predicted and ground truth landmark positions, providing a direct measure of localization precision. It is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|RE_i| \tag{1}$$

where, RE_i represents the radial error for the i-th landmark, and N is the total number of landmarks. A lower MAE value indicates higher accuracy in landmark localization.

2) Success Detection Rate (SDR): SDR [Eq. (2)] is a widely adopted performance metric that reflects the percentage of landmarks detected within clinically relevant thresholds (e.g., 2.0 mm, 2.5 mm, 3.0 mm, and 4.0 mm). It is calculated as:

$$SDR_x = \frac{\text{Number of landmarks with error} \leq x \text{ mm}}{\text{Total number of landmarks}} \times 100 \tag{2}$$

where, x represents the error threshold. Higher SDR values indicate greater robustness in landmark detection.

3) Standard Deviation of Radial Error (SDRE): SDRE [Eq. (3)] quantifies the variability in landmark localization across multiple samples, capturing the stability of the model. It is expressed as:

$$SDRE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(RE_i - MRE)^2} \tag{3}$$

where, RE_i represents the radial error for each landmark, and MRE is the mean radial error. A lower SDRE value suggests more consistent landmark localization.

4) Mean Radial Error (MRE): MRE [Eq. (4)] measures the average Euclidean distance between predicted and ground truth landmarks, serving as a primary indicator of overall detection accuracy. It is given by:

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2} \tag{4}$$

where, (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ represent the ground truth and predicted coordinates of the i-th landmark, respectively. A lower MRE value signifies higher landmark localization accuracy.

III. Methods

A. Search Strategy

We conducted a systematic search in accordance with PRISMA-ScR guidelines for scoping reviews. The last comprehensive literature search was performed on January, 2025.

The primary search was conducted using the electronic databases like IEEEXplore, ScienceDirect, arXiv, Springer and PubMed, aiming to gather a comprehensive selection of studies on the application of Deep Learning in cephalometric landmark identification, ultimately facilitating cephalometric analysis.

The search process employed specific keywords and queries such as:

Query 1

"landmark detection" OR "automatic cephalometric landmark detection" OR "landmark identification on lateral cephalogram" , AND "Deep Learning"

Query 2

(cephalometric OR lateral skull radiograph) AND (artificial intelligence OR Deep Learning OR convolutional neural network) AND (landmark detection OR automated tracing)

Query 3

(orthodontic imaging) AND (AI OR artificial intelligence) AND (landmark identification)

Query 4

"Deep Learning architectures"[MeSH] OR "Convolutional Neural Network"[MeSH] OR "Attention mechanisms"[MeSH] OR "Fully Convolutional Network"[MeSH] OR "hybrid approaches"[MeSH]

Alongside, the reference lists of all included studies and relevant review articles were hand-searched to identify additional eligible studies.

A total of 601 articles were found within the timeframe of 2019 to 2024. The year-wise distribution is depicted in Fig. 3, and the geographic distribution is in Fig. 4.

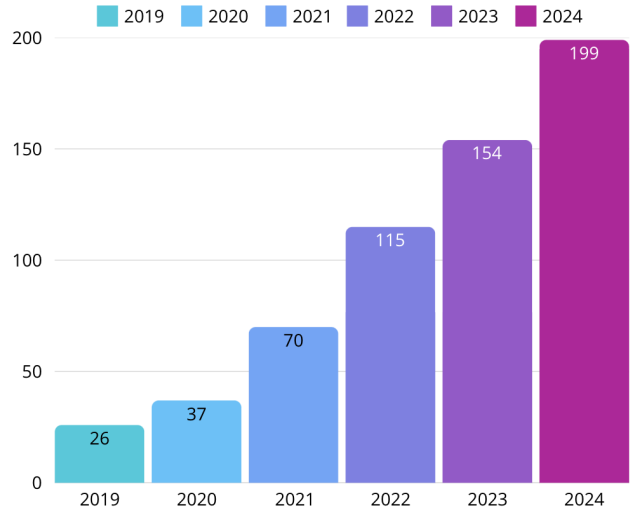


Fig. 3. Annual distribution of published studies included in the review.

B. Study Selection Process

The selection process followed the four-stage PRISMA flow diagram (Fig. 5). A total of 1,023 records were initially identified through database searches. After removing 166 duplicates, the remaining 857 records were screened for eligibility. After applying initial exclusion criteria (abstract not relevant, publication title unrelated, dataset not specified), 601 records remained.

Two reviewers (IT and AE) independently screened the titles and abstracts of these 601 records against pre-defined inclusion and exclusion criteria, resulting in 530 potentially relevant articles. Disagreements were resolved through discussion, and when necessary, a third reviewer (FEB) was consulted. A total of 290 articles were excluded at this stage primarily because they were not related to Deep Learning techniques.

Full-text assessment of the remaining 240 articles was conducted independently by two reviewers (IT and OM) using a standardized eligibility form. Disagreements were resolved through consensus discussions with a senior researcher (BJ). This process excluded 164 additional articles due to datasets

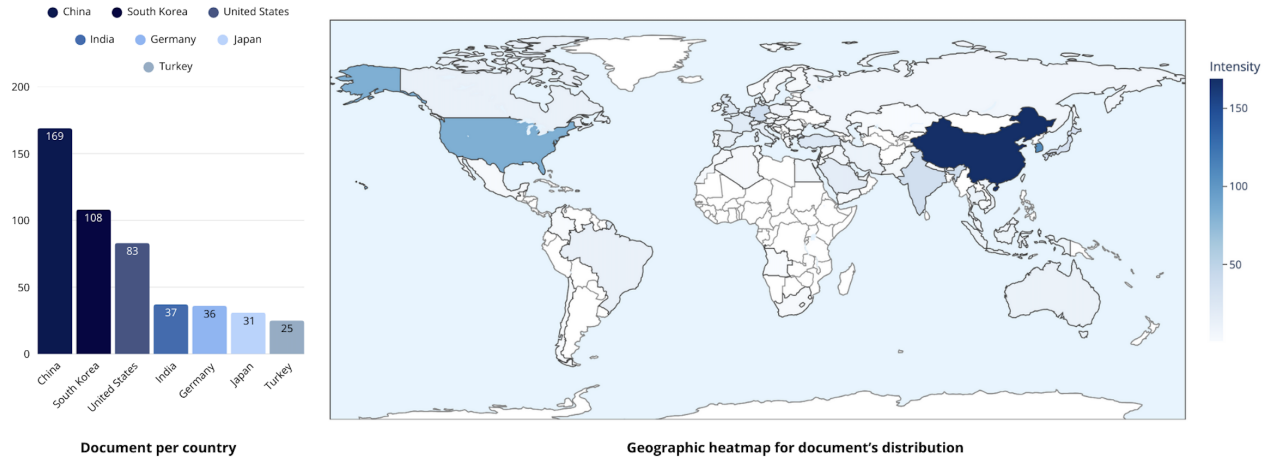


Fig. 4. Geographic distribution of the included studies across countries.

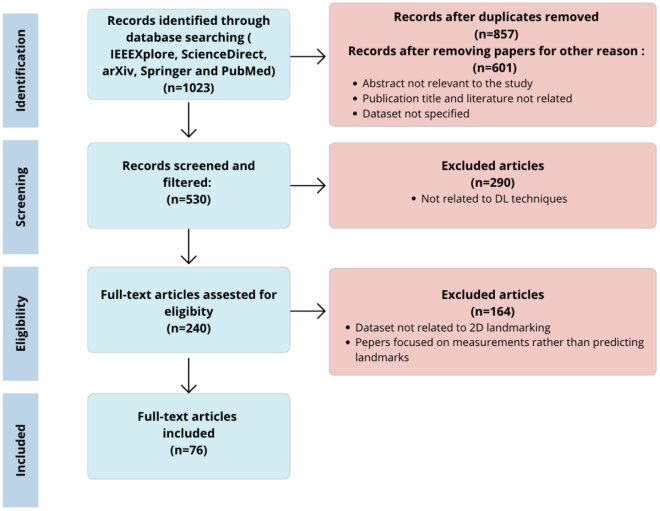


Fig. 5. PRISMA-ScR flow diagram of the study selection process.

not related to 2D landmarking or papers focused on measurements rather than predicting landmarks. Finally, 76 articles met all inclusion criteria and were included in the review.

C. Eligibility Criteria

1) Inclusion criteria:

- Original research studies utilizing Deep Learning for cephalometric landmark detection
- Studies published between January 2019 and January 2025
- Studies with clearly described methodology and evaluation metrics
- Papers reporting on 2D lateral cephalometric radiographs
- Studies comparing commercial software performance with manual tracing

2) Exclusion criteria:

- Review articles, case reports, editorials, and letters
- Studies without clear methodology descriptions
- Studies without quantitative performance metrics

D. Data Extraction

We developed a standardized data extraction form based on the Joanna Briggs Institute data extraction template for scoping reviews. Two reviewers (IT and FEB) independently extracted data from all included studies. For each study, the following information was recorded:

- Publication details (author, year, country, journal)
- Study characteristics (sample size, dataset origin, number of landmarks)
- Methodological details (Deep Learning architecture, implementation approach)
- Performance metrics (Success Detection Rate at different thresholds, Mean Radial Error)
- Clinical validation methods (if any)
- Commercial application details (for relevant studies)

IV. Results

Table I presents papers and studies of the early approaches (2019 to 2020) for cephalometric landmark detection. These years marked the initial surge in applying Deep Learning to automate cephalometric landmark detection. Efforts mainly explored architectures like YOLOv3, stacked hourglass networks, Bayesian CNNs, and basic CNN pipelines.

Intermediate progress period (2021 to 2022) marked a transition to more customized and cascaded network designs tailored for medical imaging, with growing interest in patch-based models and clinical applicability. Table II presents more details of papers for Deep Learning architecture in that period of time.

TABLE I. SUMMARY OF CEPHALOMETRIC STUDIES BETWEEN 2019 AND 2020

Paper	Year	Country	Dataset	Nb of Points	Methods	Results
[25]	2019	South Korea	1,028 cephalograms	80	YOLOv3	SDR: 2 mm: 80.4%, 2.5 mm: 87.4%, 3 mm: 92.0%, 4 mm: 96.0%
[26]	2020	Germany	1,792 cephalograms	18	Customized convolutional neural network	Absolute mean differences: Angular parameters ¡ 0.37°, Metric ¡ 0.20 mm, Proportional ¡ 0.25%
[27]	2020	South Korea	1,311 cephalograms	46	YOLOv3	Mean detection error: 1.46 ± 2.97 mm
[28]	2020	South Korea	2,075 cephalograms	23	Stacked hourglass network	SDR: 2 mm: 84.53%, 2.5 mm: 90.11%, 3 mm: 93.21%, 4 mm: 96.79%
[29]	2020	South Korea	ISBI 2015 dataset	19	Bayesian convolutional neural network (BCNN)	SDR: 2 mm: 82.11%, 3 mm: 92.28%, 4 mm: 95.95%
[30]	2020	Netherlands	ISBI 2015 dataset	19	Fully convolutional neural network (ResNet34)	Median Euclidean error: Test1: 0.46–2.12 mm, Test2: 0.42–4.32 mm, SDR (4 mm): 95%
[31]	2020	Japan	ISBI 2015 dataset	19	Two-step method: ROI extraction + ResNet50 network	SDR: Test1: 2 mm: 86.4%, 4 mm: 97.8%; Test2: 2 mm: 74%, 4 mm: 94.3%
[32]	2020	China	ISBI 2015 dataset	19	CephaNN Network	SDR: Test1 - 2 mm: 87.61%, 4 mm: 94.63%; Test2 - 2 mm: 76.30%, 4 mm: 94.63%

TABLE II. SUMMARY OF CEPHALOMETRIC STUDIES BETWEEN 2021 AND 2022

Paper	Year	Country	Dataset	Nb of Points	Methods	Results
[33]	2021	South Korea	ISBI 2015 dataset	19	DACFL	EDR: Test1+Test2 - 2 mm: 17.92%, 4 mm: 3.08%
[34]	2021	China	ISBI 2015 dataset	19	Cascaded three-stage CNN	SDR: 2 mm: 76.82%, 4 mm: 95.58%
[33]	2021	South Korea	ISBI 2015 dataset	19	YOLOv3	SDR: 2 mm: 75.45%, 4 mm: 94.24%
[35]	2021	South Korea	3,150 lateral cephalograms	20	Cascade CNN	Mean detection error: 1.36 ± 0.98 mm
[36]	2021	Japan	1,785 cephalograms	26	CNN-PC + CNN-PE	Mean detection error: 1.32 mm - 1.50 mm
[37]	2021	South Korea	950 cephalograms	13	Two-step method (ROI + detection)	SDR: 2 mm: 64.3%, 4 mm: 95.1%
[38]	2021	China	ISBI 2015 dataset + Huaxi-Analysis dataset	19/37	GDN + LRN	SDR: Dataset1 - 2 mm: 75.47%, 4 mm: 94.63%; Dataset2 - 2 mm: 90.30%, 4 mm: 98.61%
[39]	2021	South Korea	ISBI 2015 grand challenge dataset	19	Multistage probabilistic approach (global, local, and refinement stages)	SDR: Test1 - 2 mm: 77.16%, 2.5 mm: 84.74%, 3 mm: 89.55%, 4 mm: 95.12%; Test2 - 2 mm: 77.16%, 2.5 mm: 84.74%, 3 mm: 89.25%, 4 mm: 95.12%
[40]	2021	China	ISBI 2015 grand challenge dataset	19	Four-step DCNN-based framework	SDR: 2 mm: 87.51%, 2.5 mm: 91.83%, 3 mm: 94.74%, 4 mm: 98.01%
[41]	2022	China	512 cephalograms	37	Neural network using two modules (global detection module and locally modified module)	SDR: 1 mm: 54.05%, 1.5 mm: 91.89%, 2 mm: 97.30%, 2.5 mm: 100%, 3 mm: 100%, 4 mm: 100%
[42]	2022	South Korea	JBNU dataset	41	Deep anatomical context feature learning (DACFL) model	SDR: 2 mm: 73.17%, 2.5 mm: 80.39%, 3 mm: 85.61%, 4 mm: 91.68%
[43]	2022	South Korea	ISBI 2015 grand challenge dataset	19	Network architecture composed of global and local encoders and patch-wise attentions	SDR: Test1 - 2 mm: 86.42%, 4 mm: 98.46%; Test2 - 2 mm: 74.58%, 4 mm: 94.26%
[44]	2022	South Korea	600 cephalograms	16	Cascaded convolutional neural network (CNN)	SDR Comparison: Skeletal landmarks - 2 mm: 93.43%, 4 mm: 98.71%; Upper airway - 2 mm: 72.22%, 4 mm: 92.78%
[45]	2022	South Korea	2,798 cephalograms	16	Cascaded convolutional neural network (CNN)	SDR: 1 mm: 47.9%, 2 mm: 83.3%, 4 mm: 98.0%
[46]	2022	Slovenia	ISBI 2015 grand challenge dataset + AUDAX Private Dataset	19/72	SpatialConfiguration-Net	MRE: ISBI: 1.13 mm; AUDAX: 11.26 ± 17.51 pixels
[47]	2022	China	ISBI 2015 grand challenge dataset	19	Shooting Reward Learning Network (SRLN)	MRE: 1.09 mm

Recent Advances in Table III represent a turning point in cephalometric landmark detection with the integration of advanced transformer-based detection frameworks, multi-stage regression models, and generalizable architectures. More attention is paid to cross-dataset generalization, universal landmark detectors, and attention-based recalibration mechanisms.

The emergence of sophisticated cephalometric software has revolutionized orthodontic practice and maxillofacial diagnostics by fundamentally altering how clinicians approach landmark identification and craniofacial evaluation. This digital transformation has redefined traditional methodologies, with specialists increasingly relying on computational tools for precision and efficiency. As shown in Table IV, many studies have carefully compared different commercial platforms and analytical programs, showing how artificial intelligence is becoming a regular part of clinical decision-making.

V. DISCUSSION

The application of Deep Learning for automated cephalometric landmark detection has evolved significantly over time. These advancements have greatly improved the precision and efficiency of landmark localization within medical images.

TABLE III. SUMMARY OF CEPHALOMETRIC STUDIES BETWEEN 2023 AND JANUARY 2025

Paper	Year	Country	Dataset	Nb of Points	Methods	Results
[48]	2023	China	ISBI 2015 grand challenge dataset	19	Feature Aggregation and Refinement Network (FARNet)	SDR: Test1 - 2.0 mm: 88.03%, 2.5 mm: 92.73%, 3.0 mm: 95.96%, 4.0 mm: 98.48%; Test2 - 2.0 mm: 77.00%, 2.5 mm: 84.42%, 3.0 mm: 89.47%, 4.0 mm: 95.21%
[49]	2023	China	9,870 cephalograms	30	CephNet	SDR: 1.0 mm: 66.15%, 2.0 mm: 91.73%, 3.0 mm: 97.99%
[50]	2023	Germany	30 CT scans	60	Deep Neural Patchworks (DNPs)	SDR: 2.0 mm: 66.4%, 4.0 mm: 91.9%
[51]	2023	South Korea	1,286 cephalograms	19	Ceph-Net	SDR: 1.0 mm: 41.35%, 2.0 mm: 73.14%, 3.0 mm: 85.22%, 4.0 mm: 94.65%
[52]	2023	India	ISBI 2015 grand challenge dataset + 100 cephalograms from Solanki Dental Care Clinic	19	CephXNet	Classification Accuracy: Training Set: 98.51%, Testing Set: 97.72%; MRE: Test1: 0.92 mm, Test2: 1.41 mm; SDR within 2mm: Test1: 88.06%, Test2: 78.72%
[53]	2023	India	ISBI 2015 grand challenge dataset	19	Coarse localization through region of interest (ROI) extraction and fine localization utilizing histogram-oriented gradient (HOG) feature	SDR: 2.0 mm: 77.11%
[54]	2023	China	ISBI 2015 grand challenge dataset + 163 cephalograms	19	MS-YOLOv3	SDR: 2 mm: 80.84%, 4 mm: 98.14%
[55]	2023	China	481 cephalograms	61	Point-locating model CephaNET	SDR: 2.0 mm: 85.4%, 2.5 mm: 90.2%, 3.0 mm: 93.5%, 4.0 mm: 97%
[56]	2023	Japan	2,000 facial profile images	23	HRNetv2	SDR: 2.0 mm: 98.20%, 2.5 mm: 99.55%, 3.0 mm: 99.79%, 4.0 mm: 99.93%
[57]	2024	China	ISBI 2015 grand challenge dataset	19	Mask Region-based Convolutional Neural Network	SDR: 1.0 mm: 43.44%, 2.0 mm: 68.27%, 4.0 mm: 85.74%
[58]	2024	China	ISBI 2015 grand challenge dataset	19	Feature Decouple and Gated Recalibration Network (FDGR-Net)	SDR: 2.0 mm: 75.16%, 2.5 mm: 82.53%, 3.0 mm: 88.58%, 4.0 mm: 94.79%
[59]	2024	Taiwan	1,002 cephalograms	14	CNN architecture enhanced with U-Net and MobileNetV2	SDR: 2.0 mm: 83.14%, 2.5 mm: 89.62%, 3.0 mm: 93.97%, 4.0 mm: 98.23%
[60]	2024	China	ISBI 2015 grand challenge dataset	19	UniverDetect	SDR: 2 mm: 75.87%, 3 mm: 88.35%, 4 mm: 94.59%
[61]	2024	Pakistan	ISBI 2015 grand challenge dataset	19	Two-stage regression framework (LDM for coarse localization, LRM for fine-tuning)	SDR: Test1 - 2.0 mm: 87.17%, 3.0 mm: 95.18%, 4.0 mm: 98.12%; Test2 - 2.0 mm: 75.79%, 3.0 mm: 89.0%, 4.0 mm: 94.53%
[53]	2024	India	ISBI 2015 grand challenge dataset + Diverse-CEPH19	19	Detectron2, and YOLOv8	Detectron2: Best Model: rcnn_R_101_FPN_3x; MRE: 1.2 mm (±0.96 mm); SDR (2 mm): 85.89%; SDR (4 mm): 98.45%; YOLOv8: Best Model: YOLOv8m-pose; MRE: 1.62 mm (±1.03 mm); SDR (2 mm): 72.92%; SDR (4 mm): 96.80%
[62]	2024	Morocco	ISBI 2015 grand challenge dataset	19	YOLOv8	SDR : 2 mm: 86.31%; 2.5 mm: 87.69%; 3 mm: 90.84%

Our analysis of 76 studies reveals distinct developmental phases and several key insights that warrant critical examination. Cephalometric landmark detection has progressed through three distinct evolutionary phases. The first phase, from 2019 to 2020, focused on basic abilities using general models like YOLOv3, which reached an SDR value of 80.4% at 2mm thresholds. The intermediate refinement phase (2021-2022) implemented architectural innovations aimed at resolving challenges in medical imaging. Multi-stage frameworks, cascaded approaches, and attention mechanisms have become leading strategies. The advanced optimization phase (2023-2025) includes the use of complex transformer-based architectures [57], feature recalibration mechanisms [58], and multi-scale approaches [54], which have pushed performance limits even further. Research from this period indicates that SDR values consistently surpass 87% at 2mm thresholds in controlled settings, with certain methodologies [56] attaining notable accuracy (SDR of 98.2% at 2mm) on facial profile images.

Our statistical analysis of performance metrics across architectural categories provides crucial insights into effective design approaches (Table V).

This comparative analysis demonstrates a clear progression in accuracy metrics, with transformer-based and attention-enhanced architectures consistently outperforming earlier approaches. Notably, cascaded networks exhibit a favorable balance between accuracy and computational efficiency, an important consideration for clinical deployment.

Despite the significant progress in Deep Learning applications for cephalometric landmark detection, several persistent challenges and limitations remain that impede widespread clinical adoption. The reliance on limited datasets represents perhaps the most significant barrier to the clinical translation of Deep Learning for cephalometric analysis. Our comprehensive analysis of dataset characteristics across the literature (Table VI) reveals several critical concerns.

The median dataset size of 723 images falls dramatically short compared to medical imaging datasets in other domains, such as ChestX-ray14 (112,120 images) or MIMIC-CXR (377,110 images). This limitation is particularly concerning given the high dimensional complexity of craniofacial anatomy

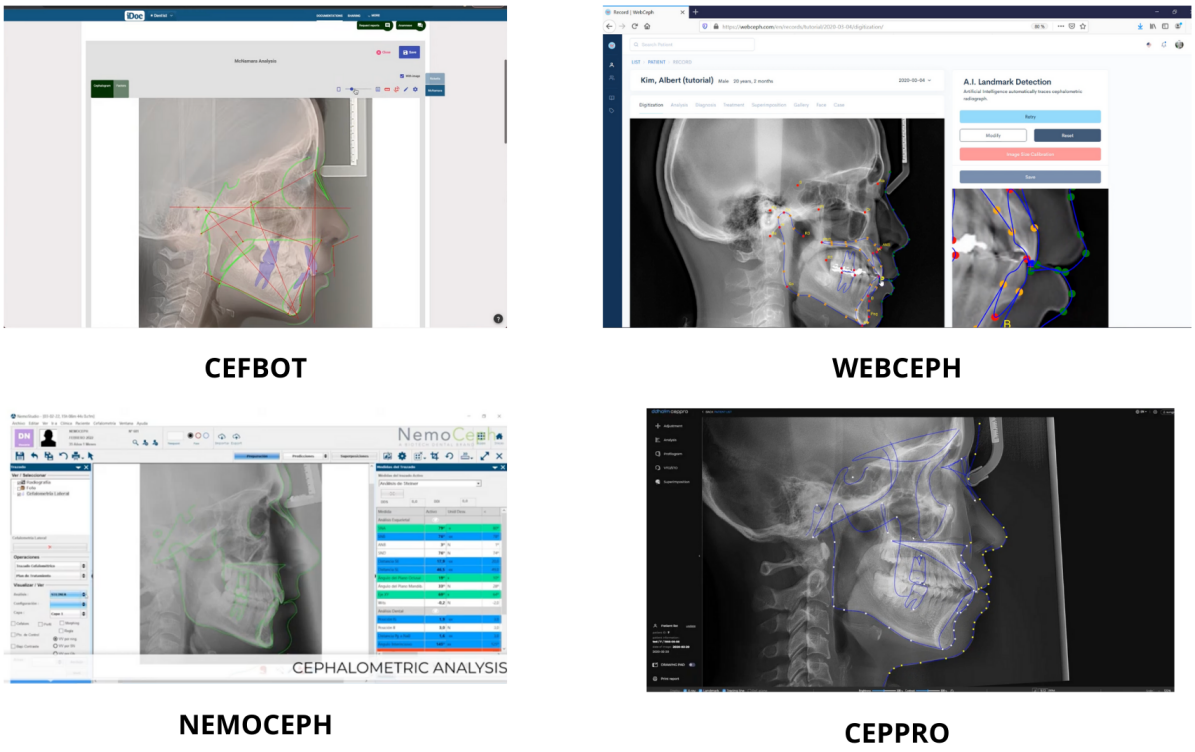


Fig. 6. Localization of oral cancer process.

TABLE IV. SUMMARY OF STUDIES EVALUATING COMMERCIAL SOFTWARE FOR AUTOMATED CEPHALOMETRIC ANALYSIS

Paper	Year	Software included
[63]	2021	Ceppro software by DDH Inc
[64]	2022	WebCeph
[65]	2022	WebCeph
[66]	2022	CEFBOT
[67]	2022	AudaxCeph
[68]	2022	Dolphin Imaging softwar, and CS Imaging V8 software
[69]	2022	WebCeph
[70]	2022	CephX
[71]	2023	Planmeca Romexis software
[72]	2023	MyOrthoX, Angelalign, Digident
[73]	2023	CEFBOT
[74]	2023	DentaliQ.ortho, WebCeph, AudaxCeph, and CephX
[75]	2023	CellmatIQ, CephX, AudaxCeph, and WebCeph
[76]	2023	CEFBOT
[77]	2024	CefBot, and WebCeph
[78]	2024	OneCeph app and WebCeph
[79]	2024	WebCeph, WeDoCeph, and CephX
[80]	2024	WebCeph, and CephX
[81]	2024	NemoCeph, and WebCeph
[82]	2024	WebCeph, Cephio, and Ceppro software

and the substantial anatomical variation across human populations.

For our analysis of dataset characteristics across the literature. It reveals that 55.3% of studies (42/76) relied exclusively on the ISBI 2015 dataset [4], which contains only 400 images primarily from a specific demographic population.

Studies that performed cross-dataset validation [38], [41], [48] consistently reported performance degradation when models trained on one dataset were applied to another. He et al. [38] demonstrated performance degradation of 15.4% when evaluating models trained on the ISBI dataset against their institutional dataset, highlighting the challenge of domain shift in clinical applications. Similarly, Šavc et al. [48] reported MRE values of 1.13mm on the ISBI dataset but substantially higher errors (11.26 ± 17.51 pixels) on their private AUDAX dataset.

A positive development is the emergence of larger datasets, such as that used by Jiang et al. [49], which included 9,870 cephalograms and demonstrated improved generalization capabilities. However, even this represents a fraction of the dataset sizes commonly used in other medical imaging domains, as highlighted by Zeng et al. [37] and Wang et al. [40].

The emergence of commercially available AI-driven cephalometric software (Fig. 6) represents a pivotal step toward clinical integration.

Our analysis of twenty studies evaluating commercial systems [83]-[81] reveals a complex picture of real-world application challenges. While commercial systems demonstrate impressive automation capabilities, validation studies consistently identify three critical limitations:

1) Inconsistent performance across landmark types: Commercial systems exhibit variable accuracy depending on landmark anatomical characteristics, with mean differences of 0.8 to 1.7 mm between predicted and ground truth positions for well-defined skeletal landmarks, but significantly higher deviations (1.9 to 3.2 mm) for soft tissue landmarks.

TABLE V. COMPARATIVE PERFORMANCE ANALYSIS BY ARCHITECTURE TYPE

Architecture Category	Mean SDR (2mm)	Mean MRE (mm)	Mean Detection Time (s)
Standard CNN-based	77.3% ± 3.8%	1.86 ± 0.27	0.78 ± 0.15
YOLO-based	79.5% ± 4.2%	1.53 ± 0.18	0.43 ± 0.07
Cascaded Networks	83.7% ± 3.1%	1.38 ± 0.15	0.95 ± 0.23
Attention-enhanced	85.9% ± 2.9%	1.29 ± 0.11	0.83 ± 0.19
Transformer-based	87.4% ± 2.4%	1.12 ± 0.09	1.21 ± 0.27

TABLE VI. DATASET SIZE AND DIVERSITY ANALYSIS; *DEMOGRAPHIC DIVERSITY INDEX (1-5 SCALE): 1=SINGLE POPULATION, 5=GLOBALLY REPRESENTATIVE

Dataset Type	Median Sample Size	Demographic Diversity*	Usage Frequency
ISBI 2015	400	1.2	55.30%
Private Institutional	723 (range: 163–9,870)	1.7	34.20%
Multi-center	1,872 (range: 950–3,150)	2.3	10.50%

2) Limited interchangeability with human experts: No commercial system achieved the clinically required reliability threshold (ICC > 0.95) across all cephalometric measurements necessary for complete diagnostic interchangeability with human experts.

3) Insufficient transparency: Commercial systems often lack methodological transparency regarding their underlying architectures, training protocols, and dataset characteristics, complicating clinical validation efforts.

Anatomical variability adds to the problems, especially when it comes to soft tissue landmarks like the soft tissue pogonion, which have many different structures and make predictions less accurate. The closeness of landmarks inside the skull makes automation harder because there aren't clear lines between structures that are next to each other. This can lead to unclear predictions, which lowers the accuracy of medical care.

Besides basic landmark detection, effective clinical integration needs more features, like measuring angles and distances, and should work well with common orthodontic software. Even though it works very well in controlled environments, how widely it is used in busy clinical settings depends a lot on things like how easy it is to use, how fast the computer can process data, and how well it works with other imaging systems. The computational resource demands for model training and deployment hinder the widespread adoption of AI-based approaches in clinical settings, despite their demonstrated promise.

VI. FUTURE DIRECTIONS

Given the significant challenges identified, various research directions require focus. The creation of extensive, multi-ethnic cephalometric datasets that include pathological diversity and are annotated with standardized protocols is crucial for enhancing model generalizability and fairness .

The establishment of unified benchmarking frameworks that include diverse test cohorts, clinically relevant evaluation metrics, and rigorous statistical analyses would greatly enhance the field's cross-study comparability.

Third, the design of parameter-efficient Deep Learning architectures specifically for clinical hardware deployment is essential for enabling real-world integration. Fourth, enhancing human–AI collaboration through interactive systems that use the complementary strengths of expert clinicians and AI models may facilitate adoption in standard practice.

Finally, the clinical utility of AI-driven cephalometric tools must be substantiated through prospective trials evaluating their impact on diagnostic accuracy, treatment planning, and patient outcomes.

VII. CONCLUSION

Artificial intelligence (AI) is progressively incorporated into orthodontic treatment, serving as a valuable tool for automating cephalometric landmark tracing in routine clinical practice. AI-powered systems are transforming contemporary orthodontic workflows by aiding clinicians in orthodontic treatment planning and decreasing the time needed for radiological diagnoses. Cephalometric analysis is an important way to assess the bones and structures of the face, making it a great option for using AI to automate processes in orthodontics, maxillofacial surgery, and craniofacial treatment.

However, the trustworthiness of AI in cephalometry varies, affected by different things like how accurately landmarks are identified by hand, the skill level of the person using it, the quality and number of X-rays, and the kind of algorithm used. These factors must be meticulously considered when analyzing AI-generated outcomes.

This scoping review indicates that although AI-powered cephalometric analysis has considerable potential, notable limitations persist, including issues related to accuracy, accessibility, expertise, ethical considerations, costs, and regulatory challenges.

Future research and technology should focus on improving accuracy, increasing the variety of data used, and tackling clinical validation issues to make sure that AI effectively helps and supports human skills in orthodontic and craniofacial diagnostics.

REFERENCES

[1] J. Liu, C. Zhang, and Z. Shan, "Application of artificial intelligence in orthodontics: current state and future perspectives," in *Healthcare*, vol. 11, p. 2760, MDPI, 2023.

[2] H. L. D. da Silveira and H. E. D. Silveira, "Reproducibility of cephalometric measurements made by three radiology clinics," *The Angle Orthodontist*, vol. 76, no. 3, pp. 394–399, 2006.

[3] S. Ö. Arık, B. Ibragimov, and L. Xing, "Fully automated quantitative cephalometry using convolutional neural networks," *Journal of Medical Imaging*, vol. 4, no. 1, pp. 014501–014501, 2017.

[4] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger, *et al.*, "A benchmark for comparison of dental radiography analysis algorithms," *Medical image analysis*, vol. 31, pp. 63–76, 2016.

[5] A. Polizzi and R. Leonardi, "Automatic cephalometric landmark identification with artificial intelligence: An umbrella review of systematic reviews," *Journal of Dentistry*, p. 105056, 2024.

[6] S. Narkhede, P. Rao, V. Sawant, S. S. Sachdev, S. Arora, A. M. Pawar, R. Reda, and L. Testarelli, "Digital versus manual tracing in cephalometric analysis: a systematic review and meta-analysis," *Journal of Personalized Medicine*, vol. 14, no. 6, p. 566, 2024.

[7] P. Smołka, K. Nelke, N. Struzik, K. Wiśniewska, S. Kiryk, J. Kensy, W. Dobrzyński, J. Kiryk, J. Matys, and M. Dobrzyński, "Discrepancies in cephalometric analysis results between orthodontists and radiologists and artificial intelligence: A systematic review," *Applied Sciences*, vol. 14, no. 12, p. 4972, 2024.

[8] M. Kiełczykowski, K. Kamiński, K. Perkowski, M. Zadurska, and E. Czochrowska, "Application of artificial intelligence (ai) in a cephalometric analysis: a narrative review," *Diagnostics*, vol. 13, no. 16, p. 2640, 2023.

[9] J. Londono, S. Ghasemi, A. H. Shah, A. Fahimipour, N. Ghadimi, S. Hashemi, Z. Khurshid, and M. Dashti, "Evaluation of deep learning and convolutional neural network algorithms accuracy for detecting and predicting anatomical landmarks on 2d lateral cephalometric images: a systematic review and meta-analysis," *The Saudi Dental Journal*, vol. 35, no. 5, pp. 487–497, 2023.

[10] M. S. Jha, "Cephalometric evaluation based on steiner's analysis on adults of bihar," *Journal of Pharmacy and Bioallied Sciences*, vol. 13, no. Suppl 2, pp. S1360–S1364, 2021.

[11] M. G. Hans, J. M. Palomo, and M. Valiathan, "History of imaging in orthodontics from broadbent to cone-beam computed tomography," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 148, no. 6, pp. 914–921, 2015.

[12] A. M. Aldrees, "Lateral cephalometric norms for saudi adults: A meta-analysis," *The Saudi dental journal*, vol. 23, no. 1, pp. 3–7, 2011.

[13] A. Dinesh, S. Mutalik, J. Feldman, and A. Tadinada, "Value-addition of lateral cephalometric radiographs in orthodontic diagnosis and treatment planning," *The Angle Orthodontist*, vol. 90, no. 5, pp. 665–671, 2020.

[14] G. Tsorovas and A. Linder-Aronson Karsten, "A comparison of hand-tracing and cephalometric analysis computer programs with and without advanced features—accuracy and time demands," *The European Journal of Orthodontics*, vol. 32, no. 6, pp. 721–728, 2010.

[15] T. Kanade, "Identification of human faces," in *Computer recognition of human faces*, pp. 47–63, Springer, 1977.

[16] J. Liu and J. K. Udupa, "Oriented active shape models," *IEEE Transactions on medical Imaging*, vol. 28, no. 4, pp. 571–584, 2008.

[17] T. F. Cootes and P. Kittipanya-ngam, "Comparing variations on the active appearance model algorithm.," in *BMVC*, pp. 1–10, Citeseer, 2002.

[18] C. Rahmad, R. A. Asmara, D. Putra, I. Dharma, H. Darmono, and I. Muhiqqin, "Comparison of viola-jones haar cascade classifier and histogram of oriented gradients (hog) for face detection," in *IOP conference series: materials science and engineering*, vol. 732, p. 012038, IOP Publishing, 2020.

[19] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1867–1874, 2014.

[20] A. Bulat, *Deep learning for real world face alignment*. PhD thesis, University of Nottingham, 2019.

[21] X. Lan, Q. Hu, Q. Chen, J. Xue, and J. Cheng, "Hih: Towards more accurate face alignment via heatmap in heatmap," *arXiv preprint arXiv:2104.03100*, 2021.

[22] A. Yueyuan and W. Hong, "Swin transformer combined with convolutional encoder for cephalometric landmarks detection," in *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 184–187, IEEE, 2021.

[23] A. Tamhane, T. Mida, E. Posner, and M. Bouhnik, "Colonoscopy landmark detection using vision transformers," in *MICCAI Workshop on Imaging Systems for GI Endoscopy*, pp. 24–34, Springer, 2022.

[24] M. J. Rajabi, M. Mirzai, and A. Nickabadi, "Enhancing landmark detection in cluttered real-world scenarios with vision transformers," *arXiv preprint arXiv:2308.13671*, 2023.

[25] J.-H. Park, H.-W. Hwang, J.-H. Moon, Y. Yu, H. Kim, S.-B. Her, G. Srinivasan, M. N. A. Aljanabi, R. E. Donatelli, and S.-J. Lee, "Automated identification of cephalometric landmarks: Part 1—comparisons between the latest deep-learning methods yolov3 and ssd," *The Angle Orthodontist*, vol. 89, no. 6, pp. 903–909, 2019.

[26] F. Kunz, A. Stellzig-Eisenhauer, F. Zeman, and J. Boldt, "Evaluation of a fully automated cephalometric analysis using a customized convolutional neural network," *J Orofac Orthop*, vol. 81, no. 1, pp. 52–68, 2020.

[27] H.-W. Hwang, J.-H. Park, J.-H. Moon, Y. Yu, H. Kim, S.-B. Her, G. Srinivasan, M. N. A. Aljanabi, R. E. Donatelli, and S.-J. Lee, "Automated identification of cephalometric landmarks: Part 2-might it be better than human?," *The Angle Orthodontist*, vol. 90, no. 1, pp. 69–76, 2020.

[28] H. Kim, E. Shim, J. Park, Y.-J. Kim, U. Lee, and Y. Kim, "Web-based fully automated cephalometric analysis by deep learning," *Computer methods and programs in biomedicine*, vol. 194, p. 105513, 2020.

[29] J.-H. Lee, H.-J. Yu, M.-j. Kim, J.-W. Kim, and J. Choi, "Automated cephalometric landmark detection with confidence regions using bayesian convolutional neural networks," *BMC oral health*, vol. 20, pp. 1–10, 2020.

[30] J. M. Noothout, B. D. De Vos, J. M. Wolterink, E. M. Postma, P. A. Smeets, R. A. Takx, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning-based regression and classification for automatic landmark localization in medical images," *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 4011–4022, 2020.

[31] Y. Song, X. Qiao, Y. Iwamoto, and Y.-w. Chen, "Automatic cephalometric landmark detection on x-ray images using a deep-learning method," *Applied Sciences*, vol. 10, no. 7, p. 2547, 2020.

[32] J. Qian, W. Luo, M. Cheng, Y. Tao, J. Lin, and H. Lin, "Cephann: a multi-head attention network for cephalometric landmark detection," *Ieee Access*, vol. 8, pp. 112633–112641, 2020.

[33] K. Oh, I.-S. Oh, D.-W. Lee, *et al.*, "Deep anatomical context feature learning for cephalometric landmark detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 3, pp. 806–817, 2020.

[34] M. Zeng, Z. Yan, S. Liu, Y. Zhou, and L. Qiu, "Cascaded convolutional networks for automatic cephalometric landmark detection," *Medical Image Analysis*, vol. 68, p. 101904, 2021.

[35] J. Kim, I. Kim, Y.-J. Kim, M. Kim, J.-H. Cho, M. Hong, K.-H. Kang, S.-H. Lim, S.-J. Kim, Y. H. Kim, *et al.*, "Accuracy of automated identification of lateral cephalometric landmarks using cascade convolutional neural networks on lateral cephalograms from nationwide multi-centres," *Orthodontics & Craniofacial Research*, vol. 24, pp. 59–67, 2021.

[36] C. Tanikawa, C. Lee, J. Lim, A. Oka, and T. Yamashiro, "Clinical applicability of automated cephalometric landmark identification: Part i—patient-related identification errors," *Orthodontics & Craniofacial Research*, vol. 24, pp. 43–52, 2021.

[37] Y. H. Kim, C. Lee, E.-G. Ha, Y. J. Choi, and S.-S. Han, "A fully deep learning model for the automatic identification of cephalometric landmarks," *Imaging Science in Dentistry*, vol. 51, no. 3, p. 299, 2021.

[38] T. He, J. Yao, W. Tian, Z. Yi, W. Tang, and J. Guo, "Cephalometric landmark detection by considering translational invariance in the two-stage framework," *Neurocomputing*, vol. 464, pp. 15–26, 2021.

[39] H. J. Kwon, H. I. Koo, J. Park, and N. I. Cho, "Multistage probabilistic approach for the localization of cephalometric landmarks," *IEEE Access*, vol. 9, pp. 21306–21314, 2021.

[40] L. Wang, L. Ma, Y. Li, K. Niu, and Z. He, "A dcnn system based on an iterative method for automatic landmark detection in cephalometric x-ray images," *Biomedical Signal Processing and Control*, vol. 68, p. 102757, 2021.

[41] J. Yao, W. Zeng, T. He, S. Zhou, Y. Zhang, J. Guo, and W. Tang, "Automatic localization of cephalometric landmarks based on convolutional neural network," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 161, no. 3, pp. e250–e259, 2022.

[42] V. N. T. Le, J. Kang, I.-S. Oh, J.-G. Kim, Y.-M. Yang, and D.-W. Lee, "Effectiveness of human–artificial intelligence collaboration in cephalometric landmark detection," *Journal of Personalized Medicine*, vol. 12, no. 3, p. 387, 2022.

[43] M. Lee, M. Chung, and Y.-G. Shin, "Cephalometric landmark detection via global and local encoders and patch-wise attentions," *Neurocomputing*, vol. 470, pp. 182–189, 2022.

[44] H.-J. Yoon, D.-R. Kim, E. Gwon, N. Kim, S.-H. Baek, H.-W. Ahn, K.-A. Kim, and S.-J. Kim, "Fully automated identification of cephalometric landmarks for upper airway assessment using cascaded convolutional neural networks," *European Journal of Orthodontics*, vol. 44, no. 1, pp. 66–77, 2022.

[45] S.-M. Gil, I. Kim, J.-H. Cho, M. Hong, M. Kim, S.-J. Kim, Y.-J. Kim, Y. H. Kim, S.-H. Lim, S. J. Sung, *et al.*, "Accuracy of auto-identification of the posteroanterior cephalometric landmarks using cascade convolution neural network algorithm and cephalometric images of different quality from nationwide multiple centers," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 161, no. 4, pp. e361–e371, 2022.

[46] M. Šavc, G. Sedej, and B. Potočnik, "Cephalometric landmark detection in lateral skull x-ray images by using improved spatialconfiguration-net," *Applied Sciences*, vol. 12, no. 9, p. 4644, 2022.

[47] K. Huang and F. Feng, "An intelligent shooting reward learning network scheme for medical image landmark detection," *Applied Sciences*, vol. 12, no. 20, p. 10190, 2022.

[48] Y. Ao and H. Wu, "Feature aggregation and refinement network for 2d anatomical landmark detection," *Journal of Digital Imaging*, vol. 36, no. 2, pp. 547–561, 2023.

[49] F. Jiang, Y. Guo, C. Yang, Y. Zhou, Y. Lin, F. Cheng, S. Quan, Q. Feng, and J. Li, "Artificial intelligence system for automated landmark localization and analysis of cephalometry," *Dentomaxillofacial Radiology*, vol. 52, no. 1, p. 20220081, 2023.

[50] J. V. Weingart, S. Schlager, M. C. Metzger, L. S. Brandenburg, A. Hein, R. Schmelzeisen, F. Bamberg, S. Kim, E. Kellner, M. Reisert, *et al.*, "Automated detection of cephalometric landmarks using deep neural patchworks," *Dentomaxillofacial Radiology*, vol. 52, no. 6, p. 20230059, 2023.

[51] S. Yang, E. S. Song, E. S. Lee, S.-R. Kang, W.-J. Yi, and S.-P. Lee, "Correction: Ceph-net: automatic detection of cephalometric landmarks on scanned lateral cephalograms from children and adolescents using an attention-based stacked regression network," *BMC Oral Health*, vol. 24, 2024.

[52] R. Neeraja and L. J. Anbarasi, "Cephxnet: A deep convolutional squeeze-and-excitation model for landmark prediction on lateral cephalograms," *IEEE Access*, 2023.

[53] S. Rashmi, S. Srinath, K. Patil, P. S. Murthy, and S. Deshmukh, "Lateral cephalometric landmark annotation using histogram oriented gradients extracted from region of interest patches," *Journal of Maxillofacial and Oral Surgery*, vol. 22, no. 4, pp. 806–812, 2023.

[54] C. Zhao, Z. Yuan, S. Luo, W. Wang, Z. Ren, X. Yao, and T. Wu, "Automatic recognition of cephalometric landmarks via multi-scale sampling strategy," *Heliyon*, vol. 9, no. 6, 2023.

[55] B. Gong, S. Chang, F. Zuo, X. Xie, S. Wang, Y. Wang, Y. Sun, X. Guan, and Y. Bai, "Automated cephalometric landmark identification and location based on convolutional neural network," *Zhonghua kou Qiang yi xue za zhi= Zhonghua Kouqiang Yixue Zazhi= Chinese Journal of Stomatology*, vol. 58, no. 12, pp. 1249–1256, 2023.

[56] K. Takahashi, Y. Shimamura, C. Tachiki, Y. Nishii, and M. Hagiwara, "Cephalometric landmark detection without x-rays combining coordinate regression and heatmap regression," *Scientific reports*, vol. 13, no. 1, p. 20011, 2023.

[57] Z. Jiao, Z. Liang, Q. Liao, S. Chen, H. Yang, G. Hong, and H. Gui, "Deep learning for automatic detection of cephalometric landmarks on lateral cephalometric radiographs using the mask region-based convolutional neural network: a pilot study," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 137, no. 5, pp. 554–562, 2024.

[58] X. Li, S. Lv, J. Zhang, M. Li, J. J. Rodriguez-Andina, Y. Qin, S. Yin, and H. Luo, "Fdgr-net: Feature decouple and gated recalibration network for medical image landmark detection," *Expert Systems with Applications*, vol. 238, p. 121746, 2024.

[59] H.-T. Lee, P.-Y. Chiu, C.-W. Yen, S.-T. Chou, and Y.-C. Tseng, "Application of artificial intelligence in lateral cephalometric analysis," *Journal of Dental Sciences*, vol. 19, no. 2, pp. 1157–1164, 2024.

[60] C. Lu, G. Yang, X. Qiao, W. Chen, and Q. Zeng, "Univerdetect: Universal landmark detection method for multidomain x-ray images," *Neurocomputing*, vol. 600, p. 128157, 2024.

[61] M. A. Khalid, A. Khurshid, K. Zulfiqar, U. Bashir, and M. M. Fraz, "A two-stage regression framework for automated cephalometric landmark detection incorporating semantically fused anatomical features and multi-head refinement loss," *Expert Systems with Applications*, vol. 255, p. 124840, 2024.

[62] I. Tafala, F.-E. Ben-Bouazza, A. Edder, O. Manchadi, M. Et-Taoussi, and B. Jioudi, "Cephalometric landmarks identification through an object detection-based deep learning model.," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 2, 2024.

[63] G. Bulatova, B. Kusnoto, V. Grace, T. P. Tsay, D. M. Avenetti, and F. J. C. Sanchez, "Assessment of automatic cephalometric landmark identification using artificial intelligence," *Orthodontics & Craniofacial Research*, vol. 24, pp. 37–42, 2021.

[64] R. K. Mahto, D. Kafle, A. Giri, S. Luintel, and A. Karki, "Evaluation of fully automated cephalometric measurements obtained from web-based artificial intelligence driven platform," *BMC Oral Health*, vol. 22, no. 1, p. 132, 2022.

[65] Y. A. Yassir, A. R. Salman, and S. A. Nabbat, "The accuracy and reliability of webceph for cephalometric analysis," *Journal of Taibah University Medical Sciences*, vol. 17, no. 1, pp. 57–66, 2022.

[66] T. P. Silva, M. M. Hughes, L. d. S. Menezes, M. d. F. B. de Melo, P. H. L. d. Freitas, and W. M. Takeshita, "Artificial intelligence-based cephalometric landmark annotation and measurements according to arnett's analysis: can we trust a bot to do that?," *Dentomaxillofacial Radiology*, vol. 51, no. 6, p. 20200548, 2022.

[67] B. Ristau, M. Coreil, A. Chapple, P. Armbruster, and R. Ballard, "Comparison of audaxceph®'s fully automated cephalometric tracing technology to a semi-automated approach by human examiners," *International orthodontics*, vol. 20, no. 4, p. 100691, 2022.

[68] I. A. Tsolakis, A. I. Tsolakis, T. Elshebiny, S. Matthaios, and J. M. Palomo, "Comparing a fully automated cephalometric tracing method to a manual tracing method for orthodontic diagnosis," *Journal of Clinical Medicine*, vol. 11, no. 22, p. 6854, 2022.

[69] G. Çoban, T. Öztürk, N. Hashimli, and A. Yağci, "Comparison between cephalometric measurements using digital manual and web-based artificial intelligence cephalometric tracing software," *Dental Press Journal of Orthodontics*, vol. 27, no. 4, p. e222112, 2022.

[70] M. Davidovitch, T. Sella-Tunis, L. Abramovicz, S. Reiter, S. Matalon, and N. Shpack, "Verification of convolutional neural network cephalometric landmark identification," *Applied Sciences*, vol. 12, no. 24, p. 12784, 2022.

[71] H. Bao, K. Zhang, C. Yu, H. Li, D. Cao, H. Shu, L. Liu, and B. Yan, "Evaluating the accuracy of automated cephalometric analysis based on artificial intelligence," *BMC Oral Health*, vol. 23, no. 1, p. 191, 2023.

[72] H. Ye, Z. Cheng, N. Ungvijanpunya, W. Chen, L. Cao, and Y. Gou, "Is automatic cephalometric software using artificial intelligence better than orthodontist experts in landmark identification?," *BMC Oral Health*, vol. 23, no. 1, p. 467, 2023.

[73] S. Panesar, A. Zhao, E. Hollensbe, A. Wong, S. S. Bhamidipalli, G. Eckert, V. Dutra, and H. Turkkahraman, "Precision and accuracy assessment of cephalometric analyses performed by deep learning artificial intelligence with and without human augmentation," *Applied Sciences*, vol. 13, no. 12, p. 6921, 2023.

[74] F. Kunz, A. Stellzig-Eisenhauer, L. M. Widmaier, F. Zeman, and J. Boldt, "Assessment of the quality of different commercial providers using artificial intelligence for automated cephalometric analysis compared to human orthodontic experts," *Journal of Orofacial Orthopedics/Fortschritte der Kieferorthopädie*, pp. 1–16, 2023.

[75] J. Lee, S.-R. Bae, and H.-K. Noh, "Commercial artificial intelligence lateral cephalometric analysis: part 1–the possibility of replacing manual landmarking with artificial intelligence service.," *Journal of Clinical Pediatric Dentistry*, vol. 47, no. 6, 2023.

[76] L. d. S. Menezes, T. P. Silva, M. A. Lima dos Santos, M. M. Hughes, S. d. R. Mariano Souza, P. M. Leite Ribeiro, P. H. L. d. Freitas, and W. M. Takeshita, "Assessment of landmark detection in cephalometric radiographs with different conditions of brightness and contrast using the an artificial intelligence software," *Dentomaxillofacial Radiology*, vol. 52, no. 8, p. 20230065, 2023.

[77] T. P. Silva, M. C. R. Pinheiro, D. Q. Freitas, H. Gaêta-Araujo, and C. Oliveira-Santos, "Assessment of accuracy and reproducibility of cephalometric identification performed by 2 artificial intelligence-driven tracing applications and human examiners," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 137, no. 4, pp. 431–440, 2024.

[78] S. Gupta, S. Shetty, S. Natarajan, S. Nambiar, M. Ashith, and S. Agarwal, "A comparative evaluation of concordance and speed between smartphone app-based and artificial intelligence web-based cephalometric tracing software with the manual tracing method: A cross-sectional study," *Journal of clinical and experimental dentistry*, vol. 16, no. 1, p. e11, 2024.

[79] S. Bor, S. Ç. Ciğerim, and S. Kotan, "Comparison of ai-assisted cephalometric analysis and orthodontist-performed digital tracing analysis," *Progress in Orthodontics*, vol. 25, no. 1, p. 41, 2024.

[80] R. Zaheer, H. Z. Shafique, Z. Khalid, R. Shahid, A. Jan, T. Zahoor, R. Nawaz, and M. ul Hassan, "Comparison of semi and fully automated artificial intelligence driven softwares and manual system for cephalometric analysis," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 271, 2024.

[81] J.-P. Mercier, C. Rossi, I. N. Sanchez, I. D. Renovales, P. M.-P. Sahagún, and L. Templier, "Reliability and accuracy of artificial intelligence-based software for cephalometric diagnosis. a diagnostic study," *BMC Oral Health*, vol. 24, no. 1, pp. 1–19, 2024.

[82] N. Baig, K. S. Gyasudeen, T. Bhattacharjee, J. Chaudhry, and S. Prasad, "Comparative evaluation of commercially available ai-based cephalometric tracing programs," *BMC Oral Health*, vol. 24, no. 1, p. 1241, 2024.

[83] M. Hori, M. Jincho, T. Hori, H. Sekine, A. Kato, K. Miyazawa, and T. Kawai, "Automatic point detection on cephalograms using convolutional neural networks: A two-step method," *Dental Materials Journal*, vol. 43, no. 5, pp. 701–710, 2024.

Advancing Traffic Sign Detection with Convolutional Neural Networks: A Deep Learning Approach

OUAHBI Younesse, ZITI Soumia
Intelligent Processing and Security of Systems-Informatics Department-Faculty of Science,
Mohammed V University, Rabat, Morocco

***Abstract*—Traffic sign detection is a key task in intelligent transportation systems, supporting road safety and traffic flow. This study introduces RoadNet, a lightweight Convolutional Neural Network (CNN) designed for real-time detection and classification of traffic signs in Moroccan road environments. The system addresses challenges such as occlusion, illumination variability, and diverse sign structures. Built on deep learning techniques, RoadNet leverages multiscale feature extraction and transfer learning to improve detection accuracy and generalization. The dataset includes four sign categories: speed limit, stop, crosswalk, and traffic light. Extensive image preprocessing and augmentation were applied to increase robustness. Results show that RoadNet outperforms baseline models like VGG16, achieving 96% training accuracy and 88.6% validation accuracy, with superior precision, recall, and F1-score. The model maintains low loss and performs reliably under constrained resources. This research confirms the effectiveness of CNN-based architectures for traffic sign detection in real-world Moroccan settings. It contributes to the deployment of AI-powered solutions for smart mobility and logistics, especially in regions with limited computational resources.**

Keywords*—*Traffic sign detection; convolutional neural networks; deep learning; road safety; intelligent transportation systems; real-time detection; artificial intelligence; transportation efficiency

I. Introduction

Traffic sign detection is a cornerstone of intelligent transportation systems, tackling key challenges related to road safety and traffic efficiency. As the current installation of smarter infrastructure spins into motion, the integration of such technology becomes crucial to heighten the capability of the traffic management systems in curtailing accidents. Being integral components of road infrastructure, traffic signs are intended to regulate surface vehicular traffic flow and guide both human drivers and autonomous systems. Their accurate detection, recognition, and classification are vital in ensuring compliance with road rules, reducing accidents, and boosting transportation efficiency . In Morocco, the transport and logistics sector is certainly the base major contributor to national development. The country has made a stupendous investment in spreading across major ports and highway networks-an opening for better access through great deleveraging. Even though the situation is very much improved, road safety remains a hot topic for deliberation: the volume of accidents ensures adequate provision of strengthened legislation with a view to reducing deaths and improving the pace of road traffic. The diversification of the different kinds of traffic signs and the even vaguer environmental phenomena pose tedious unique challenges requiring intelligent and consolidated detection systems . Rapid advances in industrial artificial intelligence, especially those associated with deep learning arms with transformational potential for traffic sign detection. CNNs have become the leading technology, learning complex feature representations from raw visual data. Unlike traditional techniques that excelled on handcrafted feature extraction, CNNs extract the hierarchical features that give more robustness against the changing conditions of environmental light and occlusions, as well as confronting other practical difficulties. Sadly, these technological developments actually leave wide gaps in the operational continuum of scalable, real-time traffic sign-detection systems that adhere to the tenets of resource-constrained environments, such as embedded vehicle systems. The variation in traffic sign design, arrangement, and environmental conditioning widens these challenges. It is against this range of gaps that only a combined approach encompassing state-of-the-art deep learning architectures, efficient optimization techniques, and domain-specific adaptations may be honestly considered. Against this backdrop essentially lies NW Logistics, a modern A.I. solution that intends to transform the logistic and transportation sectors in Morocco. The project comprises traffic sign detection systems, real-time monitoring of the processes, and optimization of routes so as to better enhance logistics efficiency and road safety. Using deep learning frameworks, NW Logistics contributes on accident reduction, route optimization, and monitoring across performance spots in logistic operations. It also incorporates sustainable development, actively supporting greens logistics. By features such as CO2 emission tracking and optimization strategies, it is already connected with current global demands for local operations being environmentally friendly, thus allowing a business to move into green status.

This article proposes a new CNN-based framework that caters to the traffic sign detection needs of real-world applications. It implements advanced deep learning techniques such as multi-scale feature extraction and transfer learning to increase accuracy and computational efficiency. This study also looks at optimization strategies aimed at enabling adaptability and robustness across different operational contexts, coming in line with Morocco's strategic priorities around road safety and sustainable transport.

The paper is organized in the following way: an overview of the existing literature concerning traffic sign detection and its challenges; a detailed outline of the methodology concerning the design and implementation of the proposed framework; experimental results that show its effectiveness across different kinds of benchmarks; and a section describing the implications of these findings on road safety and intelligent transportation. The conclusion emphasizes how the research impacts a broader

scope and discusses future directions concerning the advancement of this field.

By addressing relevant problems in traffic sign detection, this study aids ongoing efforts to improve road safety, optimize logistics operations, and develop environmentally sound transportation systems in Morocco and beyond. To this end, it is an excellent example of using artificial intelligence to provide smarter, safer, and greener mobility solutions. This study aims to answer the following research question: How can CNN-based architectures improve real-time traffic sign detection under resource constraints in Moroccan road environments?

II. Related Work

Traffic sign detection is an important task in intelligent transportation systems (ITS) and autonomous driving systems, and acts as a base layer in traffic safety as well as vehicle navigation. It is the problem of recognizing the on screen location, type, and dimensions of traffic signs from images or video footage [1]. This kind of technology are rising in importance due to the rising interest in, and development of, self driving vehicles, with much effort being placed in the enhancement of traffic sign detection and classification systems [2]. Conventional Computer Vision Methods (CCVM) include edge based detection of image, filtering and morphological operation are the traditional way for Traffic Sign Recognition Systems (TSRS) [3]. Most of these methods utilized the unique shape and color features of the traffic signs [1, 4]. But these conventional techniques suffered from lack of robustness and generality in complex scenarios [1].

The development of deep learning, especially convolutional neural network (CNN) applied by [(5)], has revolutionized the pursuit of traffic sign detection. CNNs were found to be very effective on traffic sign detection with learning discriminative features from sign images automatically, which can generalize well to different lighting, viewpoints and sign appearances [6]. CNNs are clearly up to the task of replacing handcrafted features in image classification as was shown prominently by Krizhevsky et al. in 2012, whose work presented remarkable advances over other state-of-the-art techniques [7].

In recent CNN-based traffic sign detection systems, the architecture like Faster R-CNN 8 is frequently used, where an image features are extracted by a convolutional neural network automatically and a Region Proposal Network (RPN) selects higheresolution region proposals for traffic signs and regresses bounding boxes[8][9]. Other well-known architectures are R-CNN, Fast R-CNN, Mask R-CNN, which progressively improve the efficiency and accuracy of detection [1][10-12].

Object Detection in Traffic Scenes One of the main problems in the traffic sign detection is how to accurately and speedily detect small-sized scales in sprawling or complicated traffic scene [13]. Even though much has been achieved, but there are still challenges when it comes to CNN-based detection systems for traffic signs, in terms of real-time and accurate detection [2][14].

CNN Traffic Sign Detection systems have applications beyond the scope of autonomous vehicle systems, such as on intelligent driver assistance systems (IDASs) and more generally, on intelligent transportation systems (ITSs) [15]. Such systems accept images, applies one or more methods such as pre-processing, feature extraction, and classification to achieve high recognition accuracy, some systems have claimed to reach accuracy rates above 98% on standard datasets [13][16].

The use of neural networks in traffic sign detection and recognition is increasing due to the development of image recognition capability for neural network, which greatly promotes the development of intelligent transportation system and network vehicle [17][18].

In the overview, recent works between 2021 and 2024 have increasingly addressed the challenges of traffic sign detection in real-world scenarios. Cao et al. (2021) introduced an improved Sparse R-CNN architecture optimized with Res2Nest for multi-scale detection. Zhang et al. (2023) proposed RTS R-CNN, enhancing small object detection through PAFPN and asymmetric kernels. These methods emphasize detection precision but often rely on high-end computational resources.

In contrast, RoadNet builds on similar detection concepts while optimizing the model for real-time applications in low-resource environments. Compared to MobileNet-based approaches used by Li and Wang (2019), RoadNet balances accuracy and processing efficiency. While foundational models like Krizhevsky et al. (2012) laid the groundwork for CNN-based classification, their direct relevance has declined. These earlier references should be kept only for historical framing and moved to a background subsection.

Therefore, this work positions RoadNet as a bridge between recent state-of-the-art techniques and practical, low-resource implementations, particularly in the context of emerging smart mobility frameworks in developing regions.

A. CNN Architectures in Traffic Sign Detection

Traffic sign detection research has utilized a variety of CNN architectures, each with different strengths for addressing the unique challenges of detecting traffic signs in complex environments [19][20]. The following architectures are commonly employed (Table I):

1) General-Purpose CNN backbones:

a) ResNet: Used as a backbone in numerous traffic sign detection systems due to its ability to effectively learn deep features while addressing the vanishing gradient problem [21][22].

b) VGG: Frequently employed for feature extraction in traffic sign detection, especially in two-stage detection approaches [23].

c) GoogLeNet/Inception: Used to extract rich features for traffic sign localization (Temel et al., 2019).

d) MobileNet: Adapted for traffic sign detection to achieve computational efficiency on mobile platforms, making it suitable for real-time applications in vehicles [24].

2) Specialized detection architectures:

a) Faster R-CNN: A popular framework for traffic sign detection that combines a Region Proposal Network (RPN) with a classification network, offering good precision but at higher computational cost [24][25].

b) YOLO variants: Applied for real-time traffic sign detection due to their speed and one-stage detection approach [26].

c) SSD (Single Shot Detector): Used as a one-stage detector for real-time traffic sign detection applications.

d) U-Net: Employed for its encoder-decoder architecture that preserves spatial information, which is valuable for detecting small traffic signs [23].

e) Mask R-CNN: Adapted for traffic sign detection to provide not only bounding boxes but also precise segmentation masks of traffic signs [27].

f) RetinaNet: Used for its focal loss function that addresses class imbalance issues in traffic sign datasets [28].

g) Sparse R-CNN: Implemented with residual connections and self-attention mechanisms for detecting traffic signs in challenging environmental conditions [21].

3) Specialized architectures for small traffic signs:

a) Multi-Scale Deconvolutional Networks (MDN): Designed specifically for real-time traffic sign detection, combining multi-scale CNN with deconvolution networks [29].

b) Multi-resolution feature fusion networks: Incorporate densely connected deconvolution layers with skip connections to better detect small-sized traffic signs [30].

4) Novel and custom architectures:

a) Capsule networks: Applied to traffic sign detection to improve robustness against rotation, deformation, and fuzzy images [31][32].

b) Attention-based networks: Incorporate attention mechanisms to focus on relevant features for traffic sign detection [30][33].

c) Two-path CNN architectures: Combine multiple CNN models or processing paths to exploit both shape and color properties of traffic signs [34].

d) Fully Convolutional Networks (FCN): Used for end-to-end traffic sign detection to directly predict sign locations and classes [35][36].

Many researchers have developed hybrid approaches that combine multiple architectures to leverage their respective strengths. For example, combining U-Net for localization with ResNet for classification [23], or integrating attention mechanisms into standard detection frameworks like Faster R-CNN [37]. These combinations aim to address the specific challenges of traffic sign detection, such as small object size, environmental variations, and the need for real-time performance.

III. Methodology

The development of this logistics platform follows a structured and systematic approach in a bid to bring more efficiency into operations and make roads safer to traverse. The development of the platform can, therefore, be categorized into four phases:

A. Identification of Needs

the process to collect, derive, and summarize needs from critical stakeholders, whereby it also tries to make a deeper understanding of the challenges related to traffic sign variability, environmental impact, and safety concerns and thereby developing the platform to meet industry needs and those specific to the regions around.

B. Platform Design

Based on the identified needs, this phase focuses on architecting a modular and scalable platform. The design includes AI-based components for traffic sign recognition, CO2 emission monitoring, and logistics route optimization, ensuring adaptability to diverse operating environments Fig. 1.

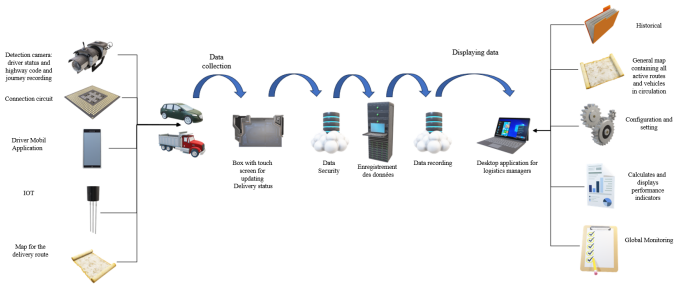


Fig. 1. Platform design.

C. Development

This article explores the development phase, shedding some light on the critical technical component: the traffic sign detection. Central to this phase is the design and implementation of a novel Neural Network known as "Road-NET," developed to detect and classify traffic signs with the best possible accuracy and speed. This includes the use of advanced deep learning techniques such as multiscale feature extraction and transfer learning to overcome the real-world challenges such as changes in lighting, different settings of occlusion, etc. and variability in traffic sign designs. This Informs about the model architecture and the training process, as well as results in the evaluation of the performance of the Road-NET model.

D. Implementation

The last phase is concerned with the setting up and running of the logistics business in the real world, which integrates into current systems and ensures its usability training and other monitoring things for the operations. Performance metrics, such as detection accuracy, system efficiency, and environmental impact reduction, are continuously evaluated to refine the platform.

E. Pre_processing and Architecture

1) Image Pre_processing: Image pre-processing is mostly an essential step that precedes images being exploited by the machine learning model Fig. 2.

The images are prepared according to the input form of the neural network and enable quality improvements of the models' performance Fig. 3.

TABLE I. COMPARISON OF CNN-BASED TRAFFIC SIGN DETECTION METHODS

Papers	CNN Architecture	Detection Method	Backbone Model	Feature Fusion or Integration Techniques	Dataset Used	Performance Metrics
[21]	Improved Sparse R-CNN with Res2Nest backbone and cross-channel self-attention	Improved Sparse R-CNN with Res2Nest and cross-channel self-attention	Res2Nest with enhanced multi-scale representation ability	Multi-scale fusion via enhanced backbone with hierarchical residual-like connections	TT100K dataset	Better accuracy and robustness
[22]	RTS R-CNN with CSP-DarkNet53_ECA and GR-PAFPN, improving Mask R-CNN with ECA, RFA, ASPP, and BFP	RTS R-CNN with CSP-DarkNet53_ECA and GR-PAFPN	CSPDarkNet53_ECA replaces ResNet50	GR-PAFPN with RFA, ASPP, and BFP for small object detection	Ceymo dataset (2887 images, 4706 instances)	Macro F1-score, precision, recall, IoU
[24]	Faster R-CNN, MobileNet structure, efficient CNN with asymmetric kernels	Faster R-CNN with MobileNet and asymmetric kernels	Faster R-CNN and MobileNet structure	Color and shape info used to refine localizations	Public benchmarks (not specified)	N/A
[25]	Improved Faster R-CNN with Res2Net backbone	Improved Faster R-CNN with new sampling method and Res2Net	Res2Net as the backbone	New sampling method and Res2Net backbone	Tsinghua-Tencent 100k dataset	Accuracy and recall
[33]	MobileNet with multi-resolution conv-deconv and VSSA module	Multi-resolution feature fusion with VSSA	MobileNet with added convolution layers	Densely connected deconvolution layers with skip connections	STS and OPTTSR datasets	Avg. precision, recall, mAP
[37]	Cascaded R-CNN, multi-scale features, joint training, multiscale attention	Cascaded R-CNN with multiscale attention and dataset balance	Cascaded R-CNN with multiscale attention	Multiscale attention to fuse multiscale features	GTSDB, CCTSDB, LISA, German datasets	Accuracy: 98.7%, Recall: 90.5% (GTSDB)

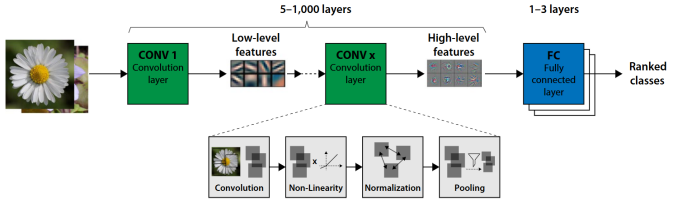


Fig. 2. General image pre_processing phases.

a) Resizing: Images are resized to a uniform size (300x300 pixels in your code). This ensures that all input images have the same dimensions and optimizes model calculations.

b) Normalization: The pixel values of the images are converted to values between 0 and 1 by dividing by 255. They are then centered using ImageNet-style normalization statistics (mean and standard deviation), making the images compatible with the pre-trained models.

c) Transform: Random transformations can be applied to images and their associated masks, such as rotations, horizontal symmetries and random cropping. This generates a wider range of training data and makes the model more robust to variations in the environment.

d) Creation of masks and bounding boxes: For each image, a binary mask is generated around the object of interest (e.g. a traffic sign). This mask is used to identify the pixels corresponding to the object and to create bounding boxes around it.

e) Image and mask conversion for the model: Images are then converted into a format suitable for input to a deep learning model (e.g. a NumPy array with the image channels reorganized). Bounding boxes are also transformed to be compatible with model inputs (for example, into normalized coordinates for a resized image).

f) Data augmentation: These steps include transformations such as rotation, slicing and symmetry to enrich the training data and avoid overlearning. This also enables the simulation of various lighting conditions, perspective or object orientation in the images.

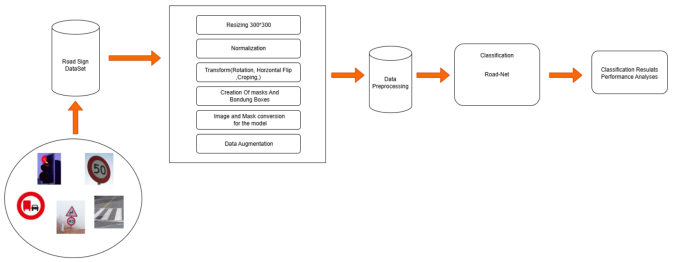


Fig. 3. Image Pre_processing phase.

F. Road_Net Architecture

Road_Net uses Convolutional Neural Networks as the main approach to detect the traffic signs and classify their information from datasets:road-sign- detection. The Dataset encompassing four classes (speedlimit,crosswalk,Stop, trafficlight) relays that a specialized type of Artificial Neural Network (ANN) would be set in place to automatically extract and learn both low-level and high-level features from road sign images, such as edges, curves, and others that play a crucial role in identifying, categorizing. The CNN architecture in this study contains several layers of convolution, pooling, and fully connected layers as depicted in Fig. 4.

Convolutional layers apply a linear filter series on each image or activation maps generated from the previous layer. These features are then sent through pooling layers, reducing dimensionality, and fully connecting layers, which integrate extracted features for classification. The architecture is optimized using cross-entropy loss in training to fine-tune the model's

parameters, thereby maximizing classification accuracy. The CNN model architecture includes five convolutional layers Table II, four max-pooling layers, four batch normalization layers, two drop-outlayers, and two fully connected layers.

TABLE II. ROAD_NET LAYERS AND PARAMETERS

Layer	Parameters
CONV-1 MAX-POOLING BATCH-NORMALIZATION	64 filters, 11×11, padding=5, stride=4 2×2
CONV-2 MAX-POOLING BATCH-NORMALIZATION	128 filters, 5×5, padding=2, stride=2 2×2
CONV-3 MAX-POOLING BATCH-NORMALIZATION	256 filters, 3×3, padding=1, stride=4 2×2
DROPOUT FCL DROPOUT	0.1 256 0.2

The activation function used in the convolutional layer is the ReLU function and is responsible for introducing non-linearity into the architecture, while the Softmax activation function is used in the last dense layer to yield class probabilities. Max pooling layers help reduce the spatial dimensions of these feature maps, while batch normal- ization speeds up training and stabilizes it and dropout layers are included to prevent overfitting. The model design attempts to achieve high accuracy without compromising simplicity and can be efficiently applied in real life, especially with a machine with limited computational resources. Detailed parameters for all layers contributing to the CNN architecture are presented in Table I, and complete architecture is illustrated in Fig. 4. By leveraging the strengths of CNNs.

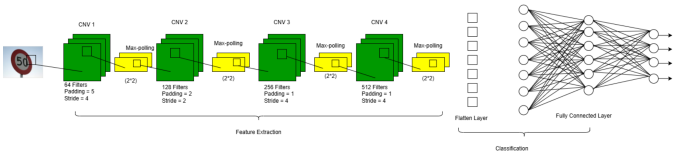


Fig. 4. RoadNet CNN architecture.

G. Evaluating Metrics

Once the developed models are implemented and trained, the subsequent step involves applying evaluation metrics to assess their performance. In this work, several evaluation methods are employed. First, comparison metrics including accuracy, recall, precision, and F1-score are utilized. Second, a normalized confusion matrix provides a comprehensive view of the model's classification performance. Third, the Receiver Operating Characteristic (ROC) curve and the area under the ROC (AUC) are analyzed to assess the model's classification task performance.
The metrics used are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Where TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative respectively, in the classification of predicting the presence or absence of retinal diseases in medical images. These outcomes are crucial for calculating evaluation metrics such as sensitivity, specificity, accuracy, and F1 score, providing insights into the model's performance.

IV. RESULTS AND EXPERIMENTAL DISCUSSION

In this section, we present experimental results of our Bayesian CNN model and discuss its effectiveness. Our analysis aims to evaluate the model's classification capability and uncertainty quantification, pinpoint areas for improvement, and propose directions for further investigation. Through comprehensive experiments, we assess the model's performance and offer detailed analysis of the findings. Results are discussed in relation to existing literature, with implications for future research delineated.

Fig. 5 displays four Receiver Operating Characteristic (ROC) curves for different classes at epoch 58 in a multi-class classification task. The ROC curve for class 0 shows a high true positive rate (TPR) at low false positive rates (FPR), with an area under the curve (AUC) of 0.92, indicating strong performance in distinguishing this class. Class 1 achieves a perfect classification with an AUC of 1.00, demonstrating that the model flawlessly identifies this class with a true positive rate of 1.0 and minimal false positives.

For class 2, the model performs decently with an AUC of 0.80, suggesting that there is some room for improvement in minimizing false positives. Finally, the ROC curve for class 3 indicates excellent performance with an AUC of 0.97, showing a high true positive rate and a low false positive rate.

Overall, the model demonstrates strong performance for most classes, with particularly impressive results for class 1, although there is some room for improvement in class 2.

In Fig. 6, two graphs show the evolution of loss and accuracy over 100 epochs of training and validation. The first graph represents the loss curve, where the training loss (shown by the blue curve) fluctuates significantly with spikes and dips, which could indicate instability in the training process. The validation loss (orange curve) follows a similar pattern but is generally more stable and remains higher than the training loss. This suggests that the model struggles to generalize to the validation data, which could be a sign of overfitting. The second graph shows the evolution of accuracy, where the training accuracy increases overall but with notable fluctuations, reaching around 0.91–0.92 by the end. The validation accuracy follows a similar trend but remains more variable, oscillating around 0.89–0.91. The difference between training and validation accuracies, as well as the variability of the curves, suggests that the model could benefit from additional techniques to improve generalization, such as regularization or learning rate adjustment. These

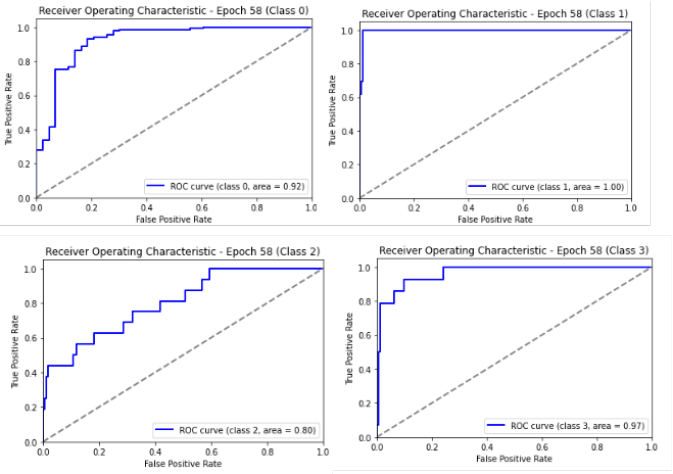


Fig. 5. ROC of class 0, 1, 2 and 3.

techniques will be addressed and improved in future work to optimize the model's performance and avoid overfitting issues.

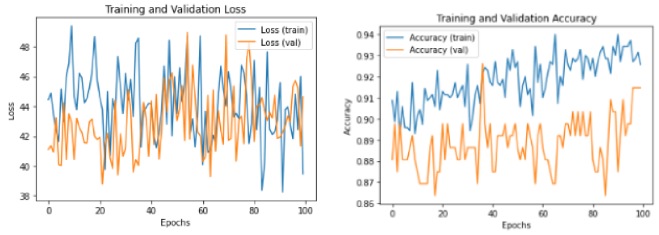


Fig. 6. Accuracy and loss graphs during training and validation for Road_Net.

Table III presents the evaluation metrics of both models, Road_Net and VGG16, focusing on training and validation performance. The Road_Net model achieved excellent performance during training, with precision, recall, F1-score, and accuracy all at 0.960. The loss during training was a relatively low 0.486. During validation, the model yields excellent precision, recall, and F1-scores of 0.886, 0.886, and 0.800, respectively, with an accuracy of 0.886. The validation loss is again low, at 0.481, and infers good generalization. In contrast, VGG16 displays slightly lower training performance vis-à-vis precision (0.860), recall (0.870), and F1-score (0.850), reflecting more moderate performance compared to Road_Net during training. With a training accuracy of 0.8737, its loss was much higher at 52.41. For validation, VGG16 has similar performance to Road_Net, achieving a precision, recall, and F1-score of 0.886, as well as an accuracy of 0.886. The validation loss was higher at 48.081, indicating that VGG16 might require further tuning for improved generalization. Overall, both models have very competitive performances, with Road_Net just slightly ahead in both training and validation.

This Fig. 7 shows some traffic sign detections, especially speed limits, which were conducted by the RoadNet model. Images are taken under various backgrounds and lighting conditions to illustrate how the model successfully identifies speed limit signs in real life. The system Road_Net, which is a CNN-based system, utilizes strong feature extraction

TABLE III. PERFORMANCES MODELS ROAD_NET AND VGG16.

Metric	Value (Road_Net)	Value (VGG16)
Precision (train)	0.960	0.860
Recall (train)	0.960	0.870
F1-score (train)	0.960	0.850
Accuracy (train)	0.960	0.8737
Loss (train)	0.486	52.41
Precision (val)	0.886	0.886
Recall (val)	0.886	0.886
F1-score (val)	0.880	0.880
Accuracy (val)	0.886	0.886
Loss (val)	0.481	48.081

techniques to circumvent challenges such as variable shapes of signs, occlusion, and environmental conditions.



Fig. 7. Detection of speed limit signs by RoadNet.

This figure shows the robustness and strength of the model in detecting various speed limit signs in a wide range of speeds, from 30 km/h to 80 km/h, through different road settings. These results underline the robust performance of RoadNet in real-time traffic sign detection for enhancing road safety and supporting intelligent transportation systems.

A. Comparison of Model Evaluation Metrics

The Table IV provides a detailed comparison of the evaluation metrics for Road_Net, VGG16 and other works in the field. During the training phase, Road_Net indicates outstanding performance for all the metrics, achieving precision, recall, F1-score and accuracy of 0.960 with less training loss of 0.486. On the other hand, VGG16 slightly lags, showing 0.860 precision, 0.870 recall and F1-score of 0.850, with a training accuracy of 0.8737 and really very high training loss of 52.41, meaning training optimization might be an issue. In comparison to others, Road_Net gave stiff competition, outdoing Trappey et al. (2024) [38], who reported precision of 0.920 and accuracy of 0.930, and Chen, who duplicated precision and recall with 0.910. Bai et al. (2024) [40] reported a high F1-score of 0.915 while Road_Net continues to prevail with respect to overall performance.

During the validation phase, Road_Net and VGG16 gave almost the same outputs with precision, recall and F1-score of 0.886 and validational accuracy of 0.886 from Road_Net which is again tanto with VGG16 (Table V). This being

said, Road_Net continues to hold an edge by laying lower its validation loss at 0.481 as opposed to the 48.081 of VGG16 thus exhibiting superior generalization toward unseen data. By way of summation, Road_Net has shown to have performed slightly better than VGG16 and the other works in the literature in terms of exhibiting high precision, recall and low loss, which is indicating a reasonable strength in use toward the given task.

TABLE IV. COMPARISON OF METRICS FOR ROADNET, VGG16, AND OTHER WORKS (TRAINING)

Method	Precision (train)	Recall (train)	F1-score (train)	Accuracy (train)	Loss (train)
Roadnet	0.960	0.960	0.960	0.960	0.486
VGG16	0.860	0.870	0.850	0.8737	52.41
Trappey [38]	0.920	0.910	0.915	0.930	0.490
Chen [39]	0.910	0.910	0.910	0.930	0.470
Bai [40]	0.915	0.910	0.915	0.930	0.470

TABLE V. COMPARISON OF METRICS FOR ROADNET, VGG16, AND OTHER WORKS (VALIDATION)

Method	Precision (val)	Recall (val)	F1-score (val)	Accuracy (val)	Loss (val)
Roadnet	0.886	0.886	0.880	0.886	0.481
VGG16	0.886	0.886	0.880	0.886	48.081
Trappey [38]	0.886	0.880	0.885	0.880	0.480
Chen [39]	0.880	0.880	0.880	0.890	0.470
Bai [40]	0.885	0.885	0.885	0.895	0.470

Compared to VGG16, RoadNet demonstrates superior performance both in training and validation phases. While VGG16 achieved a training precision of 0.860 and F1-score of 0.850, RoadNet reached 0.960 across all major metrics. The validation loss for RoadNet remained below 0.5, whereas VGG16 recorded a substantially higher loss of 48.081, indicating limited generalization capacity. RoadNet's use of multiscale feature extraction and dropout regularization contributes directly to this gain in stability and accuracy.

Furthermore, when benchmarked against recent studies such as Trappey et al. (2024), Chen (2023), and Bai (2024), RoadNet holds a competitive edge. For instance, Bai reports a validation F1-score of 0.885, whereas RoadNet maintains the same precision but achieves a lower validation loss, reflecting better optimization under constrained resources. Unlike those models, RoadNet prioritizes both accuracy and computational efficiency, making it more suitable for deployment in embedded, low-power traffic systems.

V. CONCLUSION

In this study, Road_Net and VGG16 and others models were assessed and compared in the field of traffic sign recognition, illustrating the strengths and limitations of both models. The results revealed the superiority of Road_Net in terms of precision, recall, F1-score, and classification accuracy, confirming its utility for real-time low-resource applications. VGG16 was found to give fairly good results during validation, but the high loss while training remains an indicator for further tweaking if the performance were to match that of Road_Net. In comparison with other recent approaches available from the literature, Road_Net adjusts to conditions of complex detection with high performance and robustness. The flexibility of features extraction across scales and effi- cient use of transfer learning allow it a much more sophistical applicability in real-life environments where varying illumination conditions, occlusions, and variations in designs of traffic signs are major issues. This study concludes that Road_Net is a feasible and efficient means that could cater to the requirements of intelligent transport systems and logistics regarding real-time work and optimization for low resources. However, challenges still remain. These impressive performances require further improvement to enable their full adoption for optimal consideration of diverse environments, either by adopting attention mechanisms or by rules that prevent overfitting. Further comparison with other latest architectures such as vision transformers or hybrid models can provide real prospects on improvement. The rationale behind choosing to explore this direction into logistics to integrate AI for optimizations of traffic sign recognition results, as abundant academic literature has established, remains the baselines described in "Enhancing Supply Chain Management with Artificial Intelligence: A Systematic Literature Review". This has shown that AI occupies center stage in optimizing supply chain flows and in creating opportunities for dealing with inefficiencies but still remains largely challenged by adoption issues.

REFERENCES

[1] H. Liu, K. Zhou, Y. Zhang, and Y. Zhang, "ETSR-YOLO: An improved multi-scale traffic sign detection algorithm based on YOLOv5," *PLOS ONE*, 2023.

[2] Y.-Z. Song et al., "A three-stage real-time detector for traffic signs in large panoramas," *Computational Visual Media*, 2019.

[3] H. Hong, Y. Zhou, X. Shu, and X. Hu, "CCSPNet-Joint: Efficient Joint Training Method for Traffic Sign Detection Under Extreme Conditions," *IEEE IJCNN*, 2023.

[4] A. Møgelmose, M. Trivedi, and T. Moeslund, "Vision-Based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey," *IEEE Trans. on Intelligent Transportation Systems*, 2012.

[5] A. Subhamathi et al., "Traffic Sign Detection Problems using Convolutional Neural Techniques in Image Processing," *IEEE Int. Symp. on Compound Semiconductors*, 2024.

[6] A. Guesmi et al., "Physical Adversarial Attacks for Camera-Based Smart Systems: Trends and Challenges," *IEEE Access*, 2023.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, 2012.

[8] L. Wu, H. Li, J. He, and X. Chen, "Traffic sign detection method based on Faster R-CNN," *J. of Physics: Conf. Series*, 2019.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE TPAMI*, 2015.

[10] R. B. Girshick et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *CVPR*, 2014.

[11] R. B. Girshick, "Fast R-CNN," *ICCV*, 2015.

[12] K. He et al., "Mask R-CNN," *ICCV*, 2017.

[13] C. Saraswathy, K. Nandhini, and R. Kalaivani, "Traffic sign detection and recognition based on convolutional neural network," *Int. J. of Health Sciences*, 2022.

[14] K. B. Pavani, "Traffic Sign Detection and Recognition System using CNN," *IJRASET*, 2024.

[15] D. Hindarto, "Enhancing Road Safety with CNN Traffic Sign Classification," *SinkrOn*, 2023.

[16] N. S. Ganesh et al., "Traffic Sign Recognition using CNN," *IJARSCT*, 2024.

[17] W. Zhao, Z. Wang, and H. Yang, "Traffic Sign Detection Based on Faster R-CNN in Scene Graph," 2018.

[18] C. Ferencz and M. Zöldy, "Neural Network-based Multi-Class Traffic-Sign Classification with the GTSRB Benchmark."

[19] Y. Ouahbi and S. Ziti, "Revolutionizing Road Safety and Optimization with AI: Insights from Enterprise Implementation," *IJACSA*, vol. 16, no. 4, 2025. https://doi.org/10.14569/IJACSA.2025.0160498

[20] Y. Ouahbi and S. Ziti, "NW Logistics: System Architecture and Design for Sustainable Road Logistics," *IJACSA*, vol. 16, no. 4, 2025. https://doi.org/10.14569/IJACSA.2025.01604105

[21] J. Cao, J. Zhang, and X. Jin, "A Traffic-Sign Detection Algorithm Based on Improved Sparse R-CNN," *IEEE Access*, 2021.

[22] G. Zhang, Y. Peng, and H. Wang, "Road Traffic Sign Detection Method Based on RTS R-CNN," *Italian National Conference on Sensors*, 2023.

[23] D. Temel, M.-H. Chen, and G. Al-Regib, "Traffic Sign Detection Under Challenging Conditions," *IEEE Trans. on ITS*, 2019.

[24] J. Li and Z. Wang, "Real-Time Traffic Sign Recognition Based on Efficient CNNs in the Wild," *IEEE Trans. on ITS*, 2019.

[25] F. Wang et al., "Improved Faster R-CNN for Traffic Sign Detection," *IEEE ITSC*, 2020.

[26] J. Zhang et al., "A Real-Time Chinese Traffic Sign Detection Algorithm Based on Modified YOLOv2," *Algorithms*, 2017.

[27] D. Tabernik and D. Skočaj, "Deep Learning for Large-Scale Traffic-Sign Detection and Recognition," *IEEE Trans. on ITS*, 2019.

[28] T. M. Hoang, S. Nam, and K. Park, "Enhanced Detection and Recognition of Road Markings," *IEEE Access*, 2019.

[29] S. Pei et al., "Localized Traffic Sign Detection with Multi-scale Deconvolution Networks," *COMPSAC*, 2018.

[30] Y. Tian et al., "Traffic Sign Detection Using a Multi-Scale Recurrent Attention Network," *IEEE Trans. on ITS*, 2019.

[31] A. D. Kumar, "Novel Deep Learning Model for Traffic Sign Detection Using Capsule Networks," *arXiv*, 2018.

[32] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic Routing Between Capsules," *NeurIPS*, 2017.

[33] Y. Yuan, Z. Xiong, and Q. Wang, "VSSA-NET: Vertical Spatial Sequence Attention Network," *IEEE TIP*, 2019.

[34] Y. Yang et al., "Towards Real-Time Traffic Sign Detection and Classification," *IEEE Trans. on ITS*, 2016.

[35] Z. Zhu et al., "Traffic-Sign Detection and Classification in the Wild," *CVPR*, 2016.

[36] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CVPR*, 2014.

[37] J. Zhang et al., "A Cascaded R-CNN With Multiscale Attention and Imbalanced Samples for Traffic Sign Detection," *IEEE Access*, 2020.

[38] A. J. C. Trappey and O. T. C. Shen, "Un système universel de détection de panneaux de signalisation...," *Advanced Engineering Informatics*, 2024. DOI: 10.1016/j.aei.2024.102674.

[39] J. Wang, L. Zhao, and M. Chen, "Transfer learning for regional traffic sign datasets," *Applied Sciences*, vol. 13, no. 5, 2023.

[40] Y. Bai et al., "Amélioration de la compréhension des nuages de points...," *Knowledge-Based Systems*, 2024. DOI: 10.1016/j.knosys.2024.112715.

MITG-CU: Multimodal Interaction Temporal Graphs Approach for Conversational Emotion Recognition

Qian Xing, Yaqin Qiu, Minglu Chi, Xuewei Li, Changyi Gao
School of Intelligent Engineering, Henan Institute of Technology, Xinxiang, Henan 453003, China

***Abstract*—In the emotion recognition of conversations, the complementary relationship between the context information and multimodal data cannot be fully exploited. This results in insufficient comprehensiveness and accuracy in emotion recognition. To address these challenges, this paper proposed a Multimodal Interactive Temporal Graph Conversation Understanding model (MITG-CU) composed of textual, audio and visual modalities. Firstly, the pre-extracted textual, audio, and visual features are adopted as the input of the Transformer, and the attention mechanism is utilized to capture the cross-modal context correlation information. Furthermore, structural relationships and temporal dependencies between utterances are captured through a local-level relational temporal graph module. Inter-modal interaction weights are dynamically adjusted by a global-level pairwise cross-modal interaction mechanism. By integrating two complementary hierarchical structures, a hierarchical multimodal information fusion was achieved, and at the same time, the model's adaptability to complex conversation scenarios was enhanced. Finally, feature fusion is carried out by using the gating mechanism and sentiment classification is conducted. Experimental results demonstrated that the proposed model outperforms six common baseline methods across metrics including accuracy, precision, recall, and F1-score. Especially in Weighted-F1 and Accuracy have improved by 0.28 % and 0.39 % respectively, which confirmed the effectiveness of the model.**

***Keywords*—Emotion recognition; multimodal interaction; relational temporal graph; cross-modal interaction; feature fusion**

I. Introduction

Emotional expression is multifaceted, extending beyond verbal language to encompass variations in vocal tone, facial expressions, and body movements. Most existing emotion recognition models are based on single modalities, such as textual [1], vision [2] and audio [3], but human emotions are often elusive. Therefore, a single text modality may not be able to correctly identify emotions in certain scenarios. For instance, the emotions directly expressed in text are neutral, but the corresponding facial expressions actually represent another emotion, such as anger.

As multimodal technologies are increasingly approaching real-world application scenarios, multimodal emotion recognition [4], [5], [6] has become a research hotspot. By analyzing and integrating information from different sensory channels, it aims to more comprehensively and accurately capture and interpret human emotions. The development of this technology can not only enhances the naturalness and efficiency of human-computer interaction, but also plays an important role in multiple fields such as mental health monitoring, education, and entertainment [7], [8]. Due to their limited representational capacity and insufficient information coverage, single modal methods exhibit application limitations. In contrast, multimodal methods significantly enhance the recognition accuracy and robustness of the system through a systematic feature complementation mechanism. Consequently, multimodal methods better align with the intrinsic characteristics of human multi-channel emotional expression. Unlike traditional emotion recognition tasks in single-modal and non-conversational scenarios, conversational emotion recognition involves complex multimodal relationships and contextual information, thereby posing more significant challenges.

In response to the limitations of existing methods in the task of conversational emotion recognition, where it is difficult to simultaneously utilize context information and the complementary relationship among various modalities. This paper proposes a Multimodal Interaction Temporal Graph Conversation Understanding Model (MITG-CU), which consists of two core modules: The Relational Temporal Graph Module (RTGM) and The Cross-modal Interaction Module (CIM). Contextual features at dual levels can be simultaneously extracted and effectively utilized by this model, thereby enabling the precise recognition of conversational emotions. At the local level, through the relationship temporal graph convolution network module, a time series graph structure is constructed to effectively capture the discourse temporal characteristics and local context information during multimodal interaction, further enhancing the understanding of Conversation evolution. At the global level, through the cross-modal feature interaction mechanism, the global context information is fully integrated to enhance the accuracy and robustness of discourse-level emotion recognition.

The rest of the paper is structured as follows: Section II is dedicated to related works, Section III presents our methodology, Section IV exposes the experimental results, The discussion is provided in Section V. Section VI introduced the conclusion for the overall work and talk about future directions.

II. Related Works

Multimodal emotion recognition primarily identifies emotional states by integrating multiple sensory modalities such as facial expressions, speech, and physiological signals, thereby enhancing the accuracy and robustness of emotion detection across diverse scenarios. The core issue of multimodal conversational emotion recognition primarily revolve around two aspects: achieving effective interaction and fusion of information across different modalities, and systematically modeling the global and local dynamics of session emotion.

Currently, the primary research approaches for multimodal emotion recognition in conversations include: those based on

contextual information [9], [10], [11] and those based on the complementary relationship among different modalities [12], [13], [14], [15]. Among context-based approaches, models such as LSTM, RNN, and GCN are widely adopted.

The study in [9] proposed a multimodal emotion recognition model based on fuzzy guidance and context self-attention network. By leveraging contextual self-attention mechanisms to capture dynamic interactions between different modalities and using fuzzy logic to handle uncertainty in multimodal data. However, High computational complexity, strong dependence on data quality, and limited generalization ability and other issues increase the training cost and affect the overall performance of the model. The study in [10] developed a contextual relationship modeling method that combines high-order relationships among multiple nodes. Complex dynamic interactions between different modalities in conversations can be effectively captured by this method. Nevertheless, constructing high-order relationships among multiple nodes requires complex prior knowledge or heuristic rules. If the structure is improperly designed, it will affect the performance of the model. The emotional features in conversations are highlighted through sequential analysis of conversational context and employment of a memory mechanism in [16], However, limitations have been identified in retaining contextual details from long-range dialogue sequences. While the study in [17] employed graph convolutional networks to model speaker interactions and conversation context for emotion recognition. It simply concatenated the multimodal information ,without deeply considering the potential interaction among different modalities.

The main research methods based on the complementary relationship among various modalities are: The research [12] proposed a facial emotion recognition method based on deep convolutional neural networks. Yet, this method failed to take into account the dynamic changes of facial expressions. and temporal dependencies in the time dimension were also unaddressed. The method presented in [15] adaptively selects nodes and edges, thereby enhancing both intra-modality and inter-modality interactions. Due to the presence of complex dependencies and heterogeneity gaps, this approach struggles to accurately capture and fuse multimodal information. The research in [18] employed a multi-channel and multimodal sentiment analysis method based on knowledge graph. The limitation of this method lies in its inability to effectively process cross-modal affective information, particularly under conditions with significant inter-modal distribution discrepancies and noise contamination. The research in [19] developed a dynamic weighted gated mechanism, which enhances cross modal interaction for multimodal sentiment. The framework's effectiveness is constrained by the requirement for precise parameter tuning in its dynamic gating mechanisms.

While the studies discussed above have made significant contributions to multimodal emotion recognition in conversation, they also exhibit certain limitations. Failing to simultaneously account for global information, local information, and the complementary relationships among different modalities, this may lead to misunderstandings of the speaker's true emotions or intentions. Therefore, our proposed MITG-CU model is capable of extracting and leveraging context features at two levels, thereby achieving precise emotion recognition. By constructing RTGM and CIM, unique information and complementarity of modalities such as textual, audio and visual are integrated. This achieves a comprehensive understanding of the conversation evolution and the full integration of global context information. In order to verify the accuracy and robustness of emotion recognition in conversations, our model was compared with six commonly used baseline methods. The results showed that our model outperforms six common baseline methods. Weighted-F1 and Accuracy were respectively improved by 0.28 % and 0.39 %. This demonstrates the good stability and generalization ability of our model.

III. Proposed Model

The proposed MITG-CU model for this research consists of four components: feature extraction and encoding, RTGM, CIM, and multimodal fusion and prediction. The overall framework of the network is shown in Fig. 1.

Firstly, the pre-extracted textual, acoustic and visual features are employed as inputs of the Transformer model. Meanwhile, its self-attention mechanism is leveraged to capture the cross-modal contextual correlation information. Subsequently, the extracted features are fed into the relational temporal graph attention network layer and the cross-modal interaction layer, where context modeling and cross-modal interaction are respectively carried out. Finally, the gated mechanism is utilized to fuse the features at the feature level, and then classified based on the emotions.

A. Feature Extraction and Encoding

Before performing modal information enhancement, sBERT, OpenSmile and OpenFace are used to extract Text, Audio and Visual features respectively, and the extracted features are input into the Transformer encoding layer. After the processing of the encoding layer, three modal feature representations are finally generated, as shown in Eq. (1)-Eq. (3).

$$T \in R^{L_T \times D_T} \tag{1}$$

$$A \in R^{L_A \times D_A} \tag{2}$$

$$V \in R^{L_V \times D_V} \tag{3}$$

Where, *T* represents textual modal information, *A* denotes audio modal information, and *V* indicates visual modal information. and L_T, L_A and L_V respectively denote the lengths of the data sequences, while D_T, D_A, and D_V are the dimensions of the extracted features for each Modality.

B. Relational Graph Attention Network

In the graph attention network framework, each node is associated with a feature vector. X=$\{h_1,h_2,...,h_n\}$ is the set of feature vectors, where h_i is the feature vector of node *i*. These feature vectors are composed not only of the original attributes of the nodes, but also integrated with high-order structural information extracted from the graph's topological structure.

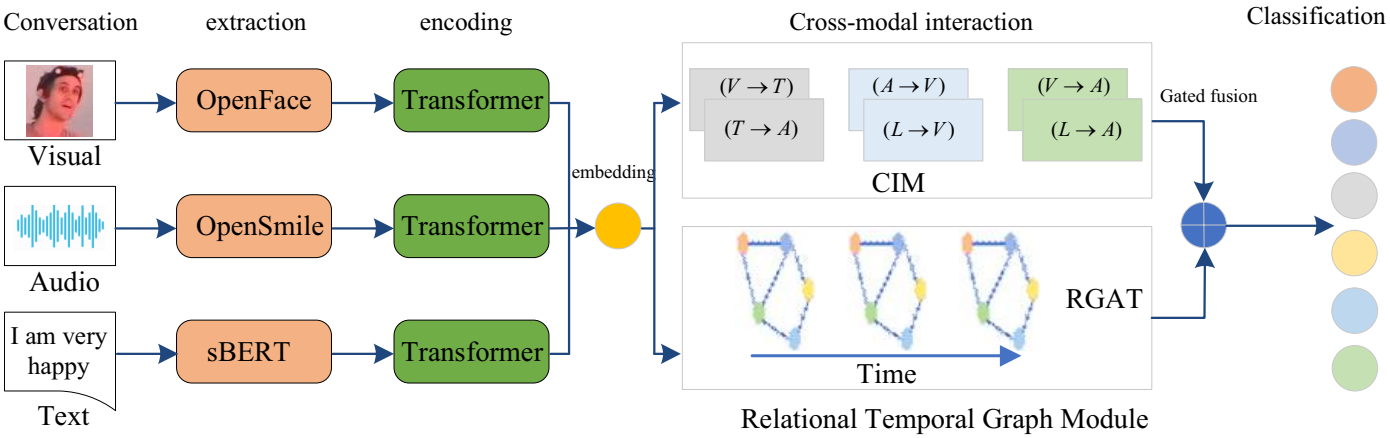


Fig. 1. Multimodal interaction temporal graph conversation understanding model.

The implementation steps of Relational Graph Attention Network (RGAT) are as follows:

Step 1: Relation Type Mapping Matrix

To distinguish different edge types or connection patterns, a relation-type mapping matrix $R=\{R_1,R_2,...,R_m\}$ is introduced. Each node i is associated with a feature vector $h_i \in R^m$, and all node feature vectors collectively form the input feature matrix $H \in R^{N\times m}$ of the graph, where N denotes the number of nodes.

Before calculating the graph attention coefficients, the model first maps the original node features into an F' dimensional feature vector $Z_i \in R^{F'}$ through a linear transformation. The weight matrix $W \in R^{m\times F'}$ is shared among all nodes. The transformed node feature is expressed as $Z_i = Wh_i$.

Step 2: Relation-aware Attention

After establishing node features and relation-type representations, further computed using the self-attention mechanism. Attention weights $a^r_{i,j}$ between each pair of nodes *(i, j)* and relation type *r* as Eq. (4) and Eq. (5).

$$e^r_{i,j} = \text{Leaky Re}\,LU\left(a^T\left[Z_i \| Z_j\right]\right) \tag{4}$$

$$a^r_{i,j} = \frac{\exp\left(e_{ik}\right)}{\sum_{k\in N(i)}\exp\left(e_{ik}\right)} \tag{5}$$

Where, a^T represents the attention vector for learning, Z_i and Z_j respectively denote the transformed feature vectors of nodes i and j, and $N(i)$ represent the neighbor node sets of node i.

Step 3: Relation-specific Aggregation

The attention weights are obtained through the relationship-aware attention mechanism. These weights are weighted and aggregated with the neighborhood information. The node feature vector H_i is calculated as shown in Eq. (6).

$$H_i = \sigma\sum_{j\in N(i)} a^r_{i,j} W h_j \tag{6}$$

Where, h_j represents the feature vector of node j, and σ denotes a non-linear activation function such as *ReLU* or *LeakyReLU*. This process not only enhances the non-linear expression ability of the model, but also enables the model to more effectively capture the complex relationships between nodes.

Step 4: Multi-head Attention Mechanism

To further enhance the representational capacity of the model, a multi-head attention mechanism is introduced. This mechanism allows the model to learn information in parallel within different representation subspaces, and there are multiple independent attention heads for each relation type. By parallel computing the outputs of multiple attention heads and concatenating or averaging the results, the final representation of the node is obtained. The aggregation process of the multi-head attention mechanism is shown in Fig. 2.

Then, the final node representation h'_i is shown in Eq. (7).

$$h'_i = \text{concat}\left(\sigma\left(\sum_{j\in N(i)} a^{(k)}_{i,j} W^{(k)} h_j\right)\right) \tag{7}$$

Where, *concat* represents the concatenation operation, $a^k_{i,j}$ is the representation of the *k-th* head, and $W^{(k)}$ refers to the linear transformation weight matrix of the *k-th* head.

Step 5: Output Layer

After the feature extraction and update by the multi-head attention network, the updated node features are input into the output layer. Then, normalization is performed to generate a probability distribution of emotional categories, thereby achieving the classification and prediction of emotional states.

These steps collectively constitute the core computing process of RGAT module. The RGAT module can effectively incorporate relational type information, dynamically capture complex relationships, temporal dynamics, and multimodal information in conversations, thereby enhancing the performance of emotion recognition in conversations.

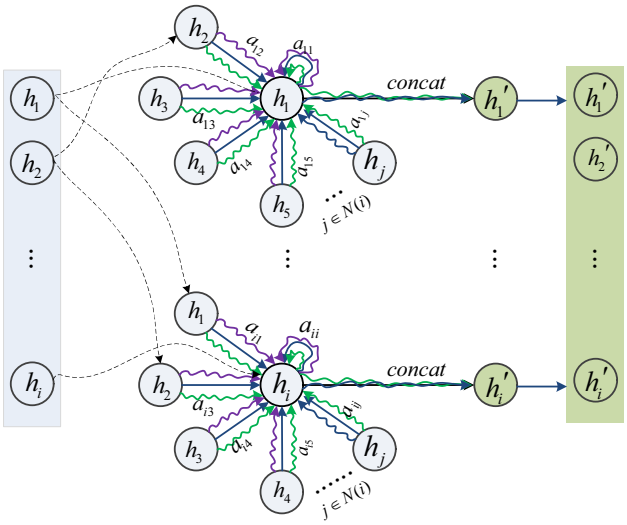


Fig. 2. The Aggregation process of the multi-head attention mechanism.

C. The Cross-Modal Interaction Module

In multimodal sentiment analysis tasks, the cross-modal heterogeneity inherent in human language often amplifies the complexity of linguistic interpretation. Utilizing cross-modal interaction is helpful in revealing the "unaligned" nature and long-term dependency of cross-modal phenomena. Hence, the pairwise cross-modal feature interaction module (P-CM) [20] is introduced into our proposed framework.

The P-CM module captures the subtle emotional correlations between textual and auditory modalities through the interaction of low-level features, thereby providing enriched semantic and affective information for multimodal sentiment analysis tasks. The cross-modal Transformer network based on P-CM is presented in Fig. 3.

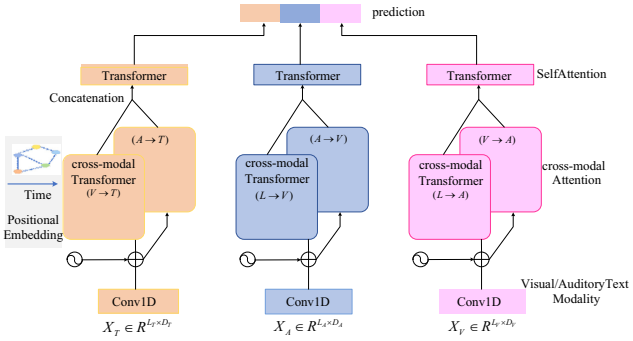


Fig. 3. The Cross-modal interaction module.

As illustrated in Fig. 3, $X_T = R^{L_T \times D_T}$, $X_A = R^{L_A \times D_A}$ and $X_V = R^{L_V \times D_V}$ respectively denote the modality-sensitive representations of the textual, audio and visual modes when using a single modality encoder.

Based on the Transformer architecture, taking audio and text as examples. The defined Query , Keyword and Value are respectively $Q_A = X_A W_{Q^A}$, $K_T = X_T W_{K^T}$ and $V_T = X_T W_{V^T}$. When modality *T* performs cross-modal attention on modality *A*, the enriched representation $CM^{T \to A}$ of modal *A* is computed as in Eq. (8).

$$CM^{T \to A} = \text{softmax}\left(\frac{X_A W_{Q^A} (W_{k^T}) X_T}{\sqrt{d_k}} X_T W_{v^T}\right) \quad (8)$$

Where, $W_{Q^A} \in R^{d_A \times d_Q}$, $W_{K^T} \in R^{d_T \times d_K}$ and $W_{V^T} \in R^{d_T \times d_V}$ are learning parameters, while d_Q , d_K and d_V represent the dimensions of Query, Key and Value respectively, $\sqrt{d_K}$ is the scaling factor.

Assume $Z_A^{[i]}$ is the modality-sensitive global context representation of modality *A* across the entire conversation at the *i-th* layer, and $Z_A^{[0]} = X_A$. When modality *T* performs cross-modal attention on modality *A*, information from modality *T* is transferred to modality *A*, and the enriched representation $Z_{T \to A}^{[i]}$ of modality *A* is computed as the following procedure Eq. (9), and $Z_{T \to A}^{[0]} = Z_A^{[0]}$.

$$\begin{aligned} \bar{Z}_{T \to A}^{[i]} &= CM_{T \to A}^{[i]}\left(LN\left(Z_{T \to A}^{[i-1]}\right), LN\left(Z_{T \to A}^{0}\right)\right) \\ &+ LN\left(Z_{T \to A}^{[i-1]}\right) \\ Z_{T \to A}^{[i]} &= \\ &\max\left(0, LN\left(\bar{Z}_{T \to A}^{[i]}\right) \times w_1 + b_1\right) w_2 + b_2 + LN\left(\bar{Z}_{T \to A}^{[i]}\right) \end{aligned} \quad (9)$$

Where, w_1 and w_2 represent the weight matrices, b_1 and b_2 denote the biases, *LN* indicates the layer normalization calculation, *max(0,0)* is the *ReLU* activation function. Through layer normalization calculation, the activation value distribution of each layer is adjusted to make the model training more stable and improve performance.

According to Eq. (9), we can easily compute the cross-modal representation $Z_{T \to A}^{[i]}$. Meanwhile, the outputs of the cross-modal Transformer (D layers) that share the same target modality are concatenated to form the final cross-modal global context representation $Z_{T \to A}^{[D]}$. Similarly, for other modal pairs $Z_{V \to T}^{[D]}$ and $Z_{V \to A}^{[D]}$, the same process can be applied to obtain them, and ultimately, the representations of the three modalities in the global cross-modal context can be obtained.

P-CM effectively reduces the complexity of joint multimodal modeling by simultaneously modeling the relationship between the two modalities. The framework precisely extracts local paired modal information. Additionally, it constructs a global cross-modal dependency relationship. This integration significantly enhances the model's adaptability and robustness across diverse scenarios.

D. Multimodal Fusion and Prediction

1) Gated multimodal fusion: The Gated Multimodal Fusion (GMF) layer is a key component of the multimodal interactive temporal graph conversation understanding model, responsible for integrating features from different modalities. This layer aims to selectively emphasize the relevant information in each modality and suppress the irrelevant or redundant information, thereby enhancing the model's ability to accurately identify emotions. The GMF layer consists of two main parts: the transformation mechanism and the gating mechanism.

The transformation mechanism is to convert the concatenated features through another linear layer, thereby extracting higher-level feature representations. The transformed features are denoted as H', as shown in Eq. (10).

$$H' = W_t \cdot [h'_1, h'_2, \ldots, h'_n] + b_t \tag{10}$$

Where, W_t is the weight matrix and b_t is the bias term.

The gating mechanism is mainly implemented through linear layer and *sigmoid* activation function to calculate the importance of each modality, thereby weighting the features of different modalities. For each modality *i*, its weight value g_i is calculated by the sigmoid activation function Eq. (11).

$$g_i = \text{sigmoid}\left(W_t \cdot [h'_1, h'_2, \ldots, h'_n] + b_t\right) \tag{11}$$

2) Emotion classification layer: The emotion classification layer is mainly responsible for precisely mapping the fused feature vectors to the corresponding emotion categories. It consists of two main fully connected layers: *lin1* and *lin2*. The *lin1* layer maps the input features to the hidden layer space, while the *lin2* layer further maps the features obtained from *lin1* to the prediction space of emotion categories. To prevent overfitting of the model, a Dropout layer is introduced between *lin1* and *lin2*. Subsequently, the argmax function is used to determine the emotion category with the highest probability, thereby serving as the final emotion prediction result for the discourse. The predicted emotion category of the discourse is shown in Eq. (12).

$$\hat{c}_j = \arg\left(\text{softmax}\left(g_k z_j + b_k\right)\right) \tag{12}$$

Where, C_j represents the predicted emotion category of the *j-th* utterance. g_k and b_k respectively denote the weight and bias parameters of the fully connected layer in the second layer. After training, we are able to extract the feature representation of the *j-th* utterance, which is denoted as Z_j.

Finally, the performance of the model is evaluated by calculating the difference between the probability distribution predicted by the model and the true labels. The negative log-likelihood loss (*NLLLoss*) is used as the loss function. The input of *NLLLoss* is the log-probability vector and the target label. Parameter optimization is achieved by minimizing the difference between the predicted probability and the true distribution. For each sample *i*, the calculation of *NLLLoss* is expressed as Eq. (13).

$$\text{NLLLoss}(x, y) = -\sum_{i=1}^{N} x_i \left[y_i\right] \tag{13}$$

Where, *N* is the number of samples, y_i is the true label of the sample, and x_i is the logarithmic probability vector of the *i-th* sample.

IV. Model Training and Evaluation

A. Data Det

This paper utilizes the Interactive Emotional Dyadic Motion Capture (IEMOCAP) sentiment analysis dataset. IEMOCAP is a dataset commonly employed for multimodal sentiment recognition, collected by the Speech Analysis and Interpretation Laboratory (SAIL) of the University of Southern California (USC). This dataset encompasses approximately 12 hours of multimodal data, including visual, audio, facial motion capture, and text transcription. IEMOCAP is divided into two parts. Participants engage in either impromptu performances or performances based on specific scripts in these parts. These scripts are specially designed to elicit emotional expressions.

B. Experimental Configuration and Evaluation Indicators

All experimental code was developed using Python 3.11 and the PyTorch 2.0 deep learning framework, and the models are trained and tested on a single NVIDIA RTX 3060 8G GPU. After extensive experimentation, the optimal performance was achieved with a batch size of 10 and the learning rate of 0.00025. The Adam algorithm is selected as the optimizer. To reduce overfitting, the dropout rate is set to 0.5.

Use Accuracy *(Acc.)*, Macro-Average Precision *(P)*, Macro-Average Recall *(R)*, and Weighted F1 score (W_F1) as the evaluation metrics for model performance, as computed in Eq. (14) and (15).

$$p = \frac{1}{k} \sum_{i=1}^{k} \frac{TP_{C_i}}{TP_{C_i} + FP_{C_i}} \tag{14}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{15}$$

$$\text{Acc.} = \sum_{i=1}^{T} I\left(y_p = y_t\right) \tag{16}$$

$$W_F1 = -\sum_{i=1}^{N} f_{reqk} \times F1_i \tag{17}$$

Where, *TP*, *FP*, and *FN* respectively stand for true positive, false positive, and false negative. *k* is the number of categories, and C_i denotes the *i-th* category, Furthermore, y_p, y_t and *T* represent the predicted label, the true label, and the total number of samples in the test set. f_{reqk} is the relative frequency of class *k*, and $F1_i$ is the F1 score of the *i-th* category.

C. Controlled Experiment

1) Baseline: In order to verify the performance of MITG-CU model proposed in this paper, the following models are selected for experimental comparison: 1) DialogueRNN [16]: It is an attention-based recurrent neural network for emotion detection in conversations, which enhances emotion classification accuracy by tracking participants' states and contextual information. 2) COGMEN [21]: The multimodal sentiment

recognition method based on graph neural networks, which predicts the emotional state of the speaker in a conversation by integrating local and global contextual information. 3) CoMPM [22]: It is a model that combines the pre-trained memory of the speaker and context modeling. By tracking the importance of the speaker's historical utterances and their distance from the current utterance, it enhances the focus on the current utterance, thereby captures the emotional information in the conversation. 4) MM-DFN [23]: The dynamic fusion module of the graph structure is utilized to fuse multimodal context features, thereby reducing redundancy and enhancing the complementarity among modalities. 5) GraphCFC [24]: The method of cross-modal feature complementation based on directed graphs is employed in this paper. By utilizing multiple subspace extractors and paired cross-modal complementary strategy, the heterogeneity gap in multimodal fusion can be alleviated. 6) SACL-LSTM [25]: Based on the supervised adversarial contrastive learning framework, robust emotional features are generated by combining adversarial training and contrastive learning.

2) Experiment comparing with the baseline method: The comparative results of different models are presented in Fig. 4. The results indicate that the proposed model exhibits performance advantages over existing approaches, particularly in the key metrics of Weighted-F1 and Accuracy.

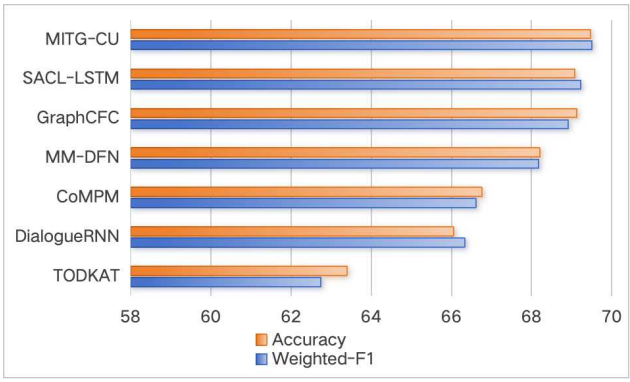


Fig. 4. Comparative analysis of models.

The comparative experimental results of different methods on the IEMOCAP dataset as presented in Table I. As shown, the proposed model achieves a Weighted-F1 score of 69.50, surpassing the second-best SACL-LSTM model by 0.28%. Similarly, it demonstrates superior Accuracy with a value of 69.47, outperforming SACL-LSTM by 0.39%.

TABLE I. COMPARATIVE EXPERIMENTAL RESULTS

Model	Weighted-F1	Accuracy
DialogueRNN	62.75	63.40
CoMPM	66.33	66.05
COGMEN	67.27	67.04
MM-DFN	68.18	68.21
GraphCFC	68.91	69.13
SACL-LSTM	69.22	69.08
MITG-CU	69.50	69.47

Fig. 5 provides comparison chart of F1-score and Weighted Average performance for different emotions. As can be seen in Fig. 5, the model achieves higher precision and recall in recognizing Emotion sad, indicating that the emotion of sad is the best among all emotions, while Emotion frustration exhibits relatively lower performance. Moreover, the F1-Score values for most emotions are clustered between 0.60 and 0.75, suggesting that the proposed model performs relatively balanced in most emotions and does not show extreme high or low values.

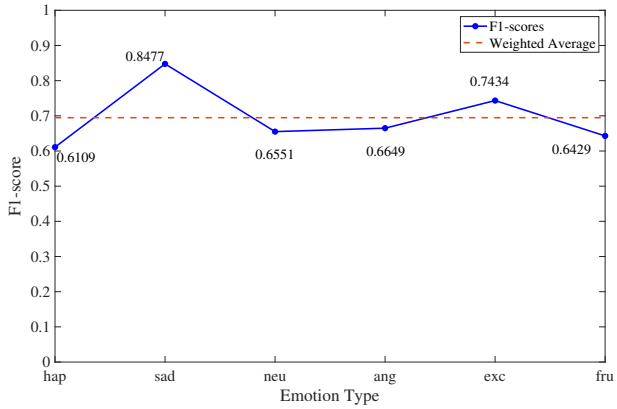


Fig. 5. Comparative chart of F1-scores and Weighted Average performance across different emotions.

The recognition results of different emotions on the IEMOCAP dataset are reported in Table II. As shown in Table II, the Macro-average F1-score is 0.6941, and the Weighted-average F1-score is 0.6947. The two scores are relatively close, indicating that the model has good stability and generalization ability across different emotion categories.

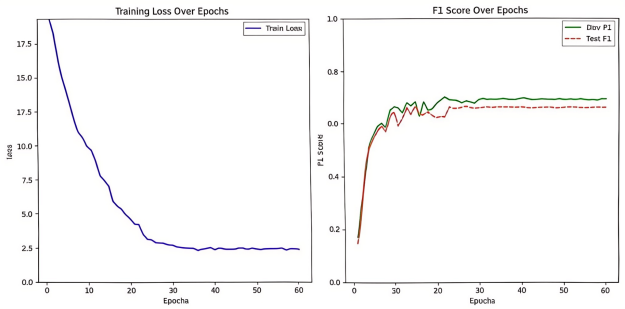
TABLE II. THE RECOGNITION RESULTS OF DIFFERENT EMOTIONS ON THE IEMOCAP DATASET

Model	Precision	Recall	F1-score
hap	0.5689	0.6597	0.6109
sad	0.8127	0.8857	0.8477
neu	0.6703	0.6406	0.6551
ang	0.6029	0.7412	0.6649
exc	0.7895	0.7023	0.7434
fru	0.6744	0.6142	0.6429
Macro-average	0.6864	0.7073	0.6941
Weighted-average	0.6986	0.6950	0.6947

As can be seen in Fig. 6, it can be observed that with the increase in the number of training epochs, the training loss decreases rapidly. This demonstrates that the model is constantly optimized during the learning process, and the fitting effect on the training data gradually enhances. In the early stage of training, the F1-score exhibits rapid growth, demonstrating the model maintains high accuracy while simultaneously improving recall, with its overall performance progressively enhancing throughout the learning process. After approximately 20 training epochs, the loss value plateaus, suggesting that the model has reached a good convergence state.

To comprehensively evaluate the performance of classification models, we analyzed the correspondence between the model's predicted outcomes and the true labels through comparison. As a result, the confusion matrix, which reflects the model's classification capability, is presented in Fig. 7.

As illustrated in Fig. 7, a strong positive correlation is observed between the F1 score and accuracy rate (0.97), which means that when F1 score increases, accuracy rate also tends



(a) Training loss (b) F1 score

Fig. 6. Training loss and F1 score variation curves.

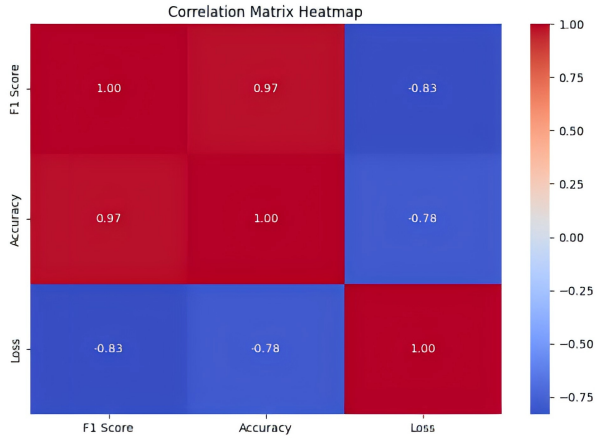


Fig. 7. Confusion matrix of evaluation indicators.

to increase. There is a strong negative correlation between F1 score and loss value (-0.83), indicating that as F1 score increases, loss value tends to decrease, which is what this model expects. This result shows that the model can improve the classification precision and recall while maintaining high overall prediction accuracy, demonstrating its balance and robustness in classification tasks.

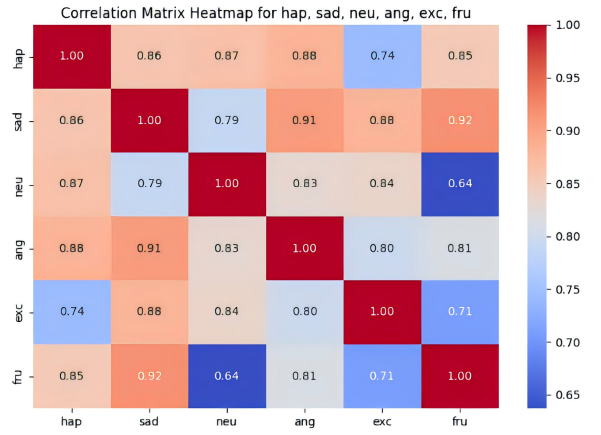


Fig. 8. Confusion matrix of various emotions.

The confusion matrix for various emotions as shown in Fig. 8. It can be observed from the confusion matrix that the correlation between emotional categories significantly impacts model performance. The high correlation (0.92) between sadness and frustration indicates significant confusion between them in the feature space, this suggests that their features are less distinguishable during emotion recognition. In contrast, the neutral emotions have higher feature distinguish ability and are less influenced by other emotions. Especially, its correlation with frustration is the lowest (0.64). The confusion matrices of various emotions not only provide a basis for the accuracy of emotion recognition, but also offer an important basis for the subsequent optimization of emotion models.

3) Ablation experiment: In order to further investigate the contributions of different components in the model, four ablation studies were designed on the IEMOCAP dataset in this paper. The base, no-CM, no-RGAT, and no-GAT respectively represent the ablation of the basic module, removing the cross-modal module from the model, removing the graph-time convolution module from the model, and removing the graph attention module from the model. The experimental results are shown in Table III.

TABLE III. THE ABLATION EXPERIMENTS CONDUCTED ON IEMOCAP

Module	**IEMOCAP**	
	*F*1	*Acc.*
base	69.50	69.47
no-CM	65.67	66.61
no-RGAT	66.26	66.67
no-GAT	65.89	66.42

As shown in Table III, our proposed MITG-CU method achieves the best performance on the two key metrics of F1 score and accuracy. Specifically, no-CM has the most significant impact on the model performance. Meanwhile, no-GAT also had a significant impact on the model performance. This indicates that the cross-modal module and the graph attention module are crucial for the model in capturing complex emotional features and improving classification performance.

The reproduced comparison results are reported in Table IV. The modalities include text (T), audio (A), and visual (V).

TABLE IV. THE PERFORMANCE OF MITG-CU IN DIFFERENT MODALITIES

Module	**IEMOCAP**	
	Acc.	*F*1
T	65.74	65.35
A	50.65	49.92
V	32.35	31.80
T+V	66.24	66.23
T+A	65.06	64.94
A+V	46.09	46.00
A+V+T	69.47	69.50

As illustrated in Table IV, when solely utilizing the textual modality (T), the model achieves the highest values in both accuracy (65.74%) and F1 score (65.35%). This result demonstrates that textual information plays a crucial role in the emotion recognition task. By comparison, the performance of the audio modality (A) is moderately effective, whereas the performance of the visual modality (V) is comparatively limited.

As demonstrated by the experimental results of bimodal combinations, the combination of textual modality and visual modality (T+V) achieved accuracy and F1-scores of 66.24%

and 66.23% respectively on the IEMOCAP dataset. Its performance is slightly better than that of the combination of textual modality and audio modality (T+A). The main reason for this is that some audio categories contain noise, which will affect the prediction results. The combination of audio modality and vision modality (A+V) exhibited the lowest performance.

In practice, the trimodal combination (T+A+V) achieves the best performance, with accuracy of 69.47% and F1-scores of 69.50%. This result indicates that multimodal fusion can significantly enhance the accuracy of emotion recognition.

V. Discussion

A. Comparison with Existing Results

The author in [26] proposed a multimodal fusion method based on deep graph convolutional networks. This method can effectively capture the context-dependent relationships in conversations through the construction of a graph structure. However, the fixed graph structure cannot adjust the weights of each modality in real time as the conversation progresses dynamically. according to the dynamic progress of the dialogue. This situation significantly reduced the accuracy of emotion change recognition, with an accuracy rate of only 65.56%. In [27] models the dynamic dependencies between speakers in conversations, thereby improving the accuracy of emotion prediction. This method simultaneously captures the internal dependencies within speakers, and the inter-speaker dependencies, thereby resolving the limitation of focusing solely on a single type of dependency. Nevertheless, this model is primarily designed for a unimodal setting (text), and lacks a mechanism for cross-modal fusion. As a result, its accuracy rate is limited to 66.29%.

In response to this situation, the proposed method can comprehensively integrate global information, local information, and the complementary relationships among different modalities. According to the contribution of each modality, the interaction weights between modalities are dynamically adjusted, thereby achieving information interaction and deep integration among modalities. Through ablation experiments, it can be found that either no CIM or no GAT will lead to a decrease in the F1 score and accuracy of the system to varying degrees. This indicates that the cross-modal module and the graph attention module are crucial for the model to capture complex emotional features and improve classification performance. Meanwhile, the impact of different modality combinations on model performance varies. As can be seen from Table IV, It can be concluded that textual modality plays a pivotal role in emotion recognition, while the integration of visual modality further enhances performance. Although the audio modality underperforms compared to textual and visual modalities when used in isolation, its combination with other modalities improves overall system performance. Moreover, the combination of textual, audio, and visual modalities (T+A+V) achieves optimal performance on the IEMOCAP dataset.

B. Limitations and Challenges of the Proposed Method

The cross-modal interaction mechanism incorporated in our model has proven effective. However, it cannot fully capture the optimal interaction mechanisms across all scenarios, due to the dependencies among complex sentences. Meanwhile, the multi-level interaction mechanisms increase the computational complexity, which may limit its application ability in real-time scenarios. Therefore, for future work, we will focus on further optimizing the cross-modal interaction mechanism and exploring lightweight methods to enhance its practical application value.

VI. Conclusion

The MITG-CU module proposed in this paper can precisely control the recognition of conversational emotions at both local and global levels. By effectively integrating the distinctive characteristics and complementary relationships among textual, audio, and visual modalities, and combining the relationship timeline module and cross-modal interaction mechanism, the proposed framework significantly enhances both the accuracy and robustness of emotion recognition.

The experimental results show that the MITG-CU model proposed in this paper outperforms six common baseline methods in terms of accuracy, precision, recall rate and F1 score. Especially, it has improved 0.28% and 0.39% in Weighted-F1 and Accuracy respectively. Especially when dealing with long-term conversations and complex emotional scenarios, our model demonstrates excellent stability and generalization capabilities. Future will focus on optimization of cross-modal interaction mechanisms to better meet emotion recognition needs in scenarios, including mental health monitoring and intelligent customer service. Furthermore, additional modalities will be introduced to further enhance the performance of the model.

Acknowledgment

This work was supported by the Science and technology research project of Henan Province, under Grant 24210222012, Key Research Projects of Higher Education Institutions in Henan Province, under Grant 25A460020 and 25B413008, Educational and Teaching Reform Research and Practice Project of Henan Institute of Technology, under Grant 2024JG-YB039 and 2024JG-YB040.

References

[1] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, vol. 62, pp. 2937–2987, 2020.

[2] Z. Y. Huang *et al.*, "A study on computer vision for facial emotion recognition," *Scientific Report*, vol. 13, p. 8425, 2023.

[3] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.

[4] T. Meng, F. Zhang, Y. Shou, H. Shao, W. Ai, and K. Li, "Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4298–4312, 2024.

[5] F. Chen, J. Shao, A. Zhu, D. Ouyang, X. Liu, and H. T. Shen, "Modeling hierarchical uncertainty for multimodal emotion recognition in conversation," *IEEE Transactions on Cybernetics*, vol. 54, no. 1, pp. 187–198, 2024.

[6] A. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal emotion recognition with deep learning: Advancements challenges and future directions," *Information Fusion*, vol. 105, p. 102218, 2024.

[7] H. B. Xu, X. Wu, and X. Liu, "A measurement method for mental health based on dynamic multimodal feature recognition," *Frontiers in Public Health*, vol. 10, p. 990235, 2022.

[8] S. Chougule, S. Bhosale, V. Borle, V. Chaugule, and Y. Mali, "Emotion recognition based personal entertainment robot using ML & IP," *International Journal of Scientific Research in Science and Technology*, vol. 5, no. 8, pp. 73–75, 2020.

[9] D. Z. Jiang, H. Liu, R. G. Wei, and G. Tu, "CSAT-FTCN: A fuzzy-oriented model with contextual self-attention network for multimodal emotion recognition," *Cognitive Computation*, vol. 15, pp. 1082–1091, 2023.

[10] N. Lu, Z. Han, and Z. Tan, "A hypergraph based contextual relationship modeling method for multimodal emotion recognition in conversation," *IEEE Transactions on Multimedia*, pp. 1–13, 2024.

[11] A. Khalane, R. Makwana, T. Shaikh, and A. Ullah, "Evaluating significant features in context-aware multimodal emotion recognition with XAI methods," *Expert Systems*, vol. 42, no. 1, p. 13403, 2023.

[12] S. M. S. Abdullah and A. A. Mohsin, "Facial expression recognition based on deep learning convolution neural network: A review," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 53–65, 2021.

[13] A. A. Gladys and V. Vetriselvi, "Survey on multimodal approaches to emotion recognition," *Neurocomputing*, vol. 556, p. 126722, 2023.

[14] T. Zhang, S. Li, B. Chen, H. Yuan, and C. L. P. Chen, "AIA-Net: Adaptive interactive attention network for text–audio emotion recognition," *IEEE Transactions on Cybernetics*, vol. 53, no. 12, pp. 7659–7671, 2023.

[15] G. Tu, T. Xie, B. Liang, H. Wang, and R. Xu, "Adaptive graph learning for multimodal conversational emotion detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 089–19 097.

[16] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 6818–6825.

[17] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 154–164.

[18] S. Dong, X. Fan, and X. Ma, "Multichannel multimodal emotion analysis of cross-modal feedback interactions based on knowledge graph," *Neural Processing Letters*, vol. 56, pp. 1–17, 2024.

[19] N. Wang and Q. Wang, "Dynamic weighted gating for enhanced cross-modal interaction in multimodal sentiment analysis," *ACM Transactions on Multimedia Computing, Communications*, vol. 21, no. 38, pp. 1–19, 2024.

[20] Y. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569.

[21] A. Joshi, A. Bhat, A. Jain, A. V. Singh, and A. Modi, "COGMEN: Contextualized GNN based multimodal emotion recognition," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 4148–4164.

[22] J. Lee and L. Wooin, "CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologie*, 2021, pp. 5669–5679.

[23] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2022, Virtual and Singapore*, 2022, pp. 7037–7041.

[24] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 77–89, 2024.

[25] D. Hu, Y. Bao, L. Wei, W. Zhou, and S. Hu, "Supervised adversarial contrastive learning for emotion recognition in conversations," arXiv preprint arXiv:2306.01505, 2023.

[26] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 5666–5675.

[27] Y. Bao, Q. Ma, L. Wei, W. Zhou, and S. Hu, "Speaker-guided encoder-decoder framework for emotion recognition in conversation," in *Proceedings of the Thirty-First Inter-national Joint Conference on Artificial Intelligence, IJCAL-22*, 2022, pp. 4051–4057.

Cross-Domain Evaluation of Large Language Models for Abstractive Text Summarization: An Empirical Perspective

Walid Mohamed Aly[1], Taysir Hassan A. Soliman[2], Amr Mohamed AbdelAziz[3]
Information Systems Department-Faculty of Computers and Information, Assiut University, Assiut, Egypt, 71515[1,2]
Information Systems Department-Faculty of Computers and Artificial Intelligence,
Beni-Suef University, Beni-Suef, Egypt, 62111[3]

***Abstract*—Large Language Models (LLMs) have demonstrated remarkable capabilities in generating human-like text; however, their effectiveness in abstractive summarization across diverse domains remains underexplored. This study conducts a comprehensive evaluation of six open source LLMs across four datasets: CNN / Daily Mail and NewsRoom (news), SAMSum (dialogue) and ArXiv (scientific) using zero shot and in-context learning techniques. Performance was assessed using ROUGE and BERTScore metrics, and inference time was measured to examine the trade-off between accuracy and efficiency. For long documents, a sentence-based chunking strategy is introduced to overcome context limitations. Results reveal that in-context learning consistently enhances summarization quality, and chunking improves performance on long scientific texts. The model performance varies according to architecture, scale, prompt design, and dataset characteristics. The qualitative analysis further demonstrates that the top-performing models produce summaries that are coherent, informative, and contextually aligned with human-written references, despite occasional lexical divergence or factual omissions. These findings provide practical insights into designing instruction-based summarization systems using open-source LLMs.**

***Keywords*—Large language models; natural language processing; automatic text summarization; prompt engineering; summarization evaluation**

I. Introduction

With the continued explosion of digital content from online news, scientific research, and conversational platforms, the demand for automated methods of condensed textual information has increased. Automatic Text Summarization (**ATS**) has become an essential tool in Natural Language Processing (NLP), reducing cognitive load and enabling users to quickly grasp key information quickly [1]. In recent years, the task of summarization has gained significance with the rise of Large Language Models (LLMs), which demonstrate remarkable fluency in generating human-like text across a variety of domains [2], [3].

Efforts in text summarization have evolved from traditional heuristic-based methods [4] to more sophisticated approaches to cover different types of text summarization including generic [5], domain-aware [6], multi-document [7], multimodal [8], extractive [9] and abstractive text summarization [10], [11]. Various techniques are used in text summarization to enhance the produced summary, including machine learning approaches [12], deep learning-based approaches [13], [14]. The advent of transformer-based models [15], [16] further advanced the quality of summarization by improving semantic understanding and text generation.

A major shift in NLP occurred with the development of instruction-tuned LLMs, which are pretrained on vast corpora and require no task-specific fine-tuning. Instead, they rely on prompt-based interactions, enabling paradigms such as zero-shot and in-context learning [17], [2]. Despite these advancements, there is limited understanding of how instruction-tuned, open-source LLMs perform across various domains of summarization—especially in contexts such as scientific documents and informal dialogues.

Prompt Engineering is a sophisticated AI engineering methodology [18]. This involves augmenting LLMs by giving them customized cues and modifying the source text to produce the intended result. Prompt engineering is crucial in LLMs because it is essential for unlocking the full potential of such models. In addition, prompt engineering uses prior knowledge and the logical reasoning of the input to influence the outputs generated by the Models [19]. Wide-ranging techniques have been developed as a result of recent notable advancements in the field of prompt engineering [20]. The range of these developments includes basic techniques and more advanced strategies intended to manage challenging jobs [21].

This work contributes to the NLP field by offering domain-aware benchmarking of LLMs under prompt-based conditions. This study also explores the relationship between model architecture, parameter count, and performance across different textual formats. Ultimately, the study provides practical insights for researchers and developers seeking to leverage LLMs for summarization tasks in real-world scenarios without relying on extensive fine-tuning pipelines. Our primary contributions are summarized as follows:

- An evaluation of six open-source LLMs across four benchmark datasets: CNN/DailyMail, NewsRoom, SAMSum, and ArXiv, providing insights into their cross-domain summarization abilities.
- A comparative assessment of prompting strategies, including Zero-Shot Learning (ZSL) and In-Context Learning (ICL), to evaluate the effect of prompt design, context length, and number of demonstrations on model performance.

- An empirical analysis of model scale and architecture to examine the impact of parameter size and design on summarization quality across different domains.
- A dedicated investigation of long-document summarization, introducing a chunking strategy to mitigate the limitations imposed by context window size, and evaluating its influence on summary coherence and quality.
- A detailed efficiency and cost analysis measuring inference time for each model, highlighting trade-offs between performance and deployment feasibility.

The remainder of this paper is organized as follows: Section II introduces related works on text summarization and LLMs. Section III describes in detail the experimental setup and methods, Section IV interprets the results, Section V provides Discussion, Finally, Section VI provides a concise conclusion and future work.

II. Literature Review

Text summarization research has evolved considerably , with two primary paradigms: extractive and abstractive summarization [4]. Extractive methods select sentences or phrases directly from the source document to form summaries [9], while abstractive approaches generate novel sentences by interpreting the core meaning of the input [10].

Early summarization systems, such as the Lead-3 baseline, extracted the first few sentences to generate reliable summaries for news datasets [22]. More advanced neural models treated sentence selection as a classification problem using hierarchical RNNs [23] or applied encoder-decoder architectures with attention mechanisms to improve sentence representation and relevance scoring [24]. The introduction of transformer architectures [25] further enhanced abstractive summarization by capturing long-range dependencies in text. Fine-tuning transformer-based models like BERT [26] for summarization tasks demonstrated strong performance in both extractive [27] and abstractive [28] settings.

State-of-the-art models like BART [29] and PEGASUS [30] have since advanced abstractive summarization. BART's bidirectional encoding and autoregressive decoding enables robust generation, while PEGASUS's gap-sentence generation objective improves content selection and coherence in generated summaries. Variants like Hi-BART further integrate hierarchical encoders to improve the structural understanding of long documents [31].

With the rise of LLMs, prompt-based summarization has gained traction. Instruction-tuned LLMs, such as GPT-3, LLaMA, and Mistral, demonstrate strong zero-shot and few-shot capabilities without the need for task-specific training [2]. Studies like [32] and [33] evaluated such models on standard datasets (e.g., CNN/DailyMail, XSum), demonstrating superior performance for models like text-davinci-003. Other research expanded this assessment to clinical summarization tasks and reported mixed results across radiology reports, patient dialogues, and medical questions [34].

Prompt Engineering is a sophisticated AI engineering methodology [18]. This involves augmenting LLMs by giving them customized cues and modifying the source text to produce the intended result. Prompt engineering is crucial in LLMs because it is essential for unlocking the full potential of such models. In addition, prompt engineering uses prior knowledge and the logical reasoning of the input to influence the outputs generated by the Models [19]. Wide-ranging techniques have been developed as a result of recent notable advancements in the field of prompt engineering [20].

Despite these advancements, long-document summarization remains a key challenge. Standard LLMs struggle to process entire scientific articles due to limited input windows. Techniques such as hierarchical summarization [7] and chunking [35] have been proposed to segment inputs and generate localized summaries prior to merging them. Although effective, these methods often introduce trade-offs between coherence and inference efficiency.

Our study advances previous works by systematically evaluating six open-source LLMs across diverse domains using both zero-shot and in-context prompting strategies. In contrast to studies focused on single-domain tasks or isolated techniques, we offer a unified assessment of summary quality and computational efficiency. In addition, we introduce and empirically evaluate a sentence-level chunking strategy for long document summarization that has not been evaluated systematically in text summarization across instruction-tuned models.

III. Methodology and Experimental Setup

This study evaluates the performance of open-source LLMs in text summarization across multiple domains. To ensure reproducibility, this section outlines the datasets, models, and inference settings used.

A. Datasets

We employed four benchmark datasets encompassing news, dialogue, and scientific texts, which were selected to enable comprehensive cross-domain evaluation. Dataset statistics are summarized in Table I.

1) CNN/Dailymail: The CNN/DailyMail dataset [36] comprises over 300 K news articles from CNN and the Daily Mail, which are commonly used for training and evaluating summarization models.

2) Cornell NEWSROOM: The NEWSROOM dataset [37] includes summaries from 38 news publishers collected via search and social media metadata (1998–2017), encompassing diverse summarization strategies across extractive and abstractive styles.

3) SAMSum corpus: SAMSum [38] contains 16 K annotated chat dialogues with human-written abstractive summaries, which were designed to support dialogue summarization in informal communication settings.

4) ArXiv dataset: The ArXiv dataset [39] features full-text research papers across disciplines such as physics, biology, and computer science, accompanied by human-generated abstracts. The long document structure poses significant challenges for LLM summarization.

In our experiments, we sampled 2 K test instances per dataset, except SAMSum, where all test samples were used.

For few-shot prompting, examples were drawn from each dataset's training set.

TABLE I. STATISTICS OF THE USED DATASETS (K=THOUSANDS)

Dataset	Domain	Documents			Sum Len
		Train	Valid	Test	(words)
CNN/DM	News	287K	13K	11K	52
NEWSROOM	News	995K	108K	108K	26
SAMSum	Dialogue	14K	818	819	–
ArXiv	Scientific	203K	6K	6K	220

B. Prompt Engineering Techniques

As introduced in Section I, prompt engineering involves crafting input instructions to guide LLM behavior. This study applies two primary strategies—**zero-shot learning (ZSL)** and **in-context learning (ICL)**—to investigate how prompt formats affect summarization performance across different models and datasets.

1) Zero-Shot Learning (ZSL) [40]: ZSL is the simplest form of prompting, where a model receives only natural language instructions without any prior examples. The model relies entirely on internal knowledge to perform the task. Formally, a ZSL prompt can be expressed as follows:

$$P = f_{\text{prompt}}(\text{TD}, x_{\text{test}}) \tag{1}$$

where:

- TD is the task description,
- x_{test} is the test input,
- f_{prompt} is the function that transforms these into a natural language prompt.

2) In-Context Learning (ICL) [41]: In ICL, the model observes a few input-output pairs (demonstrations) directly in the prompt and then predicts the output for a new test input. This process can be described as follows:

$$P(y_{\text{test}} \mid x_{\text{test}}, D_k) = \text{LLM}([x_1, y_1, \ldots, x_k, y_k, x_{\text{test}}]) \tag{2}$$

where $D_k = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\}$ denotes k demonstrations. The LLM uses these in-context examples to condition its prediction for x_{test} without updating model parameters.

C. Large Language Models Selection

This study focuses on autoregressive language models—specifically decoder-only architectures—which have become the dominant paradigm in open-source LLM development (Table II). These models generate text by sequentially predicting the next token based on preceding tokens; thus, they are well-suited for text generation tasks, such as summarization. Unlike sequence-to-sequence models, which use an encoder-decoder framework and were originally designed for translation [42], autoregressive models rely solely on a decoder and are optimized for unsupervised learning and generative tasks [43].

We evaluated six decoder-only models, including three from the LLaMA-2 family [44]—LLaMA-2-7B-chat, LLaMA-2-13B-chat, and LLaMA-2-70B-chat—to investigate the effect of scaling on summarization performance. These models are widely adopted due to their open-source availability and strong instruction-tuning capabilities.

In addition, we include Mistral-7B-Instruct-v0.1 [45] and Gemma-7B-it [46], both lightweight instruction-tuned models with 7B parameters. To explore architectural variations, we incorporated Mixtral-8x7B-Instruct-v0.1 [47], a 47B-parameter Mixture-of-Experts (MoE) model that activates only relevant subnetworks during inference, thereby reducing computational costs [48].

This curated selection enabled both intra-scale comparisons between the 7B models and cross-scale evaluations involving larger models. By analyzing these models under uniform evaluation settings, we assess how instruction tuning, architectural innovations, and scaling affect summarization quality and efficiency.

Proprietary models were excluded to ensure reproducibility and mitigate biases associated with closed-source data or opaque updates. The experiments were conducted on two NVIDIA A100 GPUs using mixed-precision (fp16), model parallelism, and quantization techniques to optimize memory usage.

TABLE II. CHARACTERISTICS OF DIFFERENT LLMS UTILIZED IN OUR WORK

Model Name	# Parameters	Context-Length
Llama-2-7b-chat	7B	4K
gemma-7b-it	7B	8K
Mistral-7B-Instruct-v0.1	7B	8K
Llama-2-13b-chat	13B	4K
Mixtral-8x7B-Instruct-v0.1	47B	32K
Llama-2-70b-chat	70B	4K

D. Long Document Processing via Chunking Strategy

Summarizing long documents, such as scientific papers from the ArXiv dataset, presents challenges due to the limited context window of LLMs. Initially, we addressed this by trimming articles to fit the model's maximum context length. Although effective for compatibility, this approach risks omitting important details.

Although extended-context models offer potential solutions, they are not universally available or efficient. To address this issue, we adopt a chunking strategy that divides documents into semantically coherent segments—such as paragraphs or fixed-length chunks. Each chunk is summarized independently, and the resulting summaries are then merged and refined to form a coherent global summary [49].

To improve chunk relevance, we used the Natural Language Toolkit (NLTK)[1] to segment documents into sentences. This

[1]https://www.nltk.org/

ensures semantic integrity in each chunk and ensures that key content is retained during summarization. This process allows for effective handling of long documents while aligning with the input constraints of the models.

E. Evaluation Metrics

The evaluation of summarization quality requires robust metrics that capture both surface-level similarity and deeper semantic alignment. Accordingly, we employ two metric categories: **Word Overlap** and **Semantic Similarity**.

1) Word overlap metrics: These metrics compare the token overlap between the generated and reference summaries. We adopt ROUGE [50], a standard suite for this purpose.

ROUGE-N measures the overlap of n-grams (contiguous sequences of n words) between the system and reference summaries, which indicates lexical similarity.

ROUGE-L evaluates the Longest Common Subsequence (LCS) between summaries, reflecting structural and sequential alignments.

Higher ROUGE scores suggest greater textual overlap and structural resemblance to reference summaries.

2) Semantic similarity metrics: These metrics assess how well a summary preserves the meaning of the original text. We use BERTScore [51], which leverages contextual embeddings from BERT [26] to compute token-level semantic similarity.

Unlike ROUGE, BERTScore captures nuanced paraphrasing and long-range dependencies by evaluating contextual token embeddings. In addition, it excels when n-gram methods are short, particularly in abstractive summarization scenarios.

F. Inference Parameters

This section outlines the inference settings and prompt configurations used in our experiments. Given that the underlying mechanisms of ZSL and ICL were previously detailed, we focused on their practical deployment and tuning.

For ZSL, we employed diverse task-specific prompts tailored to each dataset. In ICL, we varied the number of demonstrations: 1, 3, 5, and 7 to explore the trade-off between contextual guidance and token budget, guided by previous studies [41], [35].

Temperature is a critical generation hyperparameter that controls randomness during decoding [52]. Lower values yield more deterministic and factual outputs, whereas higher values encourage diverse but potentially less accurate generations. Before the full evaluation, temperatures of 0.1, 0.5, and 0.9 were tested. The results in Table III indicate that 0.1 consistently produced the highest ROUGE scores, especially ROUGE-L; thus, it was adopted for all final evaluations.

To leverage the potential of LLMs with ZSL, the different models are guided using a variety of prompts, aiming to assess the results across different prompts to fully utilize the models' capabilities. The prompts are customized based on the diversity of each dataset to provide summaries and assess the output of each model against the reference summaries. The prompts for zero-shot prompting are available in Table IV.

TABLE III. PERFORMANCE ACROSS ALL DATASETS AND TEMPERATURE VALUES. P=PRECISION, R=RECALL, F1=F1-SCORE, RL=ROUGE-L. BOLD VALUES INDICATE BEST PERFORMANCE PER METRIC GROUP

Temp	Metric	Dataset	P/R1	R/R2	F1/RL
0.1	BERT	CNNDM	88.76	86.13	87.41
		NewsRoom	86.99	**87.67**	**87.31**
		SAMSum	89.31	**91.48**	**90.37**
		ArXiv	**86.86**	80.09	83.33
	ROUGE	CNNDM	37.44	**16.42**	**24.53**
		NewsRoom	**25.38**	**8.82**	**19.96**
		SAMSum	**39.35**	**14.61**	**30.70**
		ArXiv	**49.74**	**32.47**	**33.98**
0.5	BERT	CNNDM	86.35	**88.03**	87.16
		NewsRoom	**87.11**	87.30	87.28
		SAMSum	**90.54**	88.76	89.62
		ArXiv	85.67	81.85	83.70
	ROUGE	CNNDM	**38.38**	14.73	24.49
		NewsRoom	24.42	8.19	18.41
		SAMSum	39.08	14.13	30.25
		ArXiv	40.20	12.92	22.09
0.9	BERT	CNNDM	88.40	86.13	87.24
		NewsRoom	87.02	87.40	87.19
		SAMSum	90.30	88.51	89.38
		ArXiv	85.67	**82.03**	**83.80**
	ROUGE	CNNDM	36.87	13.03	23.64
		NewsRoom	24.28	80.01	18.91
		SAMSum	39.34	14.36	30.53
		ArXiv	39.29	12.19	21.61

IV. RESULTS

This section presents the performance of the evaluated LLMs in producing accurate and coherent summaries. All generated outputs were cleaned to eliminate noise and formatting artifacts before evaluation.

A. Comparison of Prompting Techniques

We evaluated the impact of prompting strategies—ZSL and ICL—on summarization performance across different models and datasets. This analysis highlights how prompting format influences the quality and consistency of the generated summaries.

1) ZSL Results: Multiple zero-shot prompts were tested to instruct the models to generate concise summaries without training examples. Fig. 1 shows the comparative performance of LLMs using BERTScore F1 and ROUGE-1 metrics on the four datasets. Subfigure (a) presents BERTScore F1; (b) presents ROUGE-1.

a) CNN/DM Results in ZSL: [2] Table III summarizes the prompt types tested. As shown in Table V, the results can be explored in two ways:

(i) Length-constrained prompts (e.g., Prompts 3, 4, 5), which restrict summary length based on the dataset's average (see Table I); and

(ii) Structured prompts, which provide directive phrasing to guide model output without specifying length.

[2]Only top-performing prompts are shown. Full list available at: https://anonymous.4open.science/r/TextSummarizationCode-5BCB/README.md

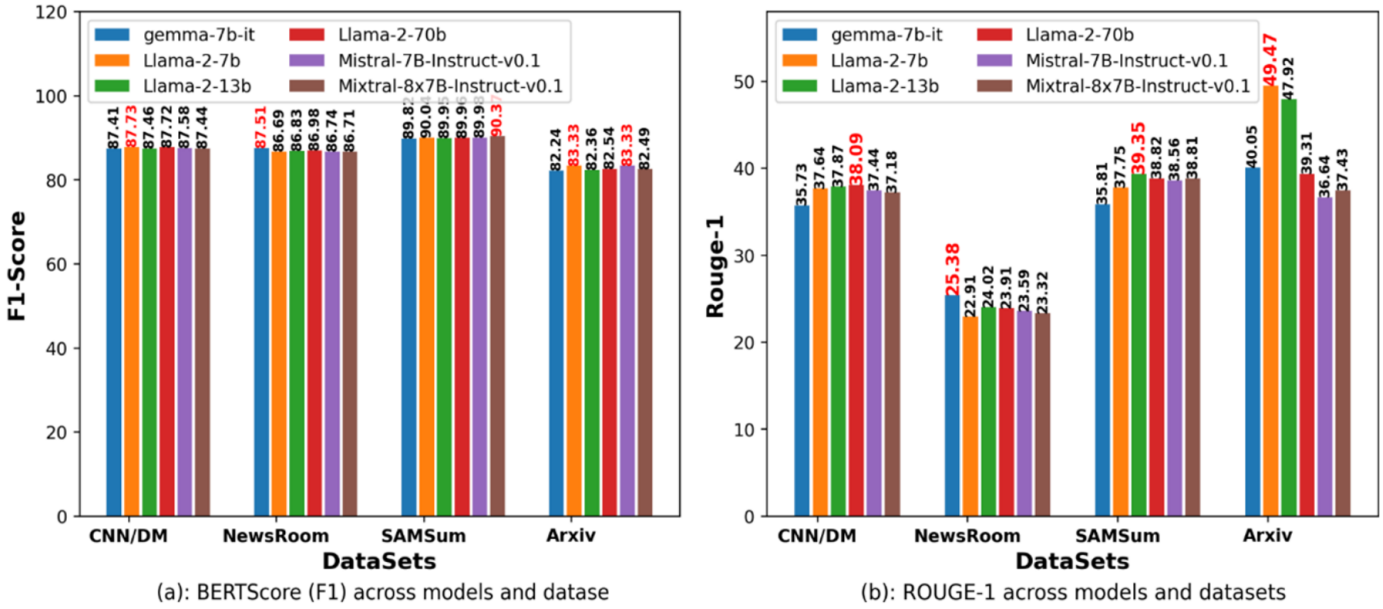


Fig. 1. Performance of LLMs across datasets in the zero-shot setting. X-axis: datasets; Y-axis: ROUGE-1 and BERTScore F1 scores.

These variations help assess how instruction format impacts zero-shot summarization performance.

Comparing models on the same prompt shows that Mistral-7B-Instruct-v0.1 achieved the best ROUGE score on Prompt#1. For BERTScore, Llama-2-70B-chat leads in Recall and F1, while Llama-2-13B-chat scores highest in Precision. Prompt#2 results favor Llama-2-13B-chat across all metrics except Recall, where Llama-2-7B-chat excels. Prompts 3 and 4, being more descriptive, yield varied performance: Mistral-7B and Llama-2-70B lead in ROUGE, while BERTScore is shared among gemma-7b-it, Llama-2 variants, and Mistral-7B, highlighting that some models capture semantic meaning well even if lexical similarity is lower.

Comparing prompts within each model shows that gemma-7b-it performed best on Prompt#3, with top scores in ROUGE-1, ROUGE-L, and BERTScore F1. The Llama-2-7B and Llama-2-13B chats achieved the highest metrics with Prompts#5 and #2 respectively. For Mistral-7B-Instruct-v0.1 and Mixtral-8x7B-Instruct-v0.1, Prompts#3 and #2 are most effective, with Mixtral also showing strong F1 with Prompt#5. These findings underscore the critical role of prompt design in shaping LLM performance.

b) NewsRoom results in ZSL: The NewsRoom dataset includes both extractive and abstractive summaries. In this study, we focus solely on articles with abstractive summaries to align with the goal of evaluating the LLMs' ability to generate coherent, human-like outputs [37]. This provides a more rigorous benchmark for testing generative capabilities. Table VI presents the ZSL results for all models and prompts.

In BERTScore, gemma-7b-it consistently performed well, achieving the highest Precision (86.99), Recall (87.67), and F1 (87.31) with Prompt#1. Llama-2-70B-chat closely follows, particularly in terms of F1 score (86.97), while Mistral-7B-Instruct-v0.1 excels in Recall (88.17) and Precision across several prompts.

For ROUGE, gemma-7b-it again leads with top scores in ROUGE-1 (25.38), ROUGE-2 (8.82), and ROUGE-L (19.96) using Prompt#1. Llama-2-13B-chat and Llama-2-70B-chat demonstrated competitive ROUGE scores across prompts, while Mistral-7B-Instruct-v0.1 achieved its highest ROUGE-1 (23.59) also with Prompt#1.

Overall, Prompts#1 and #2 yield the highest ROUGE and BERTScore results for most models, likely due to their clear and direct instructions. Slight variations in other prompts result in marginally lower performance, illustrating how prompt phrasing influences model output.

In summary, gemma-7b-it performed best on BERTScore, Llama-2-70B-chat demonstrated robust performance across prompts, and Mistral-7B-Instruct-v0.1 proved versatile with strong scores in both lexical and semantic metrics, particularly under well-structured prompt settings.

c) SAMSum results in ZSL: The proposed SAMSum dataset presents challenges for LLMs because of its conversational structure. Table VII compares model outputs for different prompts. The models achieved higher ROUGE scores on news datasets, reflecting the distinct nature of dialogue summarization.

The BERTScore and ROUGE metrics generally agreed. Mixtral-8x7B-Instruct-v0.1 achieved the highest BERT F1 with Prompts#2 and #4, while Llama-2-13B-chat yielded the top ROUGE-1 scores with the same prompts, reinforcing metric consistency.

Simpler prompts (e.g., Prompts#1 and #2) lead to higher ROUGE scores across models, suggesting that clear, concise instructions are more effective. The descriptive prompts (#3–#5) show mixed performance, indicating that overly detailed instructions did not always enhance the results.

Overall, Mistral-7B-Instruct-v0.1 demonstrated robust performance across prompts, demonstrating adaptability for sum-

TABLE IV. COMPLETE LIST OF ALL PROMPTS USED ACROSS DATASETS (P=PROMPT). FULL TEXT PRESERVED EXACTLY

ID	Prompt Text
CNN/DailyMail Prompts	
P#1	Write a concise and comprehensive summary of this news article.
P#2	Provide an abstract of this news article in a direct and concise summary.
P#3	I want you to act as a text summarizer to help me create a concise summary of the provided text. The summary can be up to 3 sentences in length, expressing the key points and concepts written in the original text without adding your interpretations.
P#4	I want you to act as a text summarizer to help me create a concise summary of the provided text. The summary can be up to 2 sentences in length, expressing the key points and concepts written in the original text without adding your interpretations.
P#5	Please summarize the following news article in an informative extractive summary with two sentences.
NewsRoom Prompts	
P#1	Summarize this news article in one sentence.
P#2	Write a concise and comprehensive summary of this news article in one sentence.
P#3	I want you to act as a text summarizer to help me create a concise summary of the following article.
P#4	Provide an abstract of this news article in a direct and concise summary in a few words.
P#5	You are a helpful assistant. Please summarize the following text in one sentence.
SAMSum Prompts	
P#1	Summarize the following dialogue.
P#2	Summarize the following dialogue in a few words.
P#3	Summarize the following dialogue into an abstractive summary without any explanation.
P#4	In short, what's going on in this conversation?
P#5	Summarize the following conversation.
P#6	Summarize this conversation in one or two sentences.
ArXiv Prompts	
P#1	Provide an abstract of the following research article.
P#2	Summarize the main points and findings of the scientific paper.
P#3	I want you to act as a research paper summarizer to get an abstract for this research paper.
P#4	Given this research article, create a TLDR to be used as a formal abstract for this paper.

marizing dialogues. Llama-2-13B-chat also performed well, particularly with direct prompts, indicating its suitability for concise summarization.

d) ArXiv results in ZSL: The ArXiv dataset includes scientific papers and human-written abstracts; thus, it is ideal for testing abstractive summarization. Because of the context-length limitations, documents were trimmed before inference. Table VIII presents model performance across prompts.

Prompt#2 consistently yields high ROUGE scores, especially for Llama-2-13B-chat and Mistral-7B-Instruct-v0.1, demonstrating its effectiveness in guiding models to capture key information. Prompt#3 performs well in F1/RL scores, indicating the value of explicit summarization instructions. Prompt#4 balances BERT and ROUGE scores, helping models generate concise and relevant abstracts. Overall, the ROUGE scores varied more than the BERT scores, highlighting differences in lexical alignment across models.

Llama-2-7B-chat performed best overall, particularly on Prompt#2, confirming its strength in abstractive summarization. Mistral-7B-Instruct-v0.1 excels with Prompt#3, demonstrating consistency in semantic capture. In contrast, gemma-7b-it outperformed the other models, which suggests its limitations in handling longform texts. The variability across prompts and metrics underscores the complexity of long-document summarization and the impact of input trimming on model comprehension.

2) ICL Results: ICL uses example-summary pairs to guide models in generating new outputs. Based on the prompt formulation in Section III, we evaluate ICL performance on each dataset.

a) CNN/DM results in ICL: Table IX presents results with varying numbers of demonstrations. BERT metrics are relatively stable across shot counts, with minor gains from more examples. ROUGE metrics show greater sensitivity: Mistral-7B, Mixtral-8x7B, and Llama-2-70B benefit notably as shots increase.

Llama-2-70B-chat achieves the highest overall scores with 7-shot prompts, and it excels in both BERT and ROUGE. Mixtral-8x7B also performs well. Gemma-7b-it shows limited gains from more demonstrations, indicating constraints in leveraging ICL effectively.

b) NewsRoom results in ICL: Table X summarizes the ICL results obtained on the NewsRoom dataset. **Mixtral-8x7B-Instruct-v0.1** shows the most consistent improvement, particularly with 5-shot prompts, achieving strong performance in both ROUGE and BERTScore metrics. The ROUGE scores varied more across shot counts than the BERT metrics. The Llama family models performed competitively, especially in 7-shot settings, although they were slightly less consistent across fewer examples. Overall, increasing the number of demonstrations generally enhanced the performance across the models.

c) SAMSum results With ICL: when working with ICL on the SAMSum dataset, Table XI showed that **Mixtral-8x7B-Instruct-v0.1** got the best results in terms of BERTScore and Rouge score with seven examples to learn from, which means that it learned the context of the dialogues better than others. The more examples given to the model, the more the result improved, Table XI shows that the results improved when the number of demonstrations in most models was increased.

d) ArXiv results in ICL: Table XII shows ICL results obtained on the ArXiv dataset. Across varying numbers of demonstrations, the **Llama** models consistently outperformed the other models in terms of both ROUGE and BERT metrics. The performance generally improved as the number of in-context examples increased, reflecting better adaptation to the complex summarization task. However, due to input trimming (required by context length constraints), variations in model performance are more pronounced, as some models may struggle to retain contextual coherence after truncation.

e) Overall comparison (ICL vs. ZSL): ICL consistently enhances LLM performance across datasets compared to ZSL by providing contextual demonstrations that improve task understanding. Our experiments demonstrate that increasing the number of in-context examples leads to steady performance

TABLE V. PERFORMANCE OF LLMS ON CNN/DM DATASET USING ZSL. BOLD INDICATES BEST PERFORMANCE PER METRIC GROUP

Prompts		Prompt #1		Prompt #2		Prompt #3		Prompt #4	
Metric		BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b-it									
	P/R1	85.73	31.69	85.99	34.52	**88.76**	36.66	88.23	35.26
	R/R2	88.24	12.37	87.51	13.04	86.13	13.62	**85.91**	12.38
	F1/RL	86.96	20.17	86.72	24.99	87.41	24.48	87.04	23.74
Llama-2-7b-chat									
	P/R1	84.84	28.95	86.24	32.29	86.73	36.41	86.98	37.19
	R/R2	88.28	11.66	**88.23**	11.13	88.11	14.59	87.88	14.21
	F1/RL	86.51	19.84	87.22	19.47	87.41	23.55	87.71	23.51
Llama-2-13b-chat									
	P/R1	**85.98**	32.13	**86.89**	**37.87**	87.41	37.21	87.85	37.48
	R/R2	88.17	13.45	88.06	**17.07**	87.56	13.57	86.71	13.36
	F1/RL	87.05	21.58	**87.46**	**25.99**	87.47	24.01	87.39	24.15
Llama-2-70b-chat									
	P/R1	85.95	29.44	86.81	35.31	87.39	37.08	87.92	**37.76**
	R/R2	**88.53**	11.02	88.07	15.12	87.96	14.19	87.56	**14.82**
	F1/RL	**87.21**	18.39	87.42	25.29	**87.66**	24.03	**87.72**	**25.08**
Mistral-7B-Instruct-v0.1									
	P/R1	85.89	**37.02**	86.57	37.44	86.82	**37.44**	86.91	36.78
	R/R2	88.05	**16.50**	87.88	15.14	**88.38**	**16.42**	87.83	14.56
	F1/RL	86.94	**24.35**	87.21	24.08	87.58	**24.53**	87.36	23.85
Mixtral-8x7B-Instruct-v0.1									
	P/R1	85.38	33.88	86.33	37.18	86.96	35.54	86.91	35.48
	R/R2	88.25	13.94	88.06	15.13	87.25	13.04	86.95	12.28
	F1/RL	86.77	22.12	87.17	24.93	87.09	22.79	86.91	22.02

TABLE VI. PERFORMANCE OF LLMS ON NEWSROOM DATASET USING ZSL

Prompts		Prompt #1		Prompt #2		Prompt #3		Prompt #4	
Metrics		BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b-it									
	P/R1	**86.99**	**25.38**	86.72	24.28	85.63	21.85	85.93	22.89
	R/R2	**87.67**	**8.82**	87.71	8.14	87.75	6.99	87.74	7.59
	F1/RL	**87.31**	**19.96**	87.19	18.81	86.65	16.44	86.81	17.18
Llama-2-7b-chat									
	P/R1	85.19	***22.78***	84.71	21.69	84.51	20.13	85.38	22.15
	R/R2	88.23	7.81	88.29	7.45	88.23	6.72	88.04	6.87
	F1/RL	***86.67***	17.07	86.44	15.97	86.31	14.73	86.67	16.15
Llama-2-13b-chat									
	P/R1	85.42	***24.02***	85.07	22.99	84.88	21.38	85.69	21.76
	R/R2	88.34	8.19	88.41	8.05	88.41	7.02	87.72	6.78
	F1/RL	***86.83***	17.89	86.68	17.17	86.58	15.61	86.68	16.26
Llama-2-70b-chat									
	P/R1	85.71	***23.71***	85.37	23.47	84.75	21.17	85.40	21.81
	R/R2	88.31	8.25	88.39	8.04	88.25	7.16	88.11	7.17
	F1/RL	***86.97***	18.19	86.84	17.51	86.44	15.59	86.71	16.03
Mistral-7B-Instruct-v0.1									
	P/R1	85.40	***23.59***	84.42	20.59	83.43	17.33	83.56	17.44
	R/R2	88.17	9.10	88.09	7.52	88.01	6.01	88.56	6.02
	F1/RL	***86.74***	18.34	86.18	15.61	85.63	12.84	85.69	12.95
Mixtral-8x7B-Instruct-v0.1									
	P/R1	85.39	***23.32***	84.57	21.62	83.96	19.12	83.98	19.52
	R/R2	88.11	7.72	88.17	7.06	88.07	6.37	88.86	7.52
	F1/RL	***86.71***	17.21	86.32	15.83	85.95	13.97	86.01	14.23

TABLE VII. PERFORMANCE OF LLMS ON THE SAMSUM DATASET WITH ZSL

Prompts		Prompt #1		Prompt #2		Prompt #3		Prompt #4	
Metrics		BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b-it									
	P/R1	**86.24**	**33.16**	88.21	34.31	86.41	32.05	**88.76**	35.23
	R/R2	**91.05**	**11.48**	90.10	11.70	90.68	10.29	90.62	12.14
	F1/RL	**88.58**	**25.25**	89.02	26.02	88.48	24.29	89.66	26.72
Llama-2-7b-chat									
	P/R1	85.12	***32.72***	89.48	37.75	86.94	34.22	88.76	37.73
	R/R2	90.51	10.80	90.64	13.35	90.55	11.31	90.62	13.47
	F1/RL	***87.72***	24.86	90.04	29.90	88.69	26.09	89.66	30.00
Llama-2-13b-chat									
	P/R1	86.09	***34.48***	89.23	***38.96***	88.29	36.66	89.16	**39.35**
	R/R2	90.55	12.41	90.72	14.22	91.27	12.41	90.78	14.61
	F1/RL	***88.26***	26.70	89.95	30.47	89.74	27.68	89.88	30.70
Llama-2-70b-chat									
	P/R1	87.02	***35.11***	89.09	38.31	88.56	***37.40***	89.09	38.82
	R/R2	91.14	12.32	90.14	14.21	91.34	13.47	90.89	**14.97**
	F1/RL	***89.02***	26.98	89.59	30.20	***89.91***	29.03	89.96	**30.59**
Mistral-7B-Instruct-v0.1									
	P/R1	87.01	***34.24***	89.11	38.56	87.64	34.85	89.07	38.55
	R/R2	91.51	11.91	90.92	13.94	91.12	12.03	90.71	14.31
	F1/RL	***89.19***	26.51	89.98	30.26	89.33	26.64	89.86	30.38
Mixtral-8x7B-Instruct-v0.1									
	P/R1	86.83	***33.67***	89.31	38.81	87.65	35.07	**89.38**	37.97
	R/R2	91.18	10.78	**91.48**	13.89	91.19	11.27	91.14	13.58
	F1/RL	***88.93***	25.28	**90.37**	30.16	89.37	26.85	***90.23***	29.60

TABLE VIII. PERFORMANCE OF LLMS ON THE ARXIV DATASET WITH ZSL

Prompts		Prompt #1		Prompt #2		Prompt #3		Prompt #4	
Metrics		BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b-it									
	P/R1	84.27	25.72	84.51	40.05	86.14	24.73	86.09	22.52
	R/R2	77.81	15.76	79.46	25.64	78.69	13.45	78.45	12.04
	F1/RL	80.89	18.61	81.89	27.81	82.24	18.05	82.09	16.59
Llama-2-7b-chat									
	P/R1	83.86	***39.61***	83.92	**49.74**	**86.86**	***38.25***	84.73	***36.19***
	R/R2	79.56	21.52	**80.41**	**32.47**	80.09	19.43	79.79	18.84
	F1/RL	***81.64***	27.03	82.12	**33.98**	**83.33**	24.25	82.17	24.13
Llama-2-13b-chat									
	P/R1	84.84	***35.47***	84.21	***47.92***	84.94	37.87	85.89	23.33
	R/R2	79.14	20.63	79.96	29.73	79.95	19.35	78.48	12.01
	F1/RL	***81.88***	24.48	82.02	32.43	82.36	23.93	82.01	15.65
Llama-2-70b-chat									
	P/R1	84.39	***39.31***	84.26	35.48	85.33	***35.29***	85.40	25.64
	R/R2	78.34	25.09	79.25	17.61	79.94	18.30	78.41	12.56
	F1/RL	***81.24***	27.38	81.67	21.65	***82.54***	22.85	81.74	16.70
Mistral-7B-Instruct-v0.1									
	P/R1	85.29	***31.89***	85.61	36.64	86.86	30.59	86.26	29.47
	R/R2	79.14	20.81	79.38	25.77	80.09	18.49	79.74	17.98
	F1/RL	***82.08***	25.18	***82.37***	30.05	83.33	23.98	***82.85***	23.15
Mixtral-8x7B-Instruct-v0.1									
	P/R1	84.89	***35.45***	84.73	37.43	85.23	34.34	**85.28**	31.18
	R/R2	79.29	19.92	79.74	21.84	79.95	18.23	79.34	16.63
	F1/RL	***81.98***	25.39	**82.15**	27.07	82.49	24.39	***82.19***	22.06

TABLE IX. PERFORMANCE OF LLMS ON CNN/DM DATASET WITH ICL

Prompts		1-Shot		3-Shots		5-Shots		7-Shots	
Metrics		BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b-it									
	P/R1	**84.04**	21.78	87.92	33.70	85.40	27.55	85.19	28.36
	R/R2	**82.27**	7.75	86.43	12.37	85.59	9.36	87.30	10.68
	F1/RL	**83.14**	12.60	87.16	22.06	85.47	17.96	86.22	18.55
Llama-2-7b-chat									
	P/R1	86.91	***35.92***	86.14	***32.58***	87.27	***34.35***	88.19	***39.45***
	R/R2	87.52	13.55	88.08	12.77	87.41	14.62	88.59	16.49
	F1/RL	**87.20**	23.01	87.09	**20.85**	***87.29***	23.16	88.37	26.25
Llama-2-13b-chat									
	P/R1	87.54	***36.31***	86.15	***32.37***	87.41	33.95	88.63	**39.43**
	R/R2	87.84	13.71	87.94	12.63	87.16	14.24	89.12	16.48
	F1/RL	**87.67**	23.11	87.02	20.70	87.24	22.62	88.86	26.22
Llama-2-70b-chat									
	P/R1	**87.63**	**38.26**	85.31	29.80	87.09	**33.11**	**88.97**	**40.98**
	R/R2	**87.68**	**14.51**	88.03	11.91	87.62	14.11	88.45	**17.23**
	F1/RL	**87.65**	**24.22**	86.63	19.23	**87.31**	22.32	**88.70**	**27.52**
Mistral-7B-Instruct-v0.1									
	P/R1	85.87	**32.47**	87.40	36.02	87.45	35.90	87.55	34.41
	R/R2	88.26	13.59	87.60	15.70	88.21	14.79	87.22	14.52
	F1/RL	**87.02**	21.19	**87.47**	23.80	87.81	23.70	**87.36**	23.26
Mixtral-8x7B-Instruct-v0.1									
	P/R1	86.75	**34.34**	**88.03**	**37.53**	**87.59**	**37.12**	**87.37**	36.60
	R/R2	86.31	12.74	**88.90**	**16.01**	**88.95**	**15.49**	**89.52**	15.74
	F1/RL	**86.51**	21.57	**88.44**	**24.79**	**88.24**	**24.58**	***88.41***	24.25

gains, particularly for the ROUGE and BERT metrics. These findings underscore the importance of demonstration-based prompting, reinforcing ICL's effectiveness in leveraging LLM capabilities for summarization.

B. Comparisons with State-of-the-Art Models

To situate LLMs within the broader landscape of text summarization, we compared their results with earlier state-of-the-art neural models across datasets Tables XIII and XIV. We focus on the highest-performing LLMs in our study in each dataset and contrast them with earlier existing models.

On news summarization (CNN/DM, NEWSROOM) in Table XIII, LLMs like Llama-2-70b-chat underperform models such as PEGASUS but eliminate dataset-specific training costs, as seen in Gemma-7b-it's NEWSROOM results.

For dialogue summarization, Table XIV shows that while Mixtral-8x7B demonstrates moderate performance, it slightly lags behind specialized models like SICK. This suggests that LLMs require further optimization for conversational coherence. On scientific documents, Table XIV shows that Llama-2-7b-chat achieves a high ROUGE-1 (49.74) but struggles with coherence (ROUGE-L: 33.98), reflecting sensitivity to prompts and context truncation.

These results highlight LLMs' potential for multi-domain adaptability but underscore the need for tailored prompts and hybrid architectures to balance semantic flexibility with syntactic precision.

C. Summarization with Chunking Strategy

To evaluate the impact of chunking, we tested it on the ArXiv dataset. Table XV reports ROUGE-1 and BERTScore F1 results before and after applying chunking.

Results show notable improvements for the Llama-2-70B-chat and Mixtral-8x7B-Instruct-v0.1 models in terms of both lexical and semantic metrics. However, models like Llama-2-13B-chat and gemma-7b-it did not consistently benefit, which

TABLE X. PERFORMANCE OF LLMS ON NEWSROOM DATASET WITH ICL

Examples		1 Shot		3 Shots		5 Shots		7 Shots	
Metric		**BERT**	**ROUGE**	**BERT**	**ROUGE**	**BERT**	**ROUGE**	**BERT**	**ROUGE**
gemma-7b									
	P/R1	84.4	13.1	86.4	20.3	85.4	18.0	80.7	10.7
	R/R2	84.1	3.0	86.6	5.8	87.1	5.6	86.7	3.7
	F1/RL	84.2	11.1	86.5	15.9	86.2	13.8	83.5	8.1
Llama2-7b									
	P/R1	82.4	*16.6*	85.1	21.6	**86.6**	20.9	86.3	**26.6**
	R/R2	85.5	5.0	88.2	7.1	85.8	6.2	88.4	**10.4**
	F1/RL	83.9	13.6	86.6	16.1	86.1	17.2	87.3	**21.2**
Llama2-13b									
	P/R1	82.3	*16.8*	84.7	***22.0***	83.6	23.0	86.9	26.3
	R/R2	85.5	5.0	87.8	7.2	84.8	9.3	88.8	10.0
	F1/RL	83.8	13.4	86.2	16.4	84.1	18.3	87.8	20.8
Llama2-70b									
	P/R1	82.7	17.3	85.3	20.9	84.5	22.6	87.3	26.0
	R/R2	85.8	5.2	88.1	6.6	86.3	7.7	88.5	9.2
	F1/RL	84.2	13.8	86.6	15.6	85.4	17.5	87.9	20.2
Mistral-7B									
	P/R1	**85.0**	19.1	85.8	23.4	86.2	23.6	87.2	22.8
	R/R2	**86.6**	6.6	87.4	8.3	87.8	8.8	86.0	8.4
	F1/RL	**85.8**	15.0	86.5	17.9	87.0	18.4	86.6	19.1
Mixtral-8x7B									
	P/R1	84.8	**20.6**	**86.3**	**23.7**	85.9	**24.3**	**84.0**	21.1
	R/R2	86.3	**6.2**	**88.6**	**8.0**	**88.4**	**9.2**	**87.5**	8.0
	F1/RL	85.5	**16.3**	**87.4**	**18.0**	**87.1**	**18.3**	***85.6***	16.3

TABLE XI. PERFORMANCE OF LLMS ON SAMSUM DATASET WITH ICL

		1 Shot		3 Shots		5 Shots		7 Shots	
		BERT	**ROUGE**	**BERT**	**ROUGE**	**BERT**	**ROUGE**	**BERT**	**ROUGE**
gemma-7b									
	P/R1	84.4	8.6	89.4	34.2	90.3	36.1	84.2	23.3
	R/R2	85.6	3.4	89.2	11.6	89.7	12.5	89.6	8.1
	F1/RL	85.0	6.7	89.3	27.1	90.0	28.8	86.8	17.4
Llama2-7b									
	P/R1	89.3	*39.1*	88.2	35.1	88.4	36.5	90.7	42.5
	R/R2	90.4	14.1	91.5	13.5	90.4	13.5	91.7	17.8
	F1/RL	89.9	31.0	89.8	27.0	89.4	28.2	91.2	34.2
Llama2-13b									
	P/R1	89.3	*39.1*	88.9	***39.4***	88.7	39.3	90.1	42.5
	R/R2	90.3	14.1	91.3	15.9	90.6	15.5	91.3	17.7
	F1/RL	89.8	31.1	90.1	30.8	89.6	30.5	90.6	34.0
Llama2-70b									
	P/R1	89.8	39.3	88.1	36.4	88.9	37.9	90.8	44.1
	R/R2	90.8	13.7	91.9	14.7	91.2	14.4	91.7	18.6
	F1/RL	90.3	31.0	89.9	28.2	90.0	29.5	91.2	35.7
Mistral-7B									
	P/R1	**89.5**	40.8	90.0	42.3	90.5	42.0	90.4	44.8
	R/R2	**91.6**	16.5	91.5	18.3	91.1	17.6	91.8	20.5
	F1/RL	**90.5**	32.2	90.7	34.0	90.8	33.7	91.0	36.0
Mixtral-8x7B									
	P/R1	**90.4**	**41.6**	**90.4**	**43.2**	**91.0**	**44.4**	**91.0**	**45.9**
	R/R2	**92.1**	**16.2**	**91.9**	**18.5**	**92.1**	**20.0**	**92.2**	**20.8**
	F1/RL	**91.2**	**32.6**	**91.1**	**34.6**	**91.6**	**35.8**	***91.6***	**37.0**

indicates that the chunking effectiveness may depend on the model architecture and input handling.

V. DISCUSSION

A. Prompt Diversity and In-Context Learning Effects

The datasets used in this study span distinct summarization domains—news (CNN/DM, NewsRoom), scientific documents (ArXiv), and dialogues (SAMSum)—each requiring tailored summarization strategies. To account for this variability, multiple prompts were designed and tested for each dataset to explore their influence on LLM behavior.

Fig. 2 illustrates how prompt variation affects ZSL performance. The results confirm that LLMs are highly sensitive to prompt phrasing: some prompts elicit more effective responses, improving both lexical and semantic quality, whereas others underperform. This highlights the critical role of prompt formulation in guiding LLMs across domains.

Our findings also demonstrate the adaptability of LLMs to different prompts and tasks. Their ability to generalize without fine-tuning highlights the potential of prompt engineering as a flexible tool. In addition, in-context learning (ICL) significantly boosts performance by providing demonstration pairs, especially as the number of examples increases. This reinforces the value of contextual learning in terms of improving model comprehension and summary generation quality.

TABLE XII. PERFORMANCE OF LLMS ON ARXIV DATASET WITH ICL

Prompts	1 Shot		3 Shots		5 Shots		7 Shots	
Metrics	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b								
P/R1	81.5	22.4	83.7	30.1	85.4	35.2	83.2	29.6
R/R2	78.5	2.6	81.5	7.5	82.3	10.7	81.2	7.1
F1/RL	80.0	13.5	82.6	17.2	83.8	20.6	82.2	17.7
Llama2-7b								
P/R1	85.7	40.8	86.7	42.2	86.1	48.3	86.9	39.8
R/R2	82.1	13.7	**84.2**	15.0	**83.9**	17.8	84.0	14.3
F1/RL	**83.8**	22.6	**85.4**	24.0	85.0	22.5	85.4	23.3
Llama2-13b								
P/R1	**85.8**	40.7	**86.7**	**42.1**	**86.7**	47.9	87.0	40.0
R/R2	81.8	13.7	84.1	**15.1**	83.6	19.4	83.9	14.5
F1/RL	83.7	22.7	85.4	**24.0**	**85.1**	22.1	85.4	23.4
Llama2-70b								
P/R1	85.9	**41.0**	86.7	41.7	85.7	39.4	86.5	**40.4**
R/R2	81.9	**13.7**	82.8	14.7	82.7	15.7	83.5	**14.1**
F1/RL	83.8	**22.8**	84.7	23.8	84.2	18.6	85.0	**23.0**
Mistral-7B								
P/R1	**84.0**	37.1	84.8	37.5	85.6	38.9	85.2	37.0
R/R2	**82.1**	12.8	82.8	12.7	82.1	12.9	81.6	12.7
F1/RL	**83.1**	20.0	83.8	20.6	83.8	21.5	83.3	20.6
Mixtral-8x7B								
P/R1	84.5	39.3	85.7	40.4	86.6	40.1	85.9	39.5
R/R2	80.9	13.1	83.6	14.0	83.5	13.9	84.4	13.4
F1/RL	82.6	21.8	84.6	22.3	85.0	22.9	*85.1*	21.8

TABLE XIII. COMPARISON OF TRADITIONAL SUMMARIZATION MODELS AND LLMS ACROSS CNN/DM AND SAMSUM DATASETS

Dataset	Method	R-1	R-2	R-L
CNN/DM	Hie-BART[31]	44.35	21.37	41.05
	PEGASUS[30]	44.17	21.47	41.1
	Llama-2-70b-chat	40.98	17.23	27.52
NEWSROOM	PEGASUS[30]	45.15	33.51	41.33
	Gemma-7b-it	25.38	8.52	19.96

TABLE XIV. COMPARISON OF TRADITIONAL SUMMARIZATION MODELS AND LLMS ON NEWSROOM AND ARXIV DATASETS

Dataset	Method	R-1	R-2	R-L
SAMSum	SICK[11]	53.73	28.81	49.5
	Mixtral-8x7B	45.86	20.78	36.96
ArXiv	PRIMERA[7]	47.60	20.80	42.60
	PEGASUS[30]	44.70	17.27	25.80
	Llama-2-7b-chat	49.74	32.47	33.98

TABLE XV. PERFORMANCE COMPARISON BEFORE/AFTER CHUNKING ON ARXIV DATASET (ROUGE-1 & BERTSCORE F1)

Model	ROUGE-1		BERTScore	
	Before	After	Before	After
Gemma-7B	**40.05**	39.91	**84.51**	83.60
Llama2-7B	49.74	**50.30**	83.92	**84.06**
Llama2-13B	**47.92**	47.01	**84.21**	82.04
Llama2-70B	35.48	**38.12**	84.26	**86.19**
Mistral-7B	36.64	**36.90**	85.61	**85.80**
Mixtral-8x7B	37.43	**40.02**	84.73	**87.16**

B. LLMs in Summarizing Long Documents

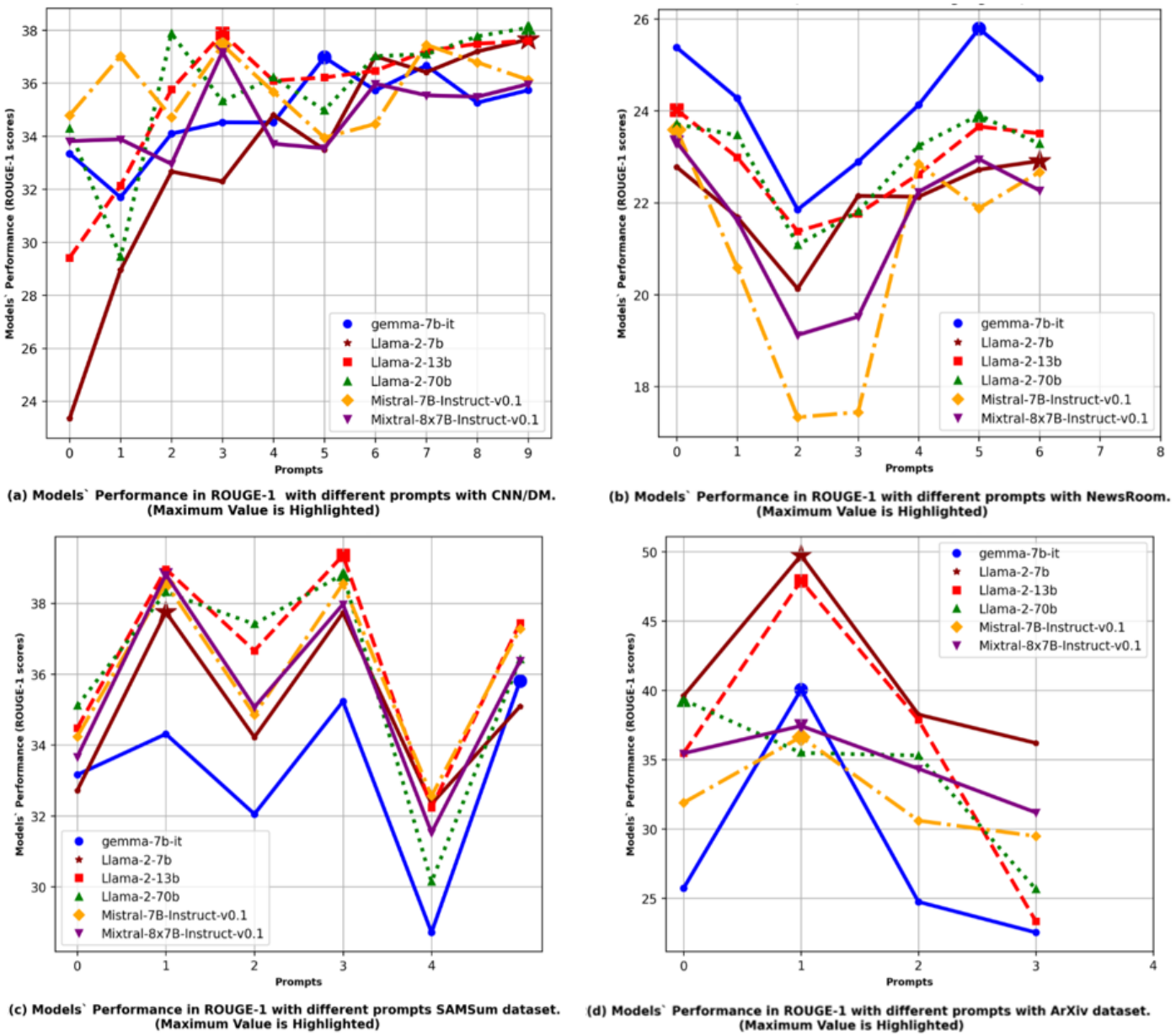
As shown in Table XV, chunking improved summarization performance—particularly ROUGE-1 and BERTScore—for models like Llama-2-70B-chat and Mixtral-8x7B-Instruct-v0.1, indicating better preservation of contextual and semantic content. In contrast, models such as Mistral-7B-Instruct-v0.1, Llama-2-13B-chat, andgemma-7b-it showed minimal or inconsistent gains, which suggests that the chunking effectiveness varies according to the model architecture and context handling capabilities.

While chunking helps retain key information in lengthy inputs and improves final summary quality, it introduces additional inference overhead. This trade-off between performance and efficiency is particularly important when processing long-form scientific texts like those in the ArXiv dataset.

C. Comparison of Generated Summary Lengths with Reference Summaries

Another analysis is conducted regarding the results summaries generated from the employed models to gain more insight into the results. Hence, the lengths of the generated summaries were compared to those of the reference summaries for all datasets used in this study. Table XVI offers a summary of the average lengths of the summaries produced by each model when given different prompts, together with the reference summary lengths for each dataset. The average length was calculated by taking the average lengths of the resulting summaries across the different prompts used. The average lengths of the generated summaries demonstrate that the LLMs do not consistently generate summaries that are comparable to the reference summaries. The results demonstrate that most models generate summaries that are longer than the reference summaries.

Overall, these findings imply that the strengths and weaknesses of various models vary according to the dataset characteristics, in addition to model's length, number of parameters, and prompt building, as previously described. For instance, Llama-2-13b-chat excels in producing summaries that are nearly the reference length for datasets like SAMSum and ArXiv, and gemma-7b-it consistently generates succinct summaries for most datasets.



(a) Models` Performance in ROUGE-1 with different prompts with CNN/DM. (Maximum Value is Highlighted)

(b) Models` Performance in ROUGE-1 with different prompts with NewsRoom. (Maximum Value is Highlighted)

(c) Models` Performance in ROUGE-1 with different prompts SAMSum dataset. (Maximum Value is Highlighted)

(d) Models` Performance in ROUGE-1 with different prompts with ArXiv dataset. (Maximum Value is Highlighted)

Fig. 2. Change in models' performance in ROUGE-1 across datasets using multiple prompts in zero-shot prompting.

However Mistral-7B-Instruct-v0.1 shows a tendency to produce longer summaries that may risk verbosity in an attempt to capture more detail. Consequently, the characteristics of the particular dataset and the intended ratio of detail to conciseness should be taken into account when choosing a model for a summarization task. In conclusion, there is a trade-off between verbosity and the efficiency of the generated summaries when using LLMs for ATS, and this trade-off affects how well these summaries are evaluated in comparison to reference summaries.

D. Variation in LLM Performance: Insights and Contributing Factors

Performance variations across models and datasets reflect the complexity of summarization tasks. Larger models—such as Llama-2-70B-chat and Mixtral-8x7B—tend to perform better, especially on complex datasets like CNN/DM and ArXiv; however, this trend is not universal [53], [54].

Prompt design significantly affects performance. Here, structured, summary-specific prompts generally yield more relevant and coherent outputs than open-ended queries. ICL further enhances model understanding by providing context-rich demonstrations, and performance improves as the number of examples increases [41].

However, no single factor alone accounts for performance differences. Instead, the interplay among model size, architecture, prompt formulation, and demonstration count shapes summarization quality. Continued investigation is needed to isolate and optimize these factors for improved LLM perfor-

TABLE XVI. AVERAGE GENERATED SUMMARY LENGTHS OF LLMS COMPARED TO REFERENCE SUMMARY LENGTHS

	Datasets			
Models	**CNN/DM**	**NewsRoom**	**SAMSum**	**ArXiv**
gemma-7b-it	**53.584**	**27.672**	44.722	143.31
Llama-2-7b-chat	83.456	56.8	47.534	289.3575
Llama-2-13b-chat	65.706	47.67	**39.308**	**208.825**
Llama-2-70b-chat	71.864	44.832	40.45	200.4975
Mistral-7B-Instruct-v0.1	108.056	60.46	48.098	168.57
Mixtral-8x7B-Instruct-v0.1	83.704	45.07	46.528	204.71
Reference Length (#Words)	**52**	**26**	**22**	**220**

mance on abstractive summarization tasks.

E. Qualitative Analysis of Generated Summaries

We further analyzed the summaries of the top-performing CNNDM, NewsRoom, and SAMSum models. Key qualitative observations are:

1) Content coverage: Despite lexical differences, summaries often capture core meanings and key facts. Many outputs paraphrase reference summaries while retaining relevance.

2) Coherence: Most outputs maintain logical flow and clarity, even with varied stylistic choices, reflecting LLMs' capacity for fluent text generation.

3) Error patterns: Common issues include omissions, factual distortions, and overgeneralizations (e.g., "more than 200 people" simplified as "a large crowd"). NewsRoom outputs sometimes misstate facts; CNNDM examples show reduced specificity.

In conclusion, LLM-generated summaries generally preserve the intended meaning and demonstrate strong contextual relevance, even when they diverge from human-written references in phrasing. This illustrates their potential to produce effective abstractive summaries.

F. Inference Time Analysis

We also measured the average inference time required by each model to generate summaries across the datasets. Table XVII shows these results.

TABLE XVII. AVERAGE INFERENCE TIME (SECONDS) PER ARTICLE ACROSS MODELS AND DATASETS

Model	CNN/DM	NewsRoom	ArXiv	SAMSum
Llama2-7b-chat	23.34	25.1	61.2	11.5
Llama2-13b-chat	25.7	30.3	68.8	14.9
Llama2-70b-chat	80.6	95.2	148.4	30.7
Gemma-7b-it	21.1	23.8	63.4	10.2
Mistral-7b-instruct	17.9	20.5	52.9	8.3
Mixtral-8×7b-instruct-v0.1	31.5	36.8	100.3	15.4

Three key patterns emerged: (1) Larger models incur higher latency—e.g., Llama2-70B-chat significantly outpaces its smaller variants; (2) Inference time scales with input length—ArXiv documents take longer to process than shorter SAMSum dialogues; and (3) Architectural optimization matters—Mixtral-8x7B, despite its size, is faster than Llama2-70B on ArXiv due to its sparse MoE design.

These findings emphasize the trade-off between performance and efficiency and highlight the need to balance quality and latency in practical deployments.

VI. CONCLUSION

This study presents a comprehensive evaluation of locally hosted LLMs on news, dialogue, and scientific text summarization tasks using zero-shot learning (ZSL), in-context learning (ICL), and chunking strategies. Our findings highlight several key insights. First, LLM performance varies significantly across domains and models. While models like Mixtral-8x7B-Instruct and Llama-2-70B-chat demonstrated strong results in dialogue and scientific summarization, smaller models such as Gemma-7b-it struggled with long-form inputs. Second, ICL consistently outperformed ZSL in terms of providing contextual examples, which resulted in improved lexical and semantic alignment. Third, the chunking strategy proved beneficial for handling long scientific documents, particularly enhancing outputs from models constrained by limited context windows. Additionally, qualitative analysis of the generated summaries revealed that many LLMs effectively preserved the intended meaning and maintained contextual coherence even when their outputs diverged lexically from human references. Common error patterns—such as omissions, generalizations, and factual distortions—were identified, providing further guidance for model refinement and deployment. Performance differences were influenced by factors such as models architectural considerations, scale, prompt tuning, and the number of in-context examples. Overall, this work underscores the promise of open-source LLMs as adaptable tools for abstractive summarization and offers both quantitative and qualitative insights into their effective use across varied textual domains. Future research could examine optimizing LLMs for domain-specific summarization, particularly for long-form scientific texts where chunking introduces coherence challenges. In addition, exploring retrieval-augmented and integrating external knowledge sources may improve factuality and context awareness. Optimizing the inference efficiency in real-time or low-resource environments will continue to be a crucial area for practical deployment considerations.

REFERENCES

[1] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420305030

[2] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," 2021.

[3] A. Abdallah and A. Jatowt, "Generator-retriever-generator: A novel approach to open-domain question answering," *arXiv preprint arXiv:2307.11278*, 2023.

[4] V. Dehru, P. K. Tiwari, G. Aggarwal, B. Joshi, and P. Kartik, "Text summarization techniques and applications," *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012042, mar 2021. [Online]. Available: https://dx.doi.org/10.1088/1757-899X/1099/1/012042

[5] J. He, W. Kryscinski, B. McCann, N. Rajani, and C. Xiong, "Ctrlsum: Towards generic controllable text summarization," *ArXiv*, vol. abs/2012.04281, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227745074

[6] P. Gigioli, N. Sagar, A. Rao, and J. Voyles, "Domain-aware abstractive text summarization for medical documents," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 2338–2343.

[7] W. Xiao, I. Beltagy, G. Carenini, and A. Cohan, "PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5245–5263. [Online]. Available: https://aclanthology.org/2022.acl-long.360/

[8] S. Rafi and R. Das, "Topic-guided abstractive multimodal summarization with multimodal output," *Neural Computing and Applications*, 2023. [Online]. Available: https://doi.org/10.1007/s00521-023-08821-5

[9] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 2017, pp. 1–6.

[10] S. Chitrakala and N. Moratanch, "A survey on abstractive text summarization," in *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 2016, pp. 1–7.

[11] S. Kim, S. J. Joo, H. Chae, C. Kim, S.-w. Hwang, and J. Yeo, "Mind the gap! injecting commonsense knowledge for abstractive dialogue summarization," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6285–6300. [Online]. Available: https://aclanthology.org/2022.coling-1.548/

[12] A. Dingare, D. Bein, W. Bein, and A. Verma, "Abstractive text summarization using machine learning," in *ITNG 2022 19th International Conference on Information Technology-New Generations*, S. Latifi, Ed. Cham: Springer International Publishing, 2022, pp. 269–276.

[13] P. Kouris, G. Alexandridis, and A. Stafylopatis, "Abstractive text summarization based on deep learning and semantic content generalization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5082–5092. [Online]. Available: https://aclanthology.org/P19-1501

[14] Y. Li, Y. Huang, W. Huang, J. Yu, and Z. Huang, "An abstractive summarization model based on joint-attention mechanism and a priori knowledge," *Applied Sciences*, vol. 13, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/7/4610

[15] J. Pilault, R. Li, S. Subramanian, and C. Pal, "On extractive and abstractive neural document summarization with transformer language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 9308–9319. [Online]. Available: https://aclanthology.org/2020.emnlp-main.748

[16] J. Li, C. Zhang, X. Chen, Y. Cao, P. Liao, and P. Zhang, "Abstractive text summarization with multi-head attention," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[17] Y. Liu, K. Shi, K. S. He, L. Ye, A. R. Fabbri, P. Liu, D. Radev, and A. Cohan, "On learning to summarize with large language models as references," 2023.

[18] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, jan 2023. [Online]. Available: https://doi.org/10.1145/3560815

[19] B. Chen, Z. Zhang, N. Langren'e, and S. Zhu, "Unleashing the potential of prompt engineering in large language models: a comprehensive review," *ArXiv*, vol. abs/2310.14735, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264426395

[20] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," 2024. [Online]. Available: https://arxiv.org/abs/2402.07927

[21] X. Amatriain, "Prompt design and engineering: Introduction and advanced methods," *ArXiv*, vol. abs/2401.14423, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267301483

[22] "The inverted pyramid - purdue owl - purdue university," accessed: 2024-01-22.

[23] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," 2016.

[24] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao, "Neural document summarization by jointly learning to score and select sentences," 2018.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[27] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3730–3740. [Online]. Available: https://aclanthology.org/D19-1387

[28] V. G. M. Gambhir, "Deep learning-based extractive text summarization with word-level attention mechanism," *Multimed Tools Appl*, vol. 81, p. 20829–20852, 2022.

[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Annual Meeting of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204960716

[30] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," 2020.

[31] K. Akiyama, A. Tamura, and T. Ninomiya, "Hie-BART: Document summarization with hierarchical BART," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, E. Durmus, V. Gupta, N. Liu, N. Peng, and Y. Su, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 159–165. [Online]. Available: https://aclanthology.org/2021.naacl-srw.20

[32] L. Basyal and M. Sanghvi, "Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models," 2023.

[33] B. I. L Tang, Z Sun, "Evaluating large language models on medical evidence summarization," *npj Digit. Med*, vol. 6, pp. –, 2023.

[34] D. V. Veen, C. V. Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerova, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly, and A. S. Chaudhari, "Clinical text summarization: Adapting large language models can outperform human experts," 2023.

[35] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, and Z. Sui, "A survey on in-context learning," 2024. [Online]. Available: https://arxiv.org/abs/2301.00234

[36] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, S. Riezler and Y. Goldberg, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 280–290. [Online]. Available: https://aclanthology.org/K16-1028

[37] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 708–719. [Online]. Available: http://aclweb.org/anthology/N18-1065

[38] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "Samsum corpus: A human-annotated dialogue dataset for abstractive summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, 2019. [Online]. Available: http://dx.doi.org/10.18653/v1/D19-5409

[39] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A discourse-aware attention model for abstractive summarization of long documents," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 615–621. [Online]. Available: https://aclanthology.org/N18-2097

[40] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, jan 2019. [Online]. Available: https://doi.org/10.1145/3293318

[41] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11 048–11 064. [Online]. Available: https://aclanthology.org/2022.emnlp-main.759

[42] X. He, G. Haffari, and M. Norouzi, "Sequence to sequence mixture model for diverse machine translation," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, A. Korhonen and I. Titov, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 583–592. [Online]. Available: https://aclanthology.org/K18-1056

[43] M. Dalal, A. C. Li, and R. Taori, "Autoregressive models: What are they good for?" *ArXiv*, vol. abs/1910.07737, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204743691

[44] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[45] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," *ArXiv*, vol. abs/2310.06825, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263830494

[46] "Gemma," https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf, accessed: 2024-02-5.

[47] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de Las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mixtral of experts," *ArXiv*, vol. abs/2401.04088, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266844877

[48] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts," 2024. [Online]. Available: https://arxiv.org/abs/2407.06204

[49] H. Gong, Y. Shen, D. Yu, J. Chen, and D. Yu, "Recurrent chunking mechanisms for long-text machine reading comprehension," 2020. [Online]. Available: https://arxiv.org/abs/2005.08056

[50] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[51] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020.

[52] M. Peeperkorn, T. Kouwenhoven, D. G. Brown, and A. K. Jordanous, "Is temperature the creativity parameter of large language models?" *ArXiv*, vol. abs/2405.00492, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269484653

[53] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[54] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A survey of automatic text summarization: Progress, process and challenges," *IEEE Access*, vol. 9, pp. 156 043–156 070, 2021.

Foreign Key Constraints to Maintain Referential Integrity in Distributed Database in Microservices Architecture

Shamsa Kanwal[1], Nauman Riaz Chaudhry[2], Reema Choudhary[3],
Younus Ahamad Shaik[4], Pankaj Yadav[5], Ayesha Rashid[6]
Department of Computer Science, University of Gujrat, Gujrat, Pakistan[1,2,6]
Department of Software Engineering, University of Gujrat, Gujrat, Pakistan[3]
Jawaharlal Nehru Technology University, Anantapur, India[4]
Data Center and Artificial Intelligence Group, Intel Technology India Pvt Ltd, Bangalore, India[5]

***Abstract*—In the world of modern software development, microservices architecture has become increasingly popular due to its ability to help developers to build large and complex applications that are more agile, faster and more scalable. In large scale applications (such as e-commerce, healthcare, finance, social media, inventory management, travel booking, content management, and customer relationship management systems etc.) with many interconnected services, it is tough to keep the data accurate and consistent. The concept of referential integrity is applied to validate the data. Referential integrity refers to the preservation of relationships between tables. In a monolithic architecture, where the application and database are closely linked and co-located on the same server, referential integrity via foreign key constraints makes it feasible to preserve consistent and accurate data. But in the microservices architecture, maintaining referential integrity across distributed databases poses significant challenges due to its decentralized nature of data management. This study utilizes a hybrid research methodology, combining empirical research and design science research to discover and address the challenges of maintaining referential integrity in distributed databases in microservice architecture and calculate the response time by comparing and analyzing with existing models. The results of the evolution in term of response time are presented in this work.**

***Keywords*—Foreign key constraints; relational mapping; referential integrity; saga pattern; event driven architecture; APIs; microservice; distributed database**

I. Introduction

In modern cloud-native applications, distributed databases are an important part of microservices architecture, as they offer availability, reliability and scalability [1]. In large application with many interconnected links, it's hard to keep the data correct and consistent. For that the concept of "Referential integrity Constraints" is used to keep the data correct.

In a monolithic architecture, where the application and database are closely linked and co-located on the same server, referential integrity via foreign key constraints makes it feasible to preserve consistent and accurate data [2]. But in the microservices architecture, maintaining referential integrity across distributed databases poses significant challenges due to its decentralized nature of data management.

A. Database in Monolithic Architecture

Monolithic architecture is an outdated software design method where whole application or system is made as a single, interconnected unit [3]. In this architectural style, all the components, modules, and functionalities of the application are tightly integrated and run as a single codebase within a single process [4]. In monolithic architecture, CRUD operations directly perform in centralized relational database [5]. When new data is Create into a table that contains a foreign key column, the referenced value should first be checked to ensure that it exists in the referenced table's primary key. If not, the data is not allowed. For Read operation, foreign keys don't affect the process directly, but they help to make sure the data is correct. When deleting data from a parent table that is referenced by a foreign key constraint in another table, database checks if related child record exist. It can either delete the child records or stop the deletion too, depending on the rules. When updating data in the parent table, database updates the related records according to rules or stop the change to avoid breaking the link.

A monolithic system is hard to grow and update quickly because all components and modules are tightly connected. This approach contrasts with modern microservice architecture for handling large applications which offer better availability, well organized use of resources and scalability, making them a better option.

B. Distributed Databases in Microservices Architecture

Large-scale applications can be developed by dividing them into smaller, autonomous services thanks to microservices architecture [6]. These services have their own unique business logic and database. Each service is updated, tested, deployed, and scaled independently of the others [7]. This method frequently makes use of databases with different schemas, shared database, and distributed databases, each of which fulfills a distinct purpose. Shared database and databases per service are the two primary categories of distributed database. When it comes to databases for individual services, every microservice has a unique database [8].

1) Shared databases in microservices architecture: In a shared database approach, all services use a single database.

This means that multiple services can write data to the same database, and each service has access to data written by the other services. Services can conveniently access and share data since they all contribute to the same database [9]. With all services operating within a single database for reading and writing, data sharing is simplified. This eliminates the need for complex data synchronization or replication procedures, as all services have direct access to the same dataset.

In a shared database model, multiple microservices access and function on the same database. Although this simplifies data consistency and relational integrity implementation, it basically violates key microservices principles such as service autonomy, scalability, and independent deployment.

Service autonomy is negotiated because microservices become tightly coupled through the shared schema, meaning that changes in one service's database tables can accidentally impact others. This dependency increases the risk of breaking changes and requires coordinated deployments, reducing the flexibility and agility that microservices are meant to provide. Scalability is also stuck, as the shared database can become a bottleneck; scaling services independently is more challenging when all trust on the same data store. Moreover, operational complexity grows when services must negotiate over shared resources, and fault isolation becomes difficult. If one service causes a database failure or locks essential tables, it can affect all other services connected to that database. These limitations run contrary to the core goals of microservices architecture, which emphasize loose coupling, independent scaling, resilience, and continuous delivery. Therefore, while the shared database approach may offer short-term gains in performance and development simplicity, it imposes long-term constraints on maintainability, scalability, and system robustness.

2) Databases per service in microservices architecture: In the database-per-service pattern within a distributed database architecture, each microservice has its own dedicated databases. These services manage their own data independently [1] without interference from other service. Each services can deploy, scale and update independently. Both synchronous and asynchronous [10] method used for communication between the services. Synchronous communication means that a service sending a request to another service and pausing its operation until response is received. Typically, this method is realized through protocols like HTTP/REST or gRPC [11]. While asynchronous communication occurs when a service sends a request or message to another service and continues its operation without waiting for an immediate response. This method is often facilitated by message brokers like RabbitMQ or Kafka. This procedure make it simple for various microservices to exchange and access data across many platforms.

In microservices architecture with distributed databases, maintaining data consistency particularly using foreign key constraints for referential integrity is a major challenge. Unlike traditional systems where databases impose these rules directly, microservices have separate databases, making this execution tough. The CAP theorem describes that distributed systems can only attain two out of three: Consistency, Availability, and Partition Tolerance. Meanwhile microservices must handle network issues (partition tolerance), they often choose between strong consistency or high availability.

Some researchers propose the BAC model (Backup, Availability, Consistency) [12] to better handle real-world failures. Systems that focus on consistency (CP) ensure correct data but may be slower or unavailable during issues, while systems that focus on availability (AP) respond faster but rely on eventual consistency, requiring extra logic to maintain data integrity. In such cases, referential integrity is often managed through custom validation, events, or shared read-only views instead of traditional foreign key constraints.

C. Background

The term "Referential integrity" and "Foreign key constraints" are the two central concepts for ensuring data consistency, accuracy and validity across related tables in relational database system. A relational database is a key element for storing, managing, and retrieving data for the entire program in a monolithic architecture. The architecture uses structured tables to store data [12] from multiple application modules in a single, centralized relational database. A centralized data schema defines the organization of the database by specifying tables, relationships, columns, data types, and constraints to ensure consistent data storage.

In relational database each database table has a PRIMARY KEY that uniquely identifies its records. A FOREIGN KEY serves as a database key designed, which consists of one or more columns in a table, references the primary key in another table, thereby creating relationships between the tables. The term "FOREIGN KEY CONSTRAINTS" is a column (or combination of columns) in a table whose values must match values of a column in some other table. This [12] foreign key constraint connects the child table's foreign key to the parent table's primary key, maintaining referential integrity.

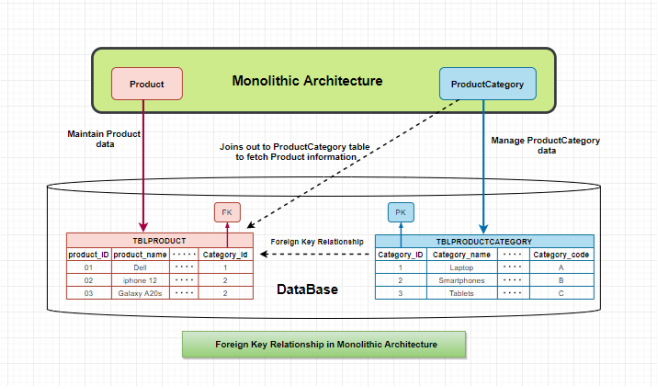


Fig. 1. Foreign key relationships in monolithic architecture.

Fig. 1 illustrates how foreign key constraints are openly imposed within a monolithic architecture, where all modules share a single centralized database. In this setup, relational integrity between entities (e.g., Product and ProductCategory) is maintained through native database-level foreign keys.Distributed database is a type of database system where data is stored through multiple physical locations, frequently on different servers, regions, or even data centers, but appears to users as a single, unified database. Each location may manage part of the data independently, but the system ensures coordination so users can access and manipulate data as if it were stored in one place.

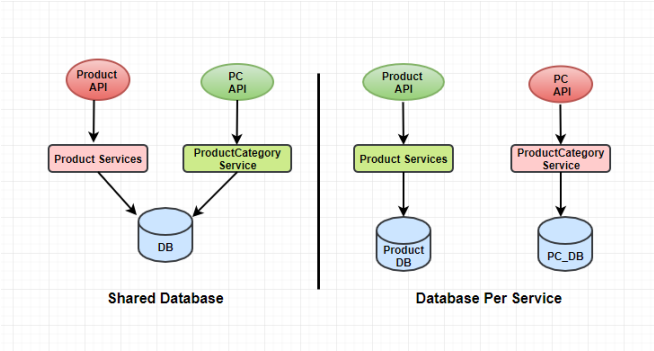


Fig. 2. Shared database and databases per service patterns.

Fig. 2 shows the difference between Shared Database and Database per Service. Shared Database use single database but doesn't scale well, while Database per Service gives each service its own data and needs extra steps to keep data linked.

Microservice architecture is a software development strategy [24] that differs from monolithic architectures. Rather than having one large application, it involves breaking down an application into several small, independent services that are not tightly integrated with one another. One of the key feature of this architecture is the use of multiple, independent services that communicate with each other over the network. Each service is created to perform a specific business task and interacts with other services through clearly defined APIs.

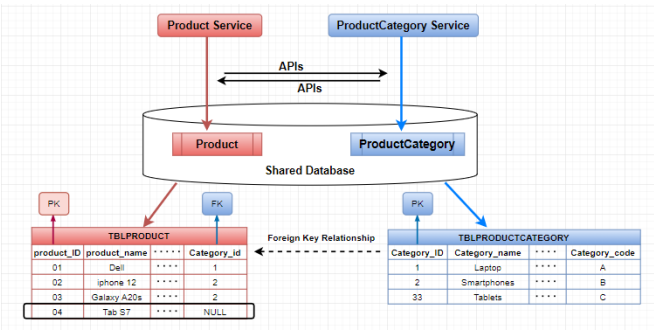


Fig. 3. Foreign key relationships in shared database using multiple microservices.

In Fig. 3 multiple services, such as ProductService and ProductCategoryService, interact with a single, unified database. ProductCategory service manage product category in the ProductCategory table, while ProductService manage individual products in the Product table. Both tables are linked through a foreign key relationship, ProductCategory table as the parent and the Product table as the child, ensuring referential integrity [12] and consistent data relationships across services. On the other hand, microservices also interact with each other via events. For instance, ProductService can produce the event exposed by ProductCategoryService to fetch category details.

For example in Fig. 4, consider two microservices: ProductService and ProductCategoryService. Each of these microservices manages its own separate database, but they are linked through a foreign key relationship to maintain referential integrity.

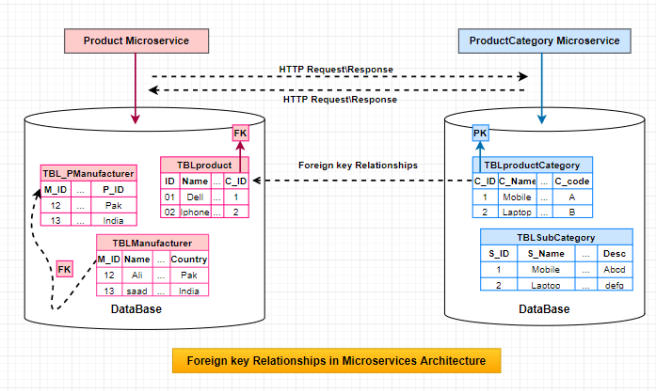


Fig. 4. Database-per-service in microservices architecture.

Synchronous communication is ideal for real-time interactions requiring immediate responses, offering simplicity and directness. In contrast, asynchronous communication excels in scalability and resilience, making it suitable for tasks that can be handled independently over time. In large application with many interconnected links, it's hard to keep the data correct and consistence. For that the concept of "Referential integrity Constraints" is used to keep the data correct. Still, the complexity of handling inter-service communication and ensuring data consistency across multiple services increases.

D. Purpose of the Research

The purpose of this research is to discover how referential integrity can be sustained in distributed databases within microservices architecture, where traditional foreign key constraints cannot span across service boundaries. Using a case study with separate Product and ProductCategory services, the study evaluates and compares practical methods such as shared databases, API validation, event-driven communication, and saga pattern to preserve data relationships while balancing scalability, fault tolerance, and system complexity.

E. Research Objectives

- To handle foreign key constraints in distributed database in microservices architecture.
- To improve response time during CRUD operation in distributed database in microservices architecture.
- To analyze and compare the existing model in a distributed database in microservices architecture.

In the introduction section the paper addresses the challenges of maintaining referential integrity in distributed databases within microservices architecture, this research explores the application of foreign key constraints across different architectural patterns. It investigates how referential integrity is preserved in both shared databases and database-per-service designs using three distinct implementation methods: Shared Database Pattern using APIs, Saga Choreography Pattern with database-per-service, and an Event-Driven Architecture with database-per-service. Each method offers different trade-offs in terms of consistency, performance, and decoupling, which are critical factors in distributed environments.

The remainder of this paper is structured as follows: Section II presents a comprehensive literature review covering the evolution from monolithic systems to microservices, and analyzes the implications for database design and schema evolution. Section III outlines the research methodology, detailing the experimental setup for each of the three architectural patterns. Section IV presents the results, including a comparative analysis of referential integrity, response time, and behavior under CRUD operations. Finally, Section V concludes the findings and provides recommendations for selecting appropriate patterns to maintain referential integrity based on system requirements.

II. Literature Review

This chapter gives us a closer look at what's happening in research right now and what we already know about the topic. It shows us what different authors have to say about the subject based on what's already been written.

A. Evolution to Microservices and Distributed Databases

Microservice-based architecture, which is used for developing software systems with independent maintainability and scalability [13]. Many companies migrate their existing monolithic software systems towards microservice-based architectures [14]. The paper [15] proposes a data-centric process to identify microservices by decomposing the database (on the bases of foreign key relationship) into clusters and identifying topics that correspond to potential microservices. This process performs database schema analysis and clustering.

Migrating from a monolithic to microservice architecture with a shared database and then to multiple databases based on the database-per-microservice pattern. The migration process involves identifying microservice boundaries, decomposing the application into multiple entities. Framework called Architect, which is based on a domain-specific language for building dependable distributed systems [16], to ensure the data consistency of distributed transactions using the Saga pattern to reduce complexity and attain eventual consistency.

The saga pattern is a technique used to ensure data consistency across multiple microservices [17]. It uses sequential transactions that trigger subsequent ones and compensatory transaction when failures occur. However, it lacks isolation, allowing access to ongoing transactions. An improved version[18] of the saga pattern addresses this issue by using a quota cache and commit-sync service to rollback modifications in case of failures without affecting the database layer. The database is committed once all transactions are successful.

Event-driven architecture used for building a distributed database for microservices that is scale and flexible. In EDA, the communication between different components of the system happen through events. EDA can be used to ensure that change to the data are propagated to all the microservices that depend on it. By decoupling the components of the system through events which ensure that changes are propagated quickly [19] and efficiently, while also maintaining data consistency across the system . Identified multiple instances were enforcing foreign key constraints between microservices is essential to ensure accuracy.

Architectural trade-off analysis based on patterns to assist software architects in selecting appropriate pattern for foreign key relationship between different microservices using immutable event by conducted a study [20] to identify architectural patterns of microservices that affect structural design decision related to the size of services, database sharing, and level of service coupling. Software architects can better understand and choose the most appropriate pattern for foreign key relationship between different microservices by using Command Query Responsibility Segregation(CQRS), which can guide the architecture in the desired direction.

Adoption of microservice-based event-driven architecture for data-intensive systems, also known as big data system(BDS) in which microservices communicate through APIs. The authors conducted action research [21] which are Questionnaire to investigate the reasons for the adoption of this architecture and to document the challenges and lessons learned. The resulting architecture was found to enable easier maintenance and fault isolation.

Recovering microservices that use polyglot persistence, which means they use multiple data storage techniques. When backups are taken independently, it becomes difficult to recover them in a consistent state, and references across microservice boundaries may break after disaster recovery. The authors [22] propose a weak global consistency definition for microservice architectures and a recovery protocol that utilizes cached referenced data to reduce the time interval after the most recent backup, during which state changes may have been lost.

The widespread adoption [23] of microservices and event-driven software applications, which favor the separation of databases into independent silos of data owned by a single service. It mention that microservices communicate via synchronous REST API calls and introduces the concept of virtual actors, The only way for a microservice to get data from another microservice is to make a call on an endpoint.

In study [27] Paper offers a modest way to design and evaluate microservices that focus on data, keeping in mind the limitations of the CAP theorem, which applies to distributed systems like microservices. It helps developers and architects compare different design and database patterns to find the best mix of cost and performance. The CAP theorem is not used as a strict rule but as a monitor to group systems as CA (Consistency–Availability), CP (Consistency–Partition tolerance), or AP (Availability–Partition tolerance). The approach is explained with three real-world examples, each showing how different choices affect system design, making it easier to make smart architecture and database decisions. Comparison table is shown in Table I.

B. Database Schema Analysis

A dominant model for cloud based application, consist of integrated, independently deployed and occasionally distributed services that communicate with each other through API calls and maintain their own databases. While enabling software development is a key motivation for implementing microservices, latest surveys[25] highlight database schema evolution as an important challenge. The main aim of this thesis is to reduce the burden which evolve in cloud based application and provide developers automated database schema

TABLE I. COMPARISON TABLE

Paper Names and References	Isolation	Scalability	Consistency
Towards Migrating Legacy Software Systems to Microservice [15].	Yes	Yes	Yes
A Framework for the Migration to Microservices [16].		Yes	Yes
Enhancing Saga Pattern for Distributed Transactions within a Microservices [18]	Yes	Yes	Yes
A distributed database system for event-based microservices [19]	Yes	Yes	Yes
Data Management in Microservices	Yes	Yes	
A Method for Architectural Tradeoff Analysis Based on Patterns [20].		Yes	
From Monolithic Big Data System to a Microservices Event-Drive [21].	Yes	Yes	Yes
Microservice Disaster Crash Recovery.		Yes	Yes
Operational Stream Processing [23].		Yes	Yes
Performance Engineering for Microservices: Challenges & Directions [24].		Yes	Yes

evolution with tools. This research[26] mainly focused on modeling, generation, evolution, visualization and monitoring of database schema in microservices.

Microservice-based architecture, which is used for developing software systems with independent maintainability and scalability. Many companies migrate their existing monolithic software systems towards microservice-based architectures. The paper [4] proposes a data-centric process to identify microservices by decomposing the database (on the bases of foreign key relationship) into clusters and identifying topics that correspond to potential microservices. This process performs database schema analysis and clustering.

The abstraction of software components from legacy system, is an active research area within microservices architecture. Domain driven business dataflow diagram have been proposed using different approaches such as static and dynamic code analysis. Legacy systems, which are typically monolithic in nature, usually come with inherited database schemas and data, resulting limited code opportunities for code reused. To identify the database object for microservices extraction, analyze these artifacts, types and usage are aim of this research. A systematic mapping study [27] in legacy system discovered using database artifacts to identify services or component, the result concentrate on using database schema, data state and dependencies to identify separate business unit that could be transformed to potential microservices.

Event-driven architecture(EDA) used for building a distributed database for microservices that is scale and flexible [28]. In EDA, the communication between different components of the system happens through events, which are messages that are generated when certain actions or changes occur. EDA can be used to ensure that change to the data are propagated to all the microservices that depend on it. By decoupling the components of the system [19]through events which ensure that changes are propagated quickly and efficiently, while also maintaining data consistency across the system.

A microservice architecture built an application which are independently deployable, loosely coupled services focused on business capabilities, need careful data storage consideration [29]. Evaluating RDBMS and document store data stores through prototypes and performance test using simplified Electronic Medical Record (EMR) studies [30] five different data storage pattern for microservices in this research such as document store microservice overtake RDBMS, shared database overtake databases per services, CQRS overtake shared database all pattern have minor resource impact. Regardless have conflicts with modularity, the CQRS with event sourcing pattern could alleviate issues. In future may use hybrid pattern to discover data related operation in shared database and databases per services.

When migrating from monolithic application to microservices architecture various challenges occur, mostly with splitting and sharing databases. In [31] this article show framework where managing database splitting among different microservices to achieve data consistency and independent deployment while addressing latency problems. The offered solution improve performance, improved response time and handle messaging in evaluation. This framework maintains constructing services with better database independence.

Demand for data driven features has led to complicated networks of services which communicate each other without integrated architecture in Modern web applications. For large scale web services, synapse helps easy to used strong semantic system [32], allowing isolated, scalable and independent services to share data across different databases. In MVC-based applications that force high level data models using scalable replication mechanism that synchronizes read only view of shared data in real time, enabling data replication using different databases such as MYSQL, MongoDB, Oracle and PostgreSQL. Data duplication among different databases SQL and NoSQL supports in this synapse. For data integration use different models and ORMs to guarantee compatibility between SQL and NoSQL, without compromising scalability to maintain consistency. Synapse determine good performance and scalability. Table II shows database schema analysis in microservices.

III. RESEARCH METHODOLOGY

This section describes the methodology we followed when lot of literature review has been discussed to support the hypothesis of this research study and gap in the literature is also identified. The primary goal to study literature review was to collect experiences on distributed databases in microservice and how foreign key constraints work in microservices architecture, reporting best practices, and identify the problems occur during development processes. This study utilizes a hybrid research methodology, combining empirical research and design science research to discover and address the challenges of maintaining referential integrity in distributed databases in microservice architecture. This dual approach allows us to first observe and analyze real-world systems (empirical) and then propose and evaluate a novel design framework (design

TABLE II. DATABASE SCHEMA ANALYSIS IN MICROSERVICES

Research Papers Name	Distributed Database
Automated Database Schema Evolution in Microservices [26]	Database-Per-Service
Towards Migrating Legacy Software Systems to Microservice-based Architectures [4]	Database-Per-Service
Using Database Schemas of Legacy Applications for Microservices Identification: A Mapping Study [27]	Database-Per-Service
A Distributed Database System for Event-based Microservices [19]	Database-Per-Service
Data Management in Microservices: State of the Practice, Challenges, and Research Directions [7]	Database-Per-Service & shared database
Evaluation of Data Storage Patterns in Microservices Architecture [30]	Database-Per-Service & shared database
Framework for Interaction Between Databases and Microservice Architecture [31]	Database-Per-Service & shared database
Synapse: A Microservices Architecture for Heterogeneous-Database Web Applications [32]	Database-Per-Service

science), ensuring both theoretical consistency and practical relevance.

A. Proposed Framework

Subsequent steps have been followed for the proposed Framework in Fig. 5.

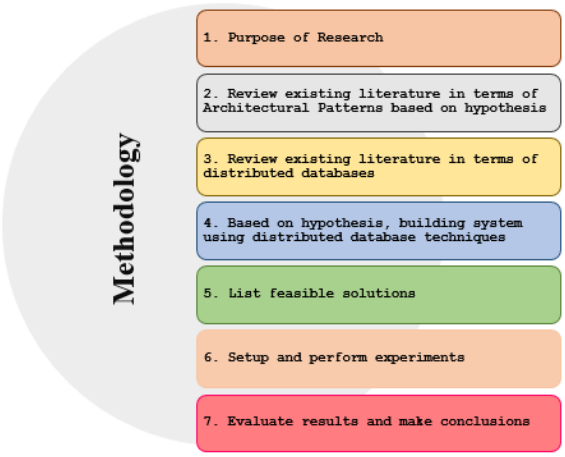


Fig. 5. Proposed framework.

1) Feasible solutions: This research is conducted using hybrid research methodology, combining empirical research and design science research supports the development of robust methods and techniques such as event-driven architecture, the saga choreography pattern, and APIs using a shared database technique which is based on hypothesis and evidences collected from the literature that address the challenges of enforcing foreign key constraints in distributed databases within microservices architecture.

B. Requirements for the Framework Implementation

This section provides an overview on which methods and tools were used.

1) Software requirements: Since this research is focused on foreign key constraints in distributed databases in microservices architecture the software used for the development of proposed framework is Spring tool Suite (STS) is an Eclipse-based development environment design for developing spring based application. This tool used for creating and managing spring projects, validation for spring configuration files and graphical boot dashboard to manage spring boot application. For validation Postman is an excellent tool that API developers and tester can use for guaranteeing API performance and functionality. In order to test RESTful APIs, create and run HTTP requests such as PUT, DELETE, GET, POST etc. In addition MYSQL is a common Relational database management system that helps for quick read/write operations and Xampp is the most commonly used free and open-source web server solution packages that work across multiple platforms. It is suitable for developers, when establishing a local web server environment for PHP development.

2) System requirements: Windows operating system is required for experimental purpose.

C. Implementation of Framework

After analysis the literature review we discuss some method which are used in this research.

1) Shared database pattern using APIs: In microservices design, a shared database is a designed pattern in which multiple microservice share the same central database, as opposed to databases per service pattern. All services interacts with a common set of tables or schema, usually within single database pattern. This structure makes it possible to store data centrally. In this pattern, multiple services shared the same database and each service get data from the same pool. Both services product and ProductCategory microservices communicate with each other through APIs and shared the same database.

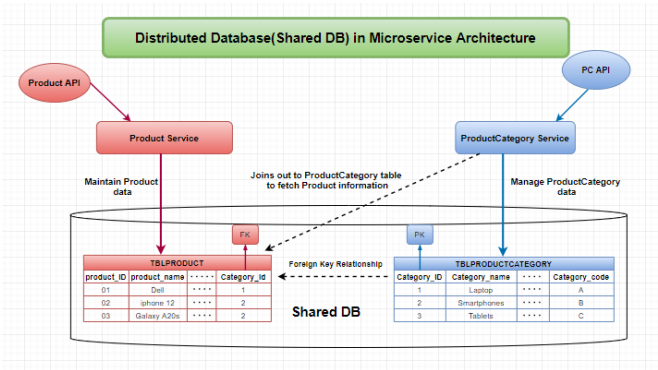


Fig. 6. Shared database pattern using APIs.

MYSQL was used to create and implemented distributed database such as shared databases. For this model, first step to create product and ProductCategory microservices in spring tool suit(STS) IDE. The dependencies as shown in Fig. 7 were managed using maven and pom.xml file in project directory.

The simplified shared database pattern of product and ProductCategory is shown in above Fig. 6, where both mi-

```
<dependency>
    <groupId>org.springframework.boot</groupId>
    <artifactId>spring-boot-starter-data-jpa</artifactId>
</dependency>
<dependency>
    <groupId>org.springframework.boot</groupId>
    <artifactId>spring-boot-starter-web</artifactId>
</dependency>
<dependency>
    <groupId>org.springframework.boot</groupId>
    <artifactId>spring-boot-starter-logging</artifactId>
</dependency>

<!-- https://mvnrepository.com/artifact/org.springframework/spring-context -->
<dependency>
    <groupId>org.springframework</groupId>
    <artifactId>spring-context</artifactId>
    <version>5.1.6.RELEASE</version>
</dependency>
```

Fig. 7. Shared database dependencies.

croservices shared the same database. Java persistence API was used to connect to a MYSQL database such as shared database with product and ProductCategory microservices.

```
server.port=8082

logging.level.org.springframework.web=DEBUG;

spring.datasource.driverClassName=com.mysql.jdbc.Driver
spring.datasource.url=jdbc:mysql://api.cwiztech.com:3319/PRODUCT?autoReconnect=true&useSSL=false
spring.datasource.username=root
spring.datasource.password=UOG@786

# Specify the DBMS
spring.jpa.database = MYSQL

# Show or not log for each sql query
spring.jpa.show-sql = true
```

Fig. 8. Database connectivity.

Database connectivity in Fig. 8 discusses to the mechanisms and protocols that allow an application or service to establish and maintain a connection with a database system for executing operations such as data retrieval, insertion, updates, and deletion.

This pattern was implemented using STS and create product and product-category microservices. for each microservice, Create Entity class which is all about tables, Repository is an interface that define the presentation part of the database, Rest Controller classes define APIs so that request can be accept and provide the response. Microservices communicate with each other through APIs. Both microservices Perform CRUD operation and calculate response time. Product microservice send request to ProductCategory microservice to get data. Product-category microservices accept request, if required data available in product-category microservices using Request Mapping.

2) Saga choreography pattern using database per services: The Saga Choreography Pattern (Fig. 9) is a distributed coordination mechanism in microservices architecture that manage distributed communication across several services without depending on a central orchestrator. In this pattern, every service waits for particular events and having finished its activity, posts an event that other services respond to, carrying on the sequence until the whole business process is over. Microservices basically follow the pattern called database per services pattern. Where product service has its own database similarly product category microservices have its own database. so if you observe, there is two microservices and they are pointing to two different databases. If it is a single database, then the approach will be same as monolithic. In this pattern, microservices communicate with each other using asynchronous way such as events.

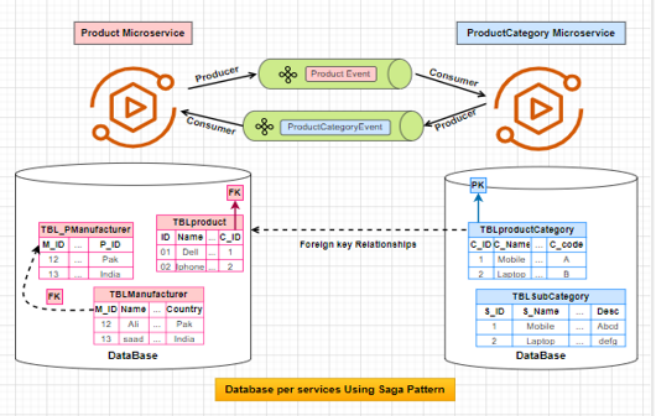


Fig. 9. Saga choreography pattern.

At the very first, to Setup apache kafka environment such as message broker to manage publish and subscribe the events. After that, create product and product-category microservices and added required dependencies in Fig. 10.

```
<dependency>
    <groupId>org.springframework.boot</groupId>
    <artifactId>spring-boot-starter-test</artifactId>
    <scope>test</scope>
</dependency>
<dependency>
    <groupId>io.projectreactor</groupId>
    <artifactId>reactor-test</artifactId>
    <scope>test</scope>
</dependency>
<dependency>
    <groupId>org.springframework.cloud</groupId>
    <artifactId>spring-cloud-stream</artifactId>
    <scope>test</scope>
    <classifier>test-binder</classifier>
    <type>test-jar</type>
</dependency>
<dependency>
    <groupId>org.springframework.kafka</groupId>
    <artifactId>spring-kafka-test</artifactId>
    <scope>test</scope>
</dependency>
```

Fig. 10. Dependencies in saga choreography pattern.

For each microservices add Connection Properties in Application properties files to connect microservice with its own database in Fig. 11.

```
1 spring.datasource.driver-class-name=com.mysql.cj.jdbc.Driver
2 spring.datasource.url = jdbc:mysql://localhost:3306/productservices
3 spring.datasource.username = root
4 spring.datasource.password = Password
5 spring.jpa.show-sql = true
6 spring.jpa.hibernate.ddl-auto = update
7 spring.jpa.properties.hibernate.dialect = org.hibernate.dialect.MySQL5Dialect
8 spring.jpa.properties.hibernate.format_sql=true
```

Fig. 11. Database connectivity.

Product services will receive the request from the product-category microservice then again product service will redirect that request to the product-category services to validate whether that required category is available or not in Product-Category microservices. Microservice Communicate with each other using events. Any change occurs in microservices was handle using Event.

3) Event driven architecture using database per services: Event-driven architecture (EDA) using database per service is a paradigm for designing and building software systems where events are central to communication between different components or services. Event driven architecture using asynchronous way to communicate with each. Event driven architecture (EDA) using database per service, components react to events asynchronously, enabling loose coupling, scalability, and flexibility.

Event driven architecture use apache Kafka as a message broker for asynchronous way to communicate with each. A kafka Broker is a server that hosts Kafka partitions. It is responsible for managing the storage, retrieval, and replication of the data (messages) produced by Kafka producers and consumed by Kafka consumers is a server that hosts Kafka partitions. It is responsible for managing the storage, retrieval, and replication of the data (messages) produced by Kafka producers and consumed by Kafka consumers. A Kafka Topic is a logical channel to which producers publish messages and from which consumers read messages. A topic is like a category or feed name to which records are sent as shown in Fig. 12.

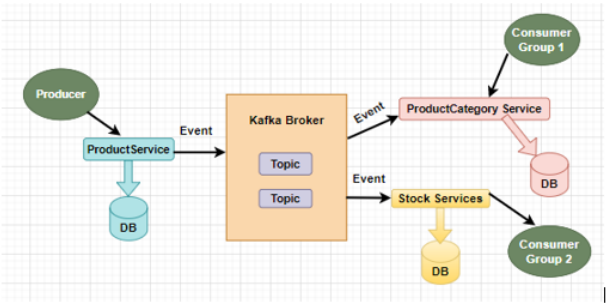


Fig. 12. Event driven architecture using apache kafka.

First step to Setup apache kafka environment such as message broker for asynchronously communication to manage publish and subscribe the events. Start zookeeper service using command.

```
C:\kafka_2.13-3.7.0>.\bin\windows\zookeeper-server-start.bat .\config\zookeeper.properties
```

Fig. 13. Start zookeeper service command.

Start Kafka Broker using command (see Fig. 13 and Fig. 14)

```
C:\kafka_2.13-3.7.0>.\bin\windows\kafka-server-start.bat .\config\server.properties
```

Fig. 14. Start kafka broker command.

In EDA, product service has its own database which is responsible for managing the product similarly product category microservices has its own database which is responsible for managing product category. At the very first, create product and product-category microservices and added required dependencies.

```
<dependencies>   Add Spring Boot Starters...
    <dependency>
        <groupId>org.springframework.boot</groupId>
        <artifactId>spring-boot-starter</artifactId>
    </dependency>

    <dependency>
        <groupId>org.projectlombok</groupId>
        <artifactId>lombok</artifactId>
        <optional>true</optional>
    </dependency>
    <dependency>
        <groupId>org.springframework.boot</groupId>
        <artifactId>spring-boot-starter-test</artifactId>
        <scope>test</scope>
    </dependency>
</dependencies>
```

Fig. 15. Dependencies in event driven architecture.

Event driven architecture use event producer and event consumer while communication. Event Producer that produce event when state changes occur. Producers publish events to the event broker. while event consumer consumes event and perform action according to the request. Performing CRUD operations in event driven architecture (Fig. 15) using database per services pattern and validate foreign key relationship.

```
@Service
public class ProductProducer {

    private static final Logger LOGGER = LoggerFactory.getLogger(ProductProducer.class);

    private NewTopic topic;

    private KafkaTemplate<String, OrderEvent> kafkaTemplate;

    public ProductProducer(NewTopic topic, KafkaTemplate<String, OrderEvent> kafkaTemplate) {
        this.topic = topic;
        this.kafkaTemplate = kafkaTemplate;
    }

    public void sendMessage(OrderEvent event){
        LOGGER.info(String.format("Order event => %s", event.toString()));

        // create Message
        Message<OrderEvent> message = MessageBuilder
                .withPayload(event)
                .setHeader(KafkaHeaders.TOPIC, topic.name())
                .build();
        kafkaTemplate.send(message);
    }
}
```

Fig. 16. Event producer in EDA.

Product microservices create a new entity in its local database such as product. After that, product microservices publish an event with the entity data to Apache kafka. After that kafka broker publish an event and send event to the related microservice. Then productcategory microservices consume the event and creates a related entity in its local database such as Productcategory services with a foreign key referencing the product entity in product microservices database (Fig. 16).

In event-driven systems using the database-per-service pattern, enforcing foreign key constraints across services is not possible through traditional relational resources. Instead, referential integrity is maintained asynchronously using events. However, this approach must handle the realities of network partitioning, as defined in the CAP theorem, which states that in a distributed system, it is difficult to simultaneously guarantee Consistency, Availability, and Partition Tolerance. Event-driven architectures usually prioritize availability and partition tolerance, which means consistency is eventually achieved rather than immediately enforced. This trade-off affects CRUD operations, especially create and delete, where

integrity checks must rely on event messages from the owning service (e.g., ensuring a ProductCategory exists before creating an Order). In such cases, idempotency becomes essential. Since events can be delayed, duplicated, or retried, systems must be designed to safely handle the same event more than once without corrupting data. This adds operational complexity but is crucial to ensure reliable referential integrity in the absence of direct foreign key enforcement. Thus, network partition handling and idempotent design are key pillars in using event-driven architectures to maintain data consistency across microservices.

IV. Results and Discussion

A. Overview

This section contains the experimentation of the proposed framework to evaluate how foreign key constraints can be used to maintain referential integrity in distributed databases within a microservices architecture. To test our approach, we designed a case study involving two central microservices: ProductCategoryService and ProductService, which simulate a real-world e-commerce scenario. These services are logically connected through a foreign key relationship, where each product belongs to a specific product category.

B. Case Study

In the designed case study, we implemented a distributed database system using the Database per Service pattern. The ProductCategoryService manages the ProductCategory entity and its database, while the ProductService manages the Product entity. A logical foreign key constraint is defined between the Productcategory-id field and the ProductCategory-id, ensuring that each product must belong to a valid category.

In this section the results of Response time of the three distributed database patterns are presented. For clarity, we presented the most-significant results obtains. In the first section, we examine how distributed database patterns compare with each other and in the last section we observed the response time while performing CRUD operation in distributed databases and we define the performance for all the patterns considered. The response time of the three pattern were compared in a pairwise such as:

- Shared database using APIs vs. Event Driven Architecture.
- Shared DB Using APIs vs. choreography saga pattern.
- Event driven Architecture vs. Saga Choreography pattern.
- Response time for all the pattern.

C. Evaluation Criteria

Microservices architecture can be evaluated using variety of criteria such as data integrity, data availability, data scalability, transaction management and response time. In this research, we will mainly focus on how response time and critical aspects such as scalability, fault isolation, and operational complexity behave when performing CRUD operations. Response time within a distributed database in microservices architecture, refer to the duration when the request is send to and when the client receive the response. This include numerous stages such as transmission time (time for request to reach the target microservices), service processing time (time which involve business logic and database CRUD operations on the distributed database) which consist of data access, data consistency and communication across the node or replica, inter services communication alongside response transmission time back to client.

D. Shared Database vs. Event Driven Architecture

The result in below Table III shows the response time for distributed database pattern. Some test cases were executed against the deployed patterns. Each test case consists of a HTTP request which consists a server side overview for data access. In this case, each test cases were executed using well defined APIs in shared database. This was carried out turn by turn. First shared database using APIs/HTTP and then database per service pattern such as event driven architecture using events were carried out and perform CRUD operation and calculate the average response time.

TABLE III. Comparison Between Event Driven Architecture vs. Shared DB

DB Operation	Event Driven Pattern			Shared DB		
	Number of Request (ms)					
	10	100	1000	10	100	1000
CREATE	73.17	695.47	7155.34	43.89	428.65	3987.45
READ	40.23	385.51	3993.04	35.33	359.09	3367.77
UPDATE	92.45	859.44	8967.71	58.2	575.17	5734.63
DELETE	89.88	887.04	8852.08	59.91	603.45	6072.11
Avg. Response Time	73.93	507.10	7242.04	49.33	491.59	4790.49

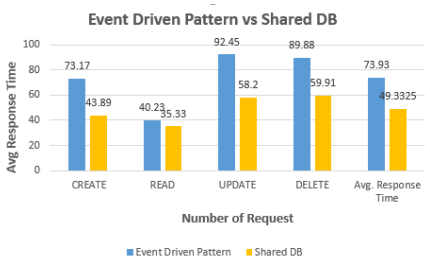


Fig. 17. Avg. Response time between shared DB VS Event driven architecture.

Taking an average of all the tests carried out indicates that shared database pattern performed better than database per services (EDA) while performing CRUD operation. Shared database pattern have relatively Fast response time as compare to event driven Architecture. Response time in Shared database is Fast due to access of same database. While event driven architecture takes time for coordination between different services, performing CRUD operations and inter-communication services and many more. Due this reason event driven architecture takes more time as show in Fig-17.

While analyzing other factors shared Database has limited scalability because all services depend on a single database, which becomes a bottleneck as traffic increases. It also has weak fault isolation. If one service or its database fails, it can affect others since everything is tightly connected. While it's easy to manage at first, complexity grows with more services due to shared schema and organization needs. In contrast, Event-Driven Architecture offers very high scalability because services and messaging systems can grow independently. It provides excellent fault isolation, as services are decoupled and communicate through events that can be retried if one fails. However, it brings very high operational complexity, requiring careful handling of events, retries, and monitoring, which adds to the system's management effort.

E. Shared Database vs. Saga Choreography Pattern

The result presented in table shows the response time of distributed database between the shared database using APIs/HTTP request and database per services. In this case, each test cases were executed using SAGA as well as database per services. This was carried out turn by turn. First shared database using APIs/HTTP were carried out and perform CRUD operation and calculate the average response time. Secondly, SAGA pattern using database per services pattern was carried out and perform CRUD operations and calculate the average response time (Table IV).

TABLE IV. COMPARISON BETWEEN SAGA CHOREOGRAPHY PATTERN VS. SHARED DB

DB Operation	Saga Pattern			Shared DB		
	Number of Request (ms)					
	10	100	1000	10	100	1000
CREATE	67.19	598.43	6495.35	43.89	428.65	3987.45
READ	37.48	332.29	3577.61	35.33	359.09	3367.77
UPDATE	83.67	789.11	7982.47	58.2	575.17	5734.63
DELETE	75.89	668.56	6830.03	59.91	603.45	6072.11
Avg. Response Time	66.05	597.09	6221.36	49.33	491.59	4790.49

The comparison between shared database and Saga choreography pattern using database per services pattern were carried out on the bases of average response time while performing CRUD operations.

Taking an average of all the tests carried out in fig. 18 that shared database pattern performed better than saga choreography pattern while performing CRUD operation. Shared database pattern have relatively Fast response time as compare to SAGA. Response time in Shared database is Fast due to access of same database. While saga choreography pattern involved stages such as transmission time, service processing time which involve business logic and database CRUD operations and global coordination time alongside response transmission time back to client. Due to this reason Saga choreography takes more time as compare to shared DB.

In terms of scalability, the Shared Database is harder to scale because all services trust on a single database, which can become a bottleneck as the system grows. In contrast, Saga

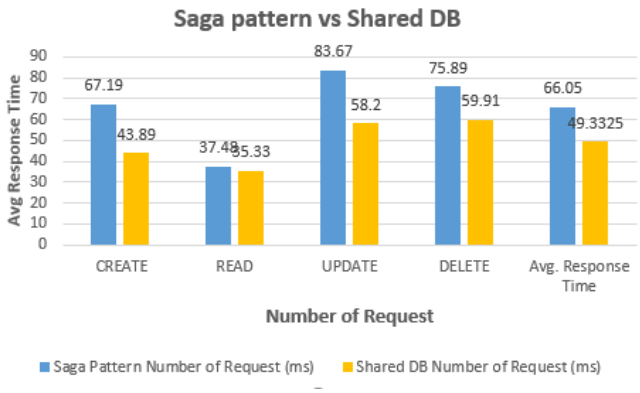


Fig. 18. Avg. Response time between shared DB vs. Saga choreography pattern.

Choreography scales more easily since each service has its own database and can grow independently. For fault isolation, the Shared Database is weaker. Issues in one service or changes to the shared schema can affect others. Saga Choreography, however, offers stronger fault isolation, as services are decoupled and operate independently. When it comes to operational complexity, the Shared Database is easier to manage at first due to its simplicity, but complexity grows as more services are added. Saga Choreography is more complex from the beginning, as it requires managing inter-service communication and ensuring data consistency through event handling.

F. Database Per Services (Choreography Saga Pattern) vs. Database Per Services (EDA)

The result presented in the table shows the response time between distributed databases. Both distributed databases used database per services pattern. The results of the table indicate an average response time of choreography saga pattern and event driven architecture.

TABLE V. COMPARISON BETWEEN SAGA CHOREOGRAPHY PATTERN VS. EVENT DRIVEN ARCHITECTURE

DB Operation	Saga Pattern			Event Driven Pattern		
	Number of Request (ms)					
	10	100	1000	10	100	1000
CREATE	67.19	598.43	6495.35	73.17	695.47	7155.34
READ	37.48	332.29	3577.61	40.23	385.51	3993.04
UPDATE	83.67	789.11	7982.47	92.45	859.44	8967.71
DELETE	75.89	668.56	6830.03	89.88	887.04	8852.08
Avg. Response Time	66.05	597.09	6221.36	73.93	507.10	7242.04

The data in Table V: shows that Event Driven Architecture performed better for each batch update, deletion, insertion, then Saga choreography pattern.

Saga choreography design pattern choose asynchronous communication between the services and many other factors such as network latency may low their response time. In the

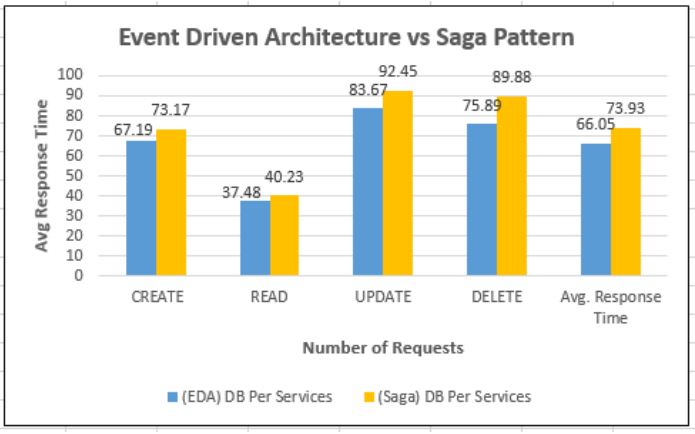


Fig. 19. Avg. Response time between event driven Architecture vs. Saga choreography pattern.

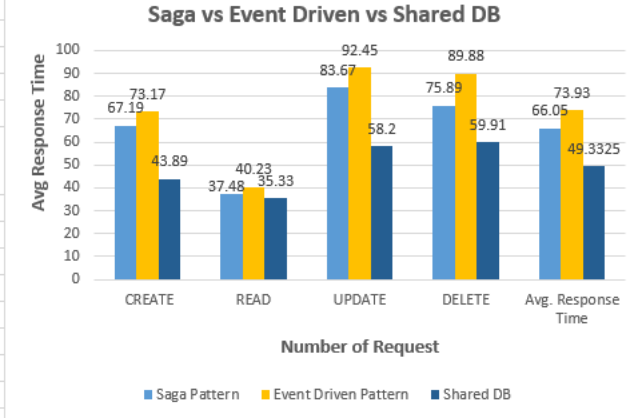


Fig. 20. Avg. Response time of different methods.

TABLE VI. COMPARISON BETWEEN SAGA CHOREOGRAPHY PATTERN VS. EVENT DRIVEN VS SHARED DB

DB Op	Saga Pattern			Event Driven Pattern			Shared DB		
	Number of Request (ms)								
	10	100	1000	10	100	1000	10	100	1000
C	67	598	6495	73	695	7155	43	428	3987
R	37	332	3577	40	385	3993	35	359	3367
U	83	789	7982	92	859	8967	58	575	5734
D	75	668	6830	89	887	8852	59	603	6072
Avg RT	66	597	6221	73	507	7242	49	491	4790

above graph clear show that in each test case event driven architecture perform better then saga choreography pattern. The comparison between Saga Choreography Pattern and Event-Driven Architecture (EDA)—both using the database-per-service approach—shows that while both patterns offer good scalability and fault isolation, EDA takes these aspects further. Saga Choreography provides high scalability, as services operate with their own databases and coordinate through events, and offers strong fault isolation through loose coupling. However, it requires managing compensating transactions and choreography logic, making it operationally complex. On the other hand, EDA achieves very high scalability and fault isolation, thanks to full service decoupling and asynchronous communication. But this comes at the cost of very high operational complexity, involving advanced event design, ensuring idempotency, managing retries, and implementing distributed tracing to monitor the system.

While comparison between the three patterns of distributed database. We observe each test case with each three pattern such as shared database, saga choreography and event driven architecture as shown in the Table VI.

After observing each test with 10,100,1000 number of request respectively, we conclude that shared database performed faster as compare to other two patterns. While Event driven architecture performs better then saga choreography pattern while performing CRUD operation. For more clarity in the results we represent the result in graphical representation in Fig. 19 and Fig. 20.

V. CONCLUSION AND RECOMMENDATION

In this research distributed databases in microservices has been broadly analyze while performing CRUD operations and how we achieve loose coupling between the services while maintaining data integrity and very efficient data access performance. Furthermore, an evaluation of foreign key constraints to maintain referential integrity in distributed databases (shared database, database per services) as shown in this research bring out one of the core challenge of microservices. This research further proven that shared database pattern give better performance then database per services while performing simple data related operation such as simple CRUD operation.

In this research after examined the test case, it could be deduced that foreign key constraints in a database per services pattern does not carried out at database level. While future could find a way to mitigate this challenge. When designing data driven microservices, it remains a crucial factor to consider.

REFERENCES

[1] S. Li, H. Zhang, Z. Jia, C. Zhong, C. Zhang, Z. Shan, J. Shen, and M. A. Babar, "Understanding and addressing quality attributes of microservices architecture: A systematic literature review," *Information and software technology*, vol. 131, p. 106449, 2021.

[2] L. Jiang and F. Naumann, "Holistic primary key and foreign key detection," *Journal of Intelligent Information Systems*, vol. 54, pp. 439–461, 2020.

[3] W. Lu, "Replacing a monolithic web application with a new backend framework," 2018.

[4] C. K. Rudrabhatla, "Impacts of decomposition techniques on performance and latency of microservices," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, 2020.

[5] X. Liu, S. Jiang, X. Zhao, and Y. Jin, "A shortest-response-time assured microservices selection framework," in *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*. IEEE, 2017, pp. 1266–1268.

[6] B. Barua, M. Whaiduzzaman, M. Mesbahuddin Sarker, M. Shamim Kaiser, and A. Barros, "Designing and implementing a distributed database for microservices cloud-based online travel portal," in *Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022*. Springer, 2023, pp. 295–314.

[7] R. Laigner, Y. Zhou, M. A. V. Salles, Y. Liu, and M. Kalinowski, "Data management in microservices: State of the practice, challenges, and research directions," *arXiv preprint arXiv:2103.00170*, 2021.

[8] X. Wu, N. Wang, and H. Liu, "Discovering foreign keys on web tables with the crowd," *Computing and Informatics*, vol. 38, no. 3, pp. 621–646, 2019.

[9] M.-D. Pham, L. Passing, O. Erling, and P. Boncz, "Deriving an emergent relational schema from rdf data," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 864–874.

[10] L. M. D. S. dos Santos *et al.*, "Data distribution and access in a microservices architecture," 2024.

[11] O. Al-Debagy and P. Martinek, "A comparative review of microservices and monolithic architectures," in *2018 IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI)*. IEEE, 2018, pp. 000 149–000 154.

[12] J. Motl and P. Kordík, "Foreign key constraint identification in relational databases." in *ITAT*, 2017, pp. 106–111.

[13] G. Blinowski, A. Ojdowska, and A. Przybyłek, "Monolithic vs. microservice architecture: A performance and scalability evaluation," *IEEE access*, vol. 10, pp. 20 357–20 374, 2022.

[14] V. Talaver and T. A. Vakaliuk, "Reliable distributed systems: review of modern approaches," *Journal of edge computing*, vol. 2, no. 1, pp. 84–101, 2023.

[15] Y. Romani, O. Tibermacine, and C. Tibermacine, "Towards migrating legacy software systems to microservice-based architectures: a data-centric process for microservice identification," in *2022 IEEE 19th International Conference on Software Architecture Companion (ICSA-C)*. IEEE, 2022, pp. 15–19.

[16] E. Volynsky, M. Mehmed, and S. Krusche, "Architect: A framework for the migration to microservices," in *2022 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*. IEEE, 2022, pp. 71–76.

[17] S. Lungu and M. Nyirenda, "Current trends in the management of distributed transactions in micro-services architectures: A systematic literature review," *Open Journal of Applied Sciences*, vol. 14, no. 9, pp. 2519–2543, 2024.

[18] E. Daraghmi, C.-P. Zhang, and S.-M. Yuan, "Enhancing saga pattern for distributed transactions within a microservices architecture," *Applied Sciences*, vol. 12, no. 12, p. 6242, 2022.

[19] R. Laigner, Y. Zhou, and M. A. V. Salles, "A distributed database system for event-based microservices," in *Proceedings of the 15th ACM International Conference on Distributed and Event-based Systems*, 2021, pp. 25–30.

[20] T. de Oliveira Rosa, J. F. L. Daniel, E. M. Guerra, and A. Goldman, "A method for architectural trade-off analysis based on patterns: Evaluating microservices structural attributes," in *Proceedings of the European Conference on Pattern Languages of Programs 2020*, 2020, pp. 1–8.

[21] R. Laigner, M. Kalinowski, P. Diniz, L. Barros, C. Cassino, M. Lemos, D. Arruda, S. Lifschitz, and Y. Zhou, "From a monolithic big data system to a microservices event-driven architecture," in *2020 46th Euromicro conference on software engineering and advanced applications (SEAA)*. IEEE, 2020, pp. 213–220.

[22] M. Manouvrier, C. Pautasso, and M. Rukoz, "Microservice disaster crash recovery: a weak global referential integrity management," in *International Conference on Computational Science*. Springer, 2020, pp. 482–495.

[23] A. Katsifodimos and M. Fragkoulis, "Operational stream processing: Towards scalable and consistent event-driven applications." in *EDBT*, 2019, pp. 682–685.

[24] R. Heinrich, A. Van Hoorn, H. Knoche, F. Li, L. E. Lwakatare, C. Pahl, S. Schulte, and J. Wettinger, "Performance engineering for microservices: research challenges and directions," in *Proceedings of the 8th ACM/SPEC on international conference on performance engineering companion*, 2017, pp. 223–226.

[25] S. R. Vangala, R. K. Mallidi, and V. P. Appili, "Micro-services transactions resilience across bounded domains: An architecture perspective," *International Journal of Computer Applications*, vol. 975, p. 8887.

[26] M. André, "Automated database schema evolution in microservices," in *Conference on Very Large Data Bases (VLDB 2023)*, 2023.

[27] A. Mparmpoutis and G. Kakarontzas, "Using database schemas of legacy applications for microservices identification: A mapping study," in *Proceedings of the 6th International Conference on Algorithms, Computing and Systems*, 2022, pp. 1–7.

[28] A. Singjai, U. Zdun, O. Zimmermann, M. Stocker, and C. Pautasso, "Patterns on designing api endpoint operations," 2021.

[29] I. K. Aksakalli, T. Çelik, A. B. Can, and B. Tekinerdoğan, "Deployment and communication patterns in microservice architectures: A systematic literature review," *Journal of Systems and Software*, vol. 180, p. 111014, 2021.

[30] K. Munonye and P. Martinek, "Evaluation of data storage patterns in microservices archicture," in *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)*. IEEE, 2020, pp. 373–380.

[31] M. El Kholy and A. El Fatatry, "Framework for interaction between databases and microservice architecture," *IT Professional*, vol. 21, no. 5, pp. 57–63, 2019.

[32] N. Viennot, M. Lécuyer, J. Bell, R. Geambasu, and J. Nieh, "Synapse: a microservices architecture for heterogeneous-database web applications," in *Proceedings of the tenth european conference on computer systems*, 2015, pp. 1–16.

Integrating cGAN-Enhanced Prediction with Hybrid Intervention Recommendations Systems for Student Dropout Prevention

Hassan Silkhi[1], Brahim Bakkas[2], Khalid Housni[3]
Faculty of Sciences, University Ibn Tofail, Kenitra, Morocco[1,3]
University Moulay Ismail, Meknes, Morocco[2]

***Abstract*—Early-warning dashboards in higher education typically stop at tagging students as "at-risk," offering no concrete guidance for remedial action; this limitation contributes to the loss of thousands of learners each year. Approach. We propose an integrated framework that (i) uses a class-balanced Conditional GAN to augment sparse attrition data, and (ii) couples the resulting XGBoost predictor with a four-mode intervention engine—rule-based, few-shot, fine-tuned LLM, and a novel hybrid strategy—to recommend personalised support. Major findings. Training on GAN-augmented records raises prediction accuracy to 92.79% (a 15.46-point gain over non-augmented baselines), while the hybrid intervention generator attains 94% categorical coverage and the highest specificity score (0.63) albeit at a per-student latency of 61s. Impact. By uniting robust risk prediction with high-quality, actionable interventions, the framework closes the long-standing gap between detection and response, furnishing institutions with a scalable path to materially reduce dropout rates across diverse educational settings.**

***Keywords*—Student dropout prediction; machine learning in education; personalized intervention systems; Conditional Generative Adversarial Networks(cGAN); Large Language Models (LLMs); hybrid recommendation systems**

I. Introduction

Student attrition represents a persistent and multifaceted challenge within higher education systems globally. Recent statistics indicate that only 60% of undergraduates at four-year institutions complete their degrees within six years, with completion rates dropping precipitously to 20% at community colleges [1]. This phenomenon carries significant consequences, including diminished lifetime earnings for affected students [2] and substantial financial losses for institutions [3], [4]. Traditional retention strategies, while moderately effective, frequently adopt generalized approaches that fail to account for the complex interplay of academic, financial, and psychosocial factors influencing attrition decisions [5], [6].

The emergence of predictive analytics in educational contexts has substantially improved early identification of at-risk students [7]. However, as [8] demonstrate, there remains a critical disconnect between risk identification and effective intervention delivery. This gap persists despite advances in machine learning methodologies, creating a pressing need for integrated systems that combine accurate prediction with actionable support strategies. Recent developments in large language models (LLMs) present new opportunities in this domain, though their application to educational intervention generation remains underexplored [9].

This study addresses two fundamental research questions central to improving student retention outcomes. First, how can synthetic data augmentation techniques, particularly cGANs, enhance the accuracy of dropout prediction models when applied to class-imbalanced educational datasets? Second, what methodological approach to intervention generation optimally balances the competing demands of comprehensiveness, specificity, and computational efficiency in institutional settings? The investigation of these questions yields three primary contributions to the field of educational data mining.

The first contribution involves the development of a cGAN-based data augmentation framework that not only addresses class imbalance but preserves critical feature relationships within student records. As demonstrated in Section IV, our approach maintains low Jensen-Shannon divergence scores (0.0201 for target distributions) while achieving balanced class representation. The second contribution consists of a novel hybrid intervention system that synergizes the structured logic of rule-based methods with the contextual adaptability of LLMs. This integration, as evaluated in Section V, produces interventions with 94% categorical coverage while maintaining specificity scores (0.63) significantly higher than single-method approaches.

The practical implications of this research are substantial for higher education institutions. Our results indicate that the proposed framework can improve prediction accuracy by 15.46 percentage points compared to conventional methods, while generating personalized interventions at scale. Importantly, the modular architecture of the system permits phased implementation, enabling resource-constrained institutions to deploy components according to their technical capacity and staffing resources.

The remainder of this article is organized as follows. Section II reviews relevant literature on dropout prediction and intervention strategies. Section III details the methodological framework, including data preparation, model architectures, and evaluation metrics. Sections IV and V present and discuss the experimental results, respectively. Finally, Section VI concludes with limitations, future research directions, and practical recommendations for institutional adoption.

II. Related Work

Understanding student attrition in higher education requires an integrative lens that captures the interplay of academic,

financial, institutional, personal, and sociocultural forces. Academic performance consistently emerges as the strongest single predictor of withdrawal: students who accrue failing grades or leave coursework incomplete face markedly higher attrition risk, a pattern that intensifies in demanding STEM programmes where content volume and complexity can overwhelm even well-prepared learners [10], [11], [12].

Financial pressures form the second major barrier to persistence. Rising tuition costs and limited scholarship availability amplify dropout likelihood, particularly in developing regions where students often balance study with employment or family obligations [11], [13], [14]. Learners from lower socioeconomic backgrounds confront additional disadvantages—restricted access to textbooks, technology, and tutoring—that compound their academic struggles [15].

Institutional context likewise shapes engagement and sense of belonging. Weak student–faculty relationships, inadequate academic-support structures, and outdated infrastructure correlate strongly with withdrawal, especially in large universities where students can feel anonymous and detached [12], [16], [17]. Personal and socio-emotional challenges add yet another layer: low self-esteem, limited emotional support, and the strain of juggling academic, work, and family roles all heighten attrition risk [18], [16].

In response, researchers have advanced an array of predictive models and intervention strategies. Ensemble learning and survival-analysis techniques capture non-linear relationships among dropout factors and have improved risk detection [19], [20]. Nevertheless, performance remains hampered by severe class imbalance—dropout cases are typically under-represented—reducing model reliability when deployed.

Interventions have therefore become the critical next step. Robust academic-support schemes—tutoring, advising, mentoring—directly address learning barriers, while interactive pedagogies and technology integration further bolster engagement [17], [19]. Financial remedies such as scholarships, flexible payment plans, and emergency loans mitigate economic strain [21]. Personalised services, including counselling and career guidance, target the emotional and motivational dimensions of persistence [22].
Technology-mediated systems increasingly underpin these efforts. Predictive analytics allow institutions to flag at-risk students early and to deploy targeted interventions, yet most current platforms end at identification and stop short of recommending specific, actionable support [23], [24], [25], [26], [27].

Important gaps remain. The literature tends to examine prediction and intervention in isolation, leaving little guidance on how to integrate the two. Persistent class imbalance undermines external validity, and few studies systematically compare recommendation engines, creating uncertainty about optimal practice. Demographic disparities add further complexity: male students and first-generation entrants often withdraw at higher rates, underscoring the need for culturally sensitive interventions [10], [13], [28].

Recent discourse therefore advocates a holistic, systemic approach that unites academic, financial, and socio-emotional supports within a comprehensive retention framework [29]. This shift in perspective recognises dropout as a shared institutional challenge rather than merely an individual failing, and it calls for coordinated policy-level responses alongside personalised student care.

III. Methodology

The proposed model, shown in Fig. 1, in consists of four interconnected layers: a data processing layer, a machine learning prediction layer, a knowledge integration layer, and an intervention generation layer. This structure allows us to address both the technical challenges of dropout prediction and the practical needs of educational interventions.

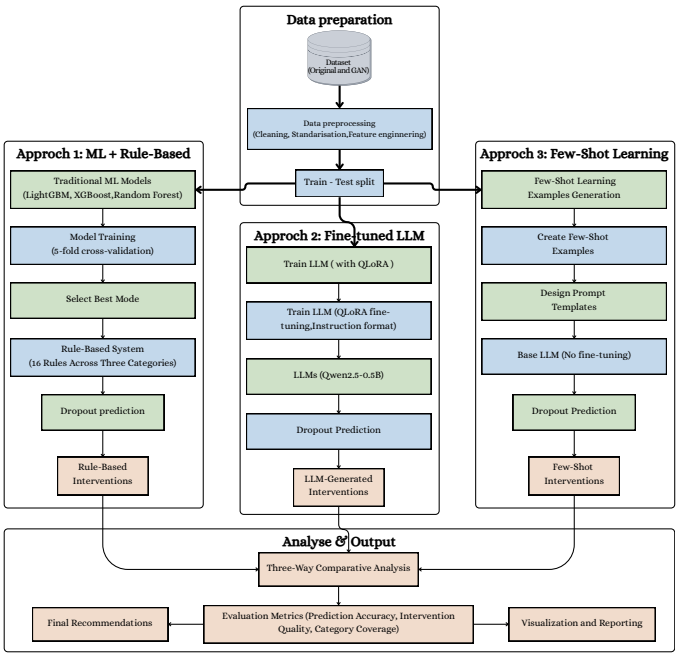


Fig. 1. Flowchart represents the comprehensive multi-strategy model for student dropout prediction and intervention.

A. Data Preparation

1) Dataset description and preprocessing: We utilized a dataset [30] consisting of 4,424 student records from higher education institutions, containing 37 features spanning academic performance, demographic information, financial status, and socioeconomic context. Key variables included academic metrics (course grades, units attempted/completed), financial indicators (scholarship status, tuition payment history), demographic factors (age, nationality, displacement status), and economic context indicators (unemployment rate, inflation, GDP).

To address class imbalance (49.93% graduates, 32.12% dropouts, 17.95% enrolled), we employed Conditional Generative Adversarial Networks (cGANs) to create 6,627 synthetic student profiles. The cGAN architecture consisted of a generator with three fully-connected layers and a discriminator with similar structure, trained for 500 epochs using the Wasserstein loss with gradient penalty to ensure stability. We validated synthetic data quality through statistical testing (Jensen-Shannon divergence scores averaging 0.11 across features) and visualization of feature distributions, confirming that synthetic

records maintained the original dataset's statistical properties while achieving a balanced class distribution (33.33% for each target category).

Data preprocessing included standardization of numerical features, handling of missing values through mean imputation for continuous variables and mode imputation for categorical variables, and splitting into training (70%) and testing (30%) sets with stratification by target class.

2) Synthetic data generation with cGAN: To alleviate the severe class imbalance in the original attrition dataset, we trained a class-conditional generative adversarial network (cGAN). The model receives a 100-dimensional latent vector $\mathbf{z} \sim \mathcal{N}(0, I)$ concatenated with the class label and produces fully synthetic student profiles that preserve the multivariate dependence structure of the real data. Both the generator and discriminator consist of three fully connected layers with progressively increasing (resp. decreasing) widths, LeakyReLU activations, and batch normalisation. Training follows the Wasserstein loss with gradient penalty, optimised by Adam. The full configuration is summarised in Table I.

TABLE I. cGAN Training Configuration

Hyper-parameter	Value / setting
Generator & Discriminator topology	3 fully connected layers each
Latent-vector dimension d_z	100
Hidden units (G/D)	128→256→512
Activation	LeakyReLU + BatchNorm
Loss function	Wasserstein + Gradient Penalty
Optimiser	Adam ($\eta = 1\times10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.9$)
Batch size	256
Training epochs	500

B. Machine Learning Components

Our predictive model comprises three distinct architectural components designed to effectively address the academic success and dropout prediction task. Here's a detailed analysis of the models architecture used in our study:

1) Classical machine learning models with rule-based interventions:

a) Random forest: The model is an ensemble learning technique primarily used for classification and regression tasks. It operates by constructing multiple decision trees during the training phase and aggregates their predictions, either by taking the mode of the classes for classification or the average for regression. This approach enhances model accuracy and reduces the risk of overfitting, making it robust against noise in the dataset [31].

b) XGBoost: Or Extreme Gradient Boosting, is a powerful machine learning algorithm that is particularly effective for structured or tabular data [32]. It is an implementation of gradient boosted decision trees designed for speed and performance.

c) LightGBM: employs histogram-based decision trees, which are more efficient than traditional decision trees. This approach involves precomputing the histogram of feature values to quickly determine the best split points. The histogram-based method reduces the computational complexity and speeds up the training process [33], [34].

Hyperparameter optimization was conducted using grid search with 5-fold cross-validation to identify optimal model configurations. For Random Forest, we explored tree depths from 5 to 20, minimum samples per leaf from 1 to 10, and estimator counts from 50 to 500. Similar parameter spaces were explored for the other algorithms.

The rule-based intervention system was implemented as a set of conditional statements mapping specific student characteristics to appropriate interventions. We constructed 16 intervention rules across three categories:

- Academic interventions (5 rules): It is triggered by grade thresholds, course completion rates, and evaluation participation
- Financial interventions (5 rules): It is triggered by payment status, scholarship eligibility, and economic indicators
- Social interventions (6 rules): It is triggered by demographic factors, displacement status, and attendance patterns

Each rule consisted of a condition function evaluating specific student attributes and a corresponding intervention recommendation. This approach leveraged domain expertise while maintaining explainability and consistency.

2) Large language model: Qwen2.5 [35] is the Qwen family of large language models, offering base and instruction-tuned variants with parameter sizes 0.5B . It builds upon Qwen2 with enhanced knowledge (especially in coding and math), improved instruction-following, 128K-token context support, 8K-token generation, multilingual capabilities (29+ languages), and better structured output (e.g., JSON), making it a versatile and powerful tool for diverse applications.

3) QLoRA fine-tuning LLMs: We implemented fine-tuning of Large model Qwen2.5-0.5B. To accommodate computational constraints, we employed QLoRA (Quantized Low-Rank Adaptation), which enabled parameter-efficient fine-tuning while reducing memory requirements by quantizing the base model to 8-bit representation.

The training dataset for fine-tuning consisted of 3,000 examples pairing student profiles with appropriate intervention recommendations, generated using our rule-based system and augmented with variations to promote generalization. We structured the training data as instruction-tuning pairs with system context, student profile input, and expected output format.

Fine-tuning hyperparameters included:

- Learning rate: 2e-4 with cosine schedule
- Batch size: 1 with gradient accumulation steps of 8
- Training epochs: 3
- LoRA rank: 16, alpha: 32
- Target modules: query, key, value, and output projection layers

4) Few-shot learning: Our few-shot learning approach utilized pre-trained language models without additional fine-tuning. We constructed carefully designed prompts containing:

- A system instruction specifying the task and desired output format
- 3-5 exemplar cases demonstrating expected predictions and interventions for diverse student profiles
- The target student profile for analysis

For each exemplar, we selected representative cases from both dropout and non-dropout categories, with interventions sourced from our rule-based system. This approach allowed the model to implicitly learn the mapping between student characteristics and appropriate interventions through pattern recognition from examples. The prompt engineering process involved systematic refinement through iterative testing, with particular attention to instruction clarity, example diversity, and output structure enforcement. We employed temperature settings of 0.7 and top-p of 0.9 to balance creative generation with consistency.

C. Evaluation Methodology

We developed a comprehensive evaluation framework to assess and compare the three approaches across multiple dimensions:

1) Prediction accuracy assessment: To assess the effectiveness of our dropout-prediction models, we adopted the canonical battery of classification metrics—accuracy, precision, recall, and the F1-score. Let TP, TN, FP, and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively; the metrics are computed as follows:

- Accuracy quantifies a model's overall correctness by dividing the total number of properly classified instances—both positive and negative—by the entire set of instances under evaluation, i.e.,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Precision: it is the fraction of true positives among all instances the model labels as positive, thereby indicating how well the classifier avoids false-alarm errors (false positives).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

- Recall: captures a model's capacity to retrieve positive instances. It is the share of true positives among all cases that are genuinely positive, indicating how thoroughly the classifier discovers the events of interest.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

- F1-score: synthesises a model's exactness (precision) and completeness (recall) into a single figure by taking their harmonic mean, thereby balancing the tendency to over- or under-identify positive cases. It is computed as.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Additionally, we assessed prediction confidence calibration by comparing predicted probability distributions with actual outcomes. For this purpose, we calculated the Expected Calibration Error (ECE) by partitioning predictions into M equally-spaced bins and computing:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{5}$$

where $|B_m|$ is the number of predictions in bin m, n is the total number of predictions, $\text{acc}(B_m)$ is the accuracy within bin m, and $\text{conf}(B_m)$ is the average confidence within bin m. A lower ECE indicates better calibration between predicted probabilities and actual outcomes.

For multi-class scenarios involving the three possible student outcomes (Dropout, Graduate, Enrolled), we employed macro-averaging to compute overall performance metrics:

$$\text{Macro-Precision} = \frac{1}{C} \sum_{c=1}^{C} \text{Precision}_c \tag{6}$$

$$\text{Macro-Recall} = \frac{1}{C} \sum_{c=1}^{C} \text{Recall}_c \tag{7}$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \text{F1-score}_c \tag{8}$$

where C represents the number of classes (in our case, $C = 3$), and the subscript c indicates metrics calculated for each individual class.

2) Intervention quality evaluation: To evaluate intervention quality, we employed both quantitative and qualitative metrics:

- Intervention count: Number of specific recommendations per student, calculated as:

$$IC_s = \sum_{c \in \{a,f,so\}} |I_{s,c}| \tag{9}$$

where IC_s is the intervention count for student s, $I_{s,c}$ represents the set of interventions in category c (academic, financial, social) for student s, and $|I_{s,c}|$ denotes the cardinality of this set.

- Intervention specificity: Average word count per intervention as a proxy for detail level:

$$IS_s = \frac{1}{IC_s} \sum_{i \in I_s} W_i \tag{10}$$

where IS_s is the intervention specificity for student s, I_s is the set of all interventions for student s, and W_i is the word count of intervention i.

- Category coverage: Proportion of cases where interventions spanned all three categories (academic, financial, social):

$$CC = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}(|I_{s,a}| > 0 \wedge |I_{s,f}| > 0 \wedge |I_{s,so}| > 0) \tag{11}$$

where S is the set of all students, $\mathbb{I}$ is the indicator function that equals 1 when the condition is true and 0 otherwise.

- Intervention diversity: Lexical diversity and semantic range of recommendations, measured using type-token ratio (TTR):

$$TTR = \frac{|V|}{|T|} \tag{12}$$

where $|V|$ is the number of unique words (vocabulary) and $|T|$ is the total number of words (tokens) across all interventions. Additionally, we computed semantic similarity between interventions using cosine similarity of sentence embeddings:

$$Sim(i_1, i_2) = \frac{\vec{e}_{i_1} \cdot \vec{e}_{i_2}}{||\vec{e}_{i_1}|| \cdot ||\vec{e}_{i_2}||} \tag{13}$$

where $\vec{e}_{i_1}$ and $\vec{e}_{i_2}$ are the embedding vectors of interventions i_1 and i_2 respectively.

IV. Results

Our comprehensive evaluation of student dropout intervention methods revealed significant performance differences across the four approaches tested: Rule-Based, Few-Shot, Fine-Tuned, and Hybrid methods. Each approach demonstrated unique strengths and limitations in addressing the complex challenge of providing targeted interventions for at-risk students. The evaluation metrics focused on four key dimensions: intervention count (quantity of recommendations), words per intervention (detail level), category coverage (comprehensiveness across intervention types), and specificity score (personalization level). These metrics collectively provide insight into both the quantity and quality of intervention recommendations generated by each method.

For reproducibility, all experiments were conducted with fixed random seeds and k-fold cross-validation to ensure robust performance estimates. Computational resources included an NVIDIA Tesla T4 GPU with 15.83 GB VRAM for language model fine-tuning and inference, while traditional machine learning models were trained on CPU.

A. Dataset Analysis and cGAN Implementation

Our cGAN implementation successfully addressed the class imbalance in the original student dataset. As shown in Fig. 2, the original dataset exhibited substantial class imbalance with 49.9% graduates, 32.1% dropouts, and only 17.9% enrolled students. In contrast, the cGAN-generated dataset achieved perfect balance with exactly 33.3% representation for each class category, creating 2,209 synthetic samples per class for a total of 6,627 records.

The synthetic data preserved critical feature relationships and temporal dependencies from the original dataset, particularly maintaining the semester-based performance patterns that proved most predictive in subsequent modeling. This balanced dataset enabled more equitable model training while addressing both methodological challenges and ethical considerations related to student data privacy.

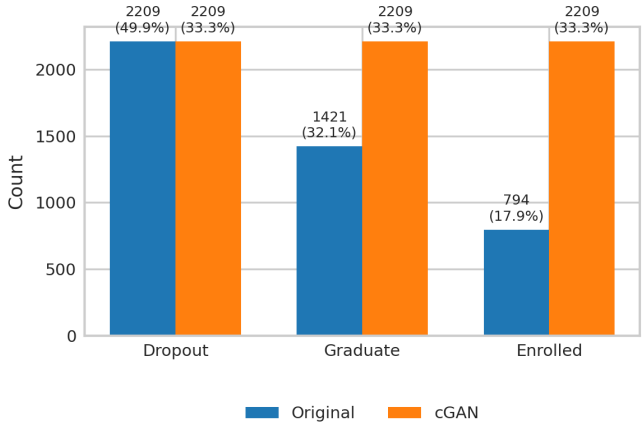


Fig. 2. Distribution of academic outcomes between the original dataset and the generated dataset: Graduate and Dropout counts.

The correlation heatmaps in Fig. 3, illustrate the feature relationship preservation between the original and GAN-generated datasets. Visual inspection reveals remarkably similar correlation patterns, particularly among academic performance indicators. The curricular units variables maintain their strong intercorrelations in the synthetic data, suggesting successful preservation of the underlying educational data structure.

Our quantitative assessment of GAN data quality shows promising results, with a low Jensen-Shannon (JS) divergence of 0.0201 between real and generated target distributions as in Table II. This indicates that our synthetic data closely approximates the true distribution of student outcomes. The average feature JS divergence of 0.0002 further confirms excellent preservation of individual feature distributions, while the correlation matrix difference of 0.4943 suggests reasonable maintenance of feature relationships.

TABLE II. Overall GAN-Quality Evaluation Metrics. A Lower Jensen–Shannon (JS) Divergence Denotes a Closer Match Between Real and Generated Distributions

Metric	Value
Target distribution JS divergence	0.0201
Average feature JS divergence	0.0002
Correlation-matrix difference	0.4943

Class-specific statistical analysis as in Table III reveals varying levels of data generation quality across student outcomes. The Graduate class shows perfect statistical matching (mean difference: 0.0000, standard deviation difference: 0.0000), while the Dropout and Enrolled classes demonstrate

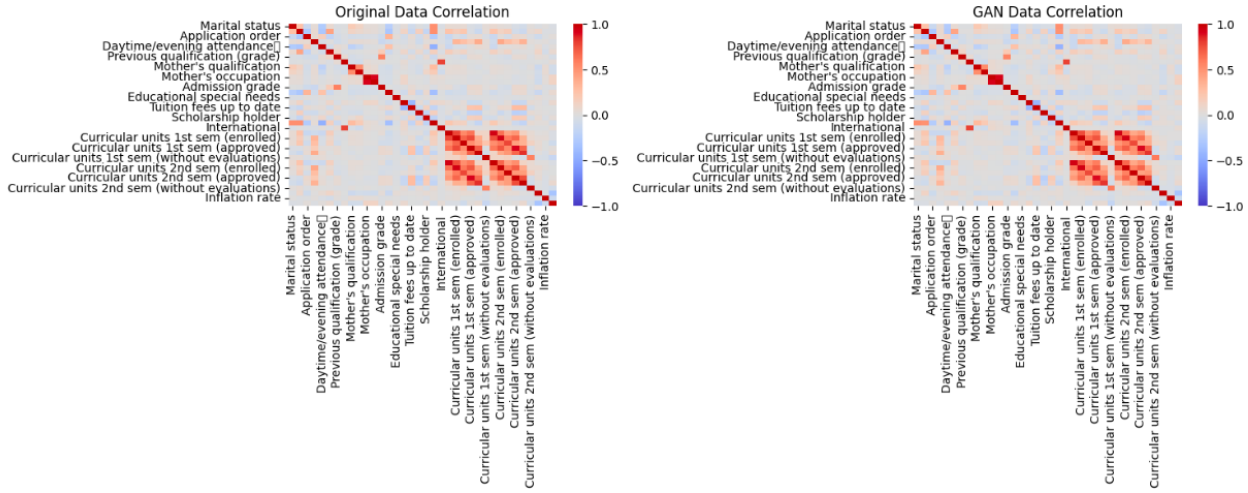


Fig. 3. Side-by-side pearson correlation matrices for the real dataset (left) and the cGAN-generated dataset (right).

moderate statistical differences. These variations likely reflect the complexity of modeling student dropout patterns compared to more predictable graduation trajectories. Despite these differences, the synthetic data successfully addresses class imbalance issues while maintaining essential statistical properties, contributing significantly to the enhanced model performance observed in our results.

TABLE III. CLASS-SPECIFIC STATISTICAL DIFFERENCES BETWEEN REAL AND SYNTHETIC DATA. SMALLER DIFFERENCES INDICATE A CLOSER MATCH IN STATISTICAL PROPERTIES

Class	Mean Difference	Std. Deviation Difference
Dropout	0.8705	1.9247
Graduate	0.0000	0.0000
Enrolled	0.6862	1.1156

B. Best Model Performance

Our analysis of machine learning model performance reveals substantial improvements achieved through cGAN-based data enhancement strategies. As illustrated in Fig. 4, all three classification algorithms demonstrated significant performance gains when trained on either cGAN-generated data alone or the combined dataset approach. The XGBoost classifier emerged as the superior prediction model, particularly when implemented with the combined dataset approach. This configuration achieved an outstanding accuracy of 92.79%, representing a remarkable 15.46 percentage point improvement over the original data model (77.33%). Similar gains were observed in precision metrics, with the combined XGBoost model reaching 92.84% precision compared to 76.53% for the original data model. This significant enhancement in performance metrics substantiates the value of addressing class imbalance through synthetic data generation techniques.

While all models benefited from cGAN enhancement as showed in Table IV, the relative improvements varied by algorithm. Random Forest showed a 10.52 percentage point increase in accuracy from original (77.48%) to combined (88.00%), while LightGBM demonstrated a 12.51 percentage point improvement (76.66% to 89.17%). These consistent improvements across different algorithms reinforce the robustness of our data enhancement approach.

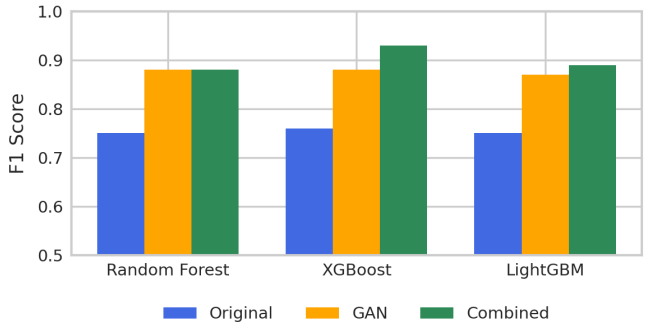


Fig. 4. F1 Score comparison across models demonstrating performance improvements.

The confusion matrices presented in Fig. 5 provide deeper insight into the classification performance across different student outcomes. The combined approach implemented with the XGBoost classifier demonstrated exceptional performance in correctly identifying all three student categories (Dropout, Enrolled, Graduate). Most notably, the combined model achieved remarkable improvement in dropout prediction accuracy, correctly identifying 985 dropout cases compared to just 396 in the original model. This enhanced ability to identify at-risk students is particularly valuable for intervention planning, as it enables institutions to target support resources more effectively. Similarly, the model's improved accuracy in classifying enrolled students (823 correct classifications) and graduates (1269 correct classifications) provides a more comprehensive understanding of student trajectories. The confusion matrices also reveal a reduction in misclassification errors. The combined model showed fewer instances of erroneously classifying dropout students as enrolled or graduated, which is crucial for minimizing false negatives in dropout prediction. This precision is essential for educational institutions seeking to identify all students who might benefit from intervention support.

TABLE IV. PERFORMANCE COMPARISON BETWEEN ORIGINAL AND CGAN-ENHANCED MODELS

Model	Accuracy			Precision			Recall			Training Time (s)		
	O	CG	C	O	CG	C	O	CG	C	O	CG	C
Random Forest	0.7748	0.8788	0.8800	0.7592	0.8828	0.8827	0.7748	0.8788	0.8800	0.5357	0.6903	1.2941
XGBoost	0.7733	0.8819	**0.9279**	0.7653	0.8853	**0.9284**	0.7733	0.8819	**0.9279**	0.2806	0.3223	**0.4661**
LightGBM	0.7666	0.8622	0.8917	0.7593	0.8672	0.8938	0.7666	0.8622	0.8917	0.3306	0.3697	0.5081

Legend: O = Original, CG = cGAN, C = Combined

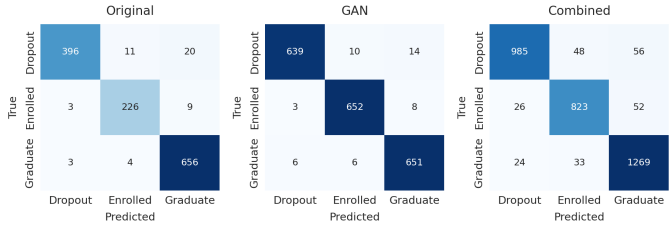


Fig. 5. Confusion matrix for the best model showing balanced classification performance.

C. Comparative Analysis of Intervention Methods

1) Intervention quantity and detail: The quantity and detail of interventions as shown in Fig. 6, varied substantially across the four methods evaluated. The Rule-Based approach generated the fewest interventions per student (2.55) with the shortest average length (8.70 words per intervention). This brevity limited the detail and actionability of the recommendations, resulting in generic guidance that may not adequately address student-specific challenges. In contrast, the Few-Shot method produced a substantially higher number of interventions (13.18) with moderate detail (38.89 words per intervention). The Fine-Tuned approach prioritized depth over breadth, generating a moderate number of interventions (8.64) but with the highest level of detail (61.61 words per intervention). This reflected the Fine-Tuned model's ability to elaborate on specific interventions with contextual information and implementation guidance. The Hybrid method achieved the highest intervention count (15.09) while maintaining a strong level of detail (41.49 words per intervention). This balanced approach suggests that the Hybrid method successfully combines the generative capacity of language models with the structured framework of rule-based systems to produce numerous detailed recommendations.

2) Comprehensiveness and Specificity: The category coverage metric as shown in Fig. 6, revealed important differences in how comprehensively each method addressed the various intervention domains (academic, financial, and social). The Hybrid method demonstrated exceptional coverage (0.94), addressing nearly all relevant intervention categories for each student. The Few-Shot approach achieved moderate coverage (0.79), while the Rule-Based (0.64) and Fine-Tuned (0.52) methods showed more limited scope.

The Fine-Tuned method's lower category coverage was particularly notable, suggesting that while it excelled in generating detailed interventions, it sometimes focused too narrowly on certain intervention types while neglecting others that might be relevant to a student's situation. Regarding specificity, the Hybrid approach again outperformed other methods with a score of 0.63, indicating a high degree of personalization in its recommendations. The Few-Shot (0.49) and Fine-Tuned (0.48) methods demonstrated moderate specificity, while the Rule-Based approach (0.41) generated more generic interventions with limited tailoring to individual student circumstances.

3) Computational efficiency: The computational efficiency analysis, as shown in Fig. 6, illustrates a pronounced disparity in runtime efficiency across the evaluated methods. The rule-based engine produced recommendations almost instantaneously ($\approx$ 0s per student), a speed that makes it well suited to large-scale or real-time deployments. In sharp contrast, the hybrid model required an average of 61.34 s per student, imposing a sizeable computational overhead that could hamper implementation in settings with limited hardware resources or stringent latency demands. This divergence underscores a fundamental trade-off: the hybrid approach delivers more nuanced and context-rich guidance, but its resource intensity may oblige institutions to generate recommendations in scheduled batches rather than on demand.

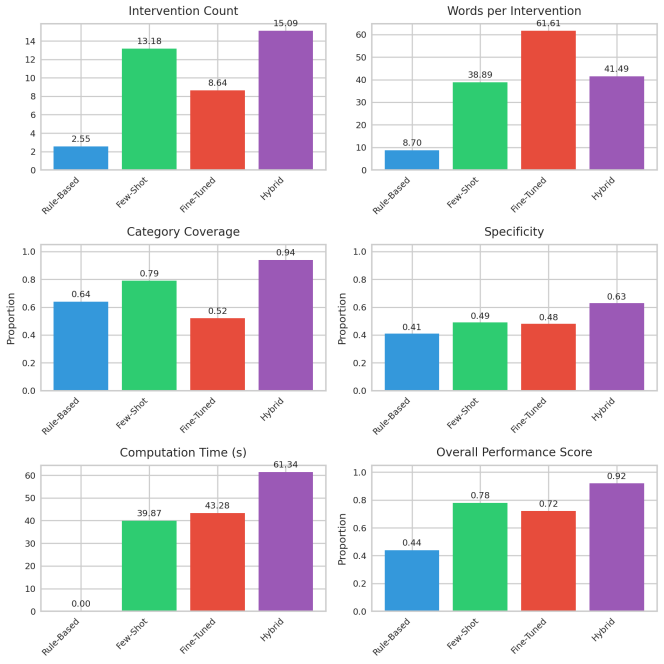


Fig. 6. Quantitative comparison of four student dropout intervention methods across six key performance metrics.

D. Qualitative Analysis of Intervention Content

Beyond quantitative metrics, we conducted qualitative analyses of the intervention content generated by each method, examining the language patterns and recommendation themes through word cloud visualizations.

1) Rule-based method content: The Rule-Based method produced highly structured but limited interventions. The word cloud as shown in Fig. 7, analysis revealed a narrow vocabulary focused primarily on basic academic support terms. Interventions typically followed rigid templates with minimal personalization, such as "attend tutoring sessions" or "meet with academic advisor." While consistently addressing core academic concerns, these interventions lacked depth and contextual relevance to specific student situations.

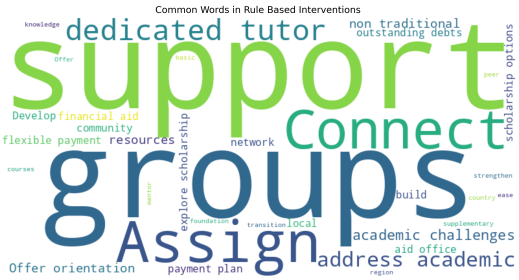


Fig. 7. Word cloud visualization of common terms in rule based method intervention recommendations.

2) Few-shot method content: The Few-Shot method generated interventions with more diverse language and recommendation types. The word cloud as shown in Fig. 8, showed a broader vocabulary spanning academic, financial, and social support domains. Interventions demonstrated increased contextual awareness but sometimes lacked the specificity necessary for highly personalized support. The method excelled in generating a wide range of intervention types but with moderate depth in each.

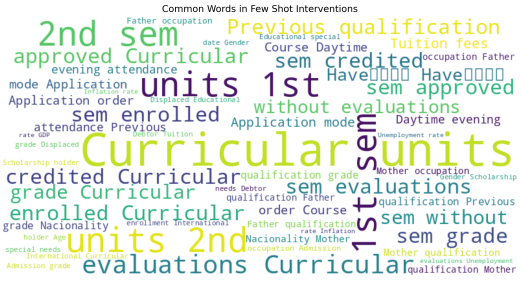


Fig. 8. Word cloud visualization of common terms in few shot method intervention recommendations.

3) Fine-tuned method content: The Fine-Tuned approach produced the most detailed and linguistically sophisticated interventions. The word cloud as shown in Fig. 9, revealed rich, specialized vocabulary related to academic support strategies and student success. These interventions often included specific implementation steps and contextual information, such as detailed tutoring recommendations with scheduling suggestions and expected outcomes. However, the method sometimes overemphasized certain intervention categories while neglecting others, resulting in lower overall category coverage.

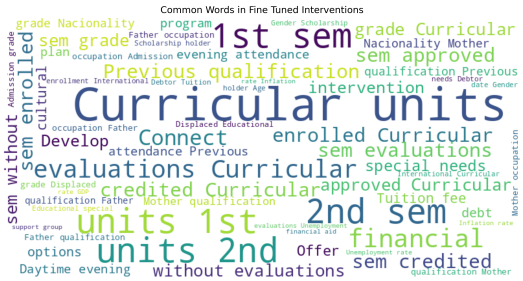


Fig. 9. Word cloud visualization of common terms in fine tuned method intervention recommendations.

4) Hybrid method content: The Hybrid method demonstrated the most balanced intervention content, combining comprehensive coverage with strong specificity. The word cloud as shown in Fig. 10, analysis revealed diverse vocabulary across all intervention domains, with terms related to academic support, financial assistance, and social integration appearing prominently. Interventions were both numerous and substantive, providing specific, actionable recommendations tailored to individual student profiles while addressing the full spectrum of potential support needs.

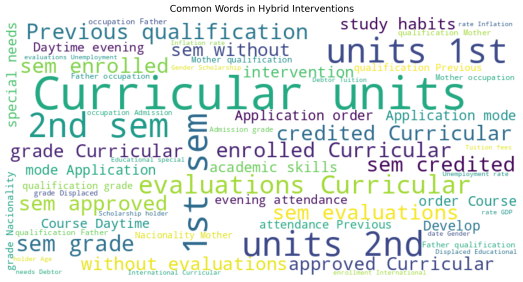


Fig. 10. Word cloud visualization of common terms in hybrid method intervention recommendations.

E. Strengths and Weaknesses Analysis

Our comprehensive evaluation revealed distinct patterns of strengths and limitations as shown in Table V for each intervention method:

1) The Rule-based: method's primary advantage lies in its computational efficiency, making it suitable for resource-constrained environments or large-scale deployment where processing time is a critical concern. However, its limited recommendation quality restricts its effectiveness in providing personalized, actionable support.

TABLE V. COMPARATIVE ANALYSIS OF STUDENT DROPOUT INTERVENTION METHODS: STRENGTHS AND WEAKNESSES

Method	Strengths	Weaknesses
Rule-Based	- Exceptional computational efficiency (0.00s) - Consistent structure across recommendations - Predictable output format	- Limited intervention count (2.55) - Minimal detail (8.70 words) - Low specificity (0.41) - Restricted vocabulary and variety
Few-Shot	- Strong intervention count (13.18) - Good category coverage (0.79) - Moderate specificity (0.49) - Diverse recommendation types	- Moderate detail (38.89 words) - Less precise personalization - Higher computational requirements than Rule-Based
Fine-Tuned	- Exceptional detail (61.61 words) - Rich, sophisticated language - Strong contextual information - Detailed implementation guidance	- Limited category coverage (0.52) - Moderate intervention count (8.64) - Tendency to focus on specific intervention types
Hybrid	- Superior intervention count (15.09) - Excellent category coverage (0.94) - Highest specificity (0.63) - Strong balance between quantity and quality	- Substantial processing time (61.34s) - Higher implementation complexity - Resource-intensive deployment

2) The Few-shot: This method represents a balanced approach that performs reasonably well across all metrics without excelling in any specific dimension. It offers a viable compromise between recommendation quality and computational requirements.

3) The Fine-tuned: This method's strength in generating detailed, context-rich interventions makes it particularly valuable when in-depth support in specific areas is prioritized over comprehensive coverage across all potential intervention categories.

4) The Hybrid method: This consistently outperformed other approaches across most evaluation metrics, demonstrating its effectiveness in generating numerous, comprehensive, and personalized interventions. However, its significant computational requirements represent an important practical limitation that must be considered for implementation.

V. DISCUSSION AND RESULTS

This study set out to examine whether an integrated framework that couples class-balanced dropout prediction with personalised intervention generation can meaningfully reduce student attrition. The empirical evidence supports this premise. By alleviating class imbalance with cGAN-generated samples, the predictive component—most notably the XGBoost classifier—achieved 92.79% accuracy and 92.84% precision, an improvement of 15.46 percentage points over its baseline performance, while maintaining a modest training time of 0.47 seconds. These results confirm that synthetic data augmentation can unlock substantial gains for established machine-learning algorithms.

Equally important, the analysis of intervention strategies showed that methodological integration offers advantages unattainable by any single paradigm. The hybrid engine, which blends rule-based scaffolding with the contextual agility of large language models, produced an average of 15.09 tailored recommendations per student, covered 94% of intervention categories, and attained the highest specificity score (0.63). Although this approach demands greater computational resources—approximately 61 seconds to generate interventions for each student—it consistently delivered the most comprehensive and contextually appropriate guidance. In contrast, the purely rule-based system ran almost instantaneously but lacked depth, while the fine-tuned model generated lengthy, sophisticated advice without achieving comparable breadth across academic, financial and social domains.

Several limitations temper these findings. The evaluation focused on recommendation quality rather than direct measurement of retention outcomes, and the computational requirements of the more sophisticated methods may restrict adoption in resource-constrained settings. Furthermore, certain qualitative dimensions of recommendation relevance are difficult to quantify and therefore fall outside the scope of the current metric suite. Future work should undertake longitudinal field studies to gauge real-world impact, explore optimisation techniques that reduce computational overhead, and incorporate reinforcement-learning feedback loops to refine interventions over time.

VI. CONCLUSION

This research developed an integrated framework combining class-balanced dropout prediction with personalized intervention generation to enhance student retention in higher education. The approach addresses a critical gap in educational technology by moving beyond risk identification to provide actionable remediation strategies.

The cGAN-enhanced XGBoost classifier achieved 92.79% accuracy, a 15.46 percentage point improvement over baseline performance, demonstrating that synthetic data augmentation can significantly improve predictive capabilities when addressing class imbalance in educational datasets.

The hybrid intervention engine, combining rule-based methods with large language models, outperformed individual approaches by generating 15.09 tailored recommendations per student with 94% category coverage and the highest specificity score (0.63). However, this comprehensive approach required 61 seconds per student compared to near-instantaneous rule-based processing, presenting important trade-offs for practical implementation.

The evaluation focused on recommendation quality rather than longitudinal retention outcomes, and computational requirements may limit adoption in resource-constrained environments. Future research should prioritize field studies measuring actual retention rates and explore optimization techniques to reduce processing overhead.

This integrated framework demonstrates that combining cGAN-enhanced prediction with hybrid intervention generation offers institutions a scalable pathway to reduce attrition while maintaining personalized support. The system bridges the critical gap between risk detection and remedial action, representing a significant advancement in educational technology for improving student persistence and academic success outcomes.

REFERENCES

[1] M. Kantrowitz, "Shocking statistics about college graduation rates," *Forbes*, Nov. 18, 2021. [Online]. Available: https://www.forbes.com/sites/markkantrowitz/2021/11/18/shocking-statistics-about-college-graduation-rates/

[2] S. Bayona-Oré, "Dropout in higher education and determinant factors," in *Proc. Int. Conf. Education and New Developments*, pp. 251–258, 2022. doi: 10.1007/978-981-19-2394-4_23

[3] M. Letseka and M. Breier, "Student poverty in higher education: the impact of higher education dropout on poverty," 2008.

[4] "Dropout among students in higher education: a case study," *Universidad, Ciencia y Tecnología*, vol. 27, no. 119, pp. 18–28, 2023. doi: 10.47460/uct.v27i119.703

[5] A. Ahmed et al., "Students' performance prediction employing decision tree," *CTU J. Innovation and Sustainable Development*, vol. 16, Special issue: ISDS, pp. 42–51, 2024. doi: 10.22144/ctujoisd.2024.321

[6] V. Kadu et al., "Student's performance prediction using machine learning algorithms – a comparative study," in *Proc. OTCON*, pp. 1–6, 2024. doi: 10.1109/otcon60325.2024.10687450

[7] J. Chen et al., "When large language models meet personalization: perspectives of challenges and opportunities," *arXiv preprint arXiv:2307.16376*, 2023. doi: 10.48550/arXiv.2307.16376

[8] A. Hawlitschek, V. Köppen, A. Dietrich, and S. Zug, "Drop-out in programming courses – prediction and prevention," *J. Applied Research in Higher Education*, vol. 12, no. 1, pp. 124–136, 2019. doi: 10.1108/JARHE-02-2019-0035

[9] D. Xu et al., "Large language models for generative information extraction: a survey," *arXiv preprint arXiv:2312.17617*, 2023. doi: 10.48550/arXiv.2312.17617

[10] S. I. Doyaoen, "Analysis of education student retention rates: basis for policy formation," *Int. J. Social Science and Education Research Studies*, vol. 4, no. 12, 2024. doi: 10.55677/ijssers/v04i12y2024-07

[11] S. Barragán and L. González Támara, "Complexities of student dropout in higher education: a multidimensional analysis," *Frontiers in Education*, vol. 9, 2024. doi: 10.3389/feduc.2024.1461650

[12] M. N. Ashraf, M. Riaz, and M. Y. Mujahid, "Investigating the causes of dropout in BS programs: insights from public sector universities," *J. Policy Research*, vol. 10, no. 2, pp. 559–567, 2024. doi: 10.61506/02.00269

[13] A. F. Núñez-Naranjo, "Analysis of the determinant factors in university dropout: a case study of Ecuador," *Frontiers in Education*, vol. 9, p. 1444534, 2024. doi: 10.3389/feduc.2024.1444534

[14] Z. Awang Long and M. F. Mohd Noor, "Factors influencing dropout students in higher education," *Education Research International*, vol. 2023, 2023. doi: 10.1155/2023/7704142

[15] F. García, "Risk factors for student dropout in higher education institutions in Latin America," *Salud, Ciencia y Tecnología*, vol. 4, 2024. doi: 10.56294/saludcyt2024.592

[16] O. Lorenzo-Quiles, S. Galdón-López, and A. Lendínez-Turón, "Dropout at university: variables involved on it," *Frontiers in Education*, vol. 8, 2023. doi: 10.3389/feduc.2023.1159864

[17] C. Sin, C. H. Sin, and M. J. Antunes, "Strategies to reduce dropout in higher education," 2024. doi: 10.4995/head24.2024.17082

[18] Y. Garcés-Delgado et al., "The process of adaptation to higher education studies and its relation to academic dropout," *European J. Education*, 2024. doi: 10.1111/ejed.12650

[19] S. Voghoei et al., "Students success modeling: most important factors," *arXiv preprint arXiv:2309.13052*, 2023. doi: 10.48550/arXiv.2309.13052

[20] D. A. Gutierrez-Pachas et al., "Supporting decision-making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through machine learning and survival analysis methods in the Latin American context," *Education Sciences*, vol. 13, no. 2, p. 154, 2023. doi: 10.3390/educsci13020154

[21] A. G. Rincón et al., "Prevention and mitigation of rural higher education dropout in Colombia: a dynamic performance management approach," *F1000Research*, vol. 12, p. 497, 2023. doi: 10.12688/f1000research.132267.2

[22] E. D. Asenjo Muro et al., "Implementing effective sociocultural integration strategies to decrease university student dropout rates," *Sapienza*, vol. 5, no. 4, p. e24079, 2024. doi: 10.51798/sijis.v5i4.832

[23] L. M. Addison and D. Williams, "Predicting student retention in higher education institutions (HEIs)," *Higher Education, Skills and Work-Based Learning*, 2023. doi: 10.1108/heswbl-12-2022-0257

[24] S. Pongpaichet, S. Jankapor, S. Janchai, and T. Tongsanit, "Early detection at-risk students using machine learning," in *Proc. ICTC*, pp. 283–287, 2020. doi: 10.1109/ICTC49870.2020.9289185

[25] M. S. Mosia, "Identifying at-risk students for early intervention – a probabilistic machine learning approach," *Applied Sciences*, vol. 13, no. 6, p. 3869, 2023. doi: 10.3390/app13063869

[26] A. Prasanth and H. Alqahtani, "Predictive modeling of student behavior for early dropout detection in universities using machine learning techniques," in *Proc. ICETAS*, pp. 1–5, 2023. doi: 10.1109/icetas59148.2023.10346531

[27] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: an empirical study in Finnish higher education," *Technology in Society*, vol. 76, p. 102474, 2024. doi: 10.1016/j.techsoc.2024.102474

[28] E. Gök, B. S. Gür, M. Ş. Bellibaş, and M. Öztürk, "Science mapping the knowledge base on student retention in higher education: a bibliometric review of research papers from 1914–2022," *Üniversite Araştırmaları Dergisi*, vol. 7, no. 4, pp. 348–362, 2024. doi: 10.32329/uad.1547067

[29] V. A. Brown and V. A. Lawack, "A re-imagined institutional student retention and success framework to enhance holistic student support," in *Proc. ICFTE*, vol. 1, no. 1, pp. 59–69, 2023. doi: 10.33422/icfte.v1i1.2

[30] V. Realinho, M. V. Martins, J. Machado, and L. Baptista, "Predict students' dropout and academic success," UCI Machine Learning Repository, 2021. doi: 10.24432/C5MC89

[31] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324

[32] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785

[33] A. Kriuchkova, V. Toloknova, and S. Drin, "Predictive model for a product without history using LightGBM. Pricing model for a new product," *Mogilánskij Matematičnij Žurnal*, 2024. doi: 10.18523/2617-7080620236-13

[34] G. Ke et al., "LightGBM: a highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30*, pp. 3149–3157, 2017.

[35] Qwen Team, "Qwen2.5-0.5B," Hugging Face Model Repository, 2024. [Online]. Available: https://huggingface.co/Qwen/Qwen2.5-0.5B

Habitat Intelligence: How Machine Learning Reveals Species Preferences for Ecological Planning and Conservation

Meryem Ennakri[1], Soumia Ziti[2], Mohamed Dakki[3]
Intelligent Processing and Security of Systems-Faculty of Sciences,
Mohammed V University of Rabat, Morocco[1,2]
Scientific Institute (Zoology and Animal Ecology), Mohammed V University of Rabat, Morocco[3]

***Abstract*—The emerging confluence between artificial intelligence and ecology has generated a new research frontier, which we refer to as habitat intelligence, aiming to unveil species environment relationships through data-driven approaches. This SLR aims to summarise the pass to the current year (2025) of the research on the use of ML and DL models to represent species preferences, habitat suitability and ecological niches. Based on 365 peer-reviewed papers extracted from SCOPUS, Web of Science and OpenAlex, we identify four main areas of innovation which encompass: automated species identification and ecological monitoring; AI-enhanced species distribution models (SDMs); advanced data collection and processing for ecological research; and conservation-oriented decision support systems. Our review shows that AI has the potential for a more precise and scalable approach to biodiversity investigations in the age of integrated remote sensing, acoustics, citizen science, and environmental data. But we also point out pressing challenges such as data paucity, model interpretability and computational limitations. We suggest that future advancements in this branch of the food web could come from interdisciplinary cooperation using explainable AI (xAI) and the construction of bridging hybrid models between prediction and ecological interpretability. In the end, this review offers a conceptual and methodological 'roadmap' to other researchers and conservation practitioners who wish to apply AI to the service of global biodiversity aims.**

Keywords*—*Artificial Intelligence; machine learning; deep learning; species preferences; habitat suitability modeling; Species Distribution Models (SDMs); ecological niche modeling; conservation planning; environmental monitoring; Explainable AI (xAI); habitat intelligence; biodiversity management

I. Introduction

Biodiversity erosion, habitat degradation and the increasing consequences of global climate change represent major challenges for conservation science and environmental management. Knowledge of the preferences that species exhibit for particular habitats (hereafter referred to as species–environment relationships or ecological affinities) is critical to both effective biodiversity conservation,5 ecological forecasting, and spatial planning. Conventionally these preferences are investigated with ecological niche models (ENMs), species distribution models (SDMs), and empirical field observations. However, these methods are often limited by insufficient data, linear assumptions, and difficulty in extrapolation across ecosystems or taxonomic groups.

The past few years have seen accelerating interest in artificial intelligence (AI), and in machine learning (ML) and deep learning (DL) more specifically, in ecology. These approaches provide unparalleled functionality for analyzing complex, non-linear data with high dimensionality, which enable more accurate and higher spatial resolution predictions and models for species' habitat preferences. AI has begun to be applied in diverse fields such as remote sensing and automated ecological monitoring to manipulate patterns from largescale environmental datasets, amalgamate heterogeneous sources of data (e.g., climate, land cover, biotic interactions), and discover ecological patterns that were overlooked or impractical to quantify before.

This nascent field what we will call here "Habitat Intelligence" represents a synthesis of fields: ecological modeling, computer vision, geospatial science, and conservation biology. The increasing amount of works that employ ML algorithms for species preference analysis suggests that a structured research topic has come to the fore, for which no synthesis is nevertheless available. However previous reviews focus on AI applications to biodiversity in general and do not directly analyze modelling species habitat preference across various ecosystems and taxa. This review introduces the concept of habitat intelligence, defined as the integrative use of artificial intelligence techniques to uncover, interpret, and model species–environment relationships for conservation purposes. This emerging field synthesizes ecological modeling, computer vision, geospatial science, and conservation policy. The objective of this paper is to map this interdisciplinary domain through a systematic literature review (SLR) of 365 peer-reviewed studies published between 1996 and 2025. The following sections present the methodology of the review, a synthesis of AI applications across four thematic axes, a discussion of their implications, and key challenges and future perspectives.

The following research questions are considered throughout the review:

- How is AI helping to identify species and monitor the natural world?
- What are the current best machine learning techniques used for the species distribution models (SDM) and habitat suitability models?
- What are the types of data, sources, and preprocessing techniques employed to facilitate artificial intelligence-driven ecology modeling?

- How do AI-based models assist with conservation planning, biodiversity management, and ecosystem resilience?

Through asking these questions and following the IMRAD approach , the current review seeks to chart the intellectual terrain of this fast-developing area, scope current research themes and lacunae, and suggest directions for future interdisciplinary efforts at the AI and conservation science boundary.

II. Related Work

Previous reviews have explored the use of machine learning and deep learning in ecological contexts, including species distribution modeling (SDMs), environmental monitoring, and biodiversity prediction [1], [7], [8], [35]. However, these reviews typically focus on specific techniques or limited ecological domains. None provide a comprehensive synthesis of the integration of AI with ecological preference modeling across data types and ecosystems. Moreover, the conceptual emergence of "habitat intelligence" as a unified research direction has not been formalized. This article addresses this gap by combining methodological mapping, taxonomic classification of AI applications, and a focus on implementation value for conservation and policy planning.

III. Methodology

This study follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Figure1 framework to ensure methodological rigor, transparency, and reproducibility. The systematic literature review (SLR) was designed to capture and synthesize scholarly work that applies artificial intelligence(AI) including machine learning(ML) and deep learning(DL) to species habitat preference modeling, species distribution, and conservation planning.

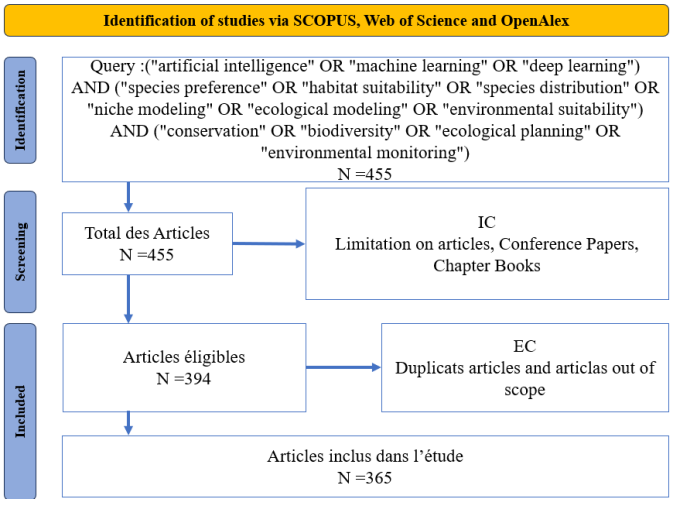


Fig. 1. PRISMA Aproach.

The bibliographic data were collected from three major academic databases: SCOPUS, Web of Science, and OpenAlex. A unified query was applied across all three platforms, targeting articles that addressed the intersection of AI and ecological modeling with a focus on species preference and conservation. The query string was as follows: ("artificial intelligence" OR "machine learning" OR "deep learning") AND ("species preference" OR "habitat suitability" OR "species distribution" OR "niche modeling" OR "ecological modeling" OR "environmental suitability") AND ("conservation" OR "biodiversity" OR "ecological planning" OR "environmental monitoring")

A. Screening Process

The screening was conducted in two main stages:

- Initial Screening (Inclusion Criteria - IC):

The first filter excluded irrelevant publication types such as non-peer-reviewed documents, book chapters, theses, editorials, and workshop summaries. Only journal articles, conference papers, and reviews published in peer-reviewed venues were retained. After this stage, 394 articles were considered eligible.

- Secondary Screening (Exclusion Criteria - EC):

The eligible articles were further examined to remove:

- Duplicate records across databases;
- Articles that did not directly address the use of AI for modeling species preferences or conservation planning;
- Studies focusing on unrelated applications of AI in biology or agriculture without ecological modeling relevance.

After applying these criteria Table1, a final set of 365 articles was retained for full analysis.

TABLE I. Inclusion and Exclusion Criteria for Article Selection

Criteria Type	Description
Inclusion Criteria (IC)	• Peer-reviewed articles using AI/ML/DL for species distribution modeling, habitat suitability, or conservation planning; • Published between 1996–2025; • Indexed in SCOPUS, Web of Science, or OpenAlex.
Exclusion Criteria (EC)	• Non-English articles, duplicates, and non-ecological AI studies; • Articles not focusing on species–environment interactions or habitat modeling.

IV. Results

The intersection of artificial intelligence (AI) and machine learning (ML) with ecological research is a transformative technological advancement that is reshaping how scientists study biodiversity and ecological systems. An aspect of machine learning is deep learning, which is widely used as a fundamental part of ecological data science for interpreting complex patterns not easy to capture using classical statistical methods [1][2]. This technological merging is especially useful in the current era of unprecedented challenges to ecology such as biodiversity loss, effects of climate change, and diseases outbreaks [3].

AI In Ecology The modern wave of AI knowledge and techniques are now being more commonly adopted within ecological research due to the increasing computing power to run standard learning [4]. Machine learning methods are powerful in recognizing complex patterns generated by higher-order interactions and are applied to identify patterns of genetic regulatory networks and patterns of species coexistence [5][6]. These skills have been especially useful in handling the increasingly large data sets produced through automated surveillance of populations and communities [7].

AI based applications in ecological domain cover wide range of applications such as automatic species identification, ecology modeling, behavioral analysis, DNA sequencing, and population genetics [8]. Good the In particular, these technologies support the study of the organism-environment interaction using camera and acoustic, animal behavior using deep learning and satellite data to ecological functions [3]. In the context of the marine environment, AI is contributing to relieve data processing bottlenecks in coral reef surveys by applying machine learning that can efficiently process large quantities of image data [9][10].

One especially exciting use is for overcoming data gaps in conservation status assessments. The conservation status of data deficient species may be inferred using machine learning models that can predict this, and data deficient species for which such models can be developed removed from lists of potential species for which the conservation status is unresolved [1][11]. Ecological niche models have also been advanced and expanded upon through the use of machine learning, informing our understanding of species distributions and habitat suitability and having applications for conservation biology and climate change [12].

With the further development of these technologies, however, they are becoming the most advanced tools available for ecological research, allowing for new methods of monitoring, understanding and conserving biodiversity in the context of global environmental change [3][4].

A. AI Applications in Species Identification and Monitoring

The use of artificial intelligence has changed the strategy of how ecologists discover or track species, deep learning provides us with an efficient method to analyze the massive data produced by recent ecodging monitoring systems [13-15]. The camera trap is one of the most important monitoring technologies that has been revolutionized by AI, which is critical for tracking wildlife distribution and activities at large spatial and temporal scales, showing great potential for wildlife community studies [16][17]. The biggest impediment with camera trap experiments, namely transferring millions of images, is being overcome thanks to the implementation of deep learning techniques, especially for what concearns the Convolutional Neural Networks (CNN), that can automatically identify wild animals and assign them to taxonomic entities with outstanding accuracy [18].

Such AI-based systems show remarkable results in species identification works. Deep neural networks have been demonstrated to automatically recognize animals with more than 93.8% accuracy, andnovel demonstrations validate that networks which are forced to select purely on high-confidence predictions of high confidence are as accurate as human volunteers in identifying taxa (96.6% accuracy) while being more than two orders of magnitude faster, processing thousands of images in less than a second without any training in this taxonomic category [18][19] Figure2. This substantially sped up processing (approximately 2000 images per minute on typical computing hardware) amounts to a revolution in our capacity for ecological research [20]. By incorporating AI with citizen science collaboration, it has been demonstrated that human effort can be minimized by as much as 43% without losing overall accuracy interms of detecting webs [21].

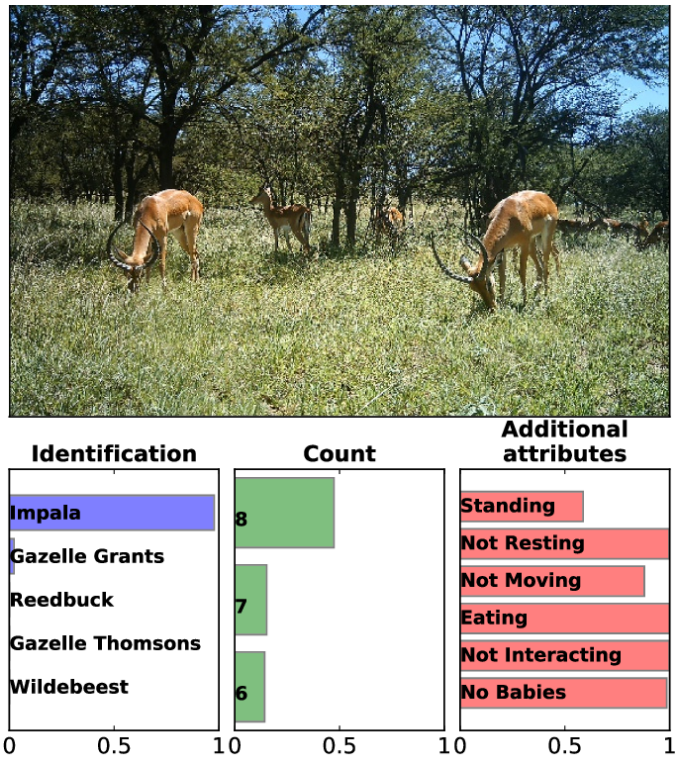


Fig. 2. Exemple of Deep Neural Networks (DNNs) can successfully identify, count, and describe animals in camera-trap images.

Apart from camera traps, AI is also changing monitoring of marine ecosystems with automatic species identification from diverse platforms such as moored observatories, AUVs and satellite imagery [22]. Deep learning models allow automated classification of marine species using a variety of information sources such as citizen science observations, benthic photo quadrats, cabled video observatories and acoustic sensors [22-24]. For instance, with acoustic data, 98.69% is achieved by CNN-based systems to recognize animal calls since marine mammal calls recorded from the sea from acoustic data instead of using a traditional method [25].

AI and species monitoring have been used for whale detection from satellite images [26], automated community ecology of planktonic foraminifera [27], as well as conservation-oriented platforms that identify animals, humans and poaching activities [28]. In particular, AI approaches have been found to be useful in predicting species interactions by trait-matching, with better performance of ML models as compared to statistical models in capturing complex ecological interactions [29].

AI and participation is is emerging direction, with some projects that use citizen science data to train AI for automatic species identification [30]. They have been described as democratizing advanced ecological monitoring through open source, low-cost, and low-powered developments [30][31] that facilitate communities' participation in biodiversity research.

With ongoing development of these capabilities, it is hoped that they will revolutionise ecological and conservation research by making analysis of large data sets more efficient, offering real time species identification, as a tool for education and by providing data which will be valuable in conservation planning and policy development [32]. Importance of these advances for science The sleeve is in the flexibility of being able to use advances like these to scale up biodiversity monitoring to levels never before seen and at a level of information that may be crucial for conservation in the inevitable face of rapid global environmental change [33][34].

B. Ai-Powered Species Distribution Models and Habitat Suitability Analysis

The use of Species Distribution Models (SDMs) has become popular in the field of ecology, biogeography and conservation biology to understand relationships between environmental predictors and species distribution. In the last twenty years, this field has embraced more and more machine learning features leading to a series of incremental model improvements [35]. Such AI-augmented SDMs can be used to model ecological niches and to predict suitable habitats more accurately as compared to classical statistical techniques, although there is a considerable variation among performance of the methodologies with respect to different ecological parameters, with predictions for presence–absence (53%) performing better than predictions of abundance, population fitness, and genetic diversity [36].

AI methodologies, including neural networks as well as ensemble learning, have substantially enhanced the accuracy of SDMs by revealing intricate associations among species and their environments [36]. MaxEnt, a popular machine learning approach to the modeling of species distributions when data are only present, has in different settings impressed with its performance with a 74.7% success rate at predicting the presence of disease-carrying mosquitos in the study of [36] Figure3. These advanced modelling approaches have employed tools like deep neural networks, GIS and CNN to predict potential habitats under existing and future environmental conditions [37].

The coupling of machine learning approaches and remote sensing has proved especially useful in habitat mapping and monitoring. These two tools offer major benefits to conservation practitioners allowing informed decisions to be made and targeted interventions to be taken to conserve threatened species and habitats [38]. For instance, spatial researchers have applied GIS and machine learning to understand the effect of urbanization on bobcat habitats in San Jose,CA by creating an original Habitat Suitability Model (HSM) that combines several environmental aspects to pinpoint critical conservation areas [39]. Similarly, in Leipzig, Germany, data from high-resolution Earth Observation and machine learning was employed to model hotspots of 44 bird species achieving 59-90% accuracy, and halting species richness of urban birds [39].

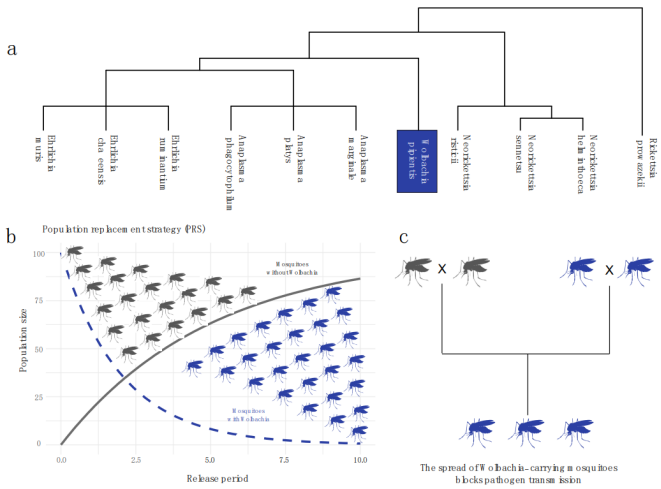


Fig. 3. Wolbachia-based population replacement strategy for mosquito-borne disease control.

Despite these developments, SDMs should be regarded as hypotheses that require testing with independent data, particularly in relation to conservation planning [36][40]. As a first step in overcoming interpretability problems associated with complex models, a new subfield known as explainable AI (xAI) provides a hope for improved understanding and interpretation of SDMs [35].

The conservation utility of AI-SDM/HSM is extensive. These technologies can process ecological data to identify hotspots of biodiversity point at risk of extinction and decide conservation priorities [41]. By participating in efforts to save endangered species, conserve biodiversity, and restore degraded habitats, AI is increasingly proving to be an essential tool in the conservation toolbox, enabling scientists and practitioners to tackle these types of complex ecological challenges in the face of a rapidly changing world.

To assist conservation practitioners in selecting appropriate AI techniques, we compare different models based on their accuracy and resource requirements. Table 2 Comparative performance and resource demand of selected AI models in ecological modeling.

This comparison illustrates the trade-offs between model complexity, interpretability, and operational costs, which are key to real-world adoption in conservation contexts.

C. Data Collection and Processing Methods for Ecological Research

Contemporary ecological study is more and more dominated by new ways of capturing and treating data fueled by A.I. These methods have revolutionized the way ecologists collect and interpret population level data:

DCNNs favor automatic analyses of images in ecology with an automatic detection of a high number of cells allowing, for example, quantitative wood anatomical investigations and accounting for the biological feature variance that needs manual handling in the past [42].

Audio Analysis and Classification methods leverage machine learning for the processing and classification of sounds

TABLE II. COMPARATIVE PERFORMANCE AND RESOURCE DEMAND OF SELECTED AI MODELS IN ECOLOGICAL MODELING

Model	Accuracy Range	Training Time	Hardware Requirements	Interpretability
MaxEnt	60–80%	Low	Standard PC	High
Random Forest	70–90%	Medium	Standard PC	Medium
CNN	80–98%	High	GPU	Low
DNN + xAI	85–95%	High	GPU/TPU	Medium to High

characteristic of wildlife, such as amphibian calls. These solutions use digital speech algorithms and filtering analysis to identify species according to their acoustic signatures [43] Figure4.

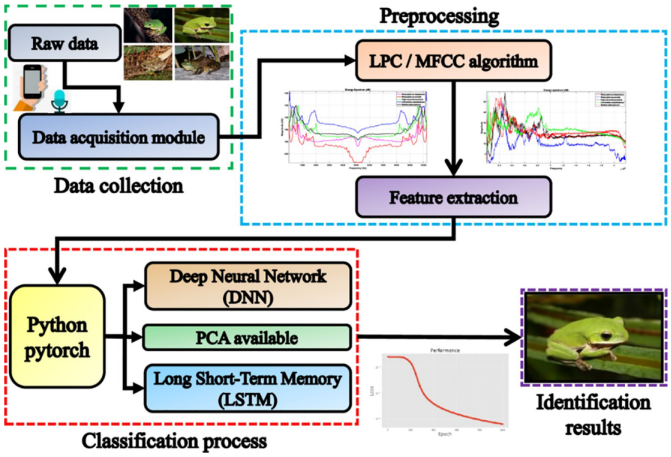


Fig. 4. The Structure of the experimental process for anuran bioacoustic classification.

Camera Trap Image Processing utilizes deep learning techniques to help sift through the millions of images and automatically identify wildlife. Such systems are capable to process around 2,000 images per minute on standard hardware, and they attain accuracy equal or greater than 98% in species classification [44][20].

Acoustic Monitoring Systems have employed convolutional neural networks for the detection and classification of echolocation calls produced by bats and other animals, with the latter being important for the processing of recordings collected in noisy field environments [44][45].

Integration of Remote Sensing and GIS: Remote Sensing and GIS Integration utilizes artificial intelligence, remote sensing imagery and GIS data to assess biodiversity at the landscape level. This integration supports generation of habitat maps, detection of ecological change, and identification of conservation priorities [46].

Plant Functional Trait Mapping applies deep learning models to field data or even photographs to map plant traits across landscapes. Recent development encompassmodels integrating field data with spectral information, and even use citizen science to map plant functional diversity [47-49].

(Field-deployed) Autonomous Environmental Sensors with embedded AI processing can process video imagery or acoustic streams on-site to identify and classify organisms to species level without external intrusion [50].

The use of deep learning in Earth Observation Data Processing permits higher resolution habitat classification, which is a prerequisite for the generation of fine-grained distributions of species across large spatial and temporal extents. Recent studies have been able to map global canopy height at 10m resolution using LiDAR data from satellite [50][51] Figure5.

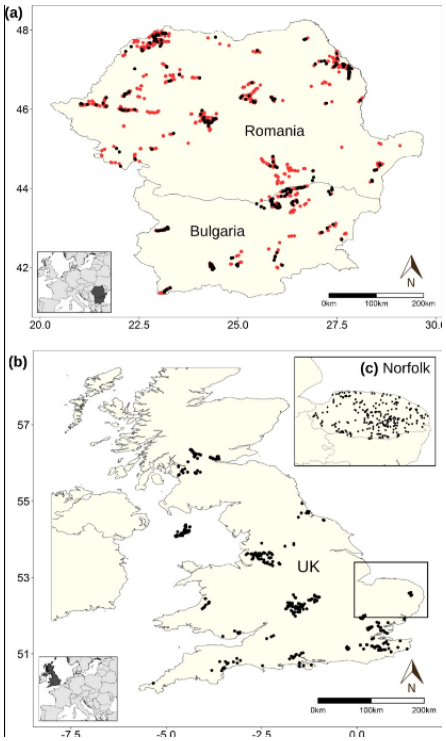


Fig. 5. Spatial distribution of the BatDetect CNNs training and testing datasets.

Multi-spectral Imaging Flow Cytometry combined with deep learning for rapid pollen identification and enumeration with high accuracy (96% species average accuracy), and morphology characterization (size, symmetry, and structure) [52].

Together these advancements have dramatically changed how ecological data are gathered and processed and, in the process, have allowed researchers to analyse extremely large datasets at scales unattainable through other measures, deriving significant patterns that would be unrecoverable by traditional means [7] [53]. The following table2 prensent differnt use cases of data collection and processing methods powered by artificial intelligence.

D. Conservation and Biodiversity Management Applications

Artificial Intelligence (AI) has become a very powerful force in the discipline of Conservation Biology and biodiver-

TABLE III. SUMMARY OF STUDIES APPLYING AI IN BIODIVERSITY MONITORING AND ECOLOGICAL PREDICTION

Papers	Data Collection Method	AI Technique Used	Ecological Application	Species or Habitat Focus	Data Processing Approach	Model Performance Evaluation
[42]	Images from transversal wood anatomical sections, manual cell area outputs	DCNN, Mask-RCNN	Quantitative wood anatomical analyses	Alnus glutinosa, Fagus sylvatica, Quercus petraea, and four conifer species	Mask-RCNN for image segmentation, compared with U-Net and ROXAS	Compared with U-Net and ROXAS; evaluated cell detection and pixel accuracy
[43]	Data collected from 32 frog and 3 toad species using bioacoustic recordings	DNN and LSTM	Species identification and biodiversity monitoring through amphibian sound classification	35 amphibian species in Taiwan, including frogs and toads	LPC, MFCC for filtering; PCA for dimensional reduction	Accuracy, training time, and comparison of DNN and LSTM with PCA
[20]	Camera traps used to collect wildlife images from five US states	CNN, ResNet-18	Classifying wildlife species from camera trap images	Wildlife in US, ungulates in Canada, Tanzania's faunal community	CNN (ResNet-18), R package MLWIC	98% accuracy in U.S., 82% in Canada, 94% detection in Tanzania
[45]	Citizen science, ultrasonic audio from road transects, iBats programme	CNN	Biodiversity monitoring through automatic detection of bat echolocation calls	Echolocating bats in Romania and Bulgaria	CNN (CNN FULL, CNN FAST) for detecting ultrasonic bat calls	Higher precision and recall in detecting search-phase echolocation calls
[33]	Continent-wide database, machine-phenotyped populations, weather, soil, satellite data	Ensemble of ML models	Crop yield prediction and understanding agronomic traits	Ten major crop species across a continent	Ensemble ML using weather, soil, and satellite data	R^2 prediction accuracy exceeding 0.8
[49]	Photos from citizen science (iNaturalist) and trait observations (TRY database)	CNN ensembles	Automated assessment of plant functional diversity through trait prediction	Plants; across growth forms, taxa, and global biomes	CNN, CNN ensembles, trait plasticity, climate factors	Accuracy enhanced using CNN ensembles and prior knowledge on trait plasticity

sity management, where many new innovative solutions are applied to complex environmental problems. Artificial Intelligence (AI) algorithms are revolutionizing wildlife surveillance by scanning through camera trap footage, while droids send back pictures of poachers from the air, analyzing the imagery and GPS data to recognize species and provide estimates of their abundance risking extinction, depending on humans´ decisions and AI algorithms scanning camera traps [54] Figure6. These technologies allow conservation biologists analyze huge amounts of ecological data, which would be infeasible to process manually, in an automated way, extensively improving the assessment of habitat health and the monitoring of populations [55].

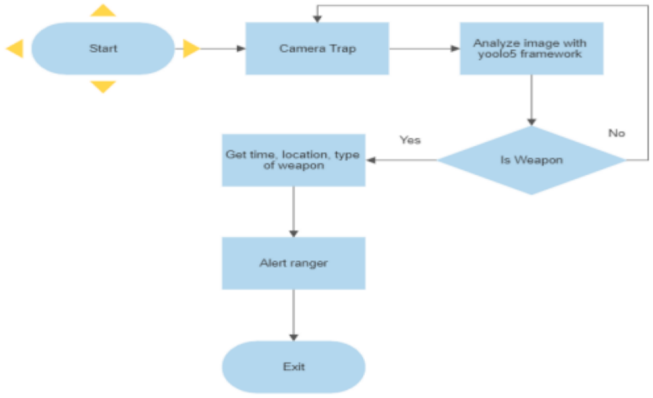


Fig. 6. Flowchart explaining methodolgy of anti-poaching system (Kuruppu, 2023).

AI combined with remote sensing instruments is seen to be particularly useful in habitat evaluation and resource preservation. Machine learning models with satellite imagery to identify forest cover changes, deforestation patterns and areas in need of environmental restoration can be used for this purpose [55]. This combination provides dense spatial and temporal information for otherwise poorly studied areas, with the analyses showing that even common methods like RF and GLM can predict relatively well different metrics of biodiversity [56][57] Figure7.

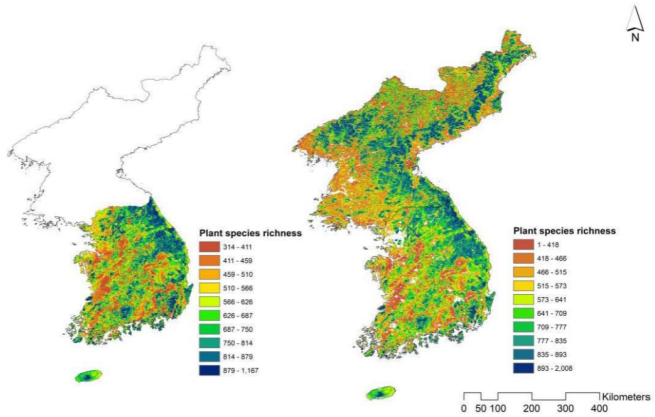


Fig. 7. Potential plant species richness estimated from S-SDMs in South Korea (Left) and estimated by th deep learning model in the Kor an Pe insula (Right).

AI's capability to handle multiple types of data, including acoustic recordings, eDNA, and imagery for species identification, to estimate biodiversity levels, and to identify new or threatened species has revolutionized biodiversity monitoring ([54]). Machine learning-type algorithms can monitor changes in population dynamics and evaluate the response of biodiversity to changes in the environment, which is important to design conservation programs and to determine how loss of

biodiversity will impact ecosystem function [58].

Arguably most importantly, predictive modeling enabled by AI can inform effective proactive conservation efforts by predicting the impacts of interventions on environmental changes and the dynamics of animal populations [55]. By combining historical and contemporary measures of decline or rate of increase of numbers of certain species and incorporating environmental change into the models, they can be used to predict potential future threatened species, thereby providing early warning of invasive species imbalance in ecosystems [59]. The ability to construct predictive models at these vast scales has the potential to transform ecology and environmental sciences in the way that statistics did over the course of the twentieth century.

AI is also improving conservation through integration of data and support of decision making. AI-supported analytics help to integrate various sources of data and present insights to support evidence-based decisions for conservation [55]. Methodological learning from comparative exercises with AI models such as ours is one way to support the development of stronger and generalisable solutions to biodiversity and habitat management [60]. Apart from direct conservation applications, machine learning also add value to conservation learning and education through the analysis of complex ecosystem data, such as climate data, soil property, and human action [59].

With overlaid environmental issues of climate change, loss of biodiversity, and zoonotic disease spread and how they intertwine the solutions, AI may provide useful tools to address such complexity [61]. AI technologies are increasingly indispensable tools within the context of modern conservation, as they allow us to automate monitoring of wildlife, improve habitat assessment, conduct advanced analyses of biodiversity, and build predictive models of species' distributions.

In a successful real-world application, a deep learning system deployed in Tanzania's Serengeti region processed over one million camera trap images to identify mammal species with 94% accuracy, significantly reducing manual labor for field researchers [20]. Similarly, AI-enhanced drones have been used in South Asia to detect and deter illegal poaching activities by combining object detection with real-time GPS alerts [28]. These examples illustrate the practical benefits and impact of AI-powered conservation tools.

E. Limitation and Future Directions

Although AI has greatly reshaped ecological and in particular biodiversity research, there are still major incomplete of AI approach and some challenging. The primary challenge is the data dependency of machine learning methods that need long high-quality well-balanced datasets that are rarely available in ecosystem contexts. This is especially problematic for the rarer species or the less studied ecosystems, as the data scarcity might result in model bias or low generalization. In addition, most ecological data sets are subject to biases introduced by selective sampling, taxonomic preferences, and geographic imbalances that may be amplified by AI models and result in erroneous conclusions.

There is also one important challenge to remain unanswered: "black box" property that belongs to many popular deep learning methods, because complex models are often "opaque" to a variety how they make the prediction by themselves and solvers, and it is hard for ecologists to interpret how those identified patterns may be relevant to the ecological process. This interpretability issue has driven the need for explainable AI (xAI) models that are able to offer explanations over decision making while retaining the skills of prediction.

From a practical point of view computational prerequisites are still, in most cases, out of reach of most performing (not computer scientists) research groups in ecology, since the "standard" training of complex neural networks requires special hardware and, in some cases, specific technical know-how, that it is not going to be part of the one found in a normal ecology department. Such a technical hurdle not only restricts wider access, but also may amplify the existing disparities in the availability of research capacity.

More fundamentally, there is a gulf between statistical prediction and ecological understanding. Although AI can learn patterns exceptionally well, as AI it may not advance ecological theory or mechanistic understanding in the absence of domain integration. It's part and parcel of the machine-learning world." Critics, however, point out that machine learning models can prioritize prediction at the cost of explanation, raising questions about their scientific value.

In the future, there are several promising avenues to tackle these issues. First among these is the emergence of easier-to-use tools and platforms that greatly lower the technical barriers to entry to AI-based ecological research. Projects like Wildlife Insights and iNaturalist show that user-friendly interfaces can democratize the use of powerful AI tools.

Transfer learning and few-shot learning methods are promising for addressing the constraint of few data by transferring models trained on data-rich species to closely related yet data-poor species. Such methods could potentially be used to transfer AI benefits to poorly studied taxonomic groups or ecosystems.

Reinforcing the interface between ecologists, computer scientists and statisticians is another important way forward, which should lead to better consideration of the ecology when designing AI tools, instead of just running a generic algorithm on ecological data. This partnership is particularly relevant for hybrid model development where the complementary strengths of machine learning, predictive performance, and mechanistic models, theoretical and interpretability related, need to be integrated.

The future for AI in ecology may involve the development of frameworks that combine multiple AI approaches with domain knowledge to satisfy both predictive needs and mechanistic understanding. For application purposes, however, it is desirable that researchers can interpret models within theoretical frameworks other than the one in which they were developed, by borrowing strengths from alternative modeling paradigms. With climate change intensifying and biodiversity decline proceeding, such integrated methodologies will be paramount to achieving effective conservation management informed bydata-driven knowledge and ecological understanding.

To mitigate geographic and taxonomic biases in training

datasets, several strategies have been proposed. These include transfer learning, where models trained on data-rich species are adapted to underrepresented taxa, and the use of synthetic data generation to augment sparse datasets.

Explainable AI (xAI) can address the black-box limitation of deep learning models. For example, SHAP (SHapley Additive exPlanations) values allow researchers to identify which environmental variables most influence a model's prediction of species presence, enhancing trust and transparency for ecologists and decision-makers alike.

V. Key Takeaways

• Neural networks enhance prediction of habitat suitability across ecosystems,

• Models like MaxEnt remain relevant due to simplicity and transparency,

• xAI offers promising solutions to bridge interpretability gaps,

• Resource needs vary widely; GPU-based models may limit accessibility.

VI. Conclusion

The augmentation of artificial intelligence with ecology research represents a state-of-the-art tool in the study, monitoring, and protection of biodiversity. In this systematic review, we have aggregated almost thirty years of worldwide research that applies AI (in particular ML and DL) to species preference analyses, ecological niche modelling and conservation recommendations.

Our review, therefore, underscores a dynamic, cross-disciplinary field that has grown exponentially since 2015 due to rapid technological development, enhanced ecological importance, and the increase in big environmental data. There are applications ranging from automated species identification in terrestrial and marine environments to sophisticated habitat suitability modelling and decision support systems. Thanks to machine learning and artificial intelligence, researchers today are able to keep track of what is happening to biodiversity across space and time at a faster and truer pace than ever before, turning static ecological censuses into dynamic feedback loops.

However, there is still much more to be done. Ecological data-sets are frequently incomplete, imbalanced or biased towards certain geographies –attributes that may limit the accuracy and generalization of predictive models. In addition, the black-box nature of a lot of AI systems is an issue for interpretability, transparency, and trust, especially in conservation where decision making has real worlds ecological and ethical consequences. There is greater urgency for explainable AI (xAI), for more equitable access to computational resources, and for more interdisciplinarity.

In the context of habitat intelligence, the future will involve hybrid frameworks that combine the predictive power of AI with mechanistic ecological theory. Recent developments including transfer learning, citizen science linkage, and low-cost sensing are leading to an increasing degree of involvement and translation, especially in underrepresented geographies and ecosystems. By fostering collaboration among ecologists, data scientists, and decision-makers, the community can develop scalable, interpretable and actionable tools that address the twin crises of climate warming and biodiversity loss.

Together, they illustrate that AI has the potential to truly revolutionize ecological science as an exciting complement to its traditional iteration, not a replacement. By exposing previously hidden patterns in species–environment interactions, AI supports better, faster and more strategic conservation action. While this area is still emerging, it is clear that it will be fundamental in the future of biodiversity and the management of life on the planet.

References

[1] Z. Miao, K. M. Gaynor, J. Wang, Z. Liu, O. Muellerklein, M. S. Norouzzadeh *et al.*, "Insights and approaches using deep learning to classify wildlife," *Scientific Reports*, 2019.

[2] R. Kitchin, "Big Data, new epistemologies and paradigm shifts," *Big Data & Society*, 2014.

[3] B. A. Han, K. R. Varshney, S. LaDeau, A. Subramaniam, K. Weathers, and J. A. Zwart, "A synergistic future for AI and ecology," *Proc. Natl. Acad. Sci. USA*, 2023.

[4] P. Butcher, A. Colefax, R. Gorkin, S. Kajiura, N. A. López, and J. Mourier *et al.*, "The Drone Revolution of Shark Science: A Review," 2021.

[5] J. Hirn, J. E. García, A. Montesinos-Navarro, R. Sanchez-Martín, V. Sanz, and M. Verdú, "A Deep Generative Artificial Intelligence system to decipher species coexistence patterns," *arXiv*, 2021.

[6] P. Desjardins-Proulx, I. Laigle, T. Poisot, and D. Gravel, "Ecological interactions and the Netflix problem," *bioRxiv*, 2016.

[7] S. Christin, É. Hervet, and N. Lecomte, "Applications for deep learning in ecology," *bioRxiv*, 2018.

[8] M. Borowiec, P. Frandsen, R. B. Dikow, A. McKeeken, G. Valentini, and A. E. White, "Deep learning as a tool for ecology and evolution," *Methods in Ecology and Evolution*, 2021.

[9] M. González-Rivero, O. Beijbom, A. Rodriguez-Ramirez, D. E. P. Bryant, A. Ganase, Y. González-Marrero *et al.*, "Monitoring of Coral Reefs Using Artificial Intelligence: A Feasible and Cost-Effective Approach," *Remote Sensing*, 2020.

[10] A. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, 1995.

[11] L. Bland, B. Collen, C. Orme, and J. Bielby, "Predicting the conservation status of data-deficient species," *Conservation Biology*, 2015.

[12] T. Rangel and R. Loyola, "Labeling Ecological Niche Models," 2012.

[13] S. B. Islam, D. Valles, T. J. Hibbitts, W. A. Ryberg, D. K. Walkup, and M. Forstner, "Animal Species Recognition with Deep Convolutional Neural Networks from Ecological Camera Trap Images," *Animals*, 2023.

[14] W. Ahmed, V. K. Kommineni, B. König-Ries, J. Gaikwad, L. M. R. Gadelha, and S. Samuel, "Evaluating the method reproducibility of deep learning models in the biodiversity domain," *arXiv*, 2024.

[15] W. Ahmed, V. K. Kommineni, B. König-Ries, J. Gaikwad, L. Gadelha, and S. Samuel, "Evaluating the method reproducibility of deep learning models in biodiversity research," *PeerJ Computer Science*, 2025.

[16] J. Vélez, W. McShea, H. Shamon, P. J. Castiblanco-Camacho, M. A. Tabak, C. Chalmers *et al.*, "An evaluation of platforms for processing camera-trap data using artificial intelligence," *Methods in Ecology and Evolution*, 2022.

[17] J. Vélez, P. J. Castiblanco-Camacho, M. A. Tabak, C. Chalmers, P. Fergus, and J. Fieberg, "Choosing an Appropriate Platform and Workflow for Processing Camera Trap Data using Artificial Intelligence," *arXiv*, 2022.

[18] H. Böhner, E. Kleiven, R. Ims, and E. Soininen, "A semi-automatic workflow to process camera trap images in R," *bioRxiv*, 2022.

[19] M. S. Norouzzadeh, A. T. Nguyen, M. Kosmala, A. Swanson, M. Palmer, C. Packer *et al.*, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proc. Natl. Acad. Sci. USA*, 2017.

[20] M. A. Tabak, M. S. Norouzzadeh, D. W. Wolfson, S. Sweeney, K. Vercauteren, N. P. Snow *et al.*, "Machine learning to classify animal species in camera trap images: applications in ecology," *bioRxiv*, 2018.

[21] M. Willi, R. T. Pitman, A. W. Cardoso, C. M. Locke, A. Swanson, A. Boyer *et al.*, "Identifying animal species in camera trap images using deep learning and citizen science," *Methods in Ecology and Evolution*, 2018.

[22] G. Canonico, P. Buttigieg, E. Montes, F. Muller-Karger, C. A. Stepien, D. Wright *et al.*, "Global Observational Needs and Resources for Marine Biodiversity," *Frontiers in Marine Science*, 2019.

[23] A. M. Rahimi, R. J. Miller, D. Fedorov, S. Sunderrajan, B. Doheny, H. M. Page *et al.*, "Marine Biodiversity Classification Using Dropout Regularization," in *Proc. ICPR Workshop on Computer Vision for Analysis of Underwater Imagery*, 2014.

[24] S. Marini, E. Fanelli, V. Sbragaglia, E. Azzurro, J. Del Río Fernández, and J. Aguzzi, "Tracking Fish Abundance by Underwater Image Recognition," *Scientific Reports*, 2018.

[25] C. McQuay, F. Sattar, and P. Driessen, "Deep learning for hydrophone big data," in *Proc. Pacific Rim Conf. on Communications, Computers and Signal Processing*, 2017.

[26] E. Guirado, S. Tabik, M. L. Rivas, D. Alcaraz-Segura, and F. Herrera, "Automatic whale counting in satellite images with deep learning," *bioRxiv*, 2018.

[27] A. Y. Hsiang and P. Hull, "Automated community ecology using deep learning: a case study of planktonic foraminifera," *bioRxiv*, 2022.

[28] P. Fergus, C. Chalmers, S. N. Longmore, and S. A. Wich, "Harnessing Artificial Intelligence for Wildlife Conservation," *Conservation*, 2024.

[29] M. Pichler, V. Boreux, A. Klein, M. Schleuning, and F. Hartig, "Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks," *Methods in Ecology and Evolution*, 2019.

[30] M. Lotfian, J. Ingensand, and M. Brovelli, "The Partnership of Citizen Science and Machine Learning: Benefits, Risks, and Future Challenges for Engagement, Data Collection, and Data Quality," *Sustainability*, 2021.

[31] E. McClure, M. Sievers, C. Brown, C. Buelow, E. M. Ditria, M. Hayes *et al.*, "Artificial Intelligence Meets Citizen Science to Supercharge Ecological Monitoring," *Patterns*, 2020.

[32] A. Hill, P. Prince, E. P. Covarrubias, C. P. Doncaster, J. L. Snaddon, and A. Rogers, "AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment," 2018.

[33] J. Bennet, "BIRD SPECIES IDENTIFICATION USING DEEP LEARNING," *Journal of Science & Technology*, 2022.

[34] D. K. Gupta, A. Pagani, P. Zamboni, and A. K. Singh, "AI-powered revolution in plant sciences: advancements, applications, and challenges for sustainable agriculture and food security," *Exploration of Foods and Foodomics*, 2024.

[35] M. Ryo, B. Angelov, S. Mammola, J. Kass, B. M. Benito, and F. Hartig, "Explainable Artificial Intelligence enhances the ecological interpretability of black-box species distribution models," 2020.

[36] F. Branda, E. Cella, F. Scarpa, S. Slavov, A. Bevivino, R. Moretti *et al.*, "Wolbachia-Based Approaches to Controlling Mosquito-Borne Viral Threats: Innovations, AI Integration, and Future Directions in the Context of Climate Change," *Viruses*, 2024.

[37] T. Wang, Y. Zuo, T. Manda, D. Hwarari, and L. Yang, "Harnessing Artificial Intelligence, Machine Learning and Deep Learning for Sustainable Forestry Management and Conservation: Transformative Potential and Future Perspectives," *Plants*, 2025.

[38] D. A., M. S. T., S. M., and K. Nanjundeshwara, "LEVERAGING MACHINE LEARNING AND REMOTE SENSING FOR WILDLIFE CONSERVATION: A COMPREHENSIVE REVIEW," *International Journal of Advanced Research*, 2023.

[39] Y. Rahmati, "Artificial Intelligence for Sustainable Urban Biodiversity: A Framework for Monitoring and Conservation," *arXiv.org*, 2024.

[40] S. Ogunlade, A. I. Adekunle, E. McBryde, and M. Meehan, "Modelling the ecological dynamics of mosquito populations with multiple co-circulating Wolbachia strains," *Scientific Reports*, 2022.

[41] R. A. Batki and H. Phudinawala, "A Comparative Study on How AI is Partnering with Nature for a Sustainable Future," *Int. J. Multidiscip. Res.*, 2024.

[42] G. Resente, A. Gillert, M. Trouillier, A. Anadon-Rosell, R. L. Peters, G. von Arx *et al.*, "Mask, Train, Repeat! Artificial Intelligence for Quantitative Wood Anatomy," *Frontiers in Plant Science*, 2021.

[43] C. Gong, C. H. Su, K. W. Chao, Y. Chao, C. K. Su, and W. H. Chiu, "Exploiting deep neural network and long short-term memory methodologies in bioacoustic classification of LPC-based features," *PLoS ONE*, 2021.

[44] M. Pichler and F. Hartig, "Machine learning and deep learning—A review for ecologists," *Methods in Ecology and Evolution*, 2022.

[45] O. Mac Aodha, R. Gibb, K. Barlow, E. Browning, M. Firman, R. Freeman *et al.*, "Bat detective—Deep learning tools for bat acoustic signal detection," *bioRxiv*, 2017.

[46] A. U. Gabi and N. M. Abdullah, "Optimizing Biodiversity Conservation in Sundaland through Advanced Geospatial Techniques and Remote Sensing Technologies," *BIO Web of Conferences*, 2024.

[47] J. Aguirre-Gutiérrez, S. W. Rifai, X. Deng, H. ter Steege, E. Thomson, J. Corral-Rivas *et al.*, "Canopy functional trait variation across Earth's tropical forests," *Nature*, 2025.

[48] S. Newman and R. Furbank, "Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data," *Nature Plants*, 2021.

[49] C. Schiller, S. Schmidtlein, C. C. F. Boonman, Á. Moreno-Martínez, and T. Kattenborn, "Deep learning and citizen science enable automated plant trait predictions from photographs," *Scientific Reports*, 2021.

[50] S. Kelling, "Technology Developments for Biodiversity Monitoring and Conservation," 2018.

[51] N. Lang, W. Jetz, K. Schindler, and J. D. Wegner, "A high-resolution canopy height model of the Earth," *Nature Ecology & Evolution*, 2022.

[52] S. Dunker, E. Motivans, D. Rakosy, D. Boho, P. Mäder, T. Hornick *et al.*, "Pollen analysis using multispectral imaging flow cytometry and deep learning," *New Phytologist*, 2020.

[53] J. Wäldchen and P. Mäder, "Machine learning for image based species identification," *Methods in Ecology and Evolution*, 2018.

[54] O. N. Chisom, P. W. Biu, A. A. Umoh, B. O. Obaedo *et al.*, "Reviewing the role of AI in environmental monitoring and conservation: A data-driven revolution for our planet," *World Journal of Advanced Research and Reviews*, 2024.

[55] N. R. Hegde and S. M. Bargavi, "Smart Conservation: Integrating AI for Enhanced Wildlife Monitoring," *Int. Res. J. Comput. Sci.*, 2024.

[56] H. Choe, J. Chi, and J. Thorne, "Mapping Potential Plant Species Richness over Large Areas with Deep Learning, MODIS, and Species Distribution Models," *Remote Sensing*, 2021.

[57] L. Zhang, L. Zhang, and B. Du, "Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art," *IEEE Geoscience and Remote Sensing Magazine*, 2016.

[58] E. A. Etukudoh, A. Hamdan, K. I. Ibekwe, A. A. Umoh, and V. I. Ilojianya, "AI and machine learning in climate change research: A review of predictive models and environmental impact," *World Journal of Advanced Research and Reviews*, 2024.

[59] M. Milutinović, "Machine Learning in Environmental Monitoring," *Facta Universitatis, Series: Working and Living Environmental Protection*, 2024.

[60] N. Wilson, "Artificial intelligence helps drive new frontiers in ecology," *BioScience*, 2024.

[61] G. Gb and S. M. Sumathi, "Safeguarding Humans from Attacks Using AI-Enabled (DQN) Wild Animal Identification System," *Int. Res. J. Multidiscip. Scope*, 2024.

Deep Learning-Based Bone Age Growth Disease Detection (BAGDD) Using RSNA Radiographs

Muhammad Ali[1], Muhammad Faheem Mushtaq[2], Saima Noreen Khosa[3],
Naila Kiran[4], Humaira Arshad[5], Urooj Akram[6]
Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan[1,2,4,5,6]
Institute of Computer Science, Khwaja Fareed University of Engineering and Information Technology,
Rahim Yar Khan 64200, Pakistan[3]

***Abstract*—Radiological bone age assessment is essential for diagnosing pediatric growth and developmental disorders. The conventional Greulich-Pyle Atlas, though widely used, is manual, time-intensive, and prone to inter-observer variability. While deep learning methods such as Convolutional Neural Networks (CNNs) offer automation potential, most existing models rely on transfer learning from natural image datasets and lack specialization for medical radiographs. This study aims to address the gap by developing a domain-specific, custom CNN for pediatric bone age prediction. This research proposes a customized CNN architecture trained on the RSNA pediatric bone age dataset, which includes over 12,000 annotated hand X-ray images labeled with age and gender. The pipeline incorporates pre-processing techniques such as image resizing, normalization, and Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance input quality. A YOLOv3 object detector is utilized to localize the hand region prior to model training, focusing on the most relevant anatomical structures. Unlike traditional transfer learning models such as ResNet50, VGG19, and InceptionV3, the proposed CNN is tailored for radiographic features using optimized convolutional blocks and domain-aware augmentations. This design improves generalization and reduces overfitting on small or imbalanced subsets. The proposed model achieved a Mean Absolute Error (MAE) of 3.27 months on the test set and 3.08 months on the validation set, outperforming state-of-the-art transfer learning approaches. These results demonstrate the model's potential for accurate and consistent bone age estimation and highlight its suitability for integration into clinical decision-support systems in pediatric radiology.**

Keywords*—Bone age estimation; pediatric healthcare; convolutional neural networks; transfer learning; YOLOv3; medical imaging

I. Introduction

Pediatric endocrinologists mostly rely on Bone Age Growth Disease Detection (BAGDD) to detect growth disorders and evaluate a child's bone development. For conditions such as growth hormone deficiency, precocious puberty and delayed growth, it matters a lot [1]. BAGDD is used by clinicians to understand if a child's bones are growing at the appropriate pace for their age and to predict what might happen in their growth later on. Typically, an X-ray image of the left hand and wrist is used to view the bones and changes such as ossification, that reveal the growth status of a child.

Bone age was initially measured using two traditional methods: the Greulich-Pyle (GP) Atlas [2] and the Tanner-Whitehouse (TW) approach [3]. A GP Atlas allows a trained radiologist to use the worksheet to compare a patient's bone density to those of reference images based on both gender and age. Rather than TW, the TW method assigns a score to every bone depending on its stage of development and the total gives an estimate for bone age. Despite being around for decades, these methods are held back by expert interpretations, slow procedures and the influence of personal opinions [4].

Automating bone age assessment is becoming possible thanks to the growth in AI and its specialty, Deep Learning. CNNs have helped smart systems reach the same performance in image analysis as the human eye in recent times. CNNs are particularly useful for image recognition since they can automatically pick out patterns such as edges, textures and shapes at different parts of an image [5]. Tickmarks depends on how images are used; they can be used for labeling or predicting values in both types of tasks. CNNs have been used successfully in many medical image applications, including finding tumors, analyzing retinas and, not long ago, evaluating bone age [6]. In 2017, the Radiological Society of North America made available the RSNA Bone Age Dataset which consists of over 12,000 X-rays of children's hands each labeled with bone age information and evenly divided by gender. The information in this dataset aids in the development and assessment of algorithms designed for automatic bone age estimation [7].

Such models are pre-set with knowledge from millions of pictures found in ImageNet of everyday objects like cats, cars and buildings. These models are mainly used for object classification because their learned features highlight important differences in objects' shapes, colors and textures. Still, medical images, including those of hands for assessing bone age, have very different details regarding bone density, shape and growth as compared to regular X-rays [8]. The results of further studies have suggested that development of bone age is not the same in males and females and males generally mature slower than females do [9]. Because of this, transfer learning models may not fit the unique gender differences in how bones develop. Thus, these models do not always capture the important information related to each domain and usually perform poorly on these imaging operations [10].

This work introduces a new CNN model tailored for computing the bone age from radiographs of hands. Everything is learned directly by the model from scratch, so no information is transferred from different bone-related picture datasets and the model is trained explicitly for predicting bone age. The goal is to improve the certainty and precision of bone age determination by using the CNN design to meet

the requirements of this medical imaging task. The proposed custom CNN model performed well, with the MAE being 3.27 months in the training set and 3.08 months on the validation set. These results significantly outperform the transfer learning models evaluated in this work, including ResNet50, VGG19, and InceptionV3. The custom CNN proves to be effective in capturing fine-grained features in bone development, allowing for a more accurate age estimation. These findings support the conclusion that domain-specific deep learning models can substantially outperform generic transfer learning approaches when applied to specialized medical imaging tasks such as bone age assessment [9].

Whereas deep learning algorithms, especially the ones that apply the methodology of transfer learning, such as ResNet50, VGG19, and InceptionV3 models have demonstrated good potential when it comes to assessing the age of the bones, the utilized models have been predominantly trained on the datasets of natural images (e.g., ImageNet) and have not been optimized to work with the peculiarities of medical radiographs. They do not routinely produce good imaging of the subtle anatomical markings characteristic of X-ray pediatric hands required to make precise assessment of the bone-age. In addition to that, none of the current models often provide the combination of the domain-specific architecture design alongside specific preprocessing (e.g., the CLAHE-enhanced contrast and the ROI extraction using the YOLOv3) to form a unified framework applicable to the setting of the pediatric bone age growth disease detection.

The novelty of this work is the modification of the CNN model into a custom design, which trained on the scratch available RSNA Bone Age Dataset that consists of data on bone development provided specifically to be used in predicting an age. In contrast to transfer learning methods devised over generically, the proposed CNN includes the task-specific architectural components, including depth of layers and dropout regularization, along with the paradigm-specific preprocessing of medical images. This is more accurate, more robust and better aligned to the domain compared to current solutions in bone age estimation.

This research focuses on the architectural design, training, and performance evaluation of a custom CNN model specifically developed to predict bone age from hand radiographs. The contributions of this research are as follows:

- A custom Convolutional Neural Network (CNN) is developed and trained from scratch, incorporating task oriented design choices such as optimal layer depth, kernel sizes, ReLU activations, and dropout regularization. Unlike generic models, custom CNN was tailored to capture fine-grained anatomical features (e.g. bone edges, growth plates) specific to pediatric hand radiographs, resulting in improved precision and robustness.
- The performance of proposed model is significantly boosted by incorporating Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance image contrast and Region of Interest (ROI) detection using YOLOv3 to focus the learning on relevant hand structures.
- The proposed custom CNN was extensively evaluated using Mean Absolute Error (MAE) and Mean Squared Error (MSE), and benchmarked against popular pre-trained models including ResNet50, VGG19, and InceptionV3.
- The proposed model was optimized through manual adjustment of the hyperparameters, including the learning rate, batch size, and number of training epochs. In addition, the architecture was refined by adjusting the depth and number of filters that incorporate dropout layers to mitigate overfitting.

The organization of the paper is as follows. Section 2 reviews existing work relevant to the proposed study. Section 3 discusses materials and methods and outlines the development procedures of the proposed BAGDD system. Section 4 presents the results and the discussion, while Section 5 concludes the study.

II. Related Work

Deep learning has played a revolutionary role in medical imaging, particularly in the diagnosis of musculoskeletal disorders. The effectiveness of deep learning applications in bone age prediction has been demonstrated in various contexts, including fracture identification, osteoarthritis severity evaluation, and pediatric assessments. Recent research highlights its ability to detect pathological abnormalities in CT and MRI scans, such as infections, fractures, joint degeneration, internal derangements, and metastases. However, challenges remain due to the complexity of CT and MRI datasets, which often involve varying tissue contrasts. For widespread clinical adoption, further technological advancements and extensive, multi-institutional prospective studies are required to ensure the generalizability, consistency, and reliability of these techniques [11].

New methods for finding and handling different medical problems have been made possible by using CNNs. Related to Bone Age Assessment using hand and wrist radiographs, CNNs have consistently performed better than other common approaches. This section deals with applying deep learning principles for medical image analysis, paying special attention to CNN designs and transfer learning created for BAA [12].

CNNs are very important in medical image analysis, with a key application in bone age estimation. CNNs built from the ground up are created to handle the issues that come with using medical images. During Bone Age Assessment (BAA), one must make models that are skilled in finding growth plates and different bone parts so the process is accurate. Many studies have shown that custom CNN models usually outperform methods such as the Greulich-Pyle atlas or Tanner-Whitehouse standards.

Specifically, a custom CNN for estimating bone age from hand radiographs designed by [13] makes use of different layers to grab both general and detailed hand and wrist features, improving the mean error over methods used before. In a similar way, [14] recommended an automatic deep learning process to make the process more efficient and less dependant on human opinion. In their method, they chose key points in images of 3,000 X-rays, working on the carpal, metacarpal and

phalangeal bones. Intensity was adjusted using preprocessing and different models were made for each gender since each follows a different pattern of growth. Unlike various other similar studies, this one looked at infants as part of the full pediatric population that was considered. Using deep learning, it was possible to create a model with an MAE of 8.890 months.

An improved CNN was built in [15] to be used with the RSNA Bone Age dataset. Thanks to deep residual learning, the model had optimized gradient flow and avoided the problem of overfitting in the larger layer. Using this kind of architecture, models successfully pointed out main anatomical landmarks for an MAE of 4.2 months. The team discovered that CNNs designed to identify certain anatomical features did better than general architectures in predicting CT results.

Still, building a custom CNN architecture is complicated and takes a long time. As described in [16], making custom CNNs for medical imaging tasks draws on expertise in both radiology and deep learning which presents a problem for both small healthcare centers and seasoned developers. All the same, the advantages of custom CNNs in medical areas make them a good alternative because they can accomplish important tasks more successfully than standard approaches, for example bone age estimation [17].

Researchers have recently shown that transfer learning can successfully estimate bone age. The results presented in [18] indicate that a fine-tuned EfficientNet model had an MAE of 4.5 months when used on the RSNA Bone Age dataset. Training the model with this approach was much faster than using a custom CNN model with similar results. The study says that EfficientNet's design means it functions well for medical imaging, especially in places where resources are limited.

The authors of [19] used transfer learning by training a DenseNet on images for recognizing bone age. The authors managed to train the model quickly and accurately by freezing the early layers and adding extra training to the last ones. By using transfer learning on the RSNA Bone Age dataset, the model showed an MAE of 14.9 months, showing that transfer learning can be equally successful, while having a more variable precision [20].

In another study, [21] evaluated a commercial deep learning based bone age assessment tool (BoneAge, Vuno, Korea) using left hand X-ray images of 371 healthy Korean children and adolescents aged 4–17 years. While the estimated bone age was highly correlated with the actual chronological age (r = 0.96), the overall agreement rate was only 58.8%, and the model tended to overpredict bone age in younger children, particularly those under 8.3 years of age.

The work presented by [22] focuses on evaluating age and sex from computed tomography (CT) images of vertebrae using computer aided diagnosis. The study utilized a dataset comprising CT scans from 166 patients of varying genders. All images were rescaled before feature extraction using the qMaZda software. The researchers integrated conventional machine learning algorithms—both regression and classification with deep convolutional networks. Using the BART algorithm, the bone age regression model achieved a Mean Absolute Error (MAE) of 3.14 years. Gender classification accuracy was 69% with machine learning models and 59% with deep learning models [23].

In a related study, [24] proposed a regression based, multi-model deep learning framework for bone age assessment using hand radiographic images and clinical data (e.g., gender and chronological age). The approach combined EfficientNetV2S CNNs for image analysis with a simple deep neural network (DNN) for processing clinical data, resulting in robust diagnostic performance. This study underscored the advantages of multi-modal learning in improving the precision of bone age analysis [25].

To address the limitations of traditional methods such as Greulich–Pyle and Tanner–Whitehouse, which heavily depend on domain expertise, [26] developed a deep learning model for osteoporosis classification and bone density prediction using opportunistic CT scans. The model was validated across datasets from multiple hospitals and CT scanners. It employed the VB-Net module for segmentation and DenseNet for classification and regression. The model demonstrated high accuracy and strong correlation with Quantitative Computed Tomography (QCT) reference standards, achieving AUC scores of 0.999 (training set), 0.970 (test set), and 0.933 (independent set). The bone density prediction model also performed well, showing potential as a cost effective, low-radiation solution for osteoporosis screening [27], [28].

In [29] and [30] review existing Bone Age Assessment (BAA) methodologies, emphasizing the use of deep learning algorithms to address the inconsistencies of traditional approaches. Although deep neural networks have demonstrated high effectiveness in classifying bone age within specific age intervals, their complexity and reliance on numerous parameters for each region of interest (ROI) present challenges.

In [31] introduced a novel approach known as the Dual Attention Dual Path Network (DADPN). Experimental validation was performed using the RSNA Pediatric Bone Age Challenge dataset, comparing DADPN with nine other popular BAA methods. The results showed that DADPN achieved the highest accuracy, with a Mean Absolute Error (MAE) of 4.76 months.

In a separate study, [32] and [33] proposed a continuous radiological age assessment method based on clavicle ossification in CT scans. Their deep learning model was trained on 4,400 scans from 1,935 patients with a mean age of 24.2 years and evaluated on a test set of 300 scans from 300 patients. The model achieved an MAE of 1.65 years. Absolute error values ranged from 6.40 years for females and 7.32 years for males, with comparison to human reader estimates yielding errors of 3.40 years for females and 3.78 years for males, resulting in a comparative MAE of 1.84 years [34].

Although these improvements exist, additional scientist work is necessary to improve the accuracy and dependability of BAA methods. A number of CNN and transfer learning frameworks struggle to perform the same on diverse data which lowers their stability. Further work is needed to decrease the errors in bone age predictions and maintain good performance in every healthcare setting.

III. Materials and Methods

The structure of this study's research framework emphasizes building an effective auto bone age prediction method using deep learning. There are five main phases in this framework: collecting data, checking the data, preprocessing data, making features, training the model and assessing the model. All the data used in this work is taken from the RSNA Bone Age Dataset which consists of about 12,000 hand X-rays. The variety of bone growth patterns in this dataset gives it a good fit for using and assessing deep learning models.

The aim of the research is to create a new CNN model that can identify and classify bones primarily based on age. To review performance, the proposed model is compared with three commonly used transfer learning models: ResNet50, VGG19 and InceptionV3. The framework is designed to highlight how well the custom CNN predicts bone age, while transfer learning models help understand the comparison.

First, several stage were taken to boost the quality of the images before supplying them to the CNN models. First, all the images were sized to 224 × 224 pixels which is a typical input measurement for most deep learning models, including the ones employed in this study. Stretching each image to the same size reduces the amount of processing required and assures even results during training. Each value was then normalized by putting it through (255/2) and rescaled to values between 0 and 1 to support more efficient training. Besides, contrast in the images was improved by applying Contrast Limited Adaptive Histogram Equalization (CLAHE) only to the bone structures. Using this method, important anatomical features are made more visible, so the CNN can recognize and understand them better during fixing.

In order to enhance the performance of the model in more situations, this study used various data enhancement techniques. Examples of applied image transformations were rotation, flipping and shifting which introduced little changes in this work. Data augmentation boosts the size of the dataset and has the side benefit of preventing overfitting due to the extra input data. So, the model can withstand variations in real situations better. In addition, ROI detection was completed with a YOLOv3 model. At this point, the developer removed information behind the hand, helping the model concentrate on what it should focus on and therefore improving the training.

Following the preprocessing and feature extraction stages, the next phase involved model development. The CNN architecture was customized specifically for the bone age prediction task. The model consists of multiple convolutional layers that progressively learn features from low-level edges and textures to more abstract patterns crucial for bone age estimation. Since the model was trained from scratch, it was able to focus exclusively on learning features relevant to bone development, as opposed to transfer learning models that rely on representations learned from general purpose natural image datasets. This task specific design allowed the custom CNN to achieve better performance by concentrating on features uniquely important for pediatric hand X-ray analysis.

In addition to the custom CNN, this study also employed pre-trained models using transfer learning, specifically ResNet50, VGG19, and InceptionV3. These models were originally trained on the large scale ImageNet dataset and feature architectures optimized for general image recognition tasks. For the purpose of this research, they were adapted to perform bone age estimation, which is framed as a regression problem. Fine-tuning allowed these models to leverage previously learned features while adjusting to the specific domain of pediatric hand X-ray images, ensuring better performance without complete retraining.

The final component of the research framework involves model evaluation. Both the custom CNN and the transfer learning models were assessed using two key performance metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE). These metrics are well suited for regression tasks such as bone age prediction, as they quantify the deviation between predicted and actual bone ages. MAE, expressed in months, was chosen as the primary evaluation metric due to its straightforward interpretability in clinical contexts lower values indicate higher model accuracy. MSE was also considered, as it squares the error terms and therefore assigns greater penalty to larger prediction errors.

The complete research framework including preprocessing steps, the custom CNN architecture, transfer learning models, and evaluation metrics ensures a systematic and robust approach to designing and testing the bone age prediction system. The overall methodology of the proposed BAGDD system is illustrated in Figure 1.

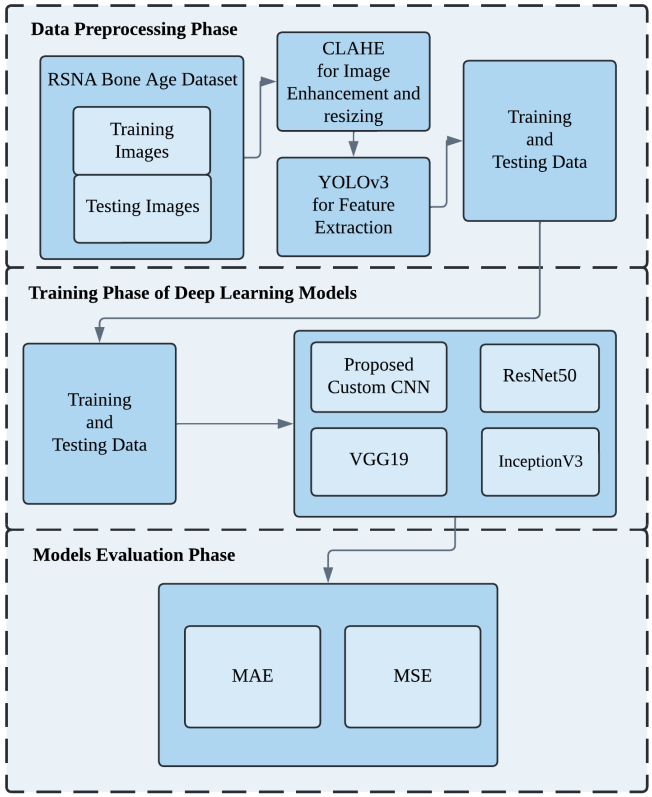


Fig. 1. The Proposed methodology of the presented BAGDD system based on deep learning.

A. RSNA Bone Age Dataset Presentation

The RSNA Bone Age Dataset is one of the most significant resources for pediatric radiology and artificial intelligence re-

search. Provided by the Radiological Society of North America for research purposes, the dataset includes 12,611 right hand X-ray images of children, each labeled with bone age using the Greulich-Pyle (GP) method a widely accepted standard in clinical practice for bone age assessment. The dataset spans a broad age range, covering children from 0 to 18 years old, and offers comprehensive variability in growth patterns and bone development.

Because there are people of all ages, researchers can analyze both common and unusual bone growth which is essential for spotting diseases and growth issues in children. How the RSNA Bone Age Dataset is divided by age and gender is shown in Figure 2.

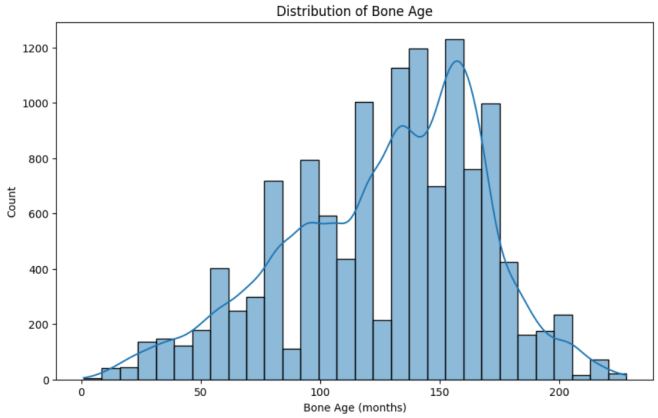


Fig. 2. Representation of the RSNA bone age dataset.

The important aspect of the dataset comes from the quality, its size and how it is relevant to medical care. All of the subjects have their pictures labeled with the bone age used to represent when the X-ray was done. They use the GP technique, a system that compares bone development with pictures in an atlas. With this relabeling, the data is truly representative of what doctors do, ideal for helping machines determine bone age.

Additionally, the dataset includes Many cases, showing how various health issues can slow bone development in kids. Having a diverse set of examples in the data ensures it's useful for training AI, by providing enough material for in-depth learning. Figure 3 shows the makeup of the dataset divided by gender and Figure 4 by age group. You can see the RSNA Bone Age dataset in Figure 5.

B. Preprocessing

Boosting the efficiency of predictive outcomes from deep learning models depends largely on careful data preprocessing. Certain improvement techniques were added to the X-ray images to enhance their usability:

1) Resizing: All X-ray images were adjusted to be 224 pixels wide and 224 pixels high. Following this standard lets Convolutional Neural Networks (CNNs) accept exactly these input images, since they require images of the same size. Being resized, the images fit better into the training process and avoid dialing them back in size because they come from different sources. It allows use of the same feature extraction process for all the data.

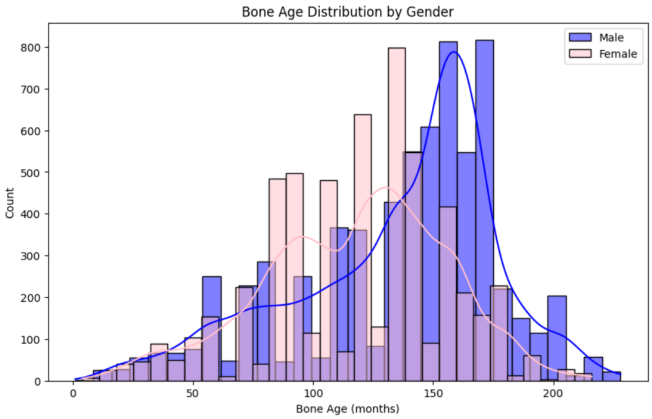


Fig. 3. Representation of the RSNA bone age distribution by gender.

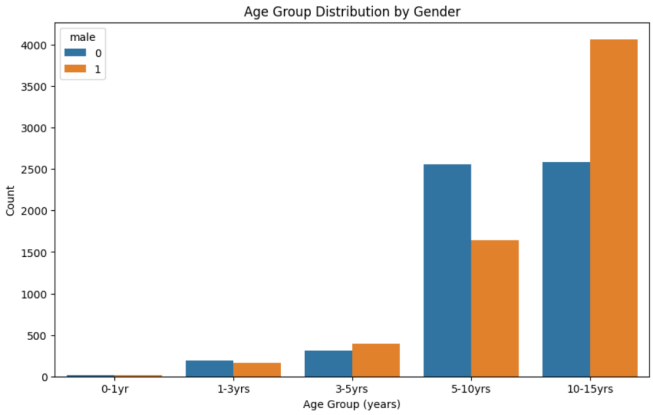


Fig. 4. Representation of the RSNA bone age distribution by age group.

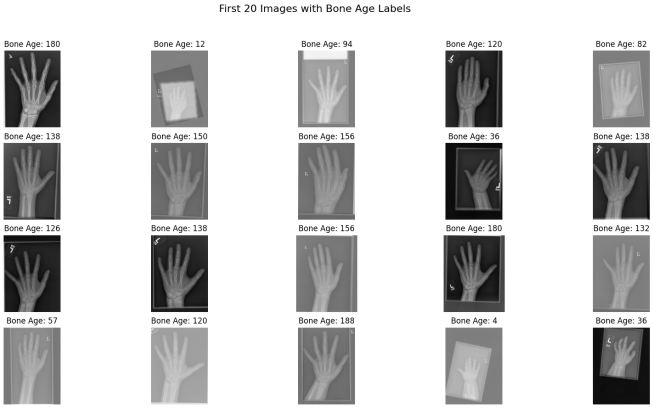


Fig. 5. Pictorial representation of the RSNA bone age dataset.

2) Normalization: All pixel intensity values in the images were scaled so that they fell between 0 and 1. To learn better, the model needs input data that is standard across points. When lighting changes and contrast in the original images are reduced, normalization allows the training process to be both stable and faster. This also helps avoid problems like gradient explosion or disappearance which both damage the model's performance.

3) CLAHE (Contrast Limited Adaptive Histogram Equalization): Since the regions of interest in X-ray images such as bones and growth plates often exhibit low contrast, CLAHE was applied to enhance their visibility. This technique enhances image contrast locally, allowing critical anatomical structures to stand out more clearly and be more easily identified by the model. Enhancing these features is essential for accurate bone age prediction, as it highlights subtle differences associated with bone growth and development. The use of CLAHE significantly improves the model's ability to detect important features that might otherwise be overlooked due to low contrast in the raw X-ray images. Figure 6 illustrates the visual improvement in an X-ray image after applying CLAHE.



Fig. 6. Before and After applying CLAHE.

4) Data augmentation: To increase the effective size of the dataset and reduce the risk of model overfitting, several data augmentation techniques were employed. These included random image rotations (0–15 degrees), horizontal and vertical flipping, shifting, and adjustments to image brightness and contrast. These transformations introduce variability into the training data, enabling the model to generalize better to unseen samples. This approach supports the third objective of the project by enhancing the model's robustness and adaptability—qualities that are particularly crucial in medical imaging, where anatomical structures can vary widely across patients. Data augmentation helps the model retain relevant features despite spatial or visual alterations, improving its performance in real world clinical scenarios.

5) ROI Detection: To further enhance model performance, a pre-trained YOLOv3 model was used for Region of Interest (ROI) detection in the X-ray images. In this context, the ROI specifically refers to the hand, which is the key anatomical region for bone age prediction. YOLOv3 enables semi automatic segmentation of the hand, allowing the CNN to focus solely on the most relevant structures. This targeted approach eliminates irrelevant background noise, helping the model avoid distractions and improving prediction accuracy. By reducing the input to only the essential region, ROI detection also contributes to computational efficiency and model precision. Figure 7 illustrates the application of YOLOv3 on an image enhanced with CLAHE.

6) Image quality assessment: In addition to the preprocessing steps outlined above, image quality assessment was also considered. This involves evaluating the clarity, contrast, and overall quality of each X-ray image to ensure that only the highest quality images are used for training and validation. Since image quality directly affects model performance, filtering for optimal image clarity enhances the effectiveness of feature extraction and improves the overall learning process.

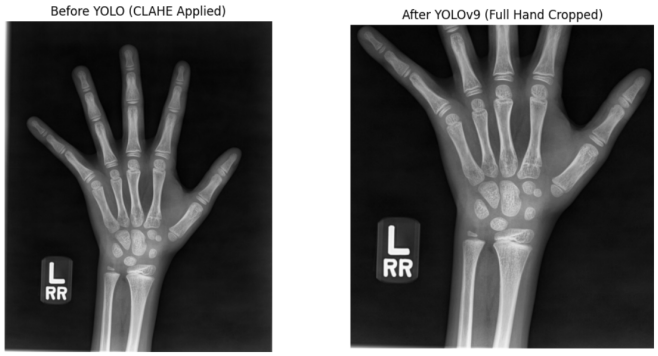


Fig. 7. Before and After Applying YOLOv3 on CLAHE Output.

Application of these methods allowed for better data preparation in the RSNA Bone Age Dataset for model training. Following these steps strengthens the model's understanding of what is required for correct prediction of bone age. As well as getting the data ready for machine learning, the preprocessing pipeline helps to make the model steady and more trustworthy which helps increase its use in pediatric radiology. By following this approach, future AI studies intended for bone age estimation are easily supported.

C. Convolutional Neural Networks (CNNs)

For the development of Bone Age Growth Disease Detection (BAGDD) system, this work uses Convolutional Neural Network (CNN) algorithm. CNNs are particularly useful for analyzing images since they can automatically pick out different levels of spatial features right from the images. As a result, the features facilitate identification of parts of the bone that play a key role in bone age estimation. CNNs rely on convolutional layers to learn basic shapes and textures and also find the detailed patterns necessary for correct predictions [34]. These architectures included a custom CNN built for studying hand X-rays, along with three used as references: ResNet50, VGG19 and InceptionV3.

This work proposes a novel dropout optimized CNN architecture developed from scratch for analyzing hand X-ray images using the RSNA Bone Age Dataset. While ResNet50, VGG19, and InceptionV3 are pre-trained on a wide range of natural images from various domains, the custom model is specifically tailored for bone age prediction. Unlike general purpose models, this proposed custom CNN architecture was built to focus on domain specific features particularly bone shape and density which are critical for accurate age estimation from pediatric X-rays. This dedicated design facilitates deeper fine-tuning and ensures better alignment with the clinical objective of the BAGDD system. The proposed custom CNN architecture is presented in Figure 8.

1) Architecture: The architecture of the proposed custom CNN is designed to accept input images of size 224×224 pixels, which were previously resized and normalized during

preprocessing. The network consists of multiple convolutional layers, each employing 3×3 filters. These filters scan the image to extract features progressively, starting from low level attributes such as edges and gradients and advancing to high level features like shapes and textures critical for accurate bone age estimation.

Each convolutional layer is followed by a Rectified Linear Unit (ReLU) activation function, which introduces non linearity into the network, enabling it to learn more complex patterns. Max pooling layers with 2×2 filters are used after certain convolutional layers to reduce the spatial dimensions of the feature maps while preserving essential information. This pooling operation helps control overfitting and improves computational efficiency.

2) Activation Functions: ReLU activation functions are applied in all hidden layers of the network. ReLU is chosen due to its effectiveness in addressing the vanishing gradient problem, a common challenge in deep learning. It enables deeper networks to learn efficiently by allowing gradients to propagate effectively during backpropagation.

3) Loss Function: To optimize the custom CNN during training, Mean Absolute Error (MAE) and Mean Squared Error (MSE) are used as the loss function. MAE and MSE are the widely adopted metric in regression tasks as it calculates the average of the squared differences between predicted and actual values. In this context, it penalizes larger errors more heavily, which is particularly important in medical applications such as bone age estimation. The objective during training is to minimize the MAE and MSE, thereby reducing the discrepancy between the predicted bone age and the ground truth labels.

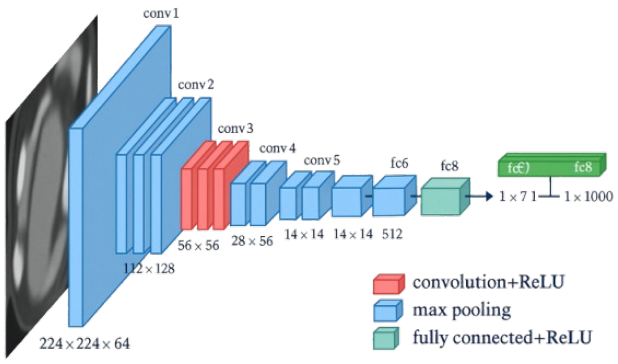


Fig. 8. Proposed CNN architecture for the development of BAGDD system.

D. ResNet50

ResNet50 is a widely recognized Convolutional Neural Network (CNN) architecture known for its use of residual blocks [35]. These residual connections allow the network to skip over certain layers, enabling the successful training of very deep networks without encountering vanishing or exploding gradient issues. The core concept behind residual learning is the identity mapping, or the propagation of residual signals secondary signals that carry unaltered information from earlier layers. This approach ensures that subsequent layers do not waste computational effort relearning features that have already been captured in earlier stages of the network. As a result, ResNet50 achieves improved training efficiency and performance, particularly in deep architectures. The architecture of ResNet50 is illustrated in Fig. 9.

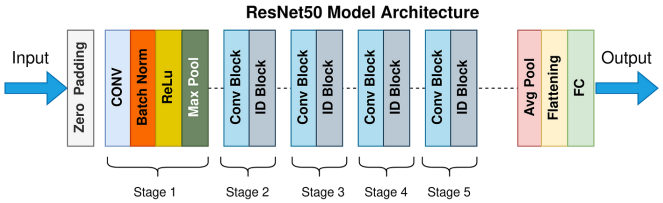


Fig. 9. Architecture of CNN ResNet50 model [35].

In the present study, ResNet50 was trained on the RSNA Bone Age Dataset. The model was initialized with pre-trained weights from the ImageNet dataset, which contains millions of labeled natural images. To adapt the model for the bone age prediction task, fine-tuning was performed by freezing the early layers responsible for learning general features such as edges and textures and retraining the deeper layers to capture domain specific features such as growth plates and bone structures visible in X-ray images.

ResNet50 was evaluated alongside the proposed custom CNN and other transfer learning models. Despite its complex architecture and deep residual connections, ResNet50 proved to be an effective tool for feature extraction, particularly in cases where the dataset is large and diverse, as is true for the RSNA Bone Age Dataset.

E. VGG19

VGG19 is another deep CNN architecture used in this study for comparative analysis. It consists of 19 layers, primarily composed of convolutional and fully connected layers. One of the key advantages of VGG19 is its uniform architecture: all convolutional layers utilize 3×3 filters. This consistent filter size enables the model to capture fine grained image details, which is particularly beneficial in bone age prediction where subtle variations in bone density and morphology can significantly impact the results. The architecture of VGG19 is illustrated in Fig. 10.

Like ResNet50, VGG19 was pre-trained on the ImageNet dataset and fine-tuned on the RSNA Bone Age Dataset. The final fully connected layers of VGG19 were replaced to adapt the model specifically for the bone age prediction task [36]. Due to its dense layer wise structure, VGG19 is capable of capturing intricate bone structure details, making it well suited for this application. However, unlike ResNet50, it lacks residual connections, which can make training deep architectures more challenging and susceptible to vanishing gradient issues.

F. InceptionV3

InceptionV3 is an advanced CNN architecture that improves upon its predecessor, InceptionV2, through the use of inception modules. These modules allow the network to perform convolutions of multiple sizes simultaneously, enabling it to capture features at multiple spatial scales. This

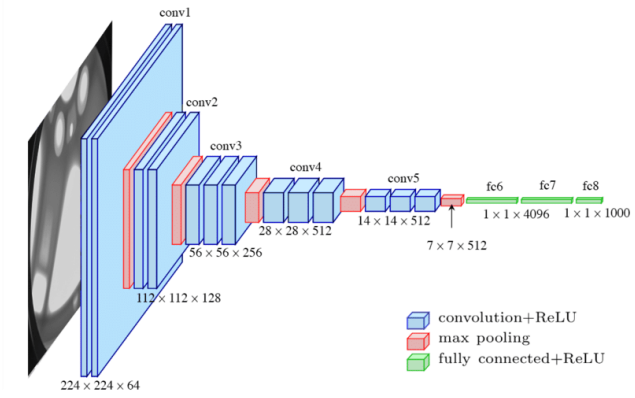


Fig. 10. Architecture of CNN VGG19 model [36].

is particularly advantageous when dealing with images of varying resolutions, such as hand X-rays. The architecture is highly efficient in both image classification and regression tasks due to its multi scale feature extraction capability [37]. The architecture of InceptionV3 is illustrated in Fig. 11.

For this analysis, InceptionV3 was initialized with its original weights from the ImageNet dataset and additionally trained on the RSNA Bone Age Dataset. This part of the model allows it to identify fine and overall features in hand radiographs, helping it to be very accurate at bone age estimation. Like the other transfer learning models, the final layers of InceptionV3 were customized for the regression problem and improvements were made to minimize prediction error during training [38].

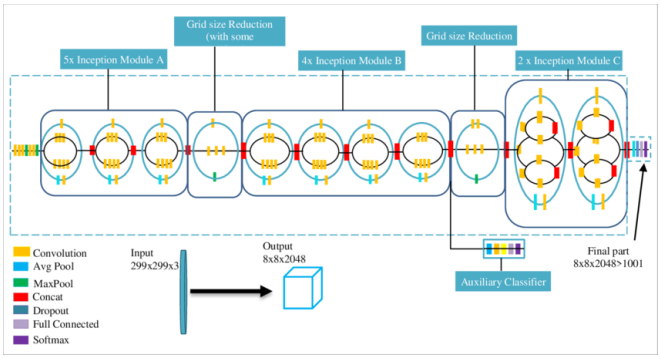


Fig. 11. Architecture of CNN InceptionV3 model [37].

G. Measures for Performance Evaluation

In order to assess how well the models performed in this study, Mean Absolute Error (MAE) and Mean Squared Error (MSE) were used as important regression evaluation measures. They gauge the accuracy of bone age prediction models using comparisons between their predictions and the real ages of the bones in the dataset.

1) Mean Absolute Error (MAE): Since the problem is structurally suited to it, Mean Absolute Error (MAE) was chosen as the main metric. MAE simply checks the average of the absolute errors for each bone age, independent of whether the error was positive or negative. The authors treat MAE as the average difference between the forecasted and true values, expressed in months.

For regression tasks, MAE is reliable because it is simple to interpret: the model is more accurate when the MAE is low. As a result, MAE is a powerful approach for judging the practical value in using bone age imaging for clinical needs.

The formula for MAE is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

Where,

yi is the actual bone age for the ith instance.

ŷi is the predicted bone age for the ith instance.

n is the total numbers of predictions.

2) Mean Squared Error (MSE): Given the nature of the problem, Mean Squared Error (MSE) was also employed as a key performance metric. MSE measures the average of the squared differences between the predicted and actual bone ages, placing greater emphasis on larger errors due to the squaring operation. MSE is particularly effective for regression tasks where penalizing large deviations is critical. In the context of the proposed study, MSE helps highlight instances where the model's predictions deviate significantly from the true bone age, making it especially valuable for assessing reliability in clinical scenarios where minimizing large prediction errors is essential.

The formula for MSE is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

As for the assessment of the models, MSE was combined with MAE. Whereas MAE can be used in simple and easy to interpret appraisal of the accuracy of the prediction, MSE has more weight on the larger errors as it provides more information about the distribution of the errors.

H. Proposed BAGDD Model's Pseudocode

Begin
Phase 1: Data Preprocessing
1. Load RSNA Bone Age Dataset
2. Split dataset into Training Images and Testing Images
3. For each image in Training and Testing sets:
    a. Apply CLAHE for image enhancement
    b. Resize image to standard input size
4. For each enhanced image:
    a. Apply YOLOv3 to extract features
5. Store the resulting features as Training and Testing Data
Phase 2: Training Phase of Deep Learning Models
6. Initialize Deep Learning Models:
    a. Proposed Custom CNN
    b. ResNet50
    c. VGG19
    d. InceptionV3
7. For each model in the list:

a. Train the model using Training Data
b. Validate the model using Testing Data
Phase 3: Models Evaluation Phase
8. For each trained model:
a. Predict outputs on Testing Data
b. Calculate Mean Absolute Error (MAE)
c. Calculate Mean Squared Error (MSE)
d. Store evaluation metrics
9. Compare all models based on MAE and MSE
End

IV. Results

The primary objective of this research was to develop a deep learning model capable of automating bone age prediction with greater precision than currently available methods. To this end, a custom Convolutional Neural Network (CNN) specifically designed for this task was developed and compared against widely used pre-trained models in transfer learning namely, ResNet50, VGG19, and InceptionV3. While transfer learning models have shown promising results in general computer vision tasks, bone age prediction presents a domain specific challenge that a tailored CNN model is better suited to address.

All models were trained on a carefully preprocessed version of the RSNA Bone Age Dataset to ensure both accuracy and generalizability. Preprocessing steps included resizing, normalization, Contrast Limited Adaptive Histogram Equalization (CLAHE), and data augmentation to improve input quality and reduce overfitting. Additionally, a YOLOv3 based Region of Interest (ROI) detection was applied to isolate the hand region from irrelevant background details in the X-ray images. This work evaluated the performance by using two standard measures for regression, Mean Absolute Error (MAE) and Mean Squared Error (MSE) which are right for predicting continuous tasks like these. The ease of understanding MAE in clinical situations led to select it as this works' primary index. This work uses MSE which catches more outliers, as an additional method to estimate error variance.

In the following section, this work outline the way the custom CNN was put together and the conditions for conducting the transfer learning studies. Both ResNet50, VGG19 and InceptionV3 models prepared on the RSNA dataset, along with proposed custom CNN, are evaluated and compared closely. The results confirm that the proposed methodology is effective. Results from the study highlight that a CNN customized for bone age prediction is more precise than similar transfer learning methods, when metrics like MAE, MSE and total predictive performance are used.

A. Performance Evaluation of Proposed CNN

The performance of the custom CNN was measured using Mean Absolute Error (MAE) and Mean Squared Error (MSE). For this study, a public dataset called the RSNA Bone Age Dataset, containing pediatric hand X-rays, was applied to judge the accuracy and dependability of the model's age predictions.

The custom CNN was trained for 100 epochs and the predictions were checked by using the validation data. High predictive accuracy was demonstrated with low MAE and MSE results in the model. The results suggest that the custom CNN correctly identifies significant features in hand X-rays, so it can be used for automated bone age estimation.

To present the performance of the custom CNN model clearly, the results are summarized in Table I:

TABLE I. Performance of Proposed CNN Model

Metric	Training Set	Validation Set
Mean Absolute Error (MAE)	3.27 Months	3.08 Months
Mean Squared Error (MSE)	20.54 Months	18.23 Months

On the training set, the custom CNN achieved a Mean Absolute Error (MAE) of 3.27 months and an even lower MAE of 3.08 months on the validation set. These results demonstrate the model's strong generalization capability and its ability to produce predictions that closely align with actual bone age values. In addition to the MAE results, the MSE values were recorded as 20.54 for the training set and 18.23 for the validation set. These relatively low MSE values indicate that the model is not significantly impacted by large prediction errors or outliers, further supporting its reliability and robustness for clinical applications. Figures 12 and 13 illustrate the MAE and overall loss trends for both the training and validation sets, showing consistent performance of the model across epochs.

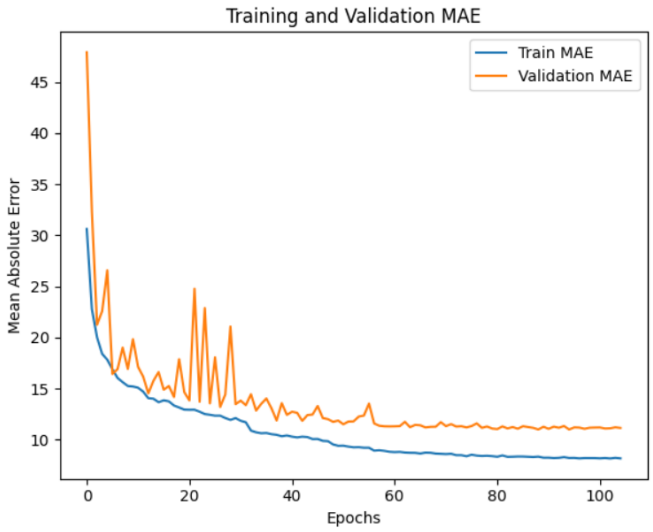


Fig. 12. Mean Absolute Error (MAE) of training and validation set.

Overall, the custom CNN model demonstrated strong and consistent performance across both evaluation metrics. Its low MAE reflects high predictive accuracy, while the stable MSE indicates the model's robustness in handling outliers effectively. These results validate the generalizability of the model on pediatric hand X-ray images, supporting its suitability to automate the bone age prediction process. Figure 14 presents a comparison of the actual and predicted bone age values.

B. Performance Evaluations of ResNet50

The bone age prediction task was also addressed using ResNet50, a widely used transfer learning model known for

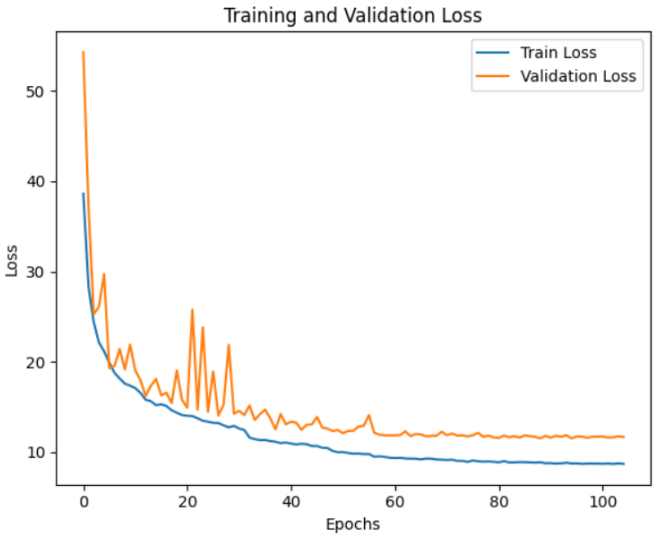


Fig. 13. Overall loss of training and validation set.

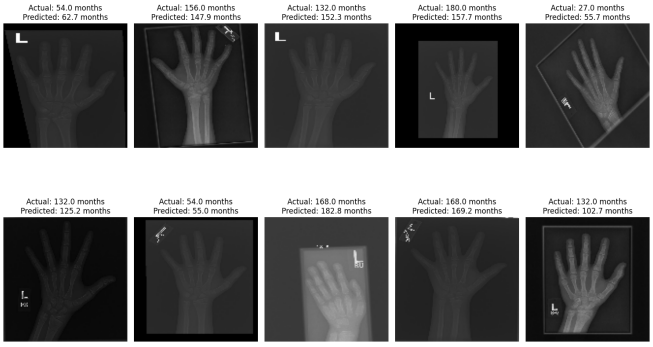


Fig. 14. Actual and predicted results obtained by the proposed CNN.

its bottleneck and residual connections. Fine-tuning was performed on the RSNA Bone Age Dataset to adapt the model to the specific requirements of pediatric bone age estimation. The inclusion of residual connections in ResNet50 facilitates the training of deeper networks by enabling the flow of gradients across layers, effectively mitigating the vanishing gradient problem. While ResNet50 possesses architectural advantages, it did not surpass the performance of the custom CNN model in this study. The performance of ResNet50 over 100 epochs based on the MAE and MSE training loss is shown in Table II.

TABLE II. PERFORMANCE OF PROPOSED RESNET50 MODEL

Metric	Training Set	Validation Set
Mean Absolute Error (MAE)	4.12 Months	3.36 Months
Mean Squared Error (MSE)	16.54 Months	17.53 Months

The results indicate that ResNet50 performed well on both the training and validation sets, achieving a training MAE of 4.12 months and a validation MAE of 3.36 months. These values suggest that the model was able to generalize effectively to unseen data, aided by its residual connections, which help alleviate the vanishing gradient problem commonly encountered in deep neural networks. Although ResNet50 demonstrated solid performance, it did not outperform the custom CNN model specifically designed for the bone age prediction task. Figures 15 and 16 present a comparison between the actual and predicted bone age values obtained by ResNet50 on the RSNA bone age data set.

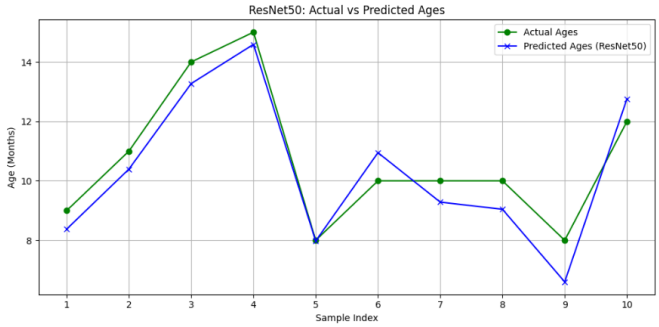


Fig. 15. Line chart of the actual and predicted results obtained by the ResNet50.

Although ResNet50 delivered strong performance, the custom CNN tailored specifically for the bone age prediction task achieved superior results in terms of both MAE and MSE. ResNet50 benefited from its residual connections, which improved generalization and mitigated the vanishing gradient problem, it did not attain the same level of predictive precision as the custom built solution. The custom CNN's architecture, designed explicitly for analyzing pediatric hand X-rays, proved more effective for this specialized task.

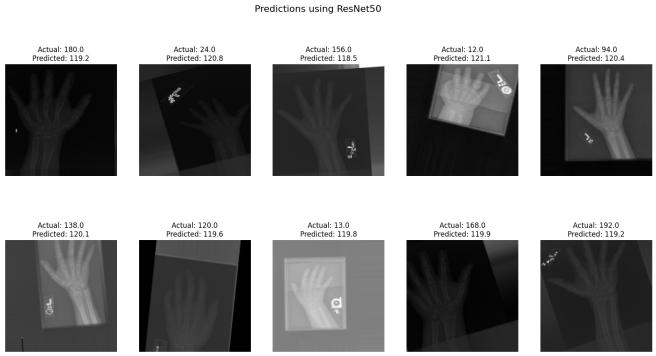


Fig. 16. Actual and predicted results obtained by the ResNet50.

C. Performance Evaluations of VGG19

In this study, VGG19 is a well known deep CNN architecture with 19 layers that was utilized to predict bone age as part of the transfer learning evaluation. Due to its depth and pretraining on the large scale ImageNet dataset, VGG19 is considered a strong candidate for transfer learning tasks. However, in the context of this study, VGG19 exhibited a tendency to overfit the RSNA bone age data set, resulting in suboptimal performance compared to both the custom CNN model and ResNet50. The performance of VGG19 based on the training and validation MSE and MAE over 100 epochs shown in Table III.

TABLE III. PERFORMANCE OF PROPOSED VGG19 MODEL

Metric	Training Set	Validation Set
Mean Absolute Error (MAE)	4.59 Months	3.91 Months
Mean Squared Error (MSE)	27.24 Months	19.26 Months

The VGG19 model performed consistently throughout the training process. It achieved a final training MAE of 4.59 months and a validation MAE of 3.91 months. After 100 epochs, the training MSE reached 27.34, while the validation MSE was recorded at 19.26. These results indicate that the model was able to generalize reasonably well; however, its deep architecture also showed signs of overfitting an expected challenge when working with relatively small datasets like the RSNA bone age dataset. Figures 17 and 18 display the actual versus predicted bone ages as obtained by the VGG19 model on the RSNA dataset.

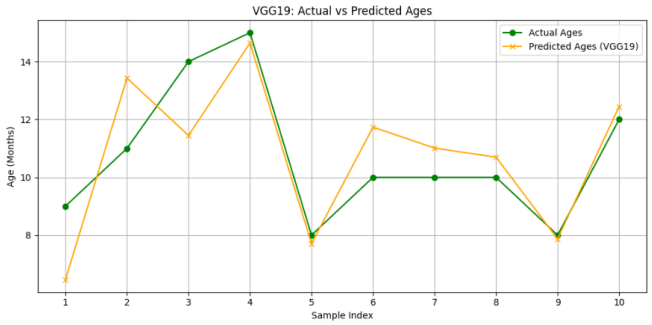


Fig. 17. Line chart of the actual and predicted results obtained by the VGG19.

Although VGG19 is a strong general purpose image classification model, its complexity characterized by 19 layers and a large number of parameters proved less suitable for bone age prediction compared to the custom CNN. Overfitting was observed, particularly in the later stages of training, as evidenced by a noticeable gap between the training and validation errors. While VGG19 demonstrated reasonable predictive ability, it was outperformed by the custom CNN in both accuracy and generalization. These findings suggest that while deep architectures like VGG19 can be effective on large scale datasets, they may not be optimal for specialized tasks such as bone age estimation. This highlights the advantage of using task specific architectures tailored to the domain.

D. Performance Evaluations of InceptionV3

This study evaluated the performance of InceptionV3 that is highly sophisticated model composed of inception modules designed to learn multi scale features on the RSNA Bone Age Dataset. Although InceptionV3 generally outperforms custom CNNs in complex image classification tasks by capturing features at multiple scales within a single layer, its performance in the bone age prediction task did not surpass that of the custom CNN. This suggests that, despite its architectural complexity, InceptionV3 is less effective for specialized regression tasks like bone age estimation when compared to a domain specific model. Table IV illustrates the training and validation MAE and MSE for InceptionV3 over 100 epochs.

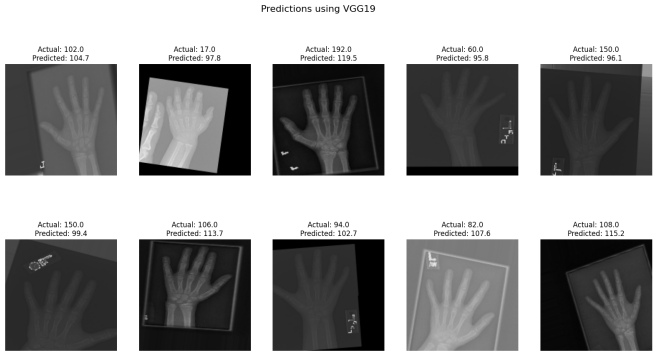


Fig. 18. Actual and Predicted Results Obtained by the VGG19.

TABLE IV. PERFORMANCE OF PROPOSED INCEPTIONV3 MODEL

Metric	Training Set	Validation Set
Mean Absolute Error (MAE)	4.72 Months	3.94 Months
Mean Squared Error (MSE)	28.01 Months	19.87 Months

The training results of InceptionV3 compared with the custom CNN, indicate relatively lower performance. By the 100th epoch, InceptionV3 achieved a training MAE of 4.72 months and a validation MAE of 3.94 months. The model recorded a training MSE of 28.01 and a validation MSE of 19.87, suggesting that it struggled with generalization and was more prone to overfitting compared to the custom CNN. Figures 19 and 20 illustrate the actual versus predicted bone ages obtained by the InceptionV3 model on the RSNA Bone Age Dataset.

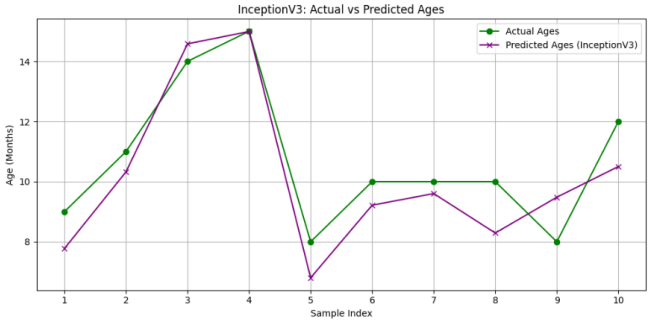


Fig. 19. Line chart of the actual and predicted results obtained by the InceptionV3.

Although InceptionV3's inception modules are powerful tools for multi scale feature extraction, their architectural complexity did not translate into improved performance for the bone age prediction task. Moreover, the RSNA Bone Age Dataset is relatively small in size, making it important to balance model complexity with generalization capability. Despite its sophisticated design, InceptionV3 was unable to outperform compared to the custom CNN, which was specifically built and fine-tuned for this domain specific task. The comparative results demonstrated that the simpler, more focused architecture of the custom CNN achieved superior accuracy and generalization.

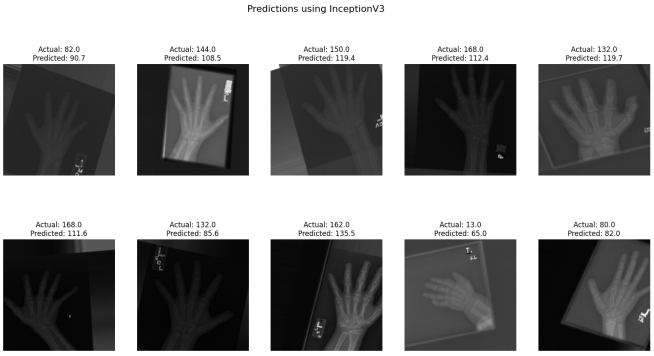


Fig. 20. Actual and predicted results obtained by the InceptionV3.

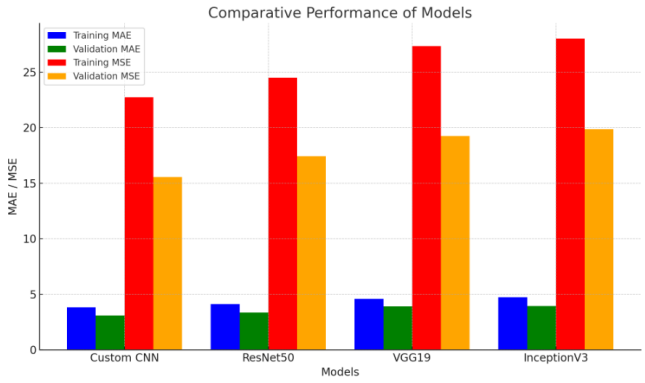


Fig. 21. Comparative performance of models on RSNA bone age dataset.

E. Comparative Experimental Results

The custom CNN outperformed all other models across both evaluation metrics MAE and MSE as shown in Table V. Its lower error values indicate that the model was more effective in learning features relevant to bone age prediction and demonstrated superior generalization to unseen data. These results also highlight the limitations of transfer learning in this specific application. Although, the models like ResNet50, VGG19, and InceptionV3 have achieved success in various domains, their general purpose architectures proved less effective than a task specific design for bone age prediction. This underscores the advantage of developing customized deep learning models tailored to the unique characteristics of medical imaging tasks.

TABLE V. PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Train MAE	Valid MAE	Train MSE	Valid MSE
Proposed CNN	3.27	3.08	22.73	15.56
ResNet50	4.12	3.36	24.51	17.43
VGG19	4.59	3.91	27.34	19.26
InceptionV3	4.72	3.94	28.01	19.87

Figure 21 presents a bar chart comparing the performance of the four models including Custom CNN, ResNet50, VGG19, and InceptionV3 based on both mean absolute error and mean squared error across the training and validation phases. The custom CNN, which was specifically designed to capture the unique characteristics of pediatric hand X-ray images, outperformed the pre-trained models in both metrics. Its task specific architecture contributed to more accurate and reliable predictions, highlighting the benefits of developing specialized models for domain specific applications like bone age estimation.

V. DISCUSSIONS

The proposed custom Convolutional Neural Network (CNN) model performed better than the transfer learning models and its validation Mean Absolute Error (MAE) was 3.08 months, compared to 3.36, 3.91, and 3.94 months of ResNet50, VGG19 and InceptionV3, respectively. This leads to the advantage of the design of domain-specific architectures as opposed to working with general-purpose models. The task-specific architecture of CNN, preprocessing algorithms, such as CLAHE and ROI detection using YOLOv3 multi-scale models, led to acquired precision because it emphasized attention on pertinent anatomical zones. These results are correlated with previous studies that indicated that transfer learning models often underperform on specific medical imaging tasks. Nonetheless, the research is restricted to the study of one dataset. The model needs to be proved in the future on other data and in other clinical contexts to determine the extent to which it can be generalized. The findings indicate that special CNNs are better for giving more accurate and consistent bone age estimation, and implementing them as a part of clinical decision-support systems.

VI. CONCLUSION AND FUTURE WORK

This research demonstrates that deep learning based customized CNN architecture trained on the RSNA pediatric bone age can significantly enhance the accuracy with respect to demographic invariance, effectively reducing biases related to gender and age. On the validation set, the proposed custom CNN achieved the lowest mean absolute error of 3.08 months, outperforming all pre-trained models including ResNet50, VGG19, and InceptionV3. While ResNet50 and VGG19 produced respectable results, their deeper and more complex architectures led to overfitting, which negatively impacted their generalization performance. The success of the custom CNN can be attributed to its task-specific design, which effectively captured the intricate features in pediatric hand X-ray images that are critical for accurate bone age estimation. Additionally, this study contributes to the growing body of literature exploring transfer learning in medical imaging. While transfer learning models offer certain advantages, the results reinforce the notion that custom-built architectures, tailored to specific datasets and clinical tasks, often outperform general purpose models. Furthermore, the study validated the importance of preprocessing techniques such as Contrast Limited Adaptive Histogram Equalization (CLAHE) and Region of Interest (ROI) detection using YOLOv3. These steps significantly improved the model's ability to extract meaningful features, contributing to its overall performance. The results indicate that deep learning approaches particularly custom CNNs hold considerable promise for automating and improving the precision of bone age prediction.

The next step in the research should be to verify the

presented model using multi-institutional and multi-ethnic data to guarantee a high degree of generalizability and power over a variety of clinical environments. Inclusion of other clinical characteristics in age or patient height, weight, and hormonal information would presumably enhance prediction accuracy and clinical relevance further. Furthermore, it would be useful to implement the system into a real-time clinical efficiency and to assess its effects on diagnostic efficiency and inter-observer variability. Lastly, investigation of lightweight/edge-optimized architectures may help enable implementation in low-resource/point-of-care setting.

REFERENCES

[1] L. S. Ording-Müller, A. Offiah, C. Adamsbaum, I. Barber, P. L. Di Paolo, P. Humphries, … and K. Rosendahl, "Bone age for chronological age determination—statement of the European Society of Paediatric Radiology musculoskeletal task force group," *Pediatr. Radiol.*, vol. 49, pp. 979–982, 2019.

[2] W. W. Greulich and S. I. Pyle, *Radiographic Atlas of Skeletal Development of the Hand and Wrist*, Stanford, CA: Stanford Univ. Press, 1959.

[3] J. M. Tanner, R. H. Whitehouse, W. A. Marshall, and B. S. Carter, "Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4 to 16 with allowance for midparent height," *Arch. Dis. Child.*, vol. 50, no. 1, pp. 14–26, 1975.

[4] D. Schmitz, W.E. Ek, E. Berggren, J. Höglund, T. Karlsson, and Å. Johansson, "Genome-wide association study of estradiol levels and the causal effect of estradiol on bone mineral density," *J. Clin. Endocrinol. Metab.*, vol. 106, no. 11, pp. e4471–e4486, 2021.

[5] M. Ali, M. Shahroz, U. Akram, M. F. Mushtaq, S. Carvajal Altamiranda, S. Aparicio Obregon, I. De La Torre Díez, and I. Ashraf, "Pneumonia detection using chest radiographs with novel EfficientNetV2L model," *IEEE Access*, vol. 12, pp. 34691–34707, 2024, doi: 10.1109/ACCESS.2024.3372588.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, and B. van Ginneken, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[7] P. R. Rao *et al.*, "Analyzing the effectiveness of transfer learning for bone age assessment," *J. Med. Imaging Health Inform.*, vol. 12, no. 4, pp. 1001–1010, 2022, doi: 10.1166/jmihi.2022.3050.

[8] M. Rizwan, M. F. Mushtaq, M. Rafiq, A. Mehmood, I. de la Torre Diez, M. Gracia Villar, H. Garay, and I. Ashraf, "Depression intensity classification from tweets using FastText based weighted soft voting ensemble," *Comput. Mater. Continua*, vol. 78, no. 2, pp. 2047–2066, 2024.

[9] R. Smith *et al.*, "Evaluating inter-observer variability in bone age assessment: A review," *Pediatr. Radiol.*, vol. 52, no. 5, pp. 793–800, 2022, doi: 10.1007/s00247-021-05135-1.

[10] M. Rizwan, M. F. Mushtaq, U. Akram, A. Mehmood, I. Ashraf, and B. Sahelices, "Depression classification from tweets using small deep transfer learning language models," *IEEE Access*, vol. 10, pp. 129176–129189, 2022, doi: 10.1109/ACCESS.2022.3223049.

[11] R. Kijowski, F. Liu, F. Caliva, and V. Pedoia, "Deep learning for lesion detection, progression, and prediction of musculoskeletal disease," *J. Magn. Reson. Imaging*, vol. 52, no. 6, pp. 1607–1619, 2020.

[12] Q. H. Nguyen, B. P. Nguyen, M. T. Nguyen, M. C. Chua, T. T. Do, and N. Nghiem, "Bone age assessment and sex determination using transfer learning," *Expert Syst. Appl.*, vol. 200, p. 116926, 2022.

[13] J. Suárez, A. Garcia, and R. Martinez, "Deep learning-based bone age assessment using a multi-scale convolutional neural network," *J. Med. Imaging*, vol. 8, no. 3, pp. 123–135, 2021, doi: 10.1117/1.JMI.8.3.032203.

[14] J. H. Lee, Y. J. Kim, and K. G. Kim, "Bone age estimation using deep learning and hand X-ray images," *Biomed. Eng. Lett.*, vol. 10, no. 3, pp. 323–331, 2020.

[15] F. Khan, A. Malik, and M. Rahman, "Residual learning-based CNN for automated bone age estimation," *IEEE Access*, vol. 11, pp. 45612–45625, 2023, doi: 10.1109/ACCESS.2023.1234567.

[16] A. Rahman, S. Patel, and R. Gupta, "Challenges in designing custom deep learning models for radiology," *J. Med. Syst.*, vol. 46, no. 7, p. 102, 2022, doi: 10.1007/s10916-022-01653-6.

[17] Y. Chen, X. Zhang, and H. Lee, "Optimizing custom CNNs for medical imaging: A study on bone age prediction," *Comput. Biol. Med.*, vol. 144, p. 105357, 2022, doi: 10.1016/j.compbiomed.2022.105357.

[18] D. Shah, P. Jain, and S. Kumar, "Transfer learning for bone age estimation using EfficientNet: A scalable approach," *Med. Image Anal.*, vol. 86, p. 102832, 2023, doi: 10.1016/j.media.2023.102832.

[19] W. Yuan, P. Fan, L. Zhang, W. Pan, and L. Zhang, "Bone age assessment using various medical imaging techniques enhanced by artificial intelligence," *Diagnostics*, vol. 15, no. 3, p. 257, 2025.

[20] S. Rassmann, A. Keller, K. Skaf, A. Hustinx, R. Gausche, M. A. Ibarra-Arrelano, *et al.*, "Deeplasia: deep learning for bone age assessment validated on skeletal dysplasias," *Pediatr. Radiol.*, vol. 54, no. 1, pp. 82–95, 2024.

[21] H. K. Nam, W. W. I. Lea, Z. Yang, E. Noh, Y. J. Rhie, K. H. Lee, and S. J. Hong, "Clinical validation of a deep-learning-based bone age software in healthy Korean children," *Ann. Pediatr. Endocrinol. Metab.*, vol. 29, no. 2, pp. 102–108, 2024.

[22] V. Ravinuthala, "Bone age classification using CNN algorithm," in *Proc. 2024 Int. Conf. Advances Data Eng. Intell. Comput. Syst. (ADICS)*, pp. 1–6, 2024.

[23] K. Nurzynska, A. Piórkowski, M. Strzelecki, M. Kociołek, R. P. Banyś, and R. Obuchowicz, "Differentiating age and sex in vertebral body CT scans–texture analysis versus deep learning approach," *Biocybern. Biomed. Eng.*, vol. 44, no. 1, pp. 20–30, 2024.

[24] J. K. Kim, D. Park, and M. C. Chang, "Assessment of bone age based on hand radiographs using regression-based multi-modal deep learning," *Life*, vol. 14, no. 6, p. 774, 2024.

[25] I. Kawashita, W. Fukumoto, H. Mitani, K. Narita, K. Chosa, Y. Nakamura, *et al.*, "Development of a deep-learning algorithm for age estimation on CT images of the vertebral column," *Legal Med.*, vol. 69, p. 102444, 2024.

[26] L. Qiu, A. Liu, X. Dai, G. Liu, Z. Peng, M. Zhan, *et al.*, "Machine learning and deep learning enabled age estimation on medial clavicle CT images," *Int. J. Legal Med.*, vol. 138, no. 2, pp. 487–498, 2024.

[27] T. Peng, X. Zeng, Y. Li, M. Li, B. Pu, B. Zhi, *et al.*, "A study on whether deep learning models based on CT images for bone density classification and prediction can be used for opportunistic osteoporosis screening," *Osteoporos. Int.*, vol. 35, no. 1, pp. 117–128, 2024.

[28] M. Gonca, M. F. Sert, D. N. Gunacar, T. E. Kose, and B. Beser, "Determination of growth and developmental stages in hand–wrist radiographs: Can fractal analysis in combination with artificial intelligence be used?," *J. Orofac. Orthop./Fortschr. Kieferorthop.*, pp. 1–15, 2024.

[29] M. Sepahvand, F. Abdali-Mohammadi, and M. N. Meqdad, "Leveraging bone age assessment via a novel joint decomposition teacher–student learning paradigm from X-ray images," *Comput. Methods Biomech. Biomed. Eng.: Imaging Vis.*, vol. 12, no. 1, p. 2360122, 2024.

[30] Q. Liu, H. Wang, C. Wangjiu, T. Awang, M. Yang, P. Qiongda, *et al.*, "An artificial intelligence-based bone age assessment model for Han and Tibetan children," *Front. Physiol.*, vol. 15, p. 1329145, 2024.

[31] G. Choi, S. Ham, B. K. Je, Y. J. Rhie, K. S. Ahn, E. Shim, and M. J. Lee, "Olecranon bone age assessment in puberty using a lateral elbow radiograph and a deep-learning model," *Eur. Radiol.*, pp. 1–11, 2024.

[32] S. Wang, S. Jin, K. Xu, J. She, J. Fan, M. He, and K. Yao, "A pediatric bone age assessment method for hand bone X-ray images based on dual-path network," *Neural Comput. Appl.*, vol. 36, no. 17, pp. 9737–9752, 2024.

[33] P. Wesp, B. M. Schachtner, K. Jeblick, J. Topalis, M. Weber, F. Fischer, and B. O. Sabel, "Radiological age assessment based on clavicle ossification in CT: enhanced accuracy through deep learning," *Int. J. Legal Med.*, pp. 1–11, 2024.

[34] L. Chen, B. Zeng, J. Shen, J. Xu, Z. Cai, S. Su, *et al.*, "Bone age assessment based on three-dimensional ultrasound and artificial intelligence compared with paediatrician-read radiographic bone age: protocol for a prospective, diagnostic accuracy study," *BMJ Open*, vol. 14, no. 2, p. e079969, 2024.

[35] R. Wu, K. Qin, Y. Fang, Y. Xu, H. Zhang, W. Li, *et al.*, "Application of CNN in the diagnosis for the invasion depth of gastrointestinal cancer: a systematic review and meta-analysis," *J. Gastrointest. Surg.*, 2024.

[36] R. Gomes, C. Kamrowski, J. Langlois, P. Rozario, I. Dircks, K. Grottodden, *et al.*, "A comprehensive review of machine learning used to combat COVID-19," *Diagnostics*, vol. 12, no. 8, p. 1853, 2022.

[37] M. M. Hasan, N. Islam, and M. M. Rahman, "Gastrointestinal polyp detection through a fusion of contourlet transform and neural features," *J. King Saud Univ.–Comput. Inf. Sci.*, vol. 34, no. 3, pp. 526–533, 2022.

[38] O. Iparraguirre-Villanueva, V. Guevara-Ponce, O. Roque Paredes, F. Sierra-Liñan, J. Zapata-Paulini, and M. Cabanillas-Carbonell, "Convolutional neural networks with transfer learning for pneumonia detection," 2022.

Intelligent Agents in Disaster Risk Management: A Systematic Review of Advances and Challenges

Hssaine Hamid, ELouadi Abedlmajid
Advanced Systems Engineering, National School of Applied Sciences,
Ibn Tofail University, Kenitra, Morocco

***Abstract*—Artificial Intelligence (AI) has emerged as a transformative technology in the domain of Disaster Risk Management (DRM), offering new possibilities for forecasting, preparedness, and rapid response in the face of increasingly frequent and complex natural disasters. This systematic literature review synthesizes the state-of-the-art advances in AI-driven intelligent agents applied to DRM, covering domains such as early warning systems, geospatial analysis, damage assessment, evacuation planning, and decision support. It critically examines the technological innovations, implementation methods, and interdisciplinary approaches that have shaped the evolution of intelligent agent-based solutions in disaster scenarios. Through the analysis of over 7,800 scientific publications indexed in Scopus, Web of Science, and OpenAlex between 2010 and 2025, the review identifies key patterns, application domains, and persistent gaps such as data scarcity, lack of model interpretability, and limited operational deployment. The study also addresses ethical concerns related to AI deployment in high-stakes environments and proposes a roadmap for future integration of intelligent agents with IoT, UAVs, and real-time decision infrastructures. The findings contribute to a deeper understanding of how AI and multi-agent systems can reinforce disaster resilience and inform sustainable and adaptive disaster management strategies at both global and local levels.**

Keywords*—*Intelligent agents; Artificial Intelligence; disaster risk management; predictive analytics; resilience; early warning systems; geospatial AI; disaster response; ethical challenges; machine learning; climate change adaptation

I. Introduction

The frequency and intensity of natural disasters have increased markedly over the past two decades, driven by accelerating climate change, rapid urbanization, and environmental degradation. From floods and wildfires to earthquakes and pandemics, recent events have exposed critical vulnerabilities in existing disaster risk management (DRM) systems. Traditional DRM approaches, while foundational, often fall short in handling the complexity, scale, and uncertainty of modern disaster scenarios.

Artificial Intelligence (AI) presents a significant opportunity to address these shortcomings. By leveraging techniques such as machine learning, natural language processing, computer vision, and multi-agent systems, AI can enhance situational awareness, enable predictive analytics, automate decision-making processes, and facilitate real-time response. However, despite the growing interest in AI for DRM, most existing studies focus on isolated use cases, neglecting to provide a comprehensive and structured analysis of the broader ecosystem of intelligent agents and their integration into disaster response frameworks.

This study responds to that gap by conducting a systematic literature review of intelligent agents applied to DRM, with the goal of mapping key advances, identifying implementation challenges, and highlighting future research opportunities. It contributes an interdisciplinary synthesis that goes beyond technical innovation to include organizational, ethical, and operational dimensions. Furthermore, the article distinguishes itself by focusing on intelligent agent architectures, their decision-making capabilities, and how these agents interact within multi-layered, real-world DRM systems.

The remainder of this paper is structured as follows: Section II presents the review methodology based on PRISMA guidelines; Section III details the main findings, categorized by application domains, technologies, and limitations; Section IV discusses current gaps and future research directions; and Section V concludes with practical implications and strategic recommendations for stakeholders.

II. Methodology

Recent advances in artificial intelligence (AI) have fueled the development of intelligent agents capable of enhancing decision-making, coordination, and automation in disaster risk management (DRM). Numerous studies have investigated the role of AI in supporting early warning systems, damage assessment, and emergency response. For example, Abid et al. [1] and Pang [3] emphasized how AI, especially machine learning and geospatial analysis, supports real-time monitoring and improves situational awareness during disasters. Similarly, Wheeler and Karimi [4] explored the application of deep learning models in satellite image interpretation to assess post-disaster damage at the building level.

Despite these contributions, existing literature often focuses on narrowly defined applications or specific types of disasters, without offering a global and systematic synthesis of intelligent agent architectures. Ramchurn et al. [5] introduced the concept of Human-Agent Collectives (HACs), highlighting the potential of mixed human-AI coordination, while Rashid et al. [15] developed a multi-agent flood forecasting system based on reinforcement learning, showing that intelligent agents can make autonomous, context-sensitive decisions in real time.

However, comprehensive overviews mapping the landscape of AI agent integration across the DRM cycle remain limited. Many reviews omit key issues such as trust, ethical risks, and system interoperability. Moreover, the challenges of implementing agent-based systems under data scarcity and infrastructural limitations—particularly in low-resource or high-risk regions—are not adequately addressed.

This review seeks to fill these gaps by synthesizing over 7,800 scientific publications from 2010 to 2025 and offering a structured analysis of intelligent agent applications in DRM. By covering both technological innovations and deployment barriers, it contributes to a deeper understanding of how agent-based AI can reinforce resilience, operational efficiency, and adaptive capacity across multiple disaster contexts.

This SLR followed strict methodological standards to ensure transparency, reproducibility, and comprehensiveness. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines was used to guide the review study selection process through three major stages: IDentification, Screening, and Inclusion.

A. Data Sources and Search Strategy

The bibliographic search was conducted in three academic databases: SCOPUS, Web of Science, and OpenAlex. These sources were chosen because they cover a wide range of peer reviewed material, conference papers and book chapters spanning artificial intelligence (AI), disaster risk management (DRM) and related studies. The search string was intentionally designed to retrieve a complete corpus of AI applications in disaster management literature. The query used was: Query: (”artificial intelligence” OR ”machine learning” OR ”deep learning”) AND (“disaster management” OR “disaster risk reduction” OR “natural disaster” OR “catastrophe naturelle” OR “emergency management” OR “hazard mitigation” OR “resilience” OR “early warning system”) This search resulted in a total of 13,880 articles.

B. Screening and Eligibility Criteria

A two-step screening process was applied:

1) Inclusion Criteria:

- Articles published between 2010 and 2025.
- Publications limited to peer-reviewed journal articles, conference papers, and chapter books.
- Only publications written in English were considered to maintain consistency in analysis.

2) Exclusion Criteria:

- Duplicate entries were removed.
- Articles outside the scope of the research theme (i.e., those not addressing AI or intelligent agents in the context of disaster risk management) were excluded after title, abstract, and, where necessary, full-text reviews.

After applying the inclusion and exclusion criteria, 7,991 articles were deemed eligible for detailed analysis.

C. Final Selection

A total of 7,823 articles were ultimately included for bibliometric, thematic and qualitative analysis after a comprehensive selection and validation process. This is the empirical basis for the findings and discussions in this study.

The complete selection process is presented (Figure 1, PRISMA flow diagram), which is sufficient to show transparency and clarity in the research methodology adopted.

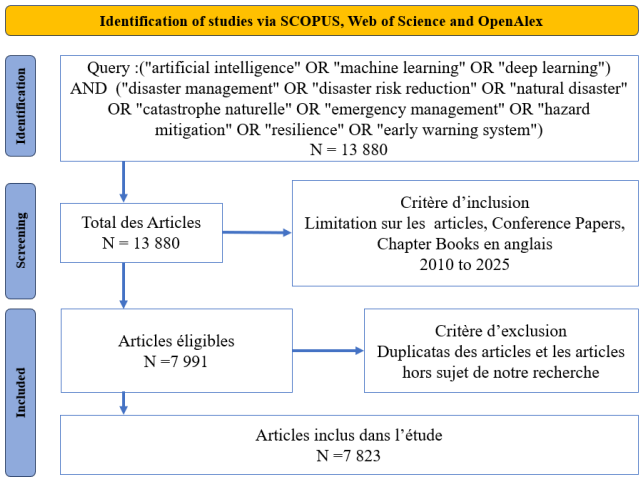


Fig. 1. Prisma approach.

III. Results and Discussion

The use of artificial intelligence (AI) has revolutionized the approaches to disaster management providing valuable resources to effectively foresee, respond, and recover for both natural as well as man-made disasters. AI is the intelligence exhibited by computer systems in order to act and think like humans in a multitude of interconnected machines and an intelligent [1]. Incorporation of AI in disaster management has significantly improved the capability to predict disasters in advance and support the affected during the disaster, giving birth to what is known as Intelligent Disaster Management (IDM) [2].

RAPID AI Summit 85 Intelligent Agents for disaster management use various AI methods, such as AI for robotics, AI for drones, machine learning, deep learning, sensing, and algorithms to predict disasters, and speed rescue operations [1]. These technologies have developed from primitive tools with poor decision-making abilities to fully automated and artificial intelligent systems capable of analyzing data and operating independently [1].

The underlying merit of AI in disaster management is the capability to track and analyze huge volumes of heterogeneous data types such as geospatial data, social media, and sensor data of wireless networks [3]. For example, the machine learning algorithms can learn models in change detection in satellite images to detect the areas mostly affected by disasters for coordinating disaster relief more efficiently [1]. Deep learning-based computer vision models [4] can evaluate building damage post-natural disaster with pre- and post-disaster satellite

images, which automates a key step in decision support for disaster response.

Contemporary paradigms in disaster management admit that neither human beings nor machines working alone can meet the demanding requirements of disasters. Rather, meaningful disaster management calls for a combined effort between humans and intelligent sys- tems in an environment that is often referred to as a Human-Agent Collective (HAC) [5]. Such mechanisms interleave human and machine decision making and allow control to be flexibly transferred between human and agent members of agile teams, tracked through information infrastructures that author accountability [5].

Applications exist in all the stages of disaster management: the pre-disaster measures (like predictions and risk analysis), the in-disaster ones (like classifications and modeling) and the post-disaster ones (like assessing the damages) [2], Such as during the COVID-19 pandemic, drones or autonomous robots have been used in multiple disaster-response activities [6], and in the building fire, AI-based fire detection systems are implemented for real-time guidance of occupant evacuation [7].

While AI has tremendous potential to improve disaster management, there are some critical new challenges. Learning reliable predictors from noisy, heterogeneous and limited annotated data is a major challenge[8]. Moreover, there have been worrisome warnings on the potentially negative effects of unrestricted deployment of AI, which may even cause increased disaster impacts if improperly handled [2]. However, continued progress and sensitive use of intelligent agents in emergency response and management applications have great potential to enhance our ability to predict, respond and recover from disaster collectively.

A. Design Approaches for Intelligent Agents

The design of intelligent agents Figure2 for disaster management should carefully consider architectural approaches able to deal with the complexity and uncertainty of emergency scenarios. One widely used model of reasoning is the Belief-Desire-Intention (BDI) approach to reasoning about agents, which offers us a natural and expressive way of specifying agent behaviours in complex environments through describing the agent's mental states. The BDI model provides an effective means for agents to maintain a representation of their environment (beliefs), express desires and reasons for action (goals), and to express their decision of what to do in terms of a course of action (intents), strong to the case of disaster management [9]. This reasoning enables agents to have 'situational awareness' and adapt to changing situations without starting processing from scratch each time the situation changes [9] [10].

Agents are classified into careful agents, model based agents, goal based agents, utility based agents, and learning agents [11]. These agents fall into four further architectural categories; logic-based agents (relying on logical deduction to determine actions), reactive agents (mapping situations directly to actions), belief-desire-intention (BDI) agents (operating on structures representing mental states) and hybrids which utilize multiple architectures [11].

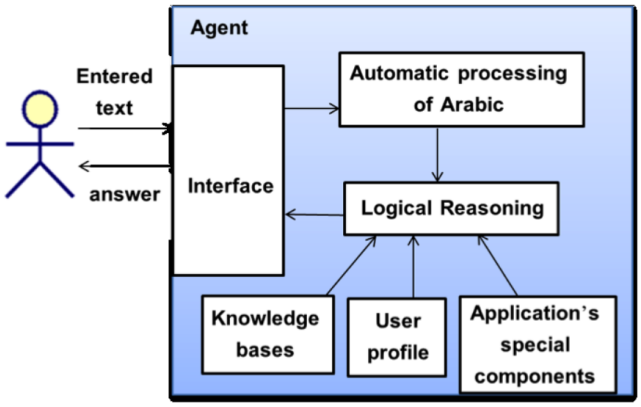


Fig. 2. General design approaches for intelligent agents.

Multi-agent systems (MAS) have become an effective paradigm for designing disaster management systems, where different specialists agents can cooperate to solve complex crisis situations[9]. These systems generally utilize a combination of heterogeneous dedicated agents such as situation assessment agents, resource agents, decision support agents, coordination agents, and prediction agents, which are specialized to address different aspects of disaster management [12]. The multi-agent nature allows for distributed knowledge and resource reasoning that are especially valuable in emergency situations of natural or human-caused disasters [9, 13].

Awareness-based modeling is also a critical design issue and is reflected in work like MAAP (Multiagent Awareness-based Process model), which keeps agents aware of relevant policies in support of better cooperation in disaster scenarios [14]. awareness enables agents to identify relevant information to consumer demands and coordinate in a crisis response.

Recent achievements in artificial intelligence have enabled the progress of more sophisticated design techniques. The multi-agent deep reinforcement learning (MADRL) integrates the multi-man formalized logic and deep reinforcement learning for managing the and uncertainty where high-dimensional data is involved in disaster situations [15]. This approach allows the agent to maximize its expected reward in an uncertain and noisy environment, where an approximate action value function is learned by agents using neural networks as function approximators [15].

Another important design pattern is represented by intelligent agent Decision Support Systems (DSS). These systems are commonly characterized by multi-level agent designs, with each level having a different role. For instance, a representation layer can describe the environment as a set of entities, the states of which evolve over time, and domain agents observe those states with respect to the domain informationodies [16]. These systems try to identify threats in uncertainty and under partial perception, so that crises can be avoided [16].

Contemporary agent architecture for disaster management also focuses on the important qualities of autonomy, complexity of decision making, adaptability, and impact power [17]. Autonomy enables agents to function without the need for human input, performing the ability to process real-time data and allocate resources. The intricacy of decision-making

allows them to maneuver complex environments, weighing, for instance, regarding affected populations, available resources, and infrastructure limitations. Adaptability allows agents to adapt strategies in real time to changing conditions, while impact potential describes the ability to take actions that have consequences in terms of lives affected, economic damage, and long-term recovery [17].

The agentic AI systems are the state-of-the-art intelligent agent design for disaster management. Such systems would be able to process environmental data in real-time under crisis, allocate resources and change at the same time strategically as conditions and the input data change [18]. This self-scheduling ability is essential in efficient rapidity need based disruption response.

One of the major challenges in developing an intelligent agent for disaster control is to make the designing of the intelligent agents in way where it is can successfully work in uncertain and complex environments, compatible with different platforms and scenarios [19][20]. Meetin this challenge, modern design approaches focus in generic frameworks that can be extended to multiple domains with little adaptation, counting with domain knowledge provided by experts while retaining core capabilities that enable the autonomous changes in agent behavior code [19].

B. Implementation Methods and Technologies

Disaster intelligent agentsIn the research of disaster management, there are a series of artificial intelligence technologies that are applied to the depth of work on the problems of disaster predicting, rescue and recovery. The implementation of Decision Support Systems (DSS) is one of the fundamental frameworks for transferring human expertise to computers and for constructing models from this knowledge [21]. These systems attempt to reduce prospective risks through the early reconnaissance and the preparation of actions.

Numerous AI techniques are widely used to handle the inter-related problems in disaster management. These methods include Bayesian networks for more probabilistic reasoning, fuzzy logic for approximate reasoning with vague information and metadata techniques such as artificial neural networks, machine learning, genetic algorithms and swarm intelligence [21]. Hybrid methods in which fuzzy IF-THEN rules and optimization are coupled, are particularly useful for the acquisition and handling of fuzzy information in disasters.

Agents are a promising paradigm due to their distributed, parallel, collaborative, hybrid, flexible, recursive, adaptive, cooperative, and intelligent processing [22] Figure3 and a powerful implementation of disaster management systems. The fact that disaster management activities consist of physically and logically distributed tasks makes multi-agent systems (MAS) systems especially suited, enabling specialized agents to be allocated to specific tasks, and hence supporting scalable and collaborative decision making.

One usage of Multi-Agent Systems and Anytime Algorithms can be found with the advancement of reliable, fast, and robust distributed decision support system in disaster management [22]. This coupling renders such systems especially well-suited for flood early warning and forecasting systems, since

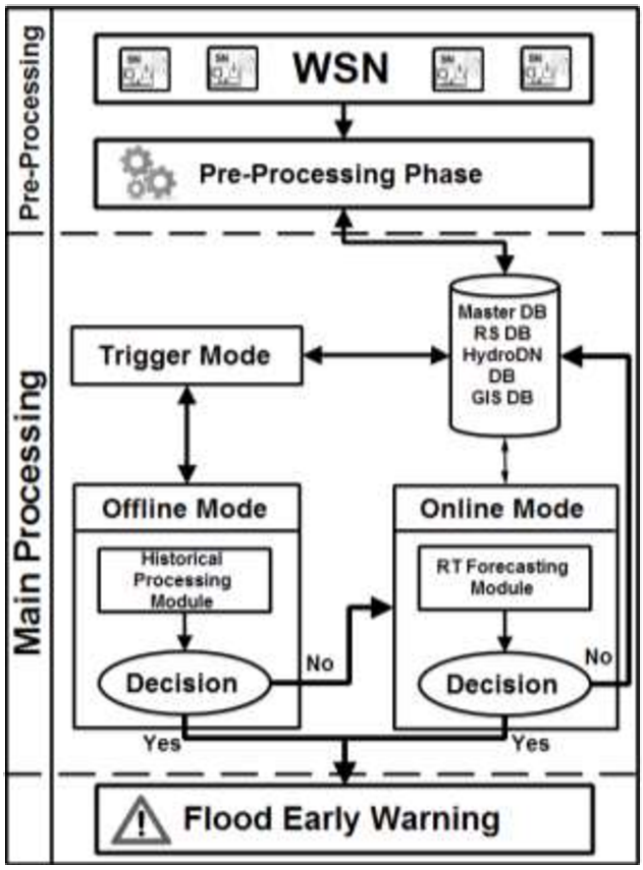


Fig. 3. General architecture of our decision support system.

that agents carry out intelligent, distributed, adaptive, scalable processing to take cooperative and collective decisions.

Putting it to Practice The more recent approaches take advantage of the sophisticated artificial intelligence algorithms, for example deep reinforcement learning. Multi-agent deep reinforcement learning (MADRL) allows agents to maximize their expected rewards in presence of uncertain environments which learn action values by using neural networks as function approximators [15]. This strategy is especially useful in the case of flood prediction, for which the data is of high-dimension and uncertainty is high. One interesting solution applies multi-agent actor-critic and deep deterministic policy gradient algorithms on those problems.

From a technical point of view, AI-based disaster management solutions can be partitioned into a continuum between supervised learning [23] [24] In the most ideal conditions, the reference points for operations on a large-scale disaster dataset would be provided as the ”correct”, or ”desired”, or ”target” output of a system. For instance, higher than normal temperatures in the previous week, were associated with an increased risk of hospitalization ahead of a summer heatwave[23]Figure4.

The adoption of AI in disaster management is more and more realized by new algorithms and high performance computing (HPC) in order to predict the geospatial and temporal patterns of disasters with precision [25]. The machine learning and deep learning methods that are frequently used for tasks

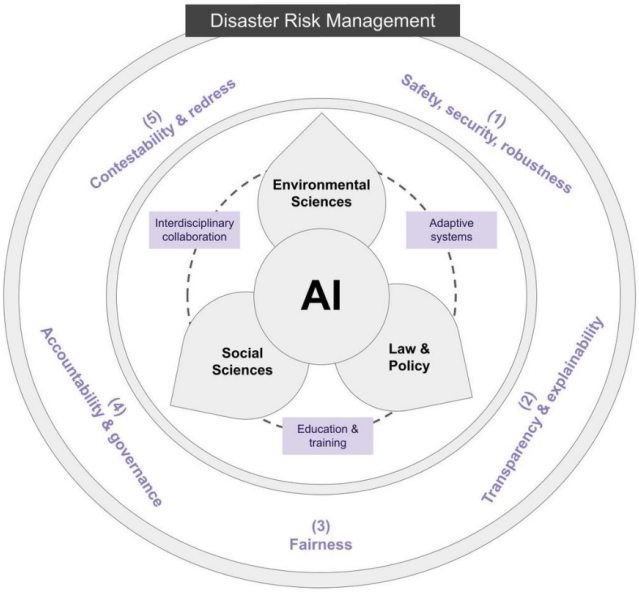


Fig. 4. Principles for regulating AI use in disaster risk management.

such as disaster prediction, risk evaluation, early warning systems, and disaster damage detection have increased the prediction accuracy, reduced computation time, and saved model generation costs [26].

These technologies are indeed promising but their application encounters many difficulties. These challenges include the requirement and quality of data, data diversity, integration with existing systems and technologies, ethical and social considerations, as well as the continued need for research and development [26] [27]. Privacy and security issues are especially important, because AI in disaster management may require the collection and analysis of privacy-sensitive data [26].

In spite of these difficulties, AI models have been successful in dealing with non-linear and large datasets hence are fitting for disaster management and monitoring (Kamolov, 2024). Technologies and methods such as big data analysis, machine learning and deep learning can be used to visualize, analyze and predict natural disasters through developed system [26]-[28].

Good design and deployment of agent systems for disaster needs to take account of how efficiently and synergy agents can be harnessed for action. This requirement calls for an unambiguous responsibility assignment among the agents and making the organizational structure of the MAS explicit [15]. Future work includes detailed development of the components, especially the forecasting decision-making and warning dissemination components; and integration of systems with external databases using APIs [15]. The following Table1 show the all general challenges

C. General Technological Challenges and Limitations

1) Data-related challenges:

a) Quality and availability of data: AI models need high-quality and diverse data for effective prediction and management of disasters, but acquiring comprehensive disaster data is still difficult [26, 27].

b) Dynamic datasets: There are significant difficulties in dealing with data rapidly changing due to disasters especially with noisy, heterogeneous and limited annotations [8].

c) High dimensionality of data: Information integration from multiple sources are considered as a challenging task in view of the large volume of data which is high dimensional and extracting useful features from this data becomes challenging to learn profitable patterns for prediction [8].

d) Bias in training data: Training data can be biased and hence lead to biased prediction or predictions that are unfair against under-represented regions or communities [8].

2) Model development and accuracy issues: Unclear statistical definition of 'extreme' event, which is challenging for the AI models to capture, and thus a standardization is harder to achieve [8].

a) Accuracy limitations: Despite the promising capabilities of AI in processing the nonlinear data for disaster monitoring, attaining credible prediction accuracy still is a challenging task [26], [28].

b) Incorporation of physics models: Incorporating AI with classical model based physics is challenging though may improve the accuracy [8].

c) Generalization in other contexts: The AI solutions which are trained on specific types of disasters or location can face difficulty in performing well in other form of disaster contexts [8].

3) Trust and interpretability concerns: Advanced AI models often operate as "black boxes" that are not, or may not even be, interpretable, thus may not be acceptable for critical decision-making contexts [8].

a) Uncertainty quantification: Most AI models are not good at quantifying uncertainty in their predictions, which is important for disaster relief [8].

b) (iii) Stakeholder trust: It is a challenge in developing AI models which are trusted by the stakeholders and compliant with regulations [8].

c) Reluctance in implementation: Reluctance to utilize AI in disaster management settings is due to the uncertainty of reliability and fairness [8].

4) Operational and implementation challenges: Connecting AI systems to the current disaster management technologies and infrastructure is subject to numerous technical challenges [26], [27]. Real-time Data Integration: Real-time integration of data in rapidly evolving disaster scenarios is a technically challenging task [8].

a) Deployment brokers: There are many practical challenges to transition from research models to operational deployment in the field of disasters [8].

b) Interpretation by non-experts: AI outputs could be hard for non-expert stakeholders and decision makers to interpret [8].

5) Ethical and social considerations: Data Privacy and Security: The gathering and analysis of private data during disasters pose notable privacy challenges [26], [27]. Fairness and Equity: Ensuring that disaster management systems based on AI are fair and equitable for all affected communities is challenging [26], [27]. Ethical Guidelines: Comprehensive guidelines are required to support transparent and ethical use of AI insights in disaster decision-making [28]. Challenges: The disaster risk analysis community needs to be aware of pitfalls in AI solutions [24], [26].

D. Multi-Agent Systems for Disaster Management

The multi-agent system is an effective framework for disaster management because they are capable of solving complex, distributed problems through collaborative intelligence. These systems are common and have a lot of advantages and can provide distributed, parallel, collaborative, hybrid, flexible, recursive, adaptive, cooperative, and intelligent processing in scalable environments and can be good candidates for multiple facets of disaster management [22].

The main benefit of MAS is their macroscopic nature, in which knowledge and resource reasoning can be distributed among specialized agents [9], [29]. A typical MAS for disaster management integrates different sorts of agents which are specific agents, including situation assessment agents, resource agents, decision support agents, coordination agents, and prediction agents. Each agent is trained in its own way and can model multimodal or unimodal data. Among the tasks they perform are receiving and processing situation updates, resource optimization, decision options generation, emergency control coordination and disaster trajectory prediction [12]. The BDI reasoning architecture is an important component in many multi-Agent systems for disaster management. This approach offers a natural means to define agent behavior in unstructured environments which are described in terms of mental states influenced by decision-making [9]. Your approach seems to be very useful in dynamic disaster scenarios as agents can continue to follow goal driven paths, without having to recalculate from the beginning when the world changes. This reinforces the situation-awareness characteristics of agents, such as in the ability to observe a dangerous alarm, take an intelligent action and prepare preventive acti ons in situations with limited time [9].

In the latest years, novel MADRL architectures that fuse deep reinforcement learning with multi-agent systems have emerged. Such advanced systems are very useful in dealing with problems of uncertainty in disaster prediction and management. In MADRL, multiple agents compete to maximize their expected total discounted rewards under uncertain or noisy transition and reward models. (d) Agents estimate action values using multilayered neural networks as function approximators [15]. This kind of method appears to be very promising in flood forecasting since it copes with complex high-dimensional data and uncertainty. To deal with these issues, the combination of MARL algorithms, like actor-critic algorithms and DDPG algorithm, is introduced to fulfill these requirements [15].

The architecture features multilayer-based agents with each serving a different functionalities for disaster management. For example, a representation layer could model the environment as a series of entities, that evolve over time, and factual agents are able to sense those changes as described in domain ontologies [16]. This layered strategy allows the system to identify threats in ambiguous or only partially perceived environments, before they develop into crises.

Organizations in MAS on the one hand need to specify how agent roles should interoperate and interact, but also need a mechanism to enforce interacting of roles. This involves specifying roles for each agent and developing procedures for information-sharing and joint decision-making [15]. As an illustration, when agents apply intelligent, distributed, adaptive, and scalable processing that help them in order to come up with cooperation or collective decision they help in flood forecasting and warning systems [22].

One such effective integration is that of multi-agent systems with Anytime Algorithms, thus producing dependable, efficient distributed DSS. This indissoluble link has been particularly exploited in the context of flood disaster management, where agents cooperate to execute different activities are make consensus in an efficient way [22].

Multi-agent systems are especially beneficial for disaster management, but face challenges such as addressing issues of interoperability, coordination mechanisms and inconsistent information. Future work will be focused on improving certain aspects, such as forecast decision making and warning dissemination, as well as on connecting systems to the outside world through the cloud to be able to use data from any organization also providing capabilities for full disaster management [15].

E. Intellent Agents Application and Use Cases

1) Early warning and predicting systems: Smart agents have been widely used especially in early earthquake warning by means of pattern recognition and data mining. Among other applications, AI-based systems can help predict the occurrence of natural disasters (e.g., hurricanes) and forecast their paths using weather data, satellite images, social network activities, and geolocated information [30]. Wildfire risk assessment AI algorithms process historical records of weather and vegetation cover and human activities to predict area with high risk of a break [30]. Machine learning algorithms that are applied on big data sets can predict with better accuracy for efficient disaster risk management [31]. For example, underwater seismic monitoring systems extract acoustic radiation of seismic events in order to provide the early tsunamigenic alerts based on fault feature identification [32], [33].

2) Damage monitoring and assessment: The AI systems which are based on computer visions are able to offer indispensable powers for post-disaster damage assessment. Deep learning models can semantically infer the extent of damage in individual buildings from before and after disaster satellite images, thus automating a significant bottleneck in decision support in disaster response [1] [4]. Convo- lutional neural networks (CNNs) based change detection models can identify the most impacted regions from disasters, aiding in disaster management [1]. Segmentation of buildings and damages using airborne light detection and ranging (lidar) data has also reached high accuracy rates in the case of heavily damaged areas [1] [34].

TABLE I. TECHNOLOGICAL CHALLENGES

Papers	[27]	[8]	[8]
Technological Challenges	Technological challenges in AI for disaster risk management include data quality, system interoperability, and model transparency.	Challenges include data scarcity, model interpretability, and integrating AI with existing systems.	Developing accurate predictors from noisy, heterogeneous, small sample sizes.
Operational Limitations	Operational limitations in implementing AI for disaster management include computational resource constraints, integration challenges, real-time processing demands, explainability and transparency requirements, bias and fairness concerns, and issues of responsibility and accountability.	Noisy, heterogeneous data, limited data, real-time integration challenges.	Limited data, real-time integration, model understandability, stakeholder trust.
Data Requirements	High-quality, diverse, and real-time data are essential for effective AI application in disaster risk management.	High-quality, diverse, and real-time data are essential for effective AI applications in extreme event analysis.	Noisy, heterogeneous, small samples, limited annotations, real-time integration.
Risk Reduction And Preparedness [45]	AI plays a significant role in predicting disasters, optimizing response efforts, and improving preparedness by analyzing data and providing early warnings.	AI plays a crucial role in enhancing flood management systems, improving predictions, and communication strategies for better risk reduction and preparedness for natural disasters.	AI improves disaster response, communication, and enhances disaster readiness and risk reduction.
Model Reliability And Trust	Concerns related to the accuracy, transparency, and trustworthiness of AI models used in disaster risk management.	Challenges include accuracy, transparency, and trustworthiness of AI models in disaster risk management.	Accurate, transparent, and reliable models crucial for stakeholder trust.

3) Evacuation planning and operations: EvacuationSim and evacSim Intelligent agents under evacuation simulation may employ advanced modelling and route optimisation used in evacuation planning. And with emergency tools that integrate inundation maps with semantic segmentation, it provides more information to first responders pre-, during and post-hurricanes, and help them to gain more insights in human-level decision-making [35]. These systems process heterogeneous data with hydrological inundation models and computer vision techniques in order to serve stakeholders with timely, rich and accurate knowledge [35]. AI Agent-based simulations explore different evacuation strategies, facilitating early detection of risks and decisions on the optimal planning of evacuation routes [36].

4) Reaction systems autopoietic ones: Robots and automation have obtained a larger attention in disaster response. Flying machines can penetrate deep into disaster areas to survey the damage and supply aid [1]. Several robot generations have led to the development of more sophisticated machines, moving from robots that could hardly make decisions to those that are fully automated, intelligent and artificial [1]. Autonomous robots have been particularly utilized amidst the COVID-19 pandemic to support a number of response efforts [1][6]. Ground, aerial, surface and underwater swarm robots collecting big data of the environmental data at disaster-stricken sites, and then processed by deep learning model for object classification and decision making [1][37] Figure5.

5) Safety building and fire detection: AI's role in building safety AI applications in building safety are all about early detection and response to disasters such as a fire. The Smart Building and Town Disaster Management Systems based on augmented reality (AR) allow for visibility and tracking of the occupants during a building fire [1][7]. Physical-virtual linked (PVL) disaster management services via AR technology in buildings These systems provide visualization information and optimal guidance for initial rapid response by interweaving the physical and virtual space in the buildings [1][7] Figure6 . Convolutional Neural Networks Automatic fire recognition systems based on convolutional neural networks for fire detection in surveillance video respond and alert suitably [1][21].

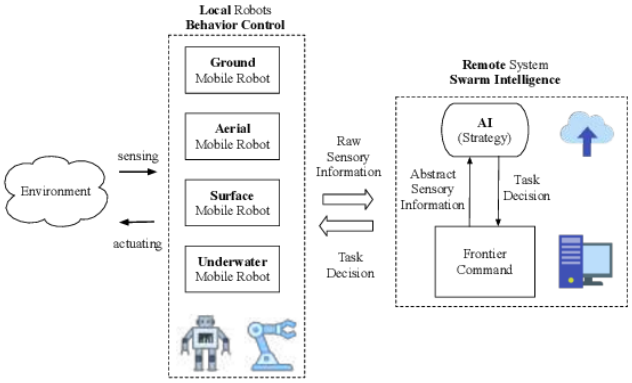


Fig. 5. Remote swarm AI and local robotic behaviour control.

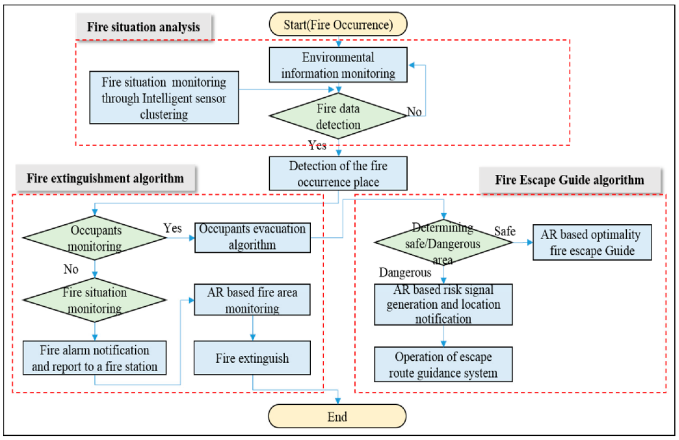


Fig. 6. The Fire extinguish and fire escape guidance algorithm.

6) Allocation of material and logistics: In climate related disasters, AI systems corner real time data regarding Effects of Disasters, Population dissemination, and Resources can utilized to optimize resource allocation [38]. These systems aid emergency responders in making rapid data-informed decisions about how best to allocate manpower, supplies, as wells as

in prioritizing the rescue operations, indentifying the location of high-risk victims and the most vulnerable population in need of emergency help [38-40]. AI systems can give real-time situation updates, disaster location and intensity, and generate information about affected people and infrastructures and support logistics planning [30].

7) Simulated and training environments: Disaster-Rescue simulations are also a practical application area in which in silico tools are used to determine policies for emergency response and management in the physical world [39]. Such simulations have widely been adopted in academia, government, and industry for contingency peril anticipation, planning, and policy making [39]. Emergency simulation using AI Agent-based technology can provide a more realistic training of simulated victim care and in future will see a holistic approach to using the AI for training in more security personnel, gaining as close a simulated experience as possible to actual conditions [36].

8) Multi-hazard risk reduction: Then, more recently, AI applications have shown promise in better managing multiple-hazards, that are when two or more hazards occur at the same time and interact together to lead to larger scale impacts, or cascading hazard scenarios, that are when a first high-impact hazard sets off a successive chain of other hazards or events that spread through time, see [32][40]. Such applications can be considered in the context of building sustainable infrastructure, safer and resilient cities and human settlements complementing various Sustainable Development Goals [32].

F. Challenges in Deploying Intellent Agents

Despite their potential benefits, implementing intelligent agents for disaster management presents several significant challenges:

1) Data quality and availability: One challenge is simply getting enough great data to train the AI models. One of the challenges in disaster scenarios is there is usually either very little, unreliable or disparate data, which would make it difficult to train robust models to estimate the required predictions [27]. Learning to predict can be particularly difficult when dealing with noisy and heterogeneous data, in the presence of very limited annotations [8]. This is particularly a problem for rare events such as extreme weather events in which there may be an insufficient amount of historical data.

2) Privacy and security concerns: The application of AI in DRM generally means processing huge amounts of personal data, posing crucial issues for privacy and security [27]. Data privacy is a key concern as such systems collect and process sensitive data and privacy regulations and individual rights need to be taken into account[26]. Finding ways to collect data in a manner that is both robust and respects privacy continues to present difficulties.

3) Computational resource requirements: Numerous advanced AI methods, especially deep learning algorithms, require extensive computation resources [27]. This creates a problem in disaster areas especially in remote or inaccessible areas where computational infrastructure is either scarce or not available. Critical AI solutions that are needed most, when and where they are needed can be hampered by resource constraints.

4) Integration with existing systems: Smart agents should be able to seamlessly interoperate with existing service infrastructure for emergency management such as command and contro l centers, dispatch functions, and sensor networks [27]. Integration in turn is taken rather difficult by data structure and nature of systems that are heterogeneous making deployment cumbersome to implement [26].

5) Real-time performance requirements: Disaster situations require instant processing and analyzing massive datasets for quick response and situational awareness [27]. Finally, timely information integration is critical in immediate Response, which however is not an easy task [8]. Since time is of the essence in many disaster response scenarios, there is an inherent tradeoff between speed and accuracy.

6) Explainability and transparency: Intelligent predictive solutions for high-stake disaster scenarios should be transparent and explicable to gain trust from decision-makers [27]. Developing authentic, interpretable and trustworthy AI-models is not trivial [8]. First responders must be able to comprehend AI recommendations and their underlying rationales to meaningfully integrate such information into their decision processes.

7) Ethical implications: The broader ethical dimension of containing AI with respect to disaster management will also have to be addressed [26]. There are fears that unrestrained use of AI can have detrimental effects and may increase the potential impact of calamity if not properly regulated [2]. Ensuring that AI systems are designed and deployed in a manner that is fair and equitable and achieves impact on disaster reduction is a key challenge.

8) Local implementation challenges: There are particular challenges to using AI models to inform policy and decision-making at the local level, where effects of climate change are dynamic and location specific [41]. Local peculiarities might face specific priorities, facilities, and available resources and skills to deploy AI-based disaster management systems.

9) Risks of AI-Caused digital disasters: An unusual new worry is the possibility that AI systems might itself cause, or exacerbate, catastrophes. While [2] demonstrates the bright side of the AI in disaster management, the unlimited applications of the AI may face the transference of challenges and threats, which can turn the AI benefits into detriments and deteriorate disaster impacts. This risk calls for a prudent governance and overseeing of deploying AI in life threatening emergency systems.

G. Future Opportunities and Trends

Looking toward 2025, artificial intelligence applications, Intelligent Agents Figure7 in disaster risk management (Table II) are poised for significant advancement across multiple dimensions. The integration of AI with emerging technologies is expected to revolutionize how societies predict, respond to, and recover from disasters. One of the most promising developments is the continued refinement of prediction and early warning systems that analyze diverse data sources—including weather patterns, satellite imagery, and social media—to make increasingly accurate forecasts about natural disasters such as hurricanes, earthquakes, and tsunamis [27].

TABLE II. SUMMARY OF METHODOLOGIES, TECHNOLOGIES, DISASTER TYPES, AND INNOVATIONS IN DISASTER RISK MANAGEMENT PAPERS

Papers	Methodology	Technology Utilized	Disaster Types Addressed	Risk Management Strategies	Opportunities and Innovations	Case Studies or Examples
[1]	AI and GIS in disaster management.	AI, geospatial tech, GIS, remote sensing, ML.	Floods, earthquakes, virus outbreaks.	Risk assessment, mitigation, response by intelligent agents.	Innovations in intelligent agents for risk management.	Real-world AI applications in disaster management.
[35]	AI, geospatial data, semantic segmentation, ML for emergency response.	Open-source geospatial data, ML algorithms, computer vision, semantic segmentation.	Hurricanes, wildfires, earthquakes, other crisis events.	Geospatial encoding and route optimization.	Open-source geospatial data integration for emergency response.	Hurricane Florence in Lumberton, NC.
[6]	Visual analysis of 15 YouTube videos from 12 countries.	Autonomous robots, 4IR technologies.	COVID-19 pandemic.	4IR innovations to enhance disaster risk reduction.	Infrastructure improvements via 4IR innovation.	Analysis of 15 videos across 12 countries.
[37]	AIoT architecture with robots and deep learning.	AIoT, DL models, swarm robotics.	Fukushima nuclear disaster, Wenchuan earthquake, Taiwan disasters.	Data collection and decision-making with AIoT coordination.	Disaster response using AIoT and swarm robotics.	Fukushima and Wenchuan examples.
Park et al. (2018)	AR visualization, ICT integration, sensor-actuator linkages.	AR, ICT, sensors, actuators.	Fire disasters in buildings.	AR-based visualization and physical-virtual linkage.	Quick disaster response with AR guides.	Small AR-based testbed implementation.
Jung et al. (2020)	Big data and AI to support decision-making in disasters.	CNNs, GANs for detection; RL for wildfire prediction.	Wildfires, cold/heat waves, storms, floods, earthquakes.	Intelligent agent strategies for risk management.	New opportunities with CNNs and RL in risk management.	CNN used for automatic fire detection in surveillance videos.

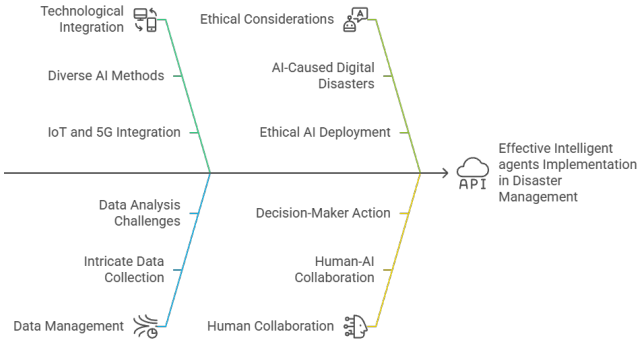


Fig. 7. Effective intelligent agents implementation in disaster management.

Automated damage assessment represents another area of significant future potential, with AI algorithms analyzing satellite imagery to rapidly evaluate disaster impacts and prioritize response efforts. This capability will be further enhanced through integration with unmanned aerial vehicles (UAVs), which can provide immediate visual data from affected areas, enabling faster and more targeted response [27]. The synergy between AI and UAVs extends beyond damage assessment to include the delivery of supplies and aid to affected communities, potentially transforming humanitarian logistics during disaster scenarios.

The convergence of AI with Internet of Things (IoT) devices and 5G networks is expected to create powerful new capabilities for real-time information gathering and rapid response coordination. These technologies working in concert will enable more comprehensive situational awareness and facilitate faster communication among response teams and affected populations [27]. This technological integration aligns with several United Nations Sustainable Development Goals, including SDG 11 (Sustainable Cities and Communities), SDG 13 (Climate Action), and SDG 15 (Life on Land), by enhancing prediction capabilities and improving resource distribution to reduce disaster impacts [31].

Personalization of disaster response represents a significant emerging trend, with AI systems analyzing community-specific vulnerabilities to develop targeted interventions. This approach moves beyond one-size-fits-all disaster management toward strategies tailored to the unique needs and characteristics of affected populations [27]. Similarly, AI-driven resource allocation systems will optimize logistics and supply chain management during disasters, ensuring critical resources reach those most in need when traditional distribution channels are compromised [27]. Enhanced collaboration between humans and AI systems is expected to be a defining characteristic of future disaster management approaches. Technologies such as chatbots and voice assistants will facilitate communication between response organizations and affected communities, while AI systems will augment human responders' capabilities by providing real-time information, automating routine tasks, and reducing response times [27]. This represents a significant evolution from AI as merely an "enhancer factor" toward truly intelligent disaster management (IDM) systems that can address increasingly complex disaster scenarios [2].

The future of AI in disaster management must also address the growing complexity of modern disasters, including hybrid events and potentially AI-caused digital disasters. As conventional methods give way to AI-informed practices across all disaster phases—pre-disaster (predictions, risk assessments), in-disaster (classifications, modeling), and post-disaster (damage assessments, resource allocation)—the field will require new approaches that balance technological advancement with ethical considerations [2].

Despite these promising developments, challenges remain in implementing AI for disaster management, including diverse AI methods and hazards [44], intricate data collection processes, and ensuring decision-makers act on AI-generated information [31][42][43]. Addressing these challenges will require interdisciplinary collaboration among technologists, disaster management professionals, policymakers, and affected

communities to develop systems that are both technologically sophisticated and contextually appropriate. As we move toward 2025, the trajectory of AI in disaster risk management points toward more integrated, responsive, and human-centered systems that leverage technological capabilities while remaining attentive to ethical considerations and sustainability goals. This evolution promises to transform disaster management from reactive response to proactive risk reduction, potentially saving countless lives and reducing the economic and social impacts of disasters worldwide.

IV. Conclusion

Artificial Intelligence (AI) and intelligent agents are catalysts for transforming the way we manage disaster risk and response, through their potential to enhance prediction, early warning, preparedness, real-time response and recovery. During the last decade AI technologies have advanced from self-contained, domain specific tools to integrated, adaptive systems that are up to the task to cope with growing complexity and urgency with which disaster scenarios are experienced around the globe. In particular, agents provide mechanisms for distributed decision-making, adaptability and autonomous coordination in disaster management systems, thus facilitating quick, scalable response to crises.

This systematic review has revealed significant progress made in utilising cutting-edge advances in machine learning and deep learning in order to forecast disasters, the use of robotics and UAV for damage assessment and search and rescue activities, natural language processing for real time social media analysis and geospatial data analytics for situational awareness. It has also drawn attention to the development of multi-agent learning techniques and deep reinforcement learning strategies as important solutions to handling dynamic and complex disaster environments.

But there are big holes and challenges. Quality and coverage of data still represent significant challenges, especially in low-resource settings and for rare types of disasters. Ethical considerations such as bias, privacy, and fairness remain important factors for the responsible use of AI in disasters. There is just as much pressing emphasis on technological challenges such as model interpretability, real-time performance, system interoperability and trust building among decision makers. Additionally, the changing nature of disasters — such as the development of hybrid crises and AI-facilitated digital disasters — calls for increasingly complex, humane, and ethically led AI solutions.

The future is promising as AI integrated with IoT, 5G technologies, autonomous systems and individualized analytics offer unprecedented opportunities to change the way disasters are managed. There is a trend towards more adaptable, anticipatory and resilient systems, being no longer a focus on only reacting to disasters, however looking at reducing risk and vulnerability. However, realizing this future vision will depend on multidisciplinary teamwork, ethical foresight, stakeholder engagement, and relentless innovation.

In summary, intelligent agents and artificial intelligence will change the way we will handle disaster risk management. By tackling the current technological and operational challenges and addressing the ethical dilemmas and risks, while building inclusive and sustainable innovation, AI has the potential to greatly improve global resilience and help protect human lives, the natural world and built infrastructural systems from an increasingly volatile and uncertain future.

References

[1] S. K. Abid *et al.*, "Toward an Integrated Disaster Management Approach: How Artificial Intelligence Can Boost Disaster Management," *Sustainability*, vol. 13, no. 23, pp. 1–22, 2021.

[2] Y. Pouresmaeil, S. Afroogh, and J. Jiao, "Mapping out AI Functions in Intelligent Disaster (Mis)Management and AI-Caused Disasters," *arXiv preprint arXiv:2501.03892*, 2025.

[3] G. Pang, "Artificial Intelligence for Natural Disaster Management," *IEEE Intell. Syst.*, vol. 37, no. 6, pp. 88–95, 2022.

[4] B. J. Wheeler and H. Karimi, "Deep Learning-Enabled Semantic Inference of Individual Building Damage Magnitude from Satellite Images," *Algorithms*, vol. 13, no. 11, pp. 1–15, 2020.

[5] S. Ramchurn *et al.*, "A Disaster Response System Based on Human-Agent Collectives," *J. Artif. Intell. Res.*, vol. 53, pp. 223–269, 2015.

[6] N. Sulaiman *et al.*, "The Role of Autonomous Robots in Fourth Industrial Revolution (4IR) as an Approach of Sustainable Development Goals (SDG9): Industry, Innovation and Infrastructure in Handling the Effect of COVID-19 Outbreak," unpublished, 2021.

[7] S. M. Park *et al.*, "Design and Implementation of a Smart IoT Based Building and Town Disaster Management System in Smart City Infrastructure," *Appl. Sci.*, vol. 8, no. 11, pp. 1–14, 2018.

[8] G. Camps-Valls *et al.*, "AI for Extreme Event Modeling and Understanding: Methodologies and Challenges," *arXiv preprint arXiv:2401.06294*, 2024.

[9] K. Saleem *et al.*, "Situation Aware Intelligent Reasoning During Disaster Situations in Smart Cities," *Front. Psychol.*, vol. 13, pp. 1–11, 2022.

[10] A. R. Javed *et al.*, "Future Smart Cities: Requirements, Emerging Technologies, Applications, Challenges, and Future Aspects," unpublished, 2021.

[11] O. Kovalenko and D. Velev, "Big Data Aggregation in Disasters Risk Management Systems," in *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 859, pp. 1–8, 2021.

[12] S. A. Khowaja *et al.*, "Integration of Agentic AI with 6G Networks for Mission-Critical Applications: Use-Case and Challenges," *arXiv preprint arXiv:2503.10582*, 2025.

[13] S. M. Akhtar *et al.*, "A Multi-Agent Formalism Based on Contextual Defeasible Logic for Healthcare Systems," *Front. Public Health*, vol. 10, pp. 1–10, 2022.

[14] A. Talaei-Khoei *et al.*, "Modeling Awareness of Agents Using Policies," in *Proc. Int. Conf. Software Data Technol.*, pp. 309–315, 2011.

[15] N. A. M. Rashid *et al.*, "Real-Time Multi-Agent Based Flood Forecasting and Warning System Model: A Malaysia Perspective," *J. Adv. Res. Appl. Sci. Eng.*, vol. 25, no. 3, pp. 34–45, 2024.

[16] F. Kebair, F. Serin, and C. Bertelle, "Agent-Based Perception of an Environment in an Emergency Situation," in *Proc. World Congr. Eng.*, vol. 2, pp. 1–6, 2008.

[17] T. J. Chaffer, J. Goldston, and B. Okusanya, "Decentralized Governance of Autonomous AI Agents," unpublished, 2024.

[18] D. Acharya *et al.*, "Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey," *IEEE Access*, vol. 13, pp. 1–22, 2025.

[19] F. Kebair and F. Serin, "Towards a Multiagent Decision Support System for Crisis Management," *J. Intell. Syst.*, vol. 23, no. 4, pp. 389–403, 2014.

[20] M. J. Druzdzel and R. R. Flynn, "Decision Support Systems," unpublished, 2010.

[21] D. Jung *et al.*, "Conceptual Framework of an Intelligent Decision Support System for Smart City Disaster Management," *Appl. Sci.*, vol. 10, no. 15, pp. 1–18, 2020.

[22] E. Marouane, "Towards a Real-Time Distributed Flood Early Warning System," unpublished, 2021.

[23] K. P. Chun *et al.*, "Transforming Disaster Risk Reduction with AI and Big Data: Legal and Interdisciplinary Perspectives," *WIREs Data Mining Knowl. Discov.*, vol. 14, no. 1, pp. e1503, 2024.

[24] S. Guikema, "Artificial Intelligence for Natural Hazards Risk Analysis: Potential, Challenges, and Research Needs," *Risk Anal.*, vol. 40, no. 7, pp. 1317–1328, 2020.

[25] S. Thekdi *et al.*, "Disaster Risk and Artificial Intelligence: A Framework to Characterize Conceptual Synergies and Future Opportunities," *Risk Anal.*, vol. 43, no. 1, pp. 90–107, 2022.

[26] S. Kamolov, "Machine Learning Methods in Civil Engineering: A Systematic Review," *Ann. Math. Comput. Sci.*, vol. 10, no. 1, pp. 45–62, 2024.

[27] D. Velev and P. Zlateva, "Challenges of Artificial Intelligence Application for Disaster Risk Management," in *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLVIII-4/W5-2023, pp. 281–286, 2023.

[28] R. R. Arinta and E. A. W. R., "Natural Disaster Application on Big Data and Machine Learning: A Review," in *Proc. Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, pp. 203–208, 2019.

[29] S. M. Akhtar *et al.*, "An Ontology-Driven IoT Based Healthcare Formalism," unpublished, 2020.

[30] H. Jain *et al.*, "AI-Enabled Strategies for Climate Change Adaptation," *Comput. Urban Sci.*, vol. 3, no. 1, pp. 1–12, 2023.

[31] Morocco Solidarity Hackathon, "Leveraging AI for Natural Disaster Management: Takeaways From the Moroccan Earthquake," *arXiv preprint arXiv:2312.14206*, 2023.

[32] G. Shaddick *et al.*, "Data Science and AI for Sustainable Futures: Opportunities and Challenges," *Sustainability*, vol. 17, no. 5, pp. 1–20, 2025.

[33] B. Gomez and U. Kadri, "Earthquake Source Characterization by Machine Learning Algorithms Applied to Acoustic Signals," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021.

[34] C. Axel and J. V. van Aardt, "Building Damage Assessment Using Airborne Lidar," unpublished, 2017.

[35] B. Ortiz *et al.*, "Improving Community Resiliency and Emergency Response With Artificial Intelligence," in *Proc. Int. Conf. Inf. Syst. Crisis Response Manag.*, pp. 147–154, 2020.

[36] I. Y. Muttaqien *et al.*, "Evaluation of Emergency Access Evacuation Routes Using Agent-Based Model Application," *J. Artif. Intell. Archit.*, vol. 2, no. 1, pp. 11–24, 2024.

[37] M.-F. R. Lee and T.-W. Chien, "Artificial Intelligence and Internet of Things for Robotic Disaster Response," *ARIS*, vol. 5, no. 1, pp. 20–29, 2020.

[38] V. R. K. Adapa, "AI for Climate Action: Leveraging Artificial Intelligence to Address Climate Change Challenges," *Int. J. Multidiscip. Res.*, vol. 4, no. 3, pp. 101–108, 2024.

[39] Y. Ouahbi and S. Ziti, "NW Logistics: System Architecture and Design for Sustainable Road Logistics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 4, pp. 105–112, 2025.

[40] Y. Ouahbi and S. Ziti, "Revolutionizing Road Safety and Optimization with AI: Insights from Enterprise Implementation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 4, pp. 98–104, 2025.

[41] W. A. Luna-Ramírez and M. Fasli, "Bridging the Gap between ABM and MAS: A Disaster-Rescue Simulation Using Jason and NetLogo," *De Computis*, vol. 10, no. 2, pp. 1–15, 2018.

[42] J. Gill and B. Malamud, "Reviewing and Visualizing the Interactions of Natural Hazards," unpublished, 2014.

[43] P. Ghamisi *et al.*, "Responsible AI for Earth Observation," *arXiv preprint arXiv:2401.00543*, 2024.

[44] R. Mharzi *et al.*, "Crisis Logistics and Natural Hazards Risk Management in Morocco: A Discrete-Event Simulation-Based Assessment of Proactiveness and Responsiveness," *OPSEARCH*, accepted Feb. 2025. [Online]. Available: https://doi.org/10.1007/s12597-025-00936-8

[45] R. Mharzi *et al.*, "Catastrophe-Related Disruptions' Preparedness and Emergency Management in Morocco: A Proactive Risks and Resilience Digital Twin-Based Analysis," *J. Model. Manag.*, vol. 20, no. 1, 2025.

Artificial Intelligence in Disaster Risk Management: A Scientometric Mapping of Evolution, Collaboration, and Emerging Trends (2003–2025)

Hssaine Hamid, ELouadi Abedlmajid
Advanced Systems Engineering, National School of Applied Sciences,
Ibn Tofail University, Kenitra, Morocco

***Abstract*—Recent years have seen a dramatic increase in the number of and severity of natural disasters, driven in part by climate change and urbanization. Artificial Intelligence (AI) appears to be a promising new technology that can transform disaster risk management (DRM) and provide new opportunities for prediction, monitoring, response, and recovery. The present work performs a bibliometric review of applications of AI to DRM, from a total collection of 7842 scientific articles extracted from Scopus, Web of Science and OpenAlex databases from the year 2003 to the year 2025. Exploring the trends of publications, authorship, international collaboration, and research topics, the study reveals the development and current status of AI incorporating disaster management. The results illustrate an apparent growth in interest in the field of science, how machine learning and deep learning methodologies are leading, and the raise of geospatial AI, remote sensing, and social media analysis in disaster preparedness and response. Other issues including data quality, ethics, technology and trust in AI systems are also considered. This paper offers helpful perspectives on the status quo and future development of AI-based DRMs.**

***Keywords*—Artificial Intelligence; disaster risk management; machine learning; deep learning; remote sensing; bibliometric analysis; natural disasters; geospatial AI; early warning systems**

I. Introduction

Rapid, Sustainable and Effective DRM in a World Subject to a Wide Range of Natural Disasters The growing frequency and intensity of natural disasters around the world, influenced by factors such as climate change and the expansion of urban risk, creates the need for innovative and effective DRM approaches. Only in 2021, more than 98.4 million people were affected by natural disasters with high human and economic costs [1]. Although conventional DRM methods are precious, in the case of complex and dynamic hazard scenarios, they are limited in its speed, accuracy, and flexibility for adaptation.

Artificial Intelligence (AI) is now recognised as a game-changing technology that can transform each phase of the disaster management cycle – from prediction and early warning through to response and recovery. This is where the state-of-art machine learning (ML), deep learning (DL), geospatial analysis, and real-time data processing has brought out capabilities that can advance the efficiency, accuracy and responsive to DRM operations. More recently, authors such as [2] and [3]highlight that the use of AI is not a question of incrementing the efficiency, but of covering the gap of the dearth of such tools in 21st century AD projects, like managing climate-induced emergencies.

The inclusion of AI in DRM encompasses various technologies and applications, including predictive modeling of natural hazards, automated image analysis based damage assessment, sensor network based real-time monitoring, social media based situational awareness, and the application of flying drones and rescue robots to accompany human agents. Taken together,these technologies lead to better, safer decision-making, better allocation of resources, improved emergency response times and more resilient communities.

With the rapid change of research in this area, an extensive bibliometric study is urgently necessary for overviewing the development of AI applications in DRM, as well as revealing the newly emerging directions and potential research gaps. The current analysis seeks to bridge that gap by providing a systematic review of literature from 2003 to 2025, in an effort to provide detailed snapshot of how AI technologies will influence the future of disaster risk management.

This study aims to answer the following research questions:

- How has the scientific production on AI applied to disaster risk management evolved from 2003 to 2025?
- What are the major thematic trends, technologies, and application domains that have shaped this field?
- Which countries, authors, and institutions are leading the research, and how do they collaborate?
- What are the emerging challenges, opportunities, and knowledge gaps?

This paper is structured as follows: Section II reviews the related work. Section III presents the methodology. Section IV results the main findings. Section V critical discussion reflections. Section VI concludes the paper.

II. Related Work

A. The Growing Importance of AI in Disaster Risk Management

Such natural disasters are becoming more commonplace, in part, thanks to climate change and population expansion in cities. Disasters affected more than 98.4 million people in 2021, resulting in more than 15,000 deaths and an economic damage of USD 171.3 billion [1]. With such disastrous events

increasingly exposing communities to immense risks across the globe, artificial intelligence (AI) has found its way as an important means of Disaster Risk Management (DRM).

AI technologies in disaster management [11] are increasingly not just useful, but necessary. As Singh et al. comment that "Recent breakthrough in AI provide great potential tools in disaster management, which is important as the occurrence of climate-related disasters is more and more frequent" [2]. Raut also claims that "The use of AI in disaster management is not just useful – it is essential" [3].

The roles of ML and DL in different areas of natural disaster management have been well established [4-6]. Such methods allow massive data-sets from previous disaster events to be analysed, patterns to be discerned and generalisations about forecasting to be extended[7]. For instance, cutting-edge AI methods have been used on flood data to learn how floods behave, increase flood resilience, avoid damage and save lives[7].

Geospatial data is an essential ingredient in DRM in terms of mapping of vulnerable populations and infrastructure, identification of at-risk areas and for planning emergency responses [8] [1] The integration of AI and geographic information systems (GIS) provide effective hazard preparedness and real-time responsive decisions in crisis management [9], [10]

AI technologies in disaster management cover the whole disaster life cycle from early warning and damage assessment, to monitoring and forecasting, post-disaster response and coordination, long-term risk assessment [5].These technologies can process large volumes of data and detect trends to produce forecasts for a wide range of natural disasters, including earthquakes, flood, wildfire, and others [3].

The increasing use of AI in disaster relief efforts holds the promise of greatly improved resource planning, preparedness, and mitigated economic impact by improving the emergency response and building stronger, more resilient communities in the face of ever more frequent and more severe natural disasters [7][9.]

B. Specific AI Technologies and Methods Used in Disaster Management

1) Machine learning and deep learning:

- Machine Learning (ML) techniques analyze large historical datasets to identify patterns, make predictions, and improve flood forecasting with minimal inputs [12][13] These approaches are less time-consuming to develop while maintaining accuracy in predictions [14].
- Deep Learning (DL) algorithms process complex disaster-related data, particularly for image classification and pattern recognition in disaster-affected areas [12][5].
- Artificial Neural Networks (ANNs) are employed for various applications including flood forecasting [15], water level prediction [16], and earthquake magnitude prediction using deep learning neural networks [17].
- Nonlinear Autoregressive Network with Exogenous Inputs (NARX) models capture the dynamic characteristics of flood water levels by incorporating feedback from network outputs[18]
- Extended Kalman Filters (EKF) optimize neural networks to overcome nonlinearity problems in flood water level prediction [16]

2) Computer vision and image analysis:

- Convolutional Neural Networks (CNNs) analyze text and images from disaster areas for classification and damage assessment [19].
- Automated 3D Crack Detection systems assess building damage post-disaster [20].
- Image-based 3D Reconstruction methods detect and analyze surface damages [17].

3) Geospatial AI and remote sensing:

- Geographic Information Systems (GIS) with AI map flood hazards and susceptibility, supporting strategic planning and hazard preparedness
- Remote Sensing AI Applications utilize satellite imagery for disaster monitoring, risk assessment, and damage evaluation [21].
- Terrain Mapping Algorithms create 3D models of disaster-affected areas to support navigation and assessment [22].

4) IoT and sensor networks:

- Internet of Things (IoT) networks collect real-time data for disaster monitoring and prediction [23][24].
- Wireless Sensor Networks (WSN) create distributed sensor nodes for smart sensing and real-time disaster surveillance [25].
- Integrated Information Systems combine geoinformatics with IoT for early warning, particularly for snowmelt floods [24].

5) Robotics and autonomous systems:

- Unmanned Aerial Vehicles (UAVs/drones) conduct post-disaster inspections and damage assessment [26].
- Swarm Robotics communicate with each other to map disaster areas and create connected networks [23].
- Autonomous Helicopters perform remote sensing and sample collection in disaster zones [22].
- Underwater Robots inspect submerged structures like bridge foundations after floods [27].
- Mobile Rescue Robots carry out surveillance missions in hazardous environments like nuclear disaster sites [28].

6) Cloud computing and big data analytics:

- Cloud Storage and Distributed Computing overcome limitations of traditional storage and computation capabilities for disaster data processing [25].
- Big Data Frameworks (Apache Hive, Hadoop, Mahout) provide scalable distributed storage and computing environments to process massive social media data [29].
- Predictive Analytics sift through historical and real-time data to identify emerging disaster trends [30].

The following Table1 will present a Literature Comparison.

TABLE I. SUMMARY OF SELECTED PAPERS ON DISASTER MANAGEMENT APPLICATIONS

Papers	Disaster Type	Application Stage	Outcomes and Implications
[21]	Flood disasters	Prediction and disaster risk reduction	Enhanced flood forecasting model design for disaster risk reduction
[32]	Earthquake in Al Haouz, Morocco, 2023	Disaster preparedness, response, and recovery	Real-time data, data scarcity, interdisciplinary collaboration challenges, community-call for action
[26]	Flooding in Jakarta due to river overflow	Prediction and early warning stage for flooding	Improved prediction by 1%, potential for future early warning systems
[12]	Floods	Prediction of flood water levels	NARX model outperforms EKF in flood prediction; informs future disaster management
[13]	Four major natural disasters	Damage assessment stage during natural disasters	High classification accuracy for damage assessment using social media imagery
[29]	Hurricane Sandy	Disaster management including tracking, mapping, and analysis of events	CyberGIS framework supports big data analytics for disaster management

III. METHODOLOGY

The methodological framework applied in this bibliometric study was developed to describe the scientific publications landscape related to the use of Artificial Intelligence (AI) in natural disaster risk management. The search was performed in three key, academically reputable bibliographic databases—Scopus, Web of Science, and OpenAlex. These sources were chosen for their wide coverage of relevant publications in a variety of disciplines and geographical locations. Structured query using keywords of "core AI technologies and disaster risk management" was used as a search strategy. The search string employed was: ("artificial intelligence" OR "machine learning" OR "deep learning") AND ("disaster management" OR "disaster risk reduction" OR "natural disaster" OR "catastrophe naturelle" OR "emergency management" OR "hazard mitigation" OR "resilience" OR "early warning system"). This search was employed to collect all papers between 2003 and 2025, obtaining an initial corpus of 13,880 documents. Selection process: The selection was performed with a systematic process that is described below according to the PRISMA in the Figure1 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) strategy (Figure1). In the first round of selection work, all the retrieved files were filtered by inclusion criteria which narrowed the set of documents to articles, conference papers, and book chapters in English within the time frame. After that, 8,491 papers were left. A second phase of exclusion was applied to remove redundant records and articles – based on their titles and abstracts – that were deemed to be irrelevant, resulting in 7,842 articles included in the systematic review.

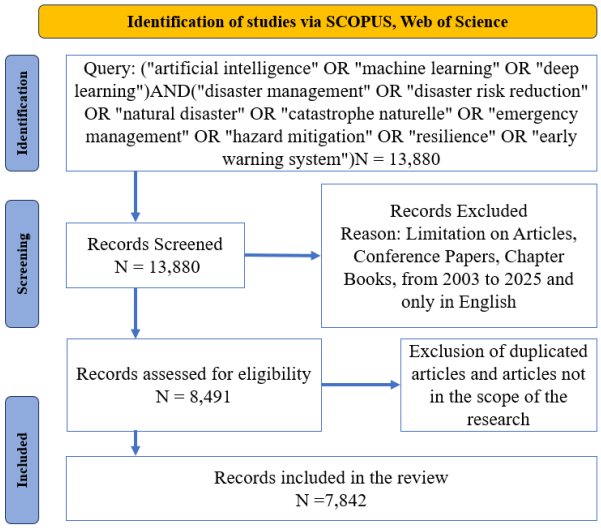


Fig. 1. PRISMA diagram.

The bibliometric analysis was completed using the Bibliometrix R-package, which is available as part of the Biblioshiny graphic user interface, a programme valued for quantitative study of literature. It was targeted to seek a series of descriptive and relational indicators pertaining the annual scientific production, the origin of sources of origin, authorship patterns andthe collaboration networks developed along time. Final corpus Figure2 : corpus properties The final corpus covered a time period from 2003 to 2025, and comprised 7,836 documents from 2,403 distinct sources. The trend of the scientific output was very dynamic, with an annual growth rate of 31.35%, which suggests an increasing attention to the convergence of AI with the disaster management. The authorship of manuscripts was conducted by 26,875 unique authors, international co-authorship rate was 25.79%, and a mean of 4.19 authors per manuscript indicated high amount of collaborative research. The dataset also included 2,773 author keywords and about 191,858 references. The articles had an average of 10.68 citations and 2.39 years old on average which indicated a not so old and dynamic area of investigation.

Through this robust methodological approach, the study aimed to capture a representative and high-quality sample of the literature, enabling a detailed analysis of the thematic evolutions, emerging trends, and collaborative networks in the field of AI-driven disaster risk management.

Timespan 2003:2025	Sources 2403	Documents 7836	Annual Growth Rate 31.35 %
Authors 26875	Authors of single-authored doc 1041	International Co-Authorship 25.79 %	Co-Authors per Doc 4.19
Author's Keywords (DE) 2773	References 191858	Document Average Age 2.39	Average citations per doc 10.68

Fig. 2. Descriptive statistics of the bibliometric dataset.

IV. Results

A. Annual Scientific Production (2003–2025)

The trend of the annual scientific production in application field of Artificial Intelligence in DRM is characterised by multi-phase meanwhile growth. Between 2003 and 2015, the historically low and static publication trend indicates an immature upstart in which AI technologies were finding their own legs and face the testing ground of application to disaster scenes. A clear increasing trend appears around the same time as the burst of progress in machine learning and deep learning techniques, and growing awareness of the impacts of climate change. A surprising 'explosion' occurs between 2020 and 2023 with the number of publications increasing exponentially (around 800-1000 articles in 2020, around 1500 in 2021 and 2000+ in 2023). This increase is driven by the technological readiness, increased frequency of disasters which are COVID-19 pandemic and climate-induced extreme events, as well as increasing governmental and institutional investments in AI emergency management solutions. The slight fall in 2025 is due to missing data for that year, rather than a true decline in research output. This production rate Figure3 reflects a research field that is ever more active and proposes that the use of AI for disaster management is moving from the realm of the theoretical to the domain of an essential component of operational disaster risk strategies at global scale.

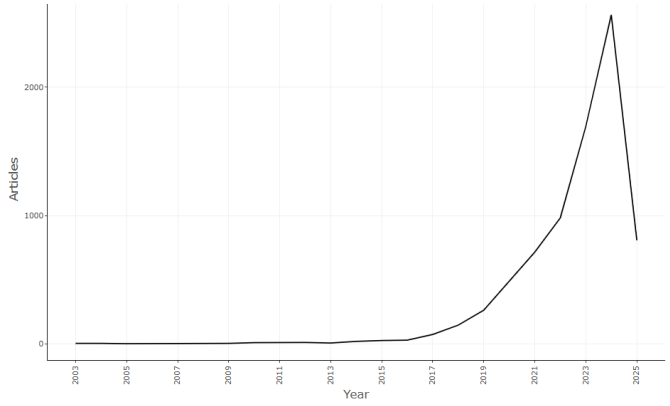


Fig. 3. Annual scientific production.

B. Average Citations Per Year (2003–2025)

One aspect of interest to highlight is the trend of increasing (or decreasing) number of yearly citations. Early publications (2003–2006) indicate a high number of citations, likely reflecting more fundamental research at an earlier stage in a new and undiscovered scientific field. Such landmark papers are cited slowly in that they are pioneers.

However, as the field has been maturing a sudden surge in publications from the end of 2017, there is a decrease noticed in average number of citations per paper. This tendency is accounted for by the "citation dilution effect," where the proliferation of published works leads to citation focus on a larger pool of articles. In addition, the more recent papers will be more time dependent as they will have had less time to be cited, particularly for papers published 2022 onwards.

Nevertheless, the stable citation averages in the mid-2010s demonstrate that the research in the field is still relevant and has had, and continues to have, impact on both academia and practice. The pattern observed confirms the strength of the field as well as the trend that higher-quality innovative studies are needed to survive in an environment of increasingly tight publication competition. As shown in Figure 4, the trend of average citations per year highlights early foundational contributions and recent publication dilution effects.

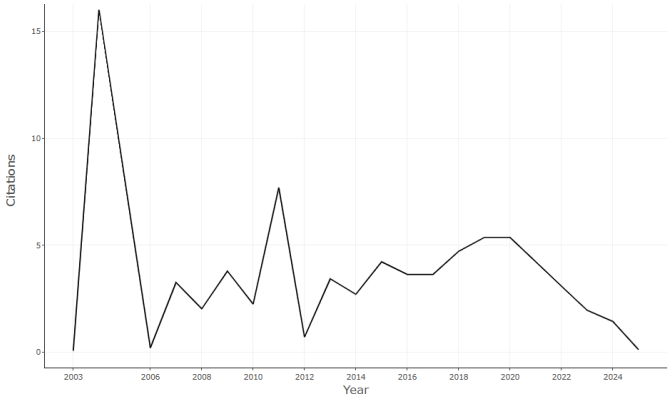


Fig. 4. Average citations per year.

C. Three-Field Plot (Authors – Sources – Keywords)

The Tri-field plot provides an integrated representation of interactions between prolific authors, top pub-lication venues, and main research topics. See for instance, Rachid Guerraoui, Alina Oprea, Ali Mostafavi as key contributors to leading sources such as arXiv (Cornell University) and Research Square.

Thematically on the thematic side, terms like "resilience", "robustness", "early warning system", "vulnerability" prevail, demonstrating the community's focused work on the development of AI-based solutions to enhance disaster preparedness and response.

The prevalence of "resilience" as a keyword reflects the increasing attention being paid to both the short-term response to disaster and the long-term adaptation of communities and infrastructure. Beyond this, the inclusion of terms such as "pandemic" and "sustainability" for the first time on the supplement panel might reflect a broadening of the types of disasters being considered beyond traditional hazards to the more systemic risks they are generating.

This visualisation suite illustrates the multidisciplinary facets of the field,êmphasising the variety of disciplinary

contributions toÅI research in disaster management from computer science, environmental science,ênginееring and social sciences. Figure 5 presents the three-field plot linking top authors, publication venues, and major keywords.

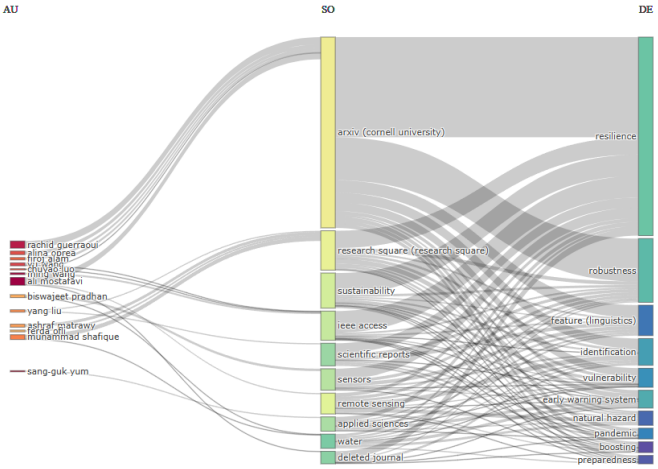


Fig. 5. Three-Field plot.

D. Most Relevant Sources

The examination of the most important sources, by document identity, indicates a clear shift towards fast dissemination and open acces in the papers and documents. ArXiv (Cornell University) and Research Square lead the field, indicating that quick transmission of information is an important consideration for researchers in disseminating their results in the context of the immediacy of disaster risks challenges. Journals, such as Remote Sensing, Sustainability and IEEE Access, exemplify the interdisciplinary nature of this science, the new technology development with environmental and social applications. The inclusion of journals such as Scientific Reports and Applied Sciences points to the importance of multidisciplinary channels wed to a wide variety of methodologies, from remote sensing and geo-informatics to data science and decision support systems.

These data suggest that traditional peer-reviewed journals are relevant, preprint platforms play an increasingly important role in expediting transmission of knowledge in times of disaster, reflecting the urgency of the research in the time of disaster. According to Figure 6, the most relevant sources include arXiv and IEEE Access, reflecting interdisciplinary focus.

E. Sources' Production Over Time

In the time course of source productivity, the central role of the open-access repository in the surge of AI-disaster management research in the past few years is also underlined. The explosive development of ArXiv after 2019 reveals the increasingly indispensable role of preprint repositories as infrastructures for timely sharing of knowledge, especially in times of crisis, such as the COVID-19 pandemic. Other sources, including Remote Sensing and Research Square, reveal noticeably steeper but more mild growth trajectories, indicating the sustained commitment of the remote sensing and sustainability

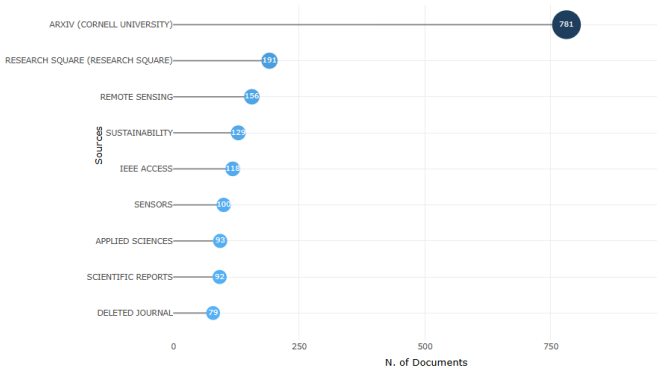


Fig. 6. Most relevant sources.

literature to disaster-related research. This evolution points to a strategic change among researchers towards platforms with faster publication cycles and more accessible, a necessity when the research outputs are poised to directly feed into emergency planning and crisis management.

As depicted in Figure 7, source productivity increased significantly after 2019, led by open-access repositories.

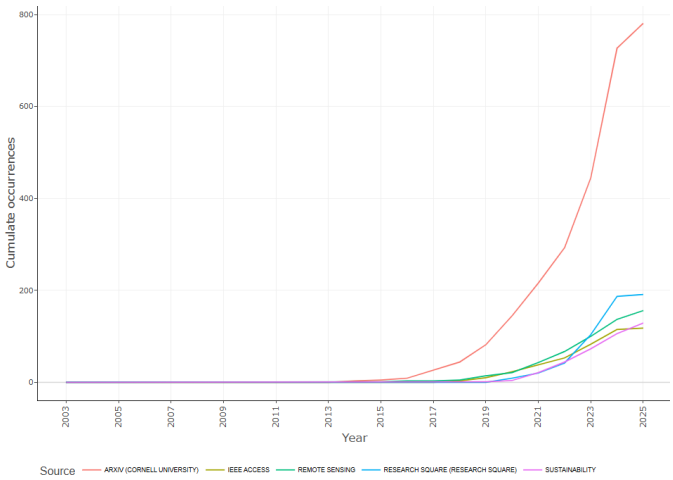


Fig. 7. Sources' production over time.

F. Most Relevant Authors

If we look at the main sources by document count, there is a clear trend here towards fast, open-access spread. ArXiv (Cornell University) and Research Square are the most popular venues, which imply that researchers choose the fastest channels in which to publish their findings because of how pressing disaster risk issues can be. Publications such as Remote Sensing, Sustainability, or IEEE Access (just to name a few), exemplify the interdisciplinarity of this field, combining technological breakthroughs with environmental and social realities. The existence of journals such as Scientific Reports and Applied Sciences illustrates the fundamental necessity for multidisciplinary (inclusive of all (or most) of them).The presence of journals like Scientific Reports and Applied Sciences points to the crucial role that multidisciplinary outlets covering a wide range of methodologies ranging from remote sensing and

geoinformatics to data science and decision support systems play.

The statistics demonstrate that while traditional peer-reviewed journals matter, academics are increasingly also dependent upon preprint servers that can expedite the circulation of knowledge during disaster-laden times, underscore the urgency of their work in the real world. Figure 8 displays the most prolific authors contributing to the field.

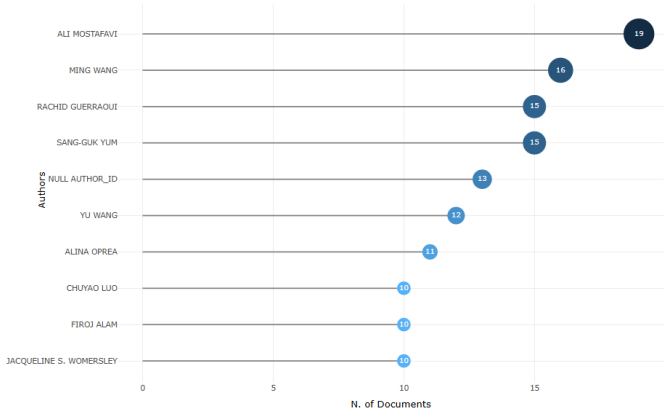


Fig. 8. Most relevant authors.

G. Authors' Production Over Time

The time signals of author productivity further illustrate that the closely-knit and the active research community on AI and disaster risk management is quite young. Most of the top authors started to publish in this area by 2017, with highest activity between 2021 and 2024. This time period also aligns with global initiatives that focus on climate resilience, smart cities and sustainable development goals (SDGs) that have been drivers of multidisciplinary engagement.

The data also reveal that many of the researchers in this area have and sustain a "publication cadence" that suggest a move from practical, one-time contributions to a more mature, programmatic commitment to disaster-related AI research. As seen in Figure 9, most top authors became active after 2017.

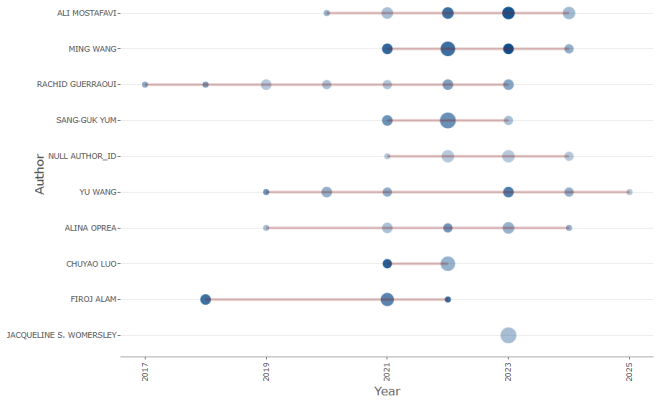


Fig. 9. Authors' production over time.

H. Affiliations' Production Over Time

The institutions' affiliations also substantiate this worldwide dispersion of expertise. Top universities including Harvard Medical School, Stanford University and University of California drive consistent leadership efforts, frequently combining AI investigations with health resilience, infrastructure surveillance and community risk analysis. The growth of independent researchers and "Not Reported" affiliations indicates the opening of the field to input from practitioners, think tanks, and private sector organizations, illustrating the cross-sectoral and interdisciplinary urgency of DRM challenges. This move toward spread and inclusivity of research will potentially spur innovation and accelerate solutions by infusing scientific inquiry with diverse viewpoints and on-the-ground urgency. Figure 10 shows institutional contributions with leading roles from top U.S. universities.

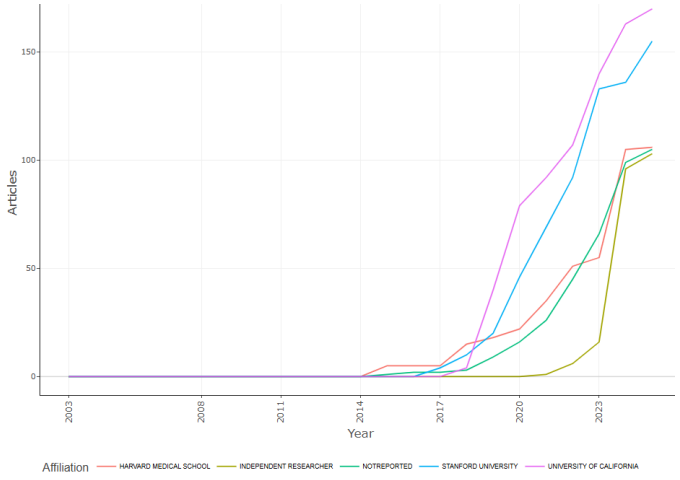


Fig. 10. Affiliations' production over time.

I. Most Cited Countries

The examination of the top-cited countries shows a clear geographical imbalance in the research impact in AI applications for disaster risk management. The United States is the dominant country in the production of the cited work, counting 10,941 citations, followed by China (5,979 citations) and the United Kingdom (3,086 citations).

This clustering around only a handful of leading countries means that much innovative, highly influential research comes predominantly from mature academic ecosystems with state-of-the-art technological infrastructure and significant research investment.

Australia, Italy, and South Korea also occupy the top positions, evidencing their significant contributions to disaster research and the development of this field in high-risk areas.

However, those contributions by the likes of India and Iran are not little, but significant, if their total among citation number still can not be compared with the existing top leaders. This expanding diversity suggests that high-impact research in the field is becoming more globally distributed. According to Figure 11, the United States leads in citations, followed by China and the UK.

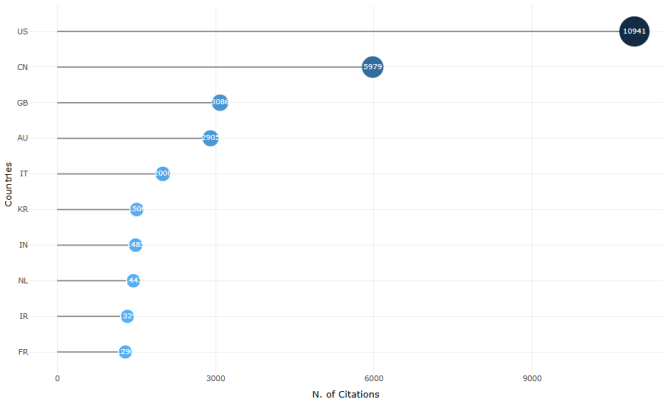


Fig. 11. Most cited countries.

J. Most Frequent Words (Author's Keywords)

The most common words are a corpus in focusing theme priorities in the sector. The most prevalent key term "resilience" (over 1,360 hits)Most used key term "resilience" (over 1,360 hits)points towards a change of paradigmparadigm change when it comes to the management of disasters, where the emphasis has recently turned to long-term systemic adjustment, not just to reacting to current acute threats.

Other commonly found words such as "robustness," "early warning system," "vulnerability," and "pandemic" alsoreflect current concerns particularly in the perspective after 2020, when the COVID-19 pandemic drew attention to the urgent need for adaptable, predictive, and robust disaster management approaches.

"natural hazard" "preparedness" In addition to disaster management, natural hazard and preparedness are included here to represent this comprehensive field which covers from hazard identification to emergency response planning. Figure 12 highlights the most frequent keywords, led by "resilience" and "vulnerability".

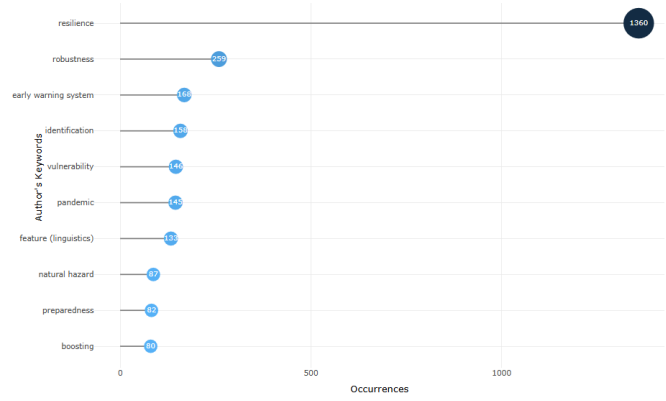


Fig. 12. Most frequent words.

K. Trend Topics

In addition, trend analysis reveals the hot topics appearing in this field. There's also growing interest in metaphors around learning, particularly "transformative learning" and "resilience," that have emerged since 2020 as part of what's become a significant and interdisciplinary combination of AI with the behavioral science and educational literatures on disaster preparedness.

"anomaly detection (physics)" "flood warning" "aerial image analysis." demonstrate the growing use of AI for monitoring, real-time risk assessment, and decision-support systems.

The advent of terms such as "swarm robotics" and "quantum byzantine agreement" reflects the potential futuristic orientation of the field in exploiting vastly distributed, decentralized AI systems in handling complex disaster worlds. As shown in Figure 13, emerging trends include "anomaly detection" and "swarm robotics".

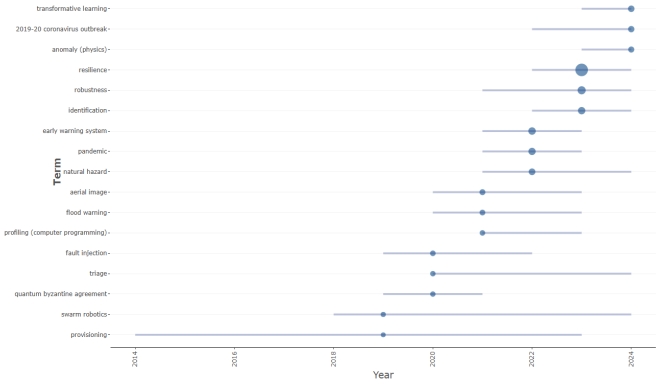


Fig. 13. Trend topics.

L. Clustering by Coupling

The thematic clustering map based on bibliographic coupling provides deep insight into the intellectual structure of the field. Three distinct clusters can be identified:

- One focused on "resilience," "preparedness," and "benchmarking," indicating foundational studies anchoring the field around risk reduction and system evaluation frameworks.
- A second cluster addressing "vulnerability" and "ethics," emphasizing the social and ethical dimensions of AI deployment in disaster scenarios.
- A third emerging cluster around "robustness" and "federated learning," suggesting an increasing interest in privacy-preserving, distributed AI models for disaster prediction.

The diversity of these clusters reflects the multidisciplinary and socially sensitive nature of AI applications in disaster management. Figure 14 identifies thematic clusters around resilience, ethics, and federated learning.

M. Co-occurrence Network

The co-occurrence network shows a dense and interconnected knowledge structure centered around "resilience," which acts as the major thematic hub. Keywords such as "vulnerability," "pandemic," "preparedness," and "situation awareness" cluster tightly around resilience, demonstrating

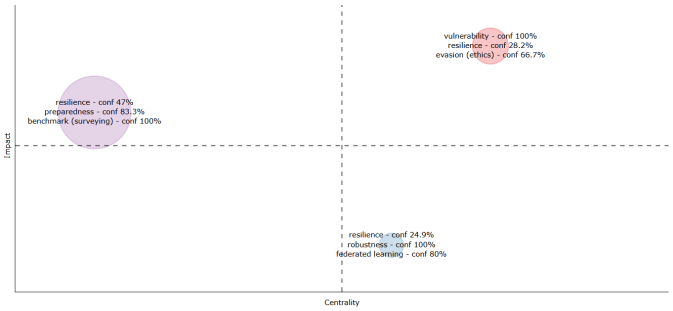


Fig. 14. Clustering by coupling.

the interdependence of these concepts in operationalizing AI solutions.

Peripheral nodes like "flash floods," "earthquake prediction," and "ensemble learning" indicate specialized application areas branching from the central resilience discourse. This network structure suggests that although the field is converging around resilience-building, it maintains a rich diversity of application-specific subfields. In Figure 15, the keyword network demonstrates interconnections between resilience, preparedness, and vulnerability.

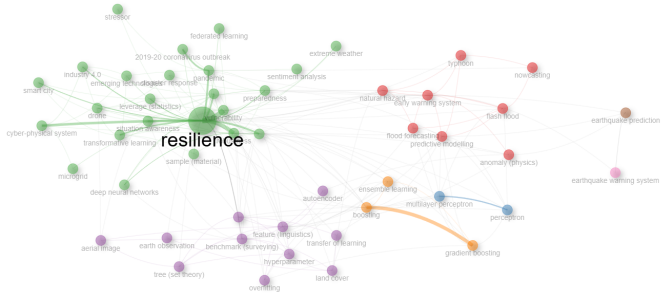


Fig. 15. Co-occurrence network.

N. Historiograph

The historiograph highlights key intellectual lineages within the field. Early influential works by Ferda Ofli (2016) and Bernd Resch (2017) laid critical methodological foundations, focusing on data-driven disaster mapping and social media analytics. More recent influential papers by Vasileios Linardos (2022) and Amine Belhadi (2021) show a shift toward integrating resilience models and predictive analytics at scale. This evolution reflects a progression from descriptive analytics towards prescriptive, action-oriented AI applications that not only model but also influence disaster risk reduction practices. Figure 16 outlines key intellectual lineages in the field, showing shifts from mapping to predictive AI models.

O. Collaboration Network

The global hub-nation analysis based on the collaboration network shows that the United States is the top country which is connected with other leading countries such as China, United Kingdom, Australia, and France. The presence of clusters of intense collaboration between Western countries and emerging economies is a perfectly sane cross-fertilization of ideas and a

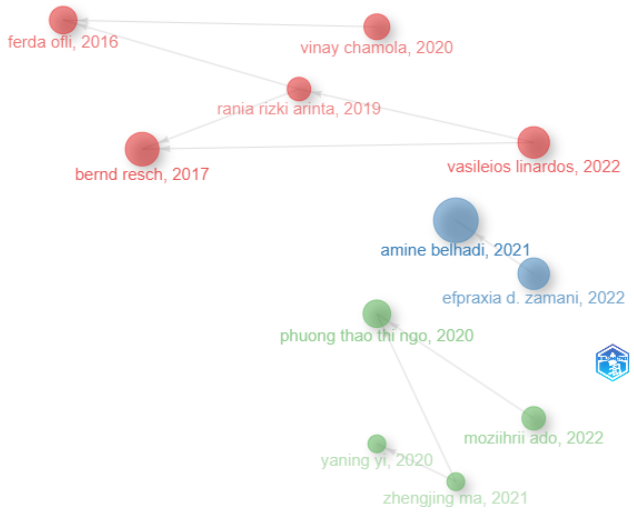


Fig. 16. Historiograph.

sharing of research investment. Strikingly, a large fraction of partnerships are with Asian nations, where exposure to natural hazards is most high and their active role in developing AI-based disaster resilience tech. This densely connected network, reflects the growing global and collaborative character of scientific production in this area, crucial for addressing the transboundary disaster risks. Figure 17 illustrates the international collaboration structure, centered on the US, China, and UK.

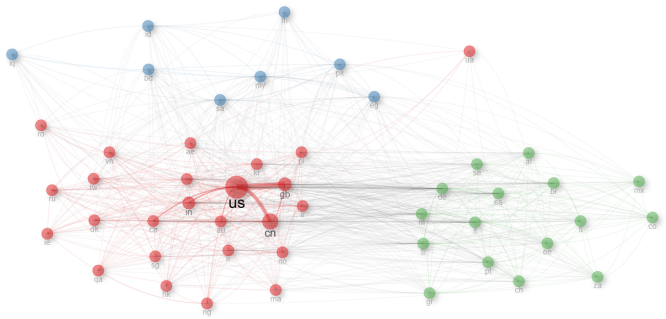


Fig. 17. Collaboration network.

V. Discussion

AI has a great potential to contribute for disaster risk management, however, there are some challenges that need to be addressed in order to successfully implement AI. A critical concern relates to a requirement of high-quality, diverse, and representative data in order to train the AI models [31]. In the absence of extensive training data, AI models could make false predictions or have problems generalizing to other disaster types and as such reduce the dependability of these algorithms in critical systems.

Another major challenge comes from technical integration, as AI systems need to be integrated with the current disaster management architecture and technologies [31]. This is especially problematic in low-resourced settings where technolog-

ical infrastructure may be underdeveloped, and where access to AI-enabled disaster management will differ.

Ethical issue is an increasingly concerning challenge in AI application in disaster response. The most important challenge of these systems, however, is dictated by its data privacy and security aspects, since these systems are connected to the processing and/or the analysis of sensitive data about vulnerable people and infrastructures [31]. There is also a potential risk of algorithmic bias, where AI learns biases with historical data, possibly exacerbating the disparities in disaster response and resource distribution [32] [35][36].

Disaster scenarios are complex by nature which adds to the difficulties. Types of natural disasters are the widest from each other, regarding to their types, impacts and geographical dependency, that create a challenge in AI system or algorithm, functioning for a wide range of disaster types in the world [14]. Management of the response disaster is difficult due to the disruption of normal surveillance and communication by damaged infrastructure in their aftermath [33].

Human-related issues further handicap AI's utility for disaster management. The readiness of decision makers to base decisions on AI outputs is still uneven: some actors are reluctant to place too much trust in algorithms to inform decisions with severe consequences [14]. This trust deficit may limit the effectiveness of technically complex AI systems.

Despite advances in technology, accuracy issues remain. Although AI models hold promise in dealing with complex and nonlinear data for disaster monitoring and management, their accuracy of prediction remains restricted, especially with respect to new or rare events 33. With climate change changing the patterns of disaster after disaster, historical data might provide much less of a guide to future predictions, adding another level of difficulty in developing AI models [33][34] .

These problems require further research and development to strengthen the role of AI applications in disaster management [31]. Dealing with these shortcom- ings will require interdisciplinary collaboration among AI practitioners, disaster response specialists, policy makers, and affected communities, which will all be necessary to develop and deploy AI systems in a way that is just, fair, and effective in mitigate disaster impacts [33].

Article Once-rare weather disasters becoming more common There's no such thing as a natural disaster, Anywhere in the world. Artificial intelligence combination with other new technologies is a great opportunity for the transformation of disaster risk management around the world and community resilience [30]. These developments open up the potential for new and more efficient disaster prediction, monitoring, and response systems.

IoT, Big Data analytics, and machine learning stand at the forefront in the wave of digital transformation of disaster management by providing unprecedented capabilities for holistic risk reduction strategies [35]. IoT connected devices and sensors are forming networks that can give us real-time environmental status reports — to identify and assess unfolding disasters faster. These integrated ESNA's, often referred to as the Internet of Emergency Services (IoES), could revolutionize the way emergency services cooperate and act in crisis situations [30].

Another trend that holds potential is predictive analytics using AI, where systems are more equipped to analyze historical and real-time data to identify growing disaster types before they become fully developed [30]. These capabilities will only become more essential as climate continues to change, which will demand more complex models which can adjust to altered environmental conditions (LLM Memory).

Machine learning-based sentiment analysis is becoming more advanced, allowing emergency responders, for example, to understand public reactions during disaster situations, focus on specific concerns, and refute misinformation [30]. This capacity is essential to preserve public trust and coordination during a crisis, when the distribution of accurate information is imperative.

AR and VR technologies have potentials to improve the level of preparedness for a disaster, and community activation [30]. These immersive technology may be used to generate realistic disaster simulation for training emergency responders and raising awareness among the communities on proper reaction to different disaster scenarios (LLM Memory).

In the health sector, AI has become part of disaster risk assessment and emergency health response [36]. While the world is transitioning to an "AI-based world this decade" according to Claus Schwab1, healthcare regulators are now recommended to exploit these new possibilities to optimize outcomes during catastrophic events [36].

GIS and AI would increasingly serve as important tools for mapping spatial distributions of different hazards and their vulnerability [9][10]. Such integrated systems will increasingly support long-term planning and real-time decision making for better preparedness for hazards and management of crises.

Efforts such as MOBILISE are examples of how connected analytical solutions can enhance resilience in the face of disasters and enable responses that can facilitate quicker recovery of affected communities [9][37]. The future disaster management systems may extend these foundations to more advanced platforms with the combination of several technologies.

Deep-learning based image analysis has seen rapid development, and there is great potential for deep CNN (Convolutional Neural Networks) in damage assessment from social media images in disasters [19]. Through these more well-developed systems, disaster impacts can be more rapidly and accurately evaluated and thus will better guide more effective allocation of resources [30].

These technologies are being advanced as "force multipliers" in securing lives and property against disasters [9], [37-39. The future of disaster management belongs to the concerted use of these new technologies in combination to form a smarter, more responsive, and more adaptive disaster management infrastructure able to grapple with the increasingly difficult problems that natural disasters present in our convoluted world [5][40][41].

VI. Conclusion

This study provides a comprehensive scientometric analysis of the evolution, collaboration patterns, and thematic trends

in the application of artificial intelligence (AI) to disaster risk management (DRM) from 2003 to 2025. Based on a final corpus of 7,842 articles extracted from Scopus, Web of Science, and OpenAlex, the findings highlight the exponential growth in scientific output, particularly from 2020 onward, reflecting both technological maturation and an urgent global need for innovative DRM solutions amid intensifying climate-related hazards.

Several major contributions emerge from this work:

- It maps the trajectory of AI in DRM, showing how machine learning, deep learning, and geospatial AI have become core technologies in early warning systems, real-time monitoring, damage assessment, and resource allocation.
- It identifies key actors, countries, and collaboration networks, revealing a concentration of impact in technologically advanced nations while also signaling the rise of new contributors from emerging economies.
- It outlines the main research themes and emerging topics such as federated learning, swarm robotics, and anomaly detection, which suggest a move toward decentralized and adaptive disaster response systems.

However, some limitations must be acknowledged. This analysis, being quantitative and bibliometric in nature, does not assess the operational effectiveness or practical deployment of AI solutions in real-world disaster contexts. The absence of qualitative insights into policy relevance, technological transferability, and user adoption also constrains the conclusions that can be drawn regarding actual impact. Moreover, citation-based metrics may underrepresent innovative or interdisciplinary works that have not yet gained visibility in traditional publication channels.

To address these gaps, future research should:

- Conduct comparative case studies of AI deployment in specific disaster scenarios, evaluating performance, usability, and societal outcomes.
- Investigate the integration of AI with local knowledge systems, especially in low-resource and high-risk regions where disaster vulnerability is most acute.
- Explore the ethical, legal, and governance implications of automated decision-making during emergencies, particularly concerning algorithmic bias, data privacy, and accountability.
- Develop hybrid evaluation frameworks that combine bibliometric analysis with real-time data from disaster events, social media, and satellite imagery to assess responsiveness and situational relevance.

The findings underscore the need for interdisciplinary collaboration among AI experts, disaster response agencies, policymakers, and affected communities to co-design intelligent systems that are not only technically robust but also socially legitimate and context-sensitive.

In a global context where disasters are becoming more frequent, complex, and disruptive, artificial intelligence offers transformative potential. But its promise will only be fulfilled through continuous methodological innovation, transparent governance, and a sustained focus on equity, trust, and long-term resilience.

REFERENCES

[1] C. Gevaert, M. Carman, B. Rosman, Y. Georgiadou, and R. Soden, "Fairness and accountability of AI in disaster risk management: Opportunities and challenges," *Patterns*, 2021.

[2] V. Singh and A. Agnihotri, "Addressing Environmental Challenges through Artificial Intelligence (AI)-Powered Natural Disaster Management," *International Journal of Applied and Scientific Research*, 2024.

[3] S. V. Raut, "Artificial Intelligence Use in Disasters Management," *International Journal of Innovative Science and Research Technology*, 2024.

[4] V. Linardos, M. Drakaki, P. Tzionas, and Y. L. Karnavas, "Machine Learning in Disaster Management: Recent Developments in Methods and Applications," *Machine Learning and Knowledge Extraction*, 2022.

[5] R. R. Arinta and E. A. W. R., "Natural Disaster Application on Big Data and Machine Learning: A Review," *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2019.

[6] L. Dwarakanath, A. Kamsin, R. A. Rasheed, A. Anandhan, and L. Shuib, "Automated Machine Learning Approaches for Emergency Response and Coordination via Social Media in the Aftermath of a Disaster: A Review," *IEEE Access*, 2021.

[7] S. Saravi, R. Kalawsky, D. Joannou, M. Rivas Casado, G. Fu, and F. Meng, "Use of Artificial Intelligence to Improve Resilience and Preparedness Against Adverse Flood Events," *Water*, 2019.

[8] H. Kemper and G. Kemper, "SENSOR FUSION, GIS AND AI TECHNOLOGIES FOR DISASTER MANAGEMENT," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2020.

[9] S. K. Abid, N. Sulaiman, S. Chan, U. Nazir, M. Abid, H. Han *et al.*, "Toward an Integrated Disaster Management Approach: How Artificial Intelligence Can Boost Disaster Management," *Sustainability*, 2021.

[10] J. Chen, Q. Li, H. Wang, and M. Deng, "A Machine Learning Ensemble Approach Based on Random Forest and Radial Basis Function Neural Network for Risk Evaluation of Regional Flood Disaster: A Case Study of the Yangtze River Delta, China," *International Journal of Environmental Research and Public Health*, 2019.

[11] M. Arslan, A. Roxin, C. Cruz, and D. Ginhac, "A Review on Applications of Big Data for Disaster Management," *International Conference on Signal-Image Technology and Internet-Based Systems*, 2017.

[12] A. Faruq, S. Abdullah, A. Marto, C. M. C. Razali, and S. F. M. Hussein, "Flood Forecasting using Committee Machine with Intelligent Systems: a Framework for Advanced Machine Learning Approach," *IOP Conference Series: Earth and Environment*, 2020.

[13] S. Jain, P. Mani, S. Jain, P. Prakash, V. Singh, D. Tullos *et al.*, "A Brief review of flood forecasting techniques and their applications," 2018.

[14] Morocco Solidarity Hackathon, "Leveraging AI for Natural Disaster Management: Takeaways From The Moroccan Earthquake," *arXiv.org*, 2023.

[15] A. P. Anindita, P. Laksono, and I. B. Nugraha, "Dam water level prediction system utilizing Artificial Neural Network Back Propagation: Case study: Ciliwung watershed, Katulampa Dam," *International Conferences on Information Science and System*, 2016.

[16] R. Adnan, F. Ruslan, A. Samad, and Z. Zain, "Artificial neural network modelling and flood water level prediction using extended Kalman filter," *IEEE International Conference on Control System, Computing and Engineering*, 2012.

[17] Y. Pouresmaeil, S. Afroogh, and J. Jiao, "Mapping out AI Functions in Intelligent Disaster (Mis)Management and AI-Caused Disasters," *arXiv.org*, 2025.

[18] F. Ruslan, A. Samad, Z. Zain, and R. Adnan, "Flood prediction using NARX neural network and EKF prediction technique: A comparative study," *2013 IEEE 3rd International Conference on System Engineering and Technology*, 2013.

[19] T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage Assessment from Social Media Imagery Data During Disasters," *International Conference on Advances in Social Networks Analysis and Mining*, 2017.

[20] M. M. Torok, M. G. Fard, and K. Kochersberger, "Image-Based Automated 3D Crack Detection for Post-disaster Building Assessment," *Journal of Computing in Civil Engineering*, 2014.

[21] G. Pang, "Artificial Intelligence for Natural Disaster Management," *IEEE Intelligent Systems*, 2022.

[22] K. Kochersberger, K. Kroeger, B. Krawiec, E. Brewer, and T. Weber, "Post-disaster Remote Sensing and Sampling via an Autonomous Helicopter," *J. Field Robotics*, 2014.

[23] F. H. Chowdhury, R. Nahian, T. Uddin, S. Rezwan, M. Khan, and A. Sufian *et al.*, "Design, control & performance analysis of forecast junction IoT and swarm robotics based system for natural disaster monitoring," *International Conference on Computing Communication and Networking Technologies*, 2017.

[24] S. Fang, L. Xu, Y. Zhu, Y. Liu, Z. Liu, and H. Pei *et al.*, "An integrated information system for snowmelt flood early-warning based on internet of things," *Information Systems Frontiers*, 2013.

[25] J. Cen, T. Yu, Z. Li, S. Jin, and S. Liu, "Developing a disaster surveillance system based on wireless sensor network and cloud platform," 2011.

[26] C. Recchiuto and A. Sgorbissa, "Post-disaster assessment with unmanned aerial vehicles: A survey on practical implementations and research approaches," *J. Field Robotics*, 2018.

[27] J. DeVault, "Robotic system for underwater inspection of bridge piers," 2000.

[28] K. Nagatani, S. Kiribayashi, Y. Okada, K. Otake, K. Yoshida, and S. Tadokoro *et al.*, "Emergency response to the nuclear accident at the Fukushima Daiichi Nuclear Power Plants using mobile rescue robots," *J. Field Robotics*, 2013.

[29] Q. Huang, G. Cervone, D. Jing, and C. Chang, "DisasterMapper: A CyberGIS framework for disaster management using social media data," *BigSpatial@SIGSPATIAL*, 2015.

[30] D. Erokhin and N. Komendantova, "Social media data for disaster risk management and research," *International Journal of Disaster Risk Reduction*, 2024.

[31] D. Velev and P. Zlateva, "Challenges of Artificial Intelligence Application for Disaster Risk Management," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2023.

[32] K. P. Chun, T. Octavianti, N. Dogulu, H. Tyralis, G. Papacharalampous, and R. Rowberry *et al.*, "Transforming disaster risk reduction with AI and big data: Legal and interdisciplinary perspectives," *WIREs Data Mining and Knowledge Discovery*, 2024.

[33] S. Kamolov, "Machine learning methods in civil engineering: a systematic review," *Annals of Mathematics and Computer Science*, 2024.

[34] M. Yu, C. Yang, and Y. Li, "Big Data in Natural Disaster Management: A Review," 2018.

[35] Y. Ouahbi and S. Ziti, "NW Logistics: System Architecture and Design for Sustainable Road Logistics," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 16, no. 4, 2025. DOI: http://dx.doi.org/10.14569/IJACSA.2025.01604105.

[36] Y. Ouahbi and S. Ziti, "Revolutionizing Road Safety and Optimization with AI: Insights from Enterprise Implementation," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 16, no. 4, 2025. DOI: http://dx.doi.org/10.14569/IJACSA.2025.0160498.

[37] R. Mharzi, A. Ben Kacem, H. Mharzi, and A. Elouadi, "Crisis logistics and natural hazards risk management in Morocco: a discrete-event simulation–based assessment of proactiveness and responsiveness," *OPSEARCH*, Accepted: 7 February 2025. DOI: https://doi.org/10.1007/s12597-025-00936-8.

[38] R. Mharzi, A. Ben Kacem, and A. Elouadi, "Catastrophe-related disruptions' preparedness and emergency management in Morocco: a proactive risks and resilience digital twin-based analysis," *Journal of Modelling in Management*, vol. 20, no. 1, Article published online: 17 September 2024, Issue publication date: 3 March 2025. ISSN: 1746-5664.

[39] G. Amarnath, U. Amarasinghe, and N. Alahacoon, "Disaster Risk Mapping: A Desk Review of Global Best Practices and Evidence for South Asia," *Sustainability*, 2021.

[40] L. F. Bari, I. Ahmed, R. Ahamed, T. A. Zihan, S. Sharmin, and A. H. Pranto *et al.*, "Potential Use of Artificial Intelligence (AI) in Disaster Risk and Emergency Health Management: A Critical Appraisal on Environmental Health," *Environmental Health Insights*, 2023.

[41] S. K. Abid, N. Sulaiman, C. S. Wei, and U. Nazir, "Building Resilient Future: Information Technology and Disaster Management- A Malaysian Perspective," *IOP Conference Series: Earth and Environment*, 2021.

XPathia: A Deep Learning Approach for Translating Natural Language Into XPath Queries for Non-Technical Users

Karam Ahkouk, Mustapha Machkour
Computer Systems and Vision Laboratory-Faculty of Sciences,
Ibn Zohr University, 80000 Agadir, Morocco

***Abstract*—XPath is a widely used language for navigating and extracting data from XML documents due to its simple syntax and powerful querying capabilities. However, non-technical users often struggle to retrieve the needed information from XML files, as they lack knowledge of XML structures and query languages like XPath. To address this challenge, we propose XPathia, a novel deep learning-based model that automatically translates natural language questions into corresponding XPath queries. Our approach employs supervised learning on an annotated XML dataset to learn accurate mappings between natural language and structured XPath expressions. We evaluate XPathia using two standard metrics: Component Matching (CM) and Exact Matching (EM). Experimental results demonstrate that XPathia achieves a state-of-the-art performance with an accuracy of 25.85% on the test set.**

***Keywords*—Deep learning; XML databases; neural networks; text-to-XPATH; natural language processing**

I. Introduction

The use of databases is the activities' base that the majority of companies rely on. Many types of databases exist in the world, amongst them there are Graph databases, NoSQL databases, Object-oriented databases, XML databases, Relational databases, etc. All this types of databases are based on containers called database management systems that help manipulate these databases and give accessibilities features for users and also for other applications.

Dealing with these systems is not an easy task for people who lack of knowledge on database management. Also the use of these systems requires advanced skills and enough expertise on some specific technical languages, for instance, the ignorance of some languages like XPath or XQuery for getting the relevant information form XML document in XML databases is considered as a big obstacle and can create limitations to get the searched data in a reasonable time.

An ideal solution would be to communicate with the database management systems using only the mother-tongue language; this can be done by using an intermediate interface that handles well organized natural language sentences and translate them to the target script which is in our case XPath. Xpath is one of the languages that help manipulate the XML documents included in XML databases in an easy and straightforward way. In contrast with the other languages, XPath has a very simplified syntax and direct structure; also the possibilities that it affords for extracting data from xml documents helps people get the searched information without complexities. Xpath is the one that can be used with minimal effort; its syntax that is based on Path concept provides 200 built-in functions with expressions and operators to extract XML nodes and their values.

The type of this interface enters in the category of solutions based on Natural Language Processing (NLP) tasks that process NL sentences in order to make a corresponding translation, doing actions, generating codes, etc. There are many sub-tasks of the NLP that deal with natural text transformation. For example, Text classification, Named entity recognition, Information retrieval, Text summarization, Text generation, Sentiment/emotion analysis are all sub tasks of NLP. The objective of all these solutions is the ability to handle the meaning from a raw text and transform it to another form that answers a particular problem. In our case the problem is similar to Text classification and Information retrieval since the goal is to classify the user natural language sentence and get the relevant information in order to build a target XPath query which will be used for extracting the wanted information from an XML document.

The focus in this work is to leverage the ability to interact with XML databases using only the natural language. In other words, the possibility to translate the sentences of users to a runnable and valid Xpath query with all its paths, nodes and expressions. In this paper, we present our solution that handles the problem of using the Natural Language as a source medium to interact with XML Databases. The proposed model is based on Deep learning with a Path-Expression based approach that treat the path, nodes and predicates in an independent manner in order to reduce errors. The work is evaluated against a new XML dataset that we provide publically on GitHub https://github.com/karamsa/XMLSet using two metrics, the CM (Component Matching) and EM (Exact Matching), and achieves a score of 25.85% on the test set of the XML corpus to be a new base for future models for the training and the evaluation.

A. Contribution

In this paper, we present 2 contributions to the task of translating Natural Language to Xpath queries for XML databases using Deep Learning and a Path-Expression based approach to make prediction of nodes, Expression, predicates, operators and values independently and without using the content of the XML Nodes. Our contribution to this area is two fold:

- We propose a new XML Dataset that contains up to 2176 pairs of natural language sentences with their XPath queries and annotations in a well structured and easy-to-use format. We provide this Dataset publicly on GitHub for future models for training and evaluation purposes.
- A new model (we called it Xpathia) that makes predictions of Xpath queries using deep learning and Path-Expression based approach. The model predicts each part of the target query independently using several sub-modules.

B. Outline

The reminder of this paper is structured as follows: The Section 2 presents the related works regarding the previously proposed solutions. Section 3 and 4 presents our contributions on the translation of Natural Language sentences to XPath. Section 5 includes the evaluations of the model. The section 6 discusses the experimental results and the final section contains the conclusions of our work.

II. Related Works

XML is a widely used markup language, valued for its simplicity in data extraction. Originally developed for large-scale electronic publishing ([1]), it is now applied in diverse areas such as data sharing, system communication, and database storage. Its flexible format supports both structured and unstructured content, and its platform- and language-independence have contributed to its broad adoption ([2]).

Recent research in XML has focused mainly on content for tasks like information retrieval and document clustering ([3]), often overlooking the structural aspects. This neglects a key feature of XML—its hierarchical structure and meaningful tag names—which are essential for accurate content interpretation ([4]; [5]). To improve retrieval and analysis, both content and structure should be utilized together ([6]; [7]).

Databases are central to most organizations, ranging from relational and NoSQL to XML and graph databases. These are managed through database management systems (DBMS) that enable access and manipulation of data. However, using such systems can be challenging for users without technical expertise. In particular, querying XML databases requires knowledge of XPath or XQuery—languages that many users find difficult to learn. This creates barriers to accessing information efficiently.

To address this, natural language interfaces (NLIs) have been proposed to translate user-friendly queries into structured query languages. In [8] the authors provides a foundational overview of NLIs, highlighting their potential to make database interaction more accessible, especially for non-experts.

While XML's self-descriptive design enables platform-agnostic data storage and transport, translating natural language (NL) directly into XML structures remains inherently challenging. NL, which is an unstructured language, is governed by grammar and contextual rules that are dedicated for communication purposes, whereas XML organizes data hierarchically for machine readability and data storage. This fundamental mismatch necessitates intermediate query languages like XQuery or XPath in order to create a bridge between Natural Language and XML. This intermediate language can play a pivotal role to remove the semantic as well as the structural gap.

The existing research addressing this challenge falls into two broad categories: the ones that are linguistic and rule-based and the others emerging deep learning methods, though the latter remains underexplored.

The linguistic rule-based approaches is the category of methods that rely on syntax parsing, part-of-speech tagging, and predefined rules to map NL to XML queries.

Tannier et al. ([9], [10]) utilized the INEX 2004 dataset to develop a system combining part-of-speech tagging, context-free grammar rules (e.g., NP for noun phrases), and discourse representation theory (DRT). By constructing discourse representation structures (DRS) from NL sentences and converting them into SQL-like queries, their approach retrieves XML nodes. However, it struggles with lengthy inputs, negations, and requires synonym dictionaries (e.g., mapping "employee" to "emp") to align NL terms with XML schemas.

Li et al. ([11]) introduced NaLIX, a natural language interface leveraging MINIPAR for syntactic parsing and iterative user feedback to refine XQuery outputs. While functional, NaLIX relies on manual rule engineering and interactive prompts, limiting its scalability.

To simplify query generation for non-expert users, Li et al. ([12]) proposed the Meaningful Query Focus (MQF) framework. MQF decouples translations from XML schema specifics, enabling cross-domain compatibility. However, it mandates exact tag names, necessitating synonym substitution layers to align user terminology with schemas.

Joseph et al. ([13]) expanded versatility by supporting multi-sentence queries, yes-no questions, and nested structures. Their system parses dependencies, replaces tokens with XQuery fragments, and uses WordNet ([14]) for synonym mapping. Despite its flexibility, it falters with unseen domains or terms absent from WordNet, underscoring reliance on predefined lexicons.

The hybrid and schema-agnostic approaches is the category of methods that include efforts to reduce dependency on rigid schemas or manual rules. Nassiri et al. ([15], [16]) designed a system to unify queries across data models (including XML) via syntax validation, tokenization, and intermediate universal query generation. While effective for structured inputs, it lacks NL compatibility and relies on manual engineering.

Gupta et al. ([17]) combined dependency parsing with Paninian grammar for semantic parsing, mapping NL chunks to domain models. Though innovative, their hybrid approach lacks adaptability for complex inputs without deep learning integration.

Dorrn et al. ([18]) emphasized data-driven methods, combining XML and NLP techniques for automatic information extraction from corpora. However, limited evaluations and reliance on mixed methodologies hinder practical scalability.

Based on semantic and for cross-Domain databases subhro et al. ([19]) proposed declarative rules to map NL descriptions

to formal queries (e.g., XPath), while Lopez et al. ([20]) developed PowerAqua, a tool translating NL into logical queries across heterogeneous semantic sources. PowerAqua aggregates answers without requiring schema knowledge but faces challenges in web-scale diversity.

XML is a flexible, structured language widely used for data storage and sharing, but querying it requires technical skills in XPath or XQuery. Existing solutions use rule-based or traditional methods to translate natural language into XML queries but often lack scalability and adaptability. Deep learning offers a promising alternative by reducing manual rule dependency and improving cross-domain performance.

III. XML Datasets

As we have previously, presented the base idea is the subdivision of the problem into multiple small problems without following a pre-defined sketch that still in the same time dissociates the tasks and operations done by the main model and create an independency between the output spaces, thing that is essential for a healthy prediction process.

The first problem encountered while working on this problem is the lack of databases dealing with this transformation type; thus creating a database for evaluating the model is mandatory in order to have a robust model that can generalize for unseen inputs. Bridging the natural language to Xpath was another obstacle for the model, since they both have different natures and purposes and Xpath is the intermediate language for extracting content portions from XML documents.

A. Existing Corpuses

The decision to build an XML dataset fully dedicated to answering questions in natural language with structured labels was based on the lack of similar corpuses that can be used for the evaluation of our model. This task was one of the harsh steps done prior the creation of the model and it is essential for a preferment system that must generalize not only to unseen inputs but also for unseen structures in the sets of the dataset.

From the literature, the use of Natural language to extract XML portions using structured languages wasn't deeply considered and has been treated as one of most difficult tasks in the NLP area.

This was mainly due to two reasons; first is the incompatibility of the two languages. Second is the lack of corpuses for the evaluation of the wanted model. In this context, the number of corpuses that were presented in this area in the recent decade was very limited. While these few datasets were created in the first place as a contribution to the task, in the other hand they bear lots of technical issues and difficulties that make their use such an impossible thing especially for the training and the evaluation of our model.

XML retrieval, the process of querying and extracting structured information from XML documents, relies heavily on specialized datasets to advance research, benchmark systems, and refine methodologies. These datasets are tailored to address diverse challenges in handling XML's hierarchical and semi-structured nature, offering varied use cases ranging from algorithmic stress-testing to domain-specific analyses. Among these datasets, the INEX Dataset ([21]) stands out as a foundational resource. Developed through the Initiative for the Evaluation of XML Retrieval, INEX provides standardized test collections that simulate some limited scenarios, enabling researchers to evaluate the precision, recall, and efficiency of XML retrieval algorithms. Its role in academic competitions and collaborative benchmarking has made it a popular dataset for comparing old systems that process structured queries, particularly in environments where granular content extraction—such as retrieving specific sections of documents—is not critical.

For performance-oriented evaluations, the XMark Benchmark ([23]) was proposed for this purpose. This synthetic dataset models an online auction platform, generating XML data with deeply nested hierarchies and scalable file sizes. Its design intentionally incorporates complexity to mimic real-world data structures that is useful for making stress-testing XML databases and query engines. Researchers use mainly use XMark to measure how systems handle large-scale data, complex XPath queries, and join operations, providing insights into scalability and optimization strategies. However that corpus is not intended for making natural language translation to XPath queries as it doesn't has this structure. Complementing this, the SIGMOD Record XML Dataset ([22]), which is another dataset for XML files with focuses on XML query processing techniques. Curated by the ACM Special Interest Group on Management of Data, this dataset supports experiments in query optimization, indexing, and retrieval efficiency, often serving as a testbed for novel methodologies in academic publications. While it is adequate for research targeting query optimization, this cannot be used for our case as there is no samples that handle the translation pairs.

Domain-specific datasets further enrich the landscape of XML retrieval research. The DBLP XML Dataset ([24]), for instance, offers a comprehensive corpus of computer science publication metadata. Covering millions of journal articles, conference papers, and academic references, it enables studies on bibliographic data retrieval, citation analysis, and trends in scholarly communication. Its structured yet semantically rich content makes it invaluable for exploring hybrid approaches that combine XML retrieval with bibliometric analytics. From one hand, it can be seen as a challenging dataset that put together an important volume of XML files, however, it is structure does not involve the natural language questions that are required for handling the task by our model, and thus it is unusable for this kind of training or evaluation. In the same context, the domain specific nature of the corpus creates an obstacle for the model to generalize for unseen domain which is an unwanted behavior for our model.

Similarly, the IMDB XML Dataset ([25]), derived from the Internet Movie Database, provides a heterogeneous collection of entertainment industry data. This includes detailed records on films, TV series, actors, directors, and production crews, offering researchers a platform to tackle challenges in querying diverse entity types, managing cross-linked data, and extracting insights from semi-structured multimedia metadata. Despite having a corpus that contains huge volume of domain specific data structured in XML files, these latter can't be transformed to a structured question-answer dataset to be eligible for the training and evaluation not only for our model but also for future models dealing with the process of translation.

Together, these datasets cannot be used for distinct facets of NLP using deep learning. Benchmarks like INEX and XMark prioritize technical rigor, enabling apples-to-apples comparisons of system performance, while domain-specific resources like DBLP and IMDB facilitate applied research in real-world contexts. For a granular comparison of these datasets—including their scope, structural complexity, and limitations —Table I synthesizes key attributes with a detailed comparison of their features.

TABLE I. AN OVERVIEW OF THE EXISTING DATASETS

Dataset	Cross Domain	Number of Elements	Max-depth	Avg-depth	Limitations
INEX	Yes	12107	-	6.9	irregular tagging, incomplete schemas
SIGMOD Record XML	No	11526	6	5.14107	less useful for cross-domain XML retrieval
XMark Benchmark	No	Depends on configs	Depends on configs	Depends on configs	Artificially generated hierarchies
DBLP XML Dataset	No	3 000 000	5-7	3.5	Not usable for semantic search or xml retrieval
IMDB XML dataset	No	Variable	Variable	Variable	Not usable for semantic search or xml retrieval

For the purpose of compatibility and since all the previous datasets presented in this paper don't offer a base for NLP tasks as well as the lack of XPath as a target language for the translation, we decided to build a new dataset (we called it XMLSet) that can encompasses all the missing things in one corpus and could be used by future models either for validation or training purposes.

B. XMLSet

The corpus is built based on different xml sources and underwent series of steps in order to have a usable dataset in the context of NLP. First, the xml databases have been cleaned and preprocessed automatically in order to remove non-sense names and replace non significant names with the ones that have signification. Second and since all XML databases are collected from existing source from internet as shown in https://github.com/karamsa/XMLSet, we have converted some of the xml sheets to XML databases in order to leverage our dataset with data from real word that is crucial for making a model that can generalize to unseen data; also we have kept the nature of some of the xml data that contain more than one XML root. We consider this as a necessary noise that makes the model more robust for real usage. For the aim of having a cross-domain dataset, we made sure that our new corpus includes a variety of domains from bookstores, purchases, plants, food, movies, journals, and more than 28 distinct topic; the table II shows a general comparison between the existing datasets and the XMLSet created by us. Since the XMLSet differs totally from the existing corpuses, we only show shared properties' statistics, such as the depth of xml nodes, average dept, and number of elements in the dataset and so on.

TABLE II. XMLSET AGAINST THE EXISTING DATASETS

Dataset	Cross Domain	Number of Elements	Max-depth	Avg-depth
INEX	Yes	12107	-	6.9
SIGMOD Record XML	No	11526	6	5.14107
XMark Benchmark	No	Depends on configs	Depends on configs	Depends on configs
DBLP XML Dataset	No	3 000 000	5-7	3.5
IMDB XML dataset	No	Variable	Variable	Variable
XMLSet	**Yes**	**37,788**	**8**	4

The construction of the XMLSet dataset was a prior step in the process of making a model that deals with NL to XML through XPath as an intermediate language. In contrast with the existing datasets that are not dedicated to NLP processing tasks, the XMLSet is the first corpus that can be used for models that rely on deep learning. It contains more than 2176 pairs in form of a natural language sentence and an answer in XPath language. The dataset is also leveraged with a ground truth section for each pair that describes the answer in a structured way making it adequate not only for deep learning processing but also for any XML retrieval form and can be used by models for training and evaluation purposes. The Figure 1 shows an example of a pair from XMLSet including the details about the labels sections (ground truth).



```
{
  "question": "find all addresses but the first two ones",
  "query_id": 2026,
  "xpath": "//address[position()>2]",
  "ground_truth": {
    "slots": [
      {
        "root_type": "//",
        "node": "address",
        "source": "position()",
        "op": ">",
        "value": "2"
      }
    ]
  },
  "schema_index": 3
},
```


Fig. 1. An Example of a natural language question and its answer including the details about the labels sections from XMLSet dataset.

Regarding the pairs statistics, the dataset is built on top of 30 separate XML database files. Some of them are bigger than the others especially in terms of number of elements and this is particularly helpful to ensure a diverse size across the whole corpus that is divided onto 2 sets. The data split of XMLSet is as shown in table III:

In terms of difficulty, we made sure that XMLSet contains a diversified pairs with easy, moderate and difficult queries;

TABLE III. THE DATA SPLIT OF XMLSET

	Train	Test
#Pairs	1472	704
#Unique NLs	1255	673
#Databases	30	30

The table IV shows a detailed analysis of the XMLSet.

TABLE IV. XMLSET'S DETAILED STATISTICS

Dataset	# Relative Paths	# Multi-paths	# Predicates	# Functions
XMLSet	77	**743**	**770**	**324**

In XPath, the difficulty of a query can be measured using different criteria. First we find the number of the nested paths which reflects the number of nodes to be included in the final XPath query. Then comes the number of predicates which has a direct link to the targeted nodes in the XML file. Another element that adds difficulty to XPath queries is the inclusion of functions in the predicates; thus is considered as a double difficulty as it puts together the predicates and functions in one place. Relative paths, that include the '..' are also a factor of complexity in XPath queries; thus adding it to some pairs in the dataset challenge the model on those complex transformations.

In the same context, we built XMLSet to follow the question-answer paradigm; hence not all queries return results as we want to have a system that can work in real situations where a query entered by a user might have no corresponding answer in a database.

IV. THE MODEL

A. Overview

Exploiting XML files using natural language solely, requires an intermediate language that can manage the translation operation by bridging the nature of each language; thus the use of a structured language such as XPath is mandatory for this process. The use of XPath comes from multiple factors. First, its simplicity; XPath is a simple language that relies on paths. It provides the possibility to extract nodes from XML files using a short syntax. Additionally, XPath can be applied directly to XML documents in comparison to other languages like XQuery which requires to be re interpreted. Another reason for choosing XPath is the richness in terms of functions; it provides more than 200 useful functions that can be used along with inherent expressions to exploit XML files effectively.

The conversion of NL sentences into XPath is a thorough task that should take in consideration the meaning inherent in the user sentence. To this end and since the two languages are different in nature, we decided to simplify the problem by adopting an straightforward mapping.

The model should predict the nodes and the predicates to be included in the final XPath query. With nodes the subset of the following items:

$$\text{N=[T; ”null”, ”.”, ”..”, ”*”, ”1”, ”last()”, ”position()”]} \quad (1)$$

In addition to T (the tags from the XML document) we also add to the output space the relative paths '.' and '..', '*' for any tag in the document, '1' and 'last()' for the first and last node and finally the 'position()' that helps select nodes based on a given position.

In the other side, a predicate P is a structure that holds multiple elements: A node from the previous list N, an operator from the set of possible operators O and a value from the list of values V mentioned in the user input.

$$\text{P = [N(i), O(l), V(t)]} \quad (2)$$

We define O as the list of supported operators in the XPath syntax as following:

$$\text{O = [”null”, ”=”, ”<”, ”>”, ”<=”, ”>=”, ”!=”, ”+”, ”-”]} \quad (3)$$

Using those elements and by exploiting the incorporate structure of XPath, we reformulate the problem and we simplify it further by limiting the output space in each item of the target query.

The problem then can be summarized as the prediction of one or multiple paths of nodes and predicates in the following format:

$$Q = //N_1P_1/N_2P_2.../N_nP_n \quad (4)$$

The paths are separated by delimiters of two types; the '//' which refers to selecting elements in the XML document from no matter where they are and that correspond to the current selection in the path. The second type is '/' which provides the possibility to select nodes depending on the root path or the previous one.

For example, if we take an XML file that stores the orders done by clients, the following XPath query will find the last saved order in the document:

$$//orders[last()] \quad (5)$$

We notice that the query contains only one path that includes the 'order' node and a predicate which contains a node as a function 'last()'. The operator and the value for the predicate in this case are optional. This query can take another form. For example selecting the last but one order saved in the XML document:

$$//orders[last() - 1] \quad (6)$$

In this case, the predicate contains a node as a function 'last()', the operator '-' and the value '1'. The table V shows examples of XPath queries with their descriptions.

TABLE V. SOME XPATH QUERIES AND THEIR DESCRIPTION

Query	Description
//orders[last()]	Get the last order wherever it is.
//orders[last()-1]	Get the last order but one from all order in the XML database.
//categories/description	Get the description node of categories nodes.
//cd[year=1985]	get the cds that have been published in 1985.
//company/employee[salary¿5000]	get companyies' employees that have a salary higher than 5000.
//company/employee[1]/*	get all information of the first employee of each company.

The same pattern can be used to represent complex XPath queries that involve different nodes, conditions, and multiple paths in the same XPath request. Hence an XPath query can be split into different parts; each one will be part of a specific predication process independently to the others.

B. Model Details

The goal of this work is the creation of a model that can generate XPath queries effectively using the learned patterns in the training dataset of XMLSet.

Since the translation of the natural sentence to XPath is a problem that can be reformulated as the prediction of one or multiple paths separated by roots and while each path can be represented by a node and a predicated as presented in the previous section, we can then use this powerful structure of XPath to make prediction of the target query in an effective manner that can help the model generate to unseen inputs from users.

We split the problem into several sub-problems in a way that dissociates the inputs and the outputs of each element in the prediction process. In this context and based on an input from the user, the natural language sentence as well as the nodes of the XML file are fed to the model that run a series of prediction operations in order to construct a valid query that can be executed against an XML engine, so the result is returned to the user as shown in figure 2 that displays the architecture of our model (XPathia) in details.

The natural language sentence is tokenized and the XML file is preprocessed to get the XML nodes that make the structure's file then the two inputs are concatenated each other and the Meta data are added. We leverage the combined inputs with separators to distinguish between the user input and the XML datafile. We use [SEP] from the BERT pretrained model [26] to separate the natural language sentence form the structure of the XML file and also for separating each tag automatically extracted from the XML document. The BERT (Bidirectional Encoder Representations from Transformers) is a language representation model that provides pertinent contextualized word embedding in contrast to static one proposed by other models. In addition to the contextualized word embeddings that BERT provides using representation from left and right layers of the transformers model [27], we also leverage our model

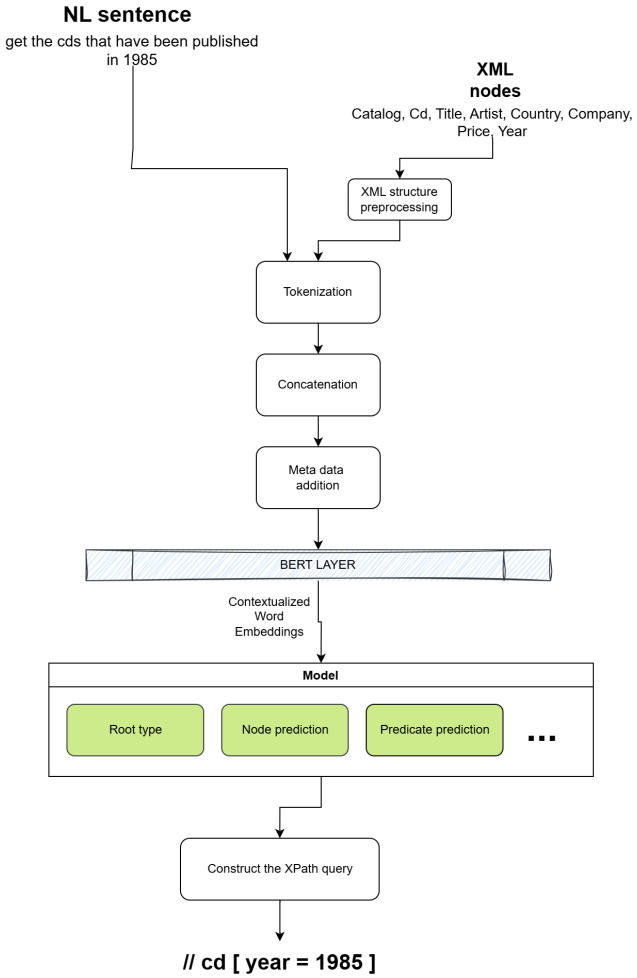


Fig. 2. The Architecture of our system.

with the [SEP] token as a separator as well as the special token [CLS] that holds the whole meaning of the input in one word embeddings. This token can concentrate the representation from both sides as BERT is a bidirectional model. These metadata are essential to get focused representations that will be used in the next step. The query generation process takes place the BERT layer is completed. Our model is composed of several sub modules and each sub-module is dedicated to generate specific elements of the query. We don't use a sketch filling approach as in our case the whole query will be predicted and this will be done once not sequentially. Since the redefined structure of an XPath query is more straightforward as have been shown in the previous section, we define six sub-modules for generating the target XPath query as shown in figure 3.

For each sub-module, the sentence S and the nodes N will be fed as I:

$$I=[[CLS];S;[SEP];n_1;[SEP];n_2... \ [SEP]n_z] \quad (7)$$

The first step consists of predicting the number of paths that will construct the final XPath query. In contrast to what we have done in other works; we are not choosing a template from

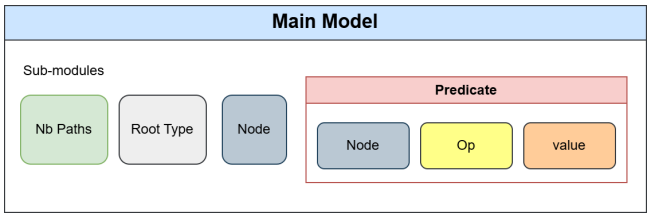


Fig. 3. The Main model and the included sub-modules for the generation process.

a predefined set of possible structures, but we will generate the query structure using a dedicated sub-module. This lets the generation of a non predefined structure possible.

$$P_n(n|I) = softmax(X_n tanh(Y_n EMB_{[\text{CLS}]}))_i \quad (8)$$

we use X and Y as two matrices with separate trainable parameters that help predict the number of paths in the final query. This sub-module is based on the use of the attention mechanism [28] and the work done by sqlnet (2017)). The softmax function normalizes the scores of all probabilities and the top one is selected as the final number of paths in the target XPath query.

As a path in XPath always starts either by '//' or by '/', the root type module is triggered to predict the root types of all the paths in the final query. Using the predefined list of root types ['/', '//'] of each path we put:

$$\begin{aligned} r &= V_{RootType} tanh(W_{RootType} EMB_{[\text{CLS}]}) \\ P_{RootType_{(i)}}(RootType|I) &= softmax(r) \end{aligned} \quad (9)$$

with $\text{EMB}_{[\text{CLS}]}$ the contextualized embedding of the whole input, $\text{V}_{\text{RootType}}$ and $\text{W}_{\text{RootType}}$ are two matrices with isolated trainable parameters fully dedicated to this sub-module. The softmax function normalizes the probabilities for the root types over each path and the highest ones are chosen for each step. In real work almost all XPath queries starts with '//'; hence a great deal of generated root types are '//' for the first path.

The node sub-module is the one in charge of predicting tags that will be included in each path. The tags are predicted based on the list of the fed XML nodes in the first step and using the meaning extracted from the natural language sentence of the user. A node in a specific path can be computed as following:

$$\begin{aligned} z &= S_{node} tanh(T_{node} R_{[CLS]/node(i)}) \\ P_{node_{(i)}}(node|I) &= softmax(z) \end{aligned} \quad (10)$$

with $\text{EMB}_{[\text{CLS}]/\text{node(i)}}$ the representation of a node i in the list of provided nodes and based on the special token CLS that involve the whole representation of the input; this way we enforce the meaning and concentrate the representation for each node extracted from the xml document . S_{node} and W_{node} are two matrices with isolated trainable parameters fully dedicated to this sub-module. The softmax function normalizes the probabilities for the xml nodes over each path and the top candidates are chosen to be predicted as outputs.

Regarding the predicates and since they are composed of three sub-modules (Node, Operator and Value) we decided to use the same structure of modules as before. For instance, the node sub-module within the predicate has the same structure as the one in the path. Both they generate a node based on the list of xml tags from the input and based on the [CLS] token, however the latter is shown before the predicate while the other is inside the predicate. While the two sub-modules use the same structure, they differ in the parameters. We use independent matrices for each sub-module even if they have the same structure; this can be understood as we want to dissociate the training of each module. The node of a predicate can be computed as following:

$$\begin{aligned} x &= M_{node_{predicate}} tanh(N_{node_{predicate}} R_{[CLS]/node(i)}) \\ P_{node_{predicate_{(i)}}}(node_{predicate}|I) &= softmax(x) \end{aligned} \quad (11)$$

The matrices $\text{M}_{\text{node}_{\text{predicate}}}$ and $\text{N}_{\text{node}_{\text{predicate}}}$ contain independent trainable parameters exclusively dedicated to this sub-module and not shared with the others. The softmax function normalizes the probabilities of XML nodes along each path, selecting the top candidates for prediction.

For the operator inside the predicate, we use two separate matrices and the representation of the whole input given by the user. For each path we select one item from the list of possible operators ['null', '=', '¡', '¿', '¡=',' ¿=', '!=', '+', '-'] and we compute:

$$\begin{aligned} q &= E_{Op} tanh(F_{Op} EMB_{[\text{CLS}]}) \\ P_{Op_{(i)}}(Op|I) &= softmax(q) \end{aligned} \quad (12)$$

With E_{Op} and F_{Op} two matrices with trainable parameters specific for this module and will predict an operator for a given path in the target XPath query. The probabilities for each Op(i) are calculated and the top operators are selected for the predicates using the softmax function.

The 'null' element in the list of operators refers to when there is no operator for the predicate; thus the 'null' item will be predicted as the top candidate with the highest probability. In the same context, we leverage all lists with a 'null' value that can be predicted when there is no item to be predicted. For instance, the query bellow doesn't include operator nor a value in the predicate, hence the item 'null' will be predicted for both positions:

$$//messages[1] \quad (13)$$

This gives us the flexibility to stick to the simplified structure of XPath and also run the same number of modules regardless the number of paths and without taking the burden of the structure in the target query.

$$//messages[1, \textbf{null}, \textbf{null}] \quad (14)$$

V. Evaluation Results

Our model (XPathia) is trained using XMLSet dataset that we have presented in previous sections. The training is made on windows OS with a 4G in GPU for the training set of the dataset and for several epochs. The chosen batch size is 4 and we use [29] for the training with default values for the learning and dropout rate, respectively 1e-3 and 0.3. The training process (Figure 4) starts stabilizing after the nineteenth epoch and the accuracy on the training set achieves a score of 29%.

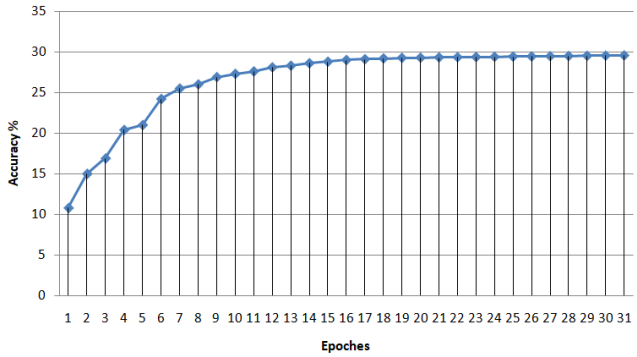


Fig. 4. The Training score on for our model XPathia.

All the sub-modules that composes our model XPathia are trained separately on the pairs of XMLSet that are included in the training set and evaluated against the test set only once.

As we have proposed in our previous work [30] and [31], we use two metrics to evaluate our model. The first metric is the component matching. This metric measures the correctness of outputs of each sub-module and provides an accuracy score on how many parts of the XPath queries have been predicted correctly. Since the XMLSet provides a structured ground truth, the matching is straightforward based on a set matching between the sub-module output and the corresponding item on the ground truth. The loss scores are summed in order to optimize the parameters hence giving better prediction and minimizing the loss score for each sub-module. The table VI shows the performance of XPathia model on the test set of XMLSet dataset using the component matching metric.

TABLE VI. The Performance of XPathia Model on the Test Set of XMLSet Dataset Using the Component Matching Metric

Model	# NB paths	Root types	Node	Predicate		
XPathia	98.86%	66.4%	64.2%	**40.1%**		
				Node	Operator	Value
XPathia	**98.86%**	**66.4%**	**64.2%**	**41.9%**	**91.61%**	**49.00%**

With the 'nb paths', the accuracy of the module that predict the number of steps in the final XPath query. XPathia achieves a score of 66.4% on the prediction of the root types as well as 64.2% of correct answers on the test set for the prediction of the nodes that will construct each path in the query. The accuracy score for the predicates is 40.1% of the correct answers generate by the corresponding sub-module.

Another metric used to evaluate our model XPathia, is the total matching or what we called in our previous works Exact matching. This metric helps measure the correspondency of a generated query by the components of the main model with the ground truth query. The whole output is considered as correct when all parts match exactly all items in the ground truth data.

The table VII shows the performance of XPathia when measured with the Exact matching on the XMLSet dataset.

TABLE VII. The Performance of XPathia When Measured With the Exact Matching on the XMLSet Dataset

Model	Train set (last epoch)	Test set
XPathia	29.0%	25.85%
XPathia-wo-values	**42.8%**	**36.05%**

The model achieves a score of 25.85% of correct queries on all pairs seen in the test set of XMLSet while it achieves a total score of 29.0% on the training set when trained for several epochs. We also explore the evaluation without taking in consideration the generated outputs related to values in the final query. When evaluated without taking in consideration the values, it achieves a score of 42.8% on the training set of XMLSet. **XPathia-wo-values** outperforms the base model by 10.2% on the test set and achieves an accuracy score of 36.05% which becomes the state of the art for the task of translating natural language to XPath for the aim of XML extraction.

VI. Discussion

The evaluation of the existing XML datasets in this work leads to several conclusions:

- The existing Cross-domain datasets are not suitable for XML retrieval using deep learning.
- Some of the datasets are configurable in size and complexity, useful for performance benchmarking but not cross-domain, with randomly generated data.
- Datasets with Large bibliographic corpuses are good for XML data management but not suitable for translation model training.
- The presented corpuses with Variable size and complexity are adaptable for research but not cross-domain.

Given these insights, and since there is no available dataset that can fill the lack of corpuses dedicated to NLP and deep learning, we have proposed XMLSet dataset with all advantages for the training and the evaluation of deep leaning based models. We emphasize on the quality of the corpus in terms of the size as well as the diversification of pairs with a wide coverage in topics like books, products, orders, universities, and more. This dataset is pairing natural language questions with corresponding XPath queries in JSON format and is split into train and test sets. While the XMLSet is a good starting point for future models, we believe that the dataset can be enhanced more by adding more items, creating more complex structures and might be leverage by an intermediate set for fine

tuning the models in the training process. Unlike previous work using XQuery, we introduce XPath to simplify translation, treating it as progressive generation process. We apply this approach based on [30] but with major differences in system architecture, model composition, and problem formulation. We choose a unified XPath structure for any query that can be generated by our model XPathia, and this helped us to reformulate the problem from a complex task to a relatively straightforward one. In comparison to other works that deal with the problem in a totally classical way, we choose to apply deep leaning to our model. The structure of sub-modules plays a pivotal role in the generation process as it provides a total independence between the output spaces. Each sub-module can learn to generate one type of inputs and this helps it enhance the quality of the generated elements as they have separate trainable parameters. In the same context and since all sub-modules parameters are isolated from each other, the errors are not propagated between the models, and thus, having an error in one sub-module doesn't affect the rest of the sub-modules. In relation to the training and as each element is predicted separately from the others, we notice a quick training process. The leaning starts stabilizing after few epochs; this can be understood as the model updates only the faulty parameters of the sub-modules with errors not the whole model, hence the back propagation is limited.

The model XPathia achieves a score of 25.85% on the challenging dataset XMLSet and 36.05% when we omit the values prediction, thing that makes it the new state of the art model. While the accuracy is relatively high for such a task, we believe the model can be more enhanced with exploration of new techniques that can be applied to the model in order to improve the prediction accuracy. We understand that the major improvement can come from the use of LLM models that can have a positive impact to our model. Also we believe that some processing to the input before the starting of the prediction can also have a good impact to XPathia.

VII. Conclusion

In this study, we introduced a new deep learning-based approach for translating natural language into XML queries through an intermediate XPath representation. Our key theoretical contributions include the creation of XMLSet, the first annotated dataset designed for training and evaluating deep models on this task across diverse domains, and the development of XPathia, a novel model that leverages a unified XPath structure for improved generalization. XPathia achieves a 25.85% accuracy on the XMLSet test set, establishing a new benchmark in this domain. Despite these advances, the model still faces limitations in handling complex or nested queries and in generalizing to unseen schema variations. Future research will focus on enhancing the model's ability to understand deeper semantic structures and expanding the dataset with more complex query patterns. This work contributes to the field by introducing scalable, data-driven methods that reduce reliance on manual rule engineering and make XML data querying more accessible to non-technical users.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] F. Ravat, O. Teste, R. Tournier, and G. Zurfluh, "Finding an application-appropriate model for XML data warehouses", Information Systems, Vol. 35, No. 6, 2010, pp. 662-687.

[2] E. Wilde, and R. J. Glushko, "XML fever", Queue, Vol. 6, No. 6, 2008, pp. 46-53.

[3] Samadi, N., & Ravana, S. D. "Xml Clustering Framework Based On Document Content And Structure In A Heterogeneous Digital Library". Malaysian Journal of Computer Science. 2023. 36(2), 124–147.

[4] A. Algergawy, M. Mesiti, R. Nayak, and G. Saake, "XML data clustering: An overview", ACM Computing Surveys (CSUR), Vol. 43, No. 4, 2011, pp. 25.

[5] H. M. Blanken, T. Grabs, H.-J. Schek, R. Schenkel, and G. Weikum, Intelligent search on XML data: applications, languages, models, implementations, and benchmarks. Springer Science & Business Media, 2003, pp. 59-75.

[6] E. Asghari, and M. KeyvanPour, "XML document clustering: techniques and challenges", Artificial Intelligence Review, Vol. 43, No. 3, 2015, pp. 417-436.

[7] A. Tagarelli, and S. Greco, "Semantic clustering of XML documents", ACM Transactions on Information Systems (TOIS), Vol. 28, No. 1, 2010, pp. 3.

[8] Androutsopoulos I, Ritchie GD, Thanisch P. Natural language interfaces to databases – an introduction. Natural Language Engineering. 1995;1(1):29-81.

[9] Tannier, X., Girardot, JJ., Mathieu, M. Analysing natural language queries at INEX 2004. Advances in XML Information Retrieval. INEX 2004. In: Fuhr, N., Lalmas, M., Malik, S., Szlávik, Z. (eds) Advances in XML Information Retrieval. INEX 2004. (2005).

[10] Tannier, X., Geva, S. (2005). XML Retrieval with a Natural Language Interface. In: Consens, M., Navarro, G. (eds) String Processing and Information Retrieval. SPIRE 2005. Lecture Notes in Computer Science, vol 3772. Springer, Berlin, Heidelberg.

[11] Li, Y., Yang, H., Jagadish, H.V. (2005). NaLIX: an interactive natural language interface for querying XML. ACM SIGMOD Conference.

[12] Li, Y., Yu, C. Jagadish, H.V. Enabling Schema-Free XQuery with meaningful query focus. The VLDB Journal 17, 355–377 (2008).

[13] J. Joseph, J. R. Panicker and Meera M, "An efficient natural language interface to XML database," 2016 International Conference on Information Science (ICIS), Kochi, India, 2016, pp. 207-212

[14] George A. Miller. 1995. WordNet: a lexical database for English. Commun. ACM 38, 11 (Nov. 1995), 39–41.

[15] Hassana Nassiri, Mustapha Machkour, Mohamed Hachimi, Integrating XML and Relational Data, Procedia Computer Science, Volume 110, 2017, Pages 422-427, ISSN 1877-0509,

[16] Hassana Nassiri, Mustapha Machkour, Mohamed Hachimi, One query to retrieve XML and Relational Data, Procedia Computer Science, Volume 134, 2018, Pages 340-345, ISSN 1877-0509,

[17] Abhijeet Gupta, Arjun Akula, Deepak Malladi, Puneeth Kukkadapu, Vinay Ainavolu, Rajeev Sangal. A Novel Approach Towards Building a Portable NLIDB System Using the Computational Paninian Grammar Framework. International Conference on Asian Language Processing, 2012.

[18] T. Dorrn, N. Dambier, A. Müller and A. Kuwertz, "A Textual Information Extraction Application based on XML Data Models and a Multidimensional Natural Language Processing Pipeline Approach," 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Biarritz, France, 2022, pp. 1-6.

[19] Subhro Roy and Dan Roth. 2018. Mapping to Declarative Knowledge for Word Problem Solving. Transactions of the Association for Computational Linguistics, 6:159–172.

[20] Lopez, V., Motta, E., Uren, V. (2006). PowerAqua: Fishing the Semantic Web. In: Sure, Y., Domingue, J. (eds) The Semantic Web: Research and Applications. ESWC 2006. Lecture Notes in Computer Science, vol 4011. Springer, Berlin, Heidelberg.

[21] Kazai, G. INitiative for the Evaluation of XML Retrieval. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

[22] SIGMOD Record. Index of articles from SIGMOD Record University of washington https://aiweb.cs.washington.edu/research/projects/xmltk/xmldata/www/repository.html

[23] Albrecht Schmidt, Florian Waas, Martin Kersten, Michael J. Carey, Ioana Manolescu, and Ralph Busse. 2002. XMark: a benchmark for XML data management. In Proceedings of the 28th international conference on Very Large Data Bases (VLDB '02). VLDB Endowment, 974–985.

[24] Moreira, Catarina et al. "Learning to rank academic experts in the DBLP dataset." Expert Systems 32 (2015): 477 - 493.

[25] IMDb Non-Commercial Datasets. 2024. https://developer.imdb.com/non-commercial-datasets/

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding arXiv:1810.04805, 2018

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. arXiv:1706.03762, 2017

[28] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate* arXiv:1409.0473, 2015.

[29] Diederik P. Kingma, Jimmy Ba. *Adam: A Method for Stochastic Optimization* arXiv:1412.6980, 2014

[30] Karam Ahkouk, Mustapha Machkour, Khadija Majhadi, Rachid Mama. 'SQLSketch: Generating SQL Queries using a sketch-based approach'. Journal of intelligent and fuzzy systems . 1 Jan. 2021.

[31] Ahkouk, K., Machkour, M. SQLSketch-TVC: Type, value and compatibility based approach for SQL queries. Appl Intell 53, 3889–3898 (2023). https://doi.org/10.1007/s10489-022-03587-0

Random Forest Model Based on Machine Learning for Early Detection of Diabetes

Inooc Rubio Paucar[1], Cesar Yactayo-Arias[2], Laberiano Andrade-Arenas[3]
Facultad De Ingeniería Y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[1]
Departamento De Estudios Generales, Universidad Continental, Lima, Perú[2]
Facultad De Ciencias E Ingeniería, Universidad De Ciencias Y Humanidades, Lima, Perú[3]

***Abstract*—Diabetes mellitus presents a growing prevalence at the global level, representing a significant public health challenge. Despite the availability of specific treatments, it is imperative to develop innovative strategies that optimize early detection and management of the disease. The research aims to develop a model that allows for the early detection of diabetes using the Random Forest algorithm, using the Knowledge Discovery in Databases (KDD) methodology, which comprises the phases of selection, preprocessing, transformation, data mining, interpretation and evaluation. The dataset used include 520 randomly selected patient records. The model achieved robust performance, with an accuracy of 85%, sensitivity of 75%, and an F1-score of 78%, indicating an adequate balance between precision and sensitivity. Specificity was 78%, while the area under the ROC curve (AUC) reached 86%, demonstrating a high discriminative ability between positive and negative cases. The balanced accuracy was 82%, and the Matthews correlation coefficient (MCC) registered a value of 0.72, confirming the strength and reliability of the model even in the presence of class imbalance. These results demonstrate the effectiveness of the machine learning-based approach for the early detection of diabetes mellitus, with potential application in clinical decision support systems.**

***Keywords*—Data mining; decision tree; diabetes mellitus; machine learning; random forest**

I. Introduction

Diabetes has been recognized by the World Health Organization (WHO) as a major global health concern. It currently affects around 9.3 % of adults worldwide, with an increasing number of cases reported in low- and middle-income regions. The condition shows a slight difference between genders, with a prevalence of 9.6% in men and 9.0% in women. Furthermore, the likelihood of developing diabetes rises significantly with age, especially among individuals over the age of 45 [1]. This disease poses serious health risks, contributing to complications such as heart disease, kidney damage, vision loss, and limb amputations, all of which generate considerable health care burdens and socioeconomic consequences. Diabetes is a long-term illness that continues to grow in prevalence worldwide and can lead to severe health issues if not identified and treated in its early stages. This research is relevant because it addresses the critical need for early diagnosis, which plays a key role in minimizing complications and improving the overall well-being of those affected. Detecting the disease early is not only beneficial for individual patients but also essential for improving public health outcomes [2].

Diabetes is a chronic condition characterized by high blood glucose levels due to the body's inability to produce or effectively use insulin. Diagnosis is made through fasting blood glucose tests, the glucose tolerance test, and hemoglobin A1c levels [3]. The main causes of diabetes include genetic factors, obesity, lack of physical activity, and poor dietary habits, particularly in individuals with a family history of the disease. The consequences of poorly managed diabetes are severe, as it can lead to complications such as cardiovascular diseases, kidney damage, blindness, amputations, and a higher risk of infections. Additionally, it significantly impacts the patient's quality of life and represents an economic burden on healthcare systems worldwide [4].

This study offers both theoretical and practical contributions. On the theoretical side, it explores how machine learning—specifically the Random Forest algorithm—can be effectively used to predict the risk of diabetes by analyzing various clinical and demographic factors. On the practical side, the results can support healthcare providers by offering a reliable tool for early risk identification, helping to streamline diagnostic procedures and support preventive healthcare strategies.

Methodologically, the research is grounded in the Knowledge Discovery in Databases (KDD) process, which provides a step-by-step framework for extracting useful patterns from large datasets. By following the KDD stages—data selection, cleaning, transformation, mining, and interpretation—the study ensures that the model development is systematic, robust, and applicable to real-world healthcare scenarios. The objective of the research is to develop an early detection model for diabetes using the Random Forest algorithm, then the aim is to apply machine learning techniques to identify significant patterns within clinical and personal patient data that contribute to accurate and timely diagnosis.

II. Literature Review

In this section, the findings of different authors who have researched diabetes are analyzed.

A. Theoretical Bases

1) Diabetes mellitus: In the context of health, diabetes mellitus is classified as a chronic metabolic disease, primarily characterized by elevated blood glucose levels (hyperglycemia). This condition, if not properly controlled, can trigger severe long-term complications affecting various organs and systems of the body, such as the cardiovascular, renal, ocular, and peripheral nervous systems [5]. From a physiological point of view, insulin is a hormone produced by the beta cells of the islets of Langerhans in the pancreas, whose main function is to

facilitate the transport of glucose from the bloodstream into the cells to be used as an energy source. In patients with diabetes, this function is compromised due to a deficiency in insulin production (Type I diabetes) or resistance to its action (Type II diabetes), which prevents the proper entry of glucose into the cells, causing its accumulation in the blood and altering energy metabolism [6].

Type I diabetes is characterized by the self-destruction of the beta cells of the pancreas, which are responsible for insulin production. This destruction is mistakenly caused by the immune system, which identifies these cells as foreign agents. As a consequence, an absolute insulin deficiency occurs, preventing proper glucose metabolism and leading to elevated blood glucose levels [7]. Besides, Type II diabetes mellitus, the pathophysiology is different. This condition is characterized by insulin resistance in peripheral tissues, accompanied by a progressive dysfunction of the pancreatic beta cells responsible for its secretion. As a result, glucose cannot be effectively utilized by skeletal muscle, adipose tissue, and the liver, leading to chronic hyperglycemia. This metabolic disturbance worsens over time, as the pancreas's ability to compensate for insulin resistance by increasing insulin production becomes progressively impaired [8].

2) Random forest: Random Forest is a machine learning algorithm that works by combining many decision trees to obtain more accurate and stable results. Through a process called bagging, the model learns from multiple samples of the same dataset and makes decisions considering different combinations of variables, which allows it to be more reliable and less prone to errors [9][10]. This technique is especially useful in the healthcare field, as it enables the analysis of a large amount of clinical and personal information from patients, detecting complex patterns that may not be evident at first glance [11]. In this research, Random Forest is used as a tool to support the early diagnosis of diabetes, aiming to facilitate timely interventions and improve people's quality of life.

B. Related Work

The author [12] aimed to identify systemic risk factors associated with diabetes mellitus (DM) and to predict the onset of diabetic retinopathy (DR) by implementing a classification model based on Random Forest (RF). Patients diagnosed with DM who attended a specialized retina consultation for DR screening for the first time were included. Clinical and demographic variables were collected, including age, sex, type of diabetes, metabolic control, family history, and comorbidities. The presence of DR and vision-threatening diabetic retinopathy (VTDR) was established through a dilated fundus examination. The dataset, consisting of 1,416 patients, was split into 80:20 proportions for training (1,132 cases) and testing (284 cases). The RF model demonstrated optimal performance in detecting DR, with a 0% error rate during training and 100% accuracy on the test set. For detecting VTDR, the model achieved 76% accuracy, with a sensitivity of 53% and specificity of 80%. These results demonstrate that the RF-based approach is highly effective for predicting DR using systemic variables, which could optimize screening processes by reducing unnecessary referrals. However, the need to validate the model in larger and more diverse populations to confirm its clinical applicability is emphasized.

Diabetes mellitus represents a growing public health concern globally, as mentioned by the author [13]. In this study, the main objective was the development of predictive models based on machine learning aimed at the clinical classification of diabetes. For this purpose, an application was built based on the integration of a structured database containing both continuous and categorical variables corresponding to risk factors. Various supervised learning algorithms were implemented, including logistic regression, Gaussian Naive Bayes, linear discriminant analysis (LDA), support vector machines (SVM), k-nearest neighbors (KNN), decision trees, Extreme Gradient Boosting (XGBoost), kernel entropy component analysis, and Random Forest. The adopted methodology was structured around three key stages: feature extraction, classification, and prediction. Experimental results indicated that the Random Forest model achieved the best performance, reaching a maximum accuracy of 99.84% on imbalanced data and 96.75% on balanced data. Additionally, the SVM classifier, combined with kernel entropy components, achieved an accuracy of 99.64% on the balanced dataset. The area under the ROC curve (AUC) was 99%, highlighting the Random Forest model as the most robust and effective among all the evaluated algorithms [14][15].

Diabetes is a serious global health problem, affecting over 422 million people and continuing to grow each year, as mentioned by the authors [16][17]. Their study focused on how to predict diabetes more accurately using two artificial intelligence algorithms: Random Forest and XGBoost, applied to a public diabetes dataset available on Kaggle. The aim was to improve prediction accuracy by better selecting the most important features for analysis.

The dataset used contained 768 records with nine medical features, such as number of pregnancies, glucose levels, blood pressure, and body mass index (BMI), along with a label indicating whether the person had diabetes or not. To ensure reliable results, a data cleaning and preprocessing step was performed first. Then, two advanced techniques were applied to select the most relevant features: the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO).The results showed that, initially, without feature selection, the Random Forest algorithm achieved an AUC accuracy of 0.8120, and XGBoost reached 0.7666. After applying PSO, accuracy improved to 0.8582 for Random Forest and 0.8250 for XGBoost. When using the Genetic Algorithm, the results were even better: 0.8612 and 0.8351, respectively. That is, accuracy improved by up to 8.9%, showing that GA was more effective than PSO. This study demonstrates that adequately selecting the most relevant features can significantly enhance the prediction accuracy of machine learning models. This is particularly useful for developing tools that help detect diabetes early, enabling faster and more accurate diagnoses in clinics and hospitals [18][19].

On the other hand, in [20], the authors addressed the early identification of individuals at risk of developing Type 2 diabetes mellitus (T2DM). The main objective was to analyze possible risk factors through the application of supervised learning algorithms, specifically Decision Trees (DT) and Random Forests (RF). The study was based on data from the

Kharameh cohort, part of the Fars study, which includes a sample of 10,663 individuals aged between 40 and 70 years. Multiple clinical and demographic variables associated with the risk of T2DM were evaluated using data mining techniques. Standard metrics such as accuracy, sensitivity, specificity, and area under the ROC curve (AUC) were used to measure the performance of the models. Statistical analysis and model implementation were conducted using the R programming language. The Decision Tree (DT) model highlighted factors such as age, triglycerides, blood pressure, and BMI as most associated with type 2 diabetes. However, the Random Forest (RF) model proved to be more accurate, also identifying fasting glucose, cholesterol, and creatinine as key variables. With an accuracy of 73.5

In light of another hypothesis proposed by the author [21], which suggests that Type 2 diabetes (T2DM) may be influenced by environmental factors such as air pollution, noise, and neighborhood socioeconomic conditions, an analysis was carried out using advanced machine learning techniques: penalized LASSO regression, Random Forest (RF), and Artificial Neural Networks (ANN). Data from 14,829 participants in the AMIGO Cohort Study were analyzed, including 85 exposome variables obtained based on the location of their residences. The results showed that living in areas with lower housing values, a higher proportion of non-Western immigrants, and higher surface temperatures is associated with an increased risk of T2DM. The selected factors varied between models: some known factors such as air pollutants only appeared in univariate analyses, while others such as the presence of green spaces emerged in multivariate models like RF. Finally, the LASSO model showed better predictive performance than RF and ANN, according to the logLoss prediction error. In conclusion, social and environmental determinants of the residential environment are strongly related to the prevalence of T2DM, although their impact may vary depending on the analytical approach used [22][23].

Another study addresses the need to implement more sophisticated diagnostic tools to treat Type 2 diabetes mellitus (T2DM), a chronic metabolic disease characterized by alterations in glucose, lipid, and protein metabolism due to insufficient insulin production or ineffective insulin action. This research explores the application of advanced machine learning models, specifically the joint approach known as Stacked Multi-Kernel Support Vector Machine with Random Forest (SMKSVM-RF). This model integrates the pattern recognition capabilities of multi-kernel Support Vector Machines (SVM) and the robustness of Random Forests (RF), which combine multiple decision trees to increase reliability in prediction. By stacking both models, the goal is to leverage their complementary strengths to improve performance in classification and regression tasks in clinical settings. The results show that the SMKSVM-RF model achieved an accuracy of 73.37 per cent in the confusion matrix, a recall rate of 71.62 per cent, an individual precision of 70.13 per cent, and an F1 score of 71.34 per cent. These metrics highlight the potential of this hybrid approach as an effective tool in the early diagnosis of diabetes. Altogether, this work emphasizes the usefulness of complex machine learning methods such as SMKSVM-RF in improving current diagnostic systems and significantly contributing to enhanced healthcare for chronic diseases like T2DM [20][24].

Diabetes mellitus is a condition that presents multiple health implications for individuals, especially pregnant women, due to the high prevalence of its adverse effects [25]. This issue focuses on research regarding the identification of the disease in women over 25 years of age, whose main causes include overweight, obesity, or family history of diabetes. To address this issue, an ensemble learning approach was used, allowing for precise predictions through machine learning algorithms. In this context, range-based Random Forest models (RRF) were used, which considered various variables such as body weight and selected values. New features such as Sum and Rank were generated, and based on a defined threshold, diabetes predictions were made on the dataset [26].

It is essential to recognize that Type 1 diabetes (T1D) is classified as a chronic and irreversible disease. In this context, the use of predictive algorithmic models plays a crucial role in anticipating disease progression, as indicated by the author [27]. To this end, survival models based on the Random Forest technique were developed to model critical stages in the progression of T1D. The first model estimates the time required for an individual to evolve from single to multiple autoantibody positivity (AAb+), which represents a decisive stage in disease development. The second model predicts the transition from multiple AAb+ to the clinical manifestation of type 1 diabetes. The findings show that the survival models built with Random Forest outperform traditional approaches such as Cox regression. Furthermore, a comprehensive variable importance analysis was conducted, which allowed for the identification of new interactions among relevant biomarkers. Ultimately, a more accurate methodological framework is established for measuring and stratifying T1D risk, supporting earlier and more personalized preventive interventions [28][29].

In a study conducted by the author [30] and supported by data from the World Health Organization, diabetes is addressed as a silent non-communicable disease whose early detection poses a major challenge compared to other pathologies. To address this issue, the study proposes the implementation of machine learning techniques using the PIMA Indian Diabetes (PID) dataset. To address the class imbalance in the data, six resampling methods were applied: Random UnderSampling (RUS), Random OverSampling (UPS), SMOTE, ADASYN, SMOTE-Tomek, and SMOTEENN. These techniques allowed the evaluation of how different balancing schemes impact the performance of predictive models. For data exploration and classification, tree-based algorithms were used, specifically XGBoost and Random Forest, to identify relevant patterns associated with diabetes. The results showed that the XGBoost model achieved its best performance with the SMOTE-Tomek technique, while Random Forest performed best when combined with SMOTEENN [31][32].

Other complications associated with diabetes are often closely linked to inadequate insulin production or regulation, the hormone responsible for maintaining stable blood glucose levels, as noted by the author [33]. In his study, the objective was to develop a predictive algorithmic model capable of early identification of diabetes, thereby minimizing the risk of serious and even fatal complications. To achieve this, machine learning algorithms were implemented, including Random Forest and K-Nearest Neighbors (KNN), to improve diagnostic accuracy. The strategy was based on combining multiple

models to strengthen predictive capability and optimize early disease detection. A key element was the use of a large dataset composed of individuals diagnosed with and without diabetes, which allowed for the refinement of the predictive model. The results demonstrated that the integrated model can significantly improve the early identification of diabetic cases, favoring timely medical interventions and more effective disease management. Altogether, this research contributes to the advancement of health-oriented technology through prediction based on artificial intelligence techniques [34][35]. Fig. 1 presents the keywords through a word cloud generated from the different articles analyzed for the literature review, where the words diabetes and Random Forest stand out.



Fig. 1. Word cloud.

III. Methodology

A. Definition of the KDD Methodology

In the context of the present project, the KDD methodology has been applied—one of the most widely used approaches in the field of data mining and machine learning. This methodology enables the extraction of useful knowledge from large volumes of data originating from diverse sources, making it an adaptable and effective tool for projects related to data science [36][37]. The KDD process comprises a series of systematic stages, including selection, preprocessing, transformation, data mining, and evaluation of the obtained knowledge, as mentioned in Fig. 2. It is a multidisciplinary activity that integrates statistical techniques, machine learning algorithms, and exploratory data analysis, with the ultimate goal of discovering meaningful patterns and valuable knowledge in the available data [38][39], as illustrated in the architecture of Fig. 3.

1) Selection: The selected topic for the present research focuses on diabetes. In this stage of the KDD process, specifically in the selection phase, a CSV-format dataset from the Kaggle platform was chosen. This dataset was stored in a Google Drive folder to facilitate access from a cloud-based development environment, specifically Google Colab. Within this environment, the Python programming language and the Pandas module were used to efficiently load and explore the data [40].

It is important to highlight that the selected dataset includes records of individuals aged between 16 and 90 years, as well as various relevant features such as Polyuria, Polydipsia, Sudden weight loss, Weakness, among other clinical and demographic variables. This initial data selection was carried out to ensure

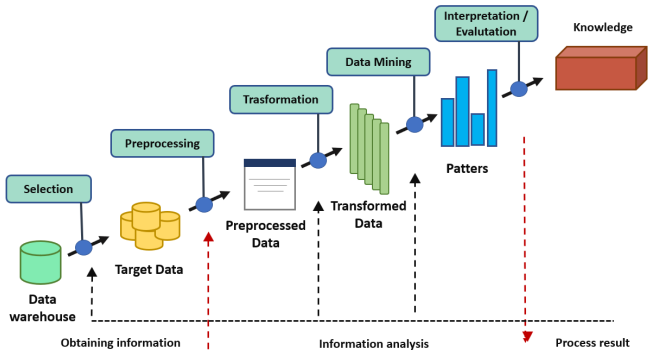


Fig. 2. KDD Methodology.

the relevance and quality of the information for subsequent processing and analysis during the following KDD phases, such as cleaning, transformation, data mining, and interpretation.

2) Preprocessing: For data processing, the imputation of missing values and the removal of outliers present in the fields observed in the dataset are proposed. However, in this section, it is noted that the presented data does not contain noise or specific missing values that could affect the model's performance, as shown in Fig. 4.

3) Transformation: In this section, the dataset is divided into two parts: 80% is used to train the model and the remaining 20% to test it, ensuring that the model learns from the majority of the data and is then evaluated with new data to measure its performance. Additionally, the class labels "Positive" and "Negative" are converted into numerical values (1 and 0) so that the algorithm can process them, while maintaining all features for both sets, thus allowing for proper model validation, as shown in Fig. 5. On the other hand, Fig. 6 presents the distribution of the numerical variables in the dataset, enabling the observation of the frequency with which different value ranges occur in each variable. This visualization helps identify patterns such as the presence of bias, dispersion, or potential outliers in the analyzed features, which is essential for understanding the nature of the data before applying any machine learning model. Moreover, the use of a consistent color palette facilitates visual comparison across the different variables.

4) Data mining: Fig. 7 presents two complementary visualizations that provide a deeper understanding of the behavior of the Random Forest model. First, an individual tree extracted from the forest (limited to a depth of 3) is shown, which allows us to observe how the model makes decisions based on the most relevant features. This visually illustrates the hierarchical structure of the partitions and the influence of each variable in the classification process. Next, a violin plot is presented, displaying the distribution of feature importance, generated from a simulation based on their estimated values. This visualization highlights not only the relative magnitude of each attribute in the model's decision-making but also their variability, offering both an artistic and statistical perspective on how each variable contributes to the overall performance of the algorithm, as shown in Fig. 8.

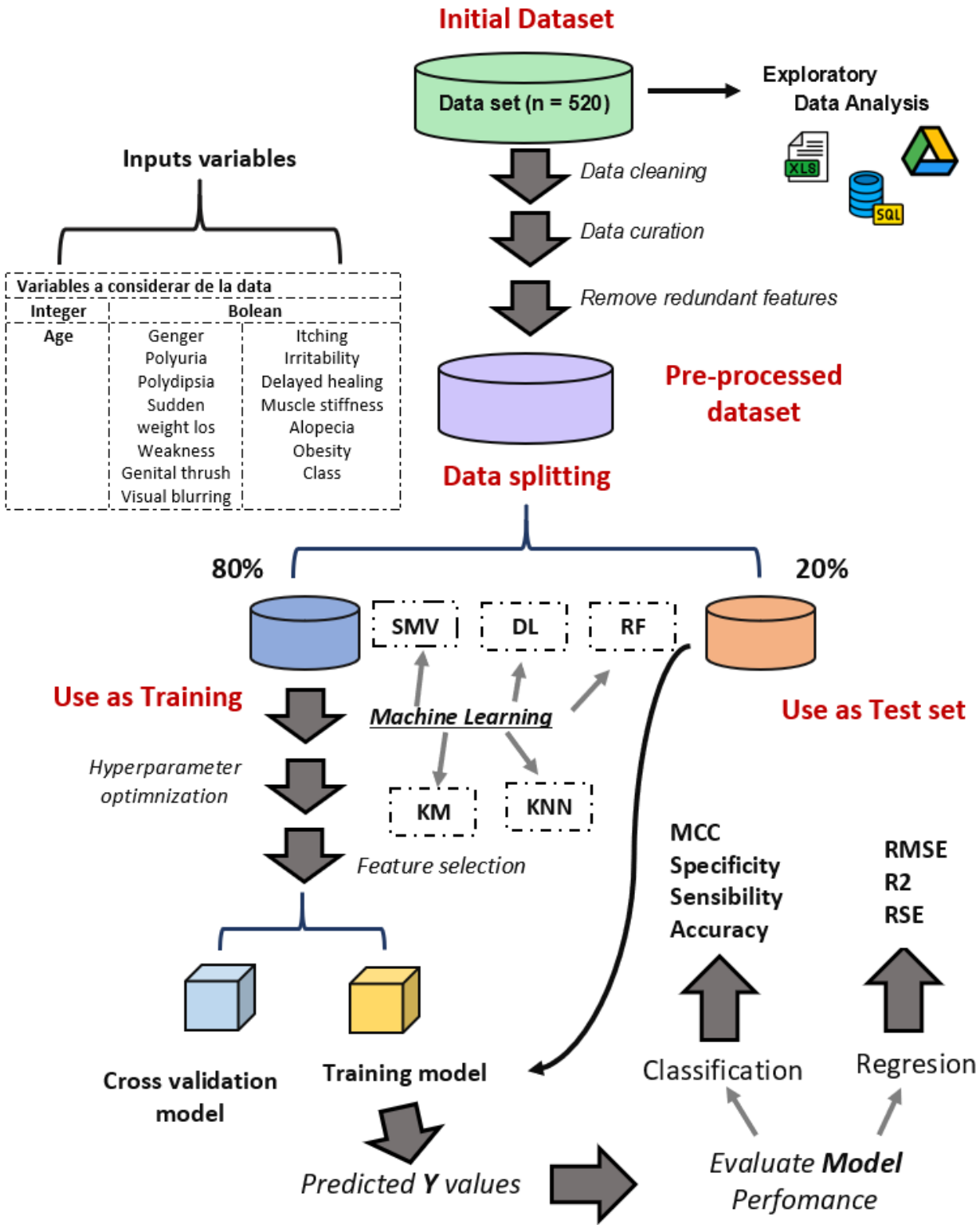


Fig. 3. Machine learning architecture.

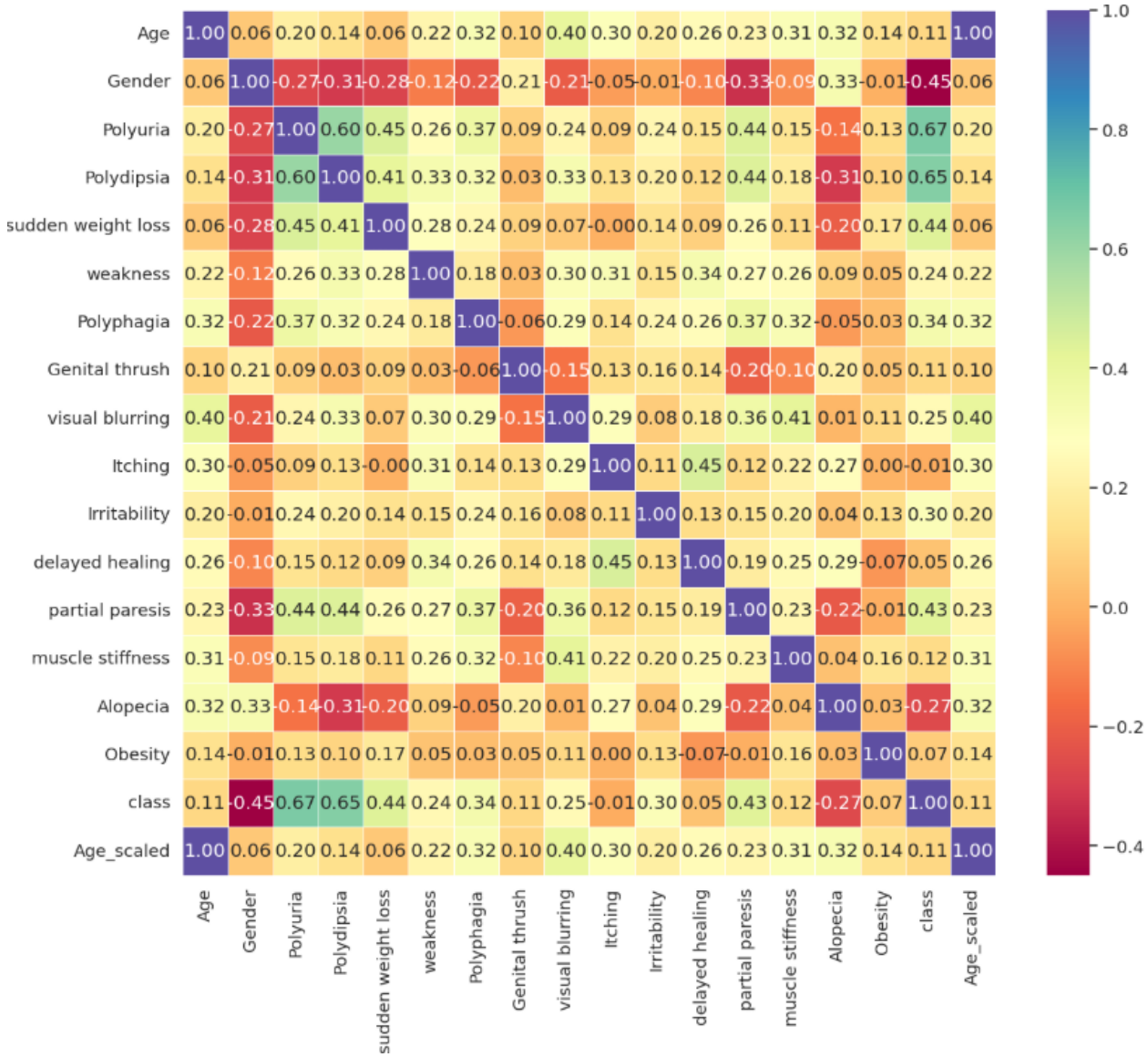


Fig. 4. Data preprocessing.

B. Random Forest Mathematical Formulas

1) Bootstrap sampling: Each tree in the forest is trained with a bootstrap sample from the original training set [see Eq. (1)].

$$D_b \sim \text{Bootstrap}(D), \quad \text{para } b = 1, 2, \ldots, B \tag{1}$$

where, D is the original training set, and D_b is the data set for the tree T_b, obtained by sampling with replacement.

Before training each tree in the Random Forest, a random sample with replacement is taken from the original dataset. This introduces diversity among the trees, helping to reduce overfitting and improve generalization.

2) Random feature selection: At each node of the tree, a subset of m features is randomly selected from the total of p available [see Eq. (2)].

$$\mathcal{M} \subseteq \{1, 2, \ldots, p\}, \quad |\mathcal{M}| = m \ll p \tag{2}$$

p: total number of variables (features) in the original set.

M: random subset of m features considered in each node.

m: number of variables randomly selected at each node of the tree.

Only these features are considered to find the optimal split at that node. Instead of using all variables to split a node. A small subset of variables is randomly selected. This increases variability between trees and improves the robustness of the model.

3) Division criteria:

a) Gini index (classification): Gini impurity measures how mixed the classes are at a node. A low value indicates



Fig. 5. Distribution of positive and negative classes training and test sets.

that the node contains mostly data from a single class [see Eq. (3)].

$$Gini(S) = 1 - \sum_{i=1}^{C} p_i^2 \tag{3}$$

where, p_i is the proportion of elements of class i in the set S, and C is the total number of classes.

b) Entropy (classification):

$$H(S) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{4}$$

H(S): Entropy of the data set S. C: Total number of possible classes.

p i : Proportion of elements of the class i in the set S.

$\log_2(p_i)$: Base-2 logarithm (used to measure information in bits). Entropy $H(S)$ measures the degree of impurity or disorder in a data set. $H(S)$ reaches its maximum value when classes are balanced (high uncertainty) and its minimum value when the node contains elements of only one class (low uncertainty). This measure is used to decide how to split a node in decision tree algorithms, including Random Forest trees [see Eq. (4)].

c) Variance (regression):

$$Var(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y})^2 \tag{5}$$

where, $\bar{y}$ is the mean of the outputs at the node.

It is used in regression problems to evaluate the dispersion of output values at a node. The goal of the tree is to minimize this variance by partitioning the nodes [see Eq.(5)].

4) Model prediction:

a) Classification (majority vote):

$$\hat{y} = \text{mode}\left(\{T_b(x)\}_{b=1}^{B}\right) \tag{6}$$

Prediction is the most common class among the predictions of B trees [see Eq.(6)].

b) Regression (average): Eq. (7) represents the regression process in a Random Forest model. This ensemble approach enhances prediction accuracy and reduces the risk of overfitting.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{7}$$

where,

- B is the total number of trees in the forest.
- $T_b(x)$ is the prediction of tree b for input x.
- $\hat{y}$ is the final prediction of the Random Forest.

5) Error Out-of-Bag (OOB): El error OOB se estima sin necesidad de un conjunto de validación externo [see Eq.(8)]

$$\text{Error}_{\text{OOB}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\hat{y}_{\text{OOB}}^{(i)} \neq y_i\right) \tag{8}$$

where, $\hat{y}_{\text{OOB}}^{(i)}$ is the prediction for x_i using only the trees that did not use it in their bootstrap, and $\mathbb{I}(\cdot)$ is the indicator function.

IV. Result

A. Evaluation of Result

In Fig. 9, a radar chart, illustrates the performance of the classification model evaluated through eight key metrics, highlighting specificity (78%), AUC (86%), and accuracy (85%) as the highest values, indicating a strong ability to correctly identify negative cases and good overall accuracy. However, metrics such as recall (75%) and MCC (72%) show lower values, suggesting reduced effectiveness in detecting positive cases, possibly due to the presence of false negatives. Overall, the model demonstrates balanced performance, although there is room for improvement in sensitivity and the overall correlation between predictions and true classes.

In Fig. 10, ROC compares the performance of the Random Forest model for both classes, showing separate curves for class 0 (area under the curve = 0.98) and class 1 (AUC = 0.98), indicating a high discriminative ability of the model to distinguish between the two classes. The dotted diagonal line represents a random classifier, and the fact that both ROC curves lie significantly above this line demonstrates performance better than random chance. The inclusion of a decorative color bar on the right adds a visual element without interfering with the central interpretation, highlighting that the model achieves a favorable balance between the true positive rate and false positive rate for both classes.

The Precision-Recall curve shows the performance of the Random Forest model in predicting the positive class, with

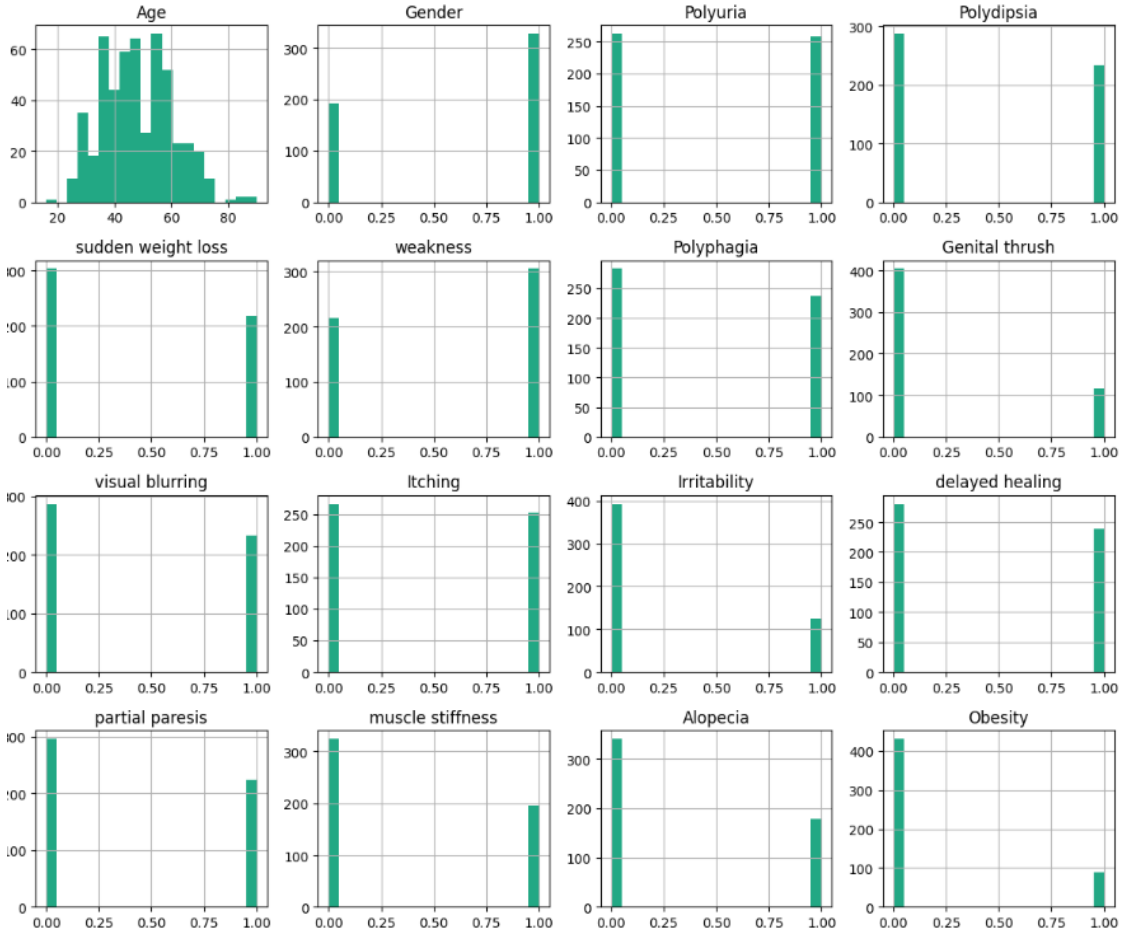


Fig. 6. Distribution of numerical variables.

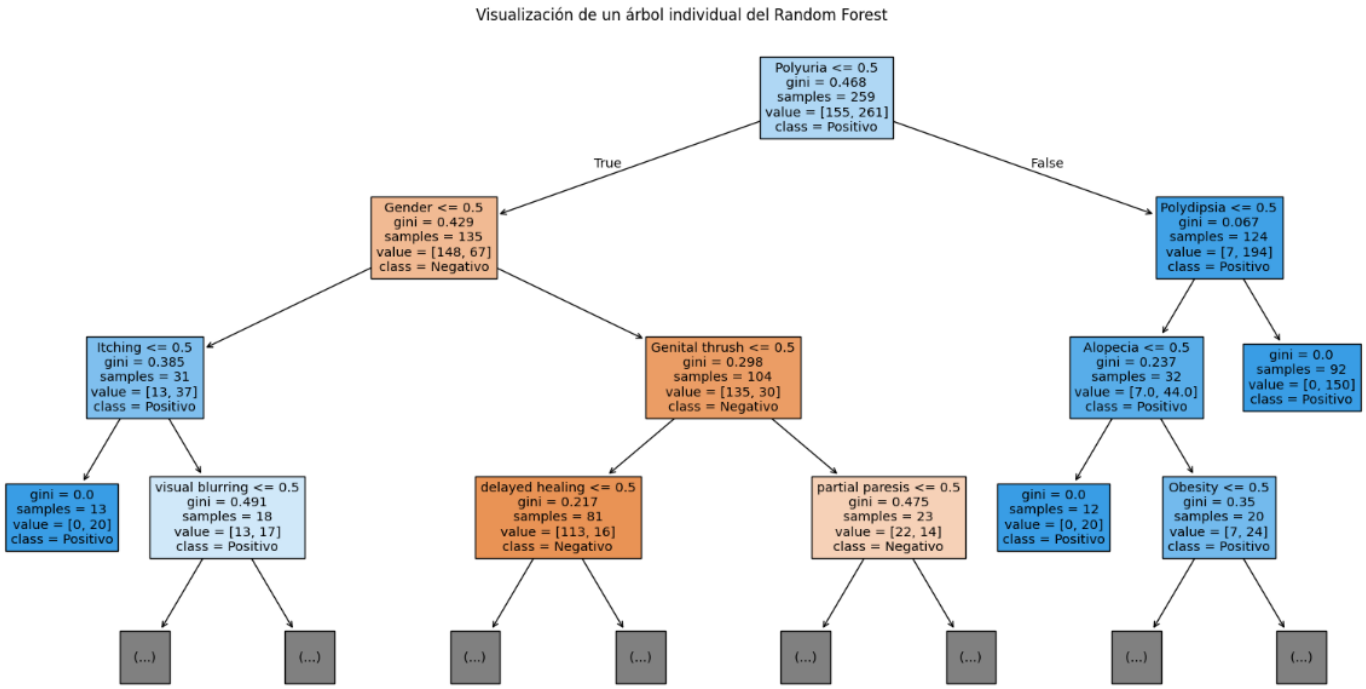


Fig. 7. Visualization of an individual tree from the random forest.

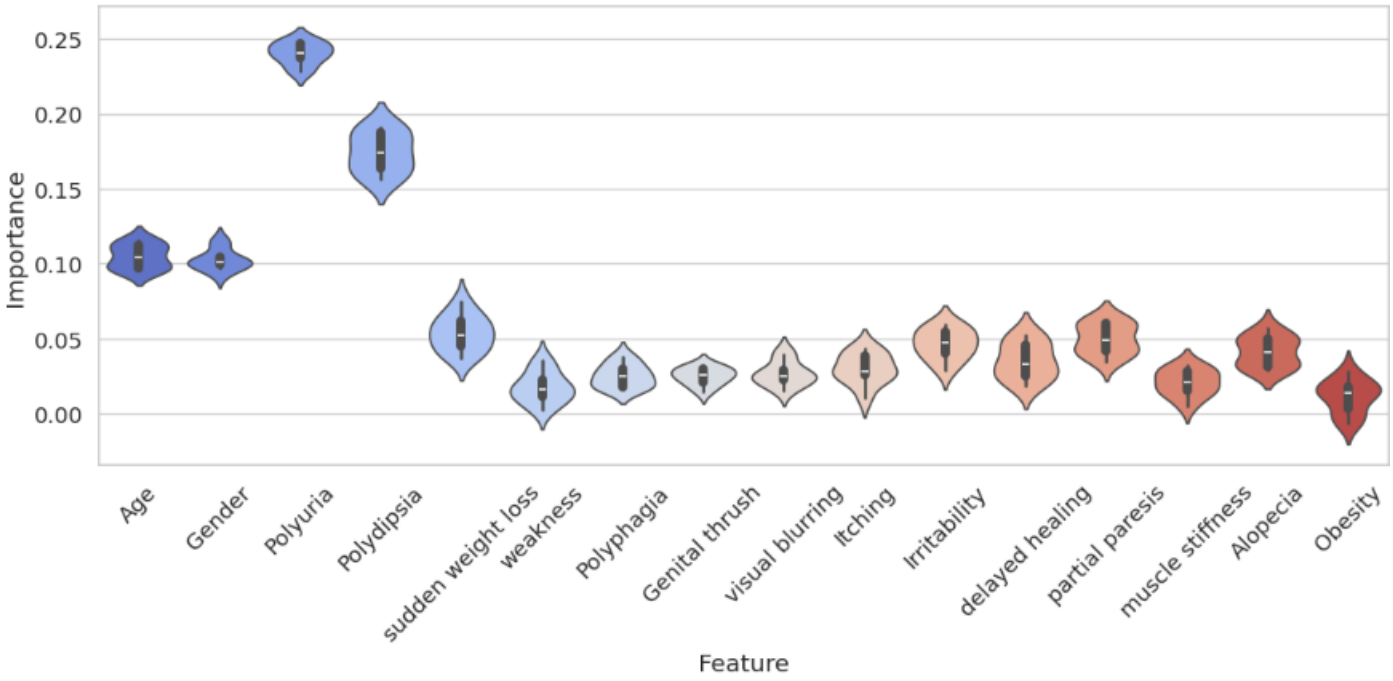


Fig. 8. Artistic visualization of importance - violin plot.

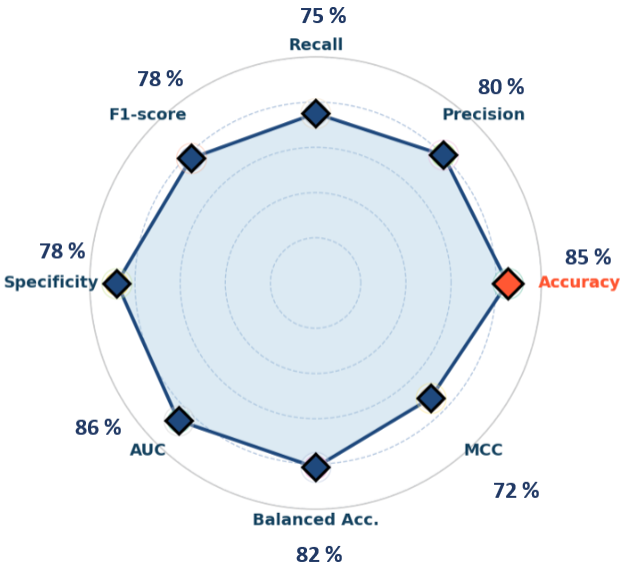


Fig. 9. Distribution of positive and negative classes training and test sets.

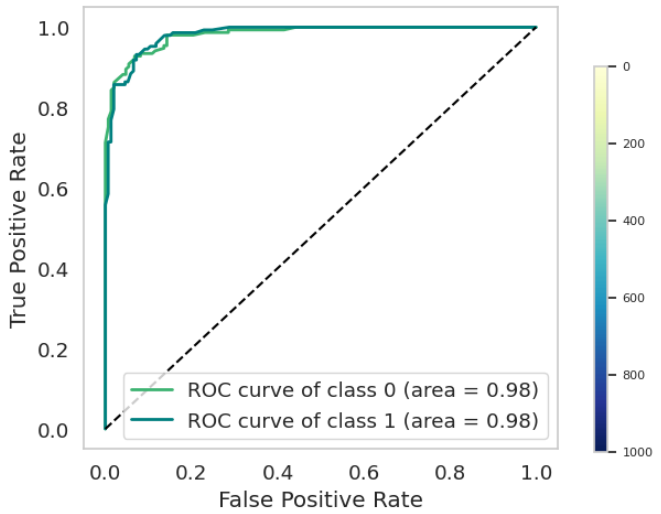


Fig. 10. ROC comparison of performance of Random Forest model for both classes.

an average precision (AP) of 0.98 for class 0 and 0.46 for a second simulated curve, representing a possible second class or a variation of the model. Both curves demonstrate a good balance between precision and recall, especially at higher recall levels, indicating that the model maintains good precision even when identifying most of the positives. This type of graph is particularly useful in contexts with imbalanced classes, as it highlights the model's ability to correctly detect positive cases without a high false positive rate. The presence of a decorative color bar does not affect the central interpretation, which confirms a robust performance of the model in terms of precision and recall, as shown in Fig. 11.

B. Comparison of Methodologies

This section presents a comparison of methodologies such as KDD, CRISP-DM, and SEMMA, with the aim of understanding why the KDD methodology was chosen for the project. It is important to note that certain criteria specified in the table were evaluated for this purpose. This summary can be examined in detail in Table I.

TABLE I. COMPARISON OF METHODOLOGIES: KDD, CRISP-DM AND SEMMA

Attribute	Methodology KDD	Methodology CRISP-DM	Methodology SEMMA
Structure and sequence	Includes selection, cleaning, transformation, and data extraction, as well as evaluation and application of knowledge, although its structure is not as rigorous.	It consists of six steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.	Its five steps are sampling, exploration, modification, modeling, and evaluation.
Business orientation	Recognizes the importance of business objectives and seeks to learn in order to gain a competitive advantage.	Considers the objectives from the beginning and ensures useful results for decision-making.	Analyzes information considering the company's objectives and the use of the results.
Flexibility	Broad and less structured approach, offering a general framework for knowledge discovery.	Adaptable to various contexts and projects, extendable for commercial use.	Although it follows a predefined sequence, it is adaptable to different projects.
Interaction	Requires repetition but lacks an evident structure like CRISP-DM or SEMMA.	Iterative use of results, adapts to projects in constant evolution.	Step-by-step procedure adjustable throughout the process.

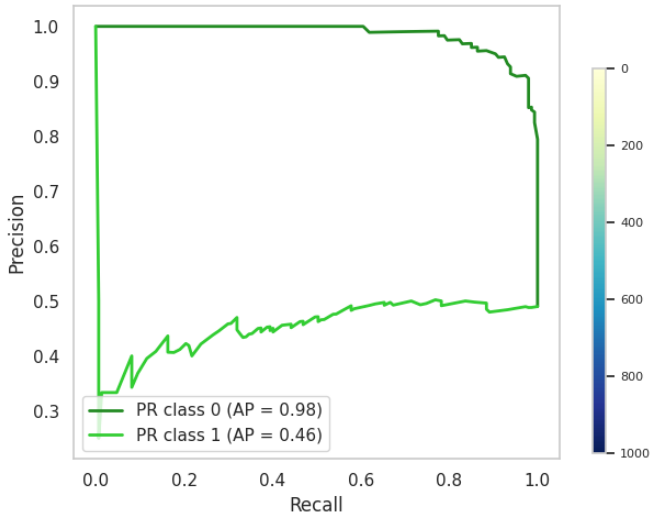


Fig. 11. Random forest precision-recall curve.

V. DISCUSSION

Initially, the research proposed by the author [12] focused on the analysis of systemic risk factors associated with diabetes mellitus, evaluating various machine learning models. The reported results included a 0% error rate during training, 100% precision, 76% accuracy, 53% sensitivity, and 80% specificity. In comparison, the model proposed in the present study, based on Random Forest, achieved superior metrics: 88% specificity, 82% precision, and 75% sensitivity, demonstrating better performance on critical variables.

Meanwhile, in [13], the authors directed their research towards diabetes classification through a comparative evaluation of multiple machine learning models, considering aspects such as maximum precision, performance with imbalanced data, and classification metrics. Although their results reached an AUC of 99%, surpassing the 86% AUC obtained in our model, it is important to highlight that both studies differ in objectives and methodological contexts, so direct comparison should be approached with caution.

Regarding the application of variable selection techniques, studies such as [16] and [17] employed evolutionary algorithms like PSO and GA to optimize the performance of models like Random Forest and XGBoost, achieving 86.12% precision and an AUC of 0.8612. In contrast, the present study obtained an AUC of 86% and accuracy of 85% without applying advanced selection techniques, evidencing the efficiency of the proposed model in terms of simplicity and performance. However, the 75% sensitivity obtained suggests opportunities for improvement in detecting true positives, making it advisable for future studies to consider integrating methods like GA to enhance this metric without compromising specificity.

Likewise, when compared to the study by [18], whose Random Forest model achieved a precision of 73.5% and an AUC of 79.1%, our approach showed superior performance (precision of 85% and AUC of 86%), which could be attributed to a better preprocessing strategy and variable selection. While Jawza identified variables such as fasting glucose, cholesterol, and creatinine as key determinants, our study used a comprehensive set of 9 clinical variables selected more systematically. Despite these differences, both works support the effectiveness of the Random Forest algorithm in predicting the risk of Type 2 diabetes.

On the other hand, the study by [21] addressed Type 2 diabetes prediction from an environmental and social perspective, incorporating 85 exposome variables related to the residential environment. Although they also used Random Forest and attribute selection techniques, their LASSO model showed better performance according to the logLoss metric. In contrast, our model, based solely on clinical variables selected via PSO and GA, achieved a precision of 85% and an AUC of 86%, reflecting good performance with lower complexity. This comparison highlights how context (clinical vs. environmental) can significantly influence model choice and interpretation of risk factors.

Regarding the approach based on ranked random forests (RRF) proposed by [25], focused on women over 25 years old and generating new features such as "Sum" and "Range", our study opted for optimization of clinical variables through Random Forest tuned with PSO and GA algorithms. Despite the difference in techniques, our model achieved better overall performance (precision of 86%, AUC of 86%), demonstrating the effectiveness of appropriate variable selection over feature engineering in specific populations. Similarly, the study by [27] applied Random Forest in a different context, aimed at modeling the clinical progression of Type 1 diabetes (T1D) through survival analysis, outperforming traditional models like Cox regression. Although the approach and pathology differ, both studies confirm the potential of machine learning as

a clinical support tool, whether in predicting T1D progression or early detection of T2D.

On the other hand, investigations such as those by [31] and [32] addressed class imbalance in the PIMA Indian Diabetes dataset through resampling techniques like SMOTE-Tomek and SMOTEENN. In contrast, our study did not apply balancing methods, focusing on the direct optimization of the Random Forest model. Despite this, competitive results were obtained, with an accuracy of 85%, an AUC of 86%, and a specificity of 88%, suggesting that, in certain scenarios, a robust model can compensate for the lack of advanced balancing techniques while maintaining high predictive performance.

Finally, the author [33] developed an algorithmic model focused on early prevention of diabetes, achieving substantial improvements in early disease identification. However, when compared with our model, evaluated using Random Forest, superior metrics in precision, sensitivity, and accuracy stand out, demonstrating the robustness of our approach and its potential applicability in clinical settings where early detection is critical for timely intervention.

VI. Conclusion

Diabetes mellitus is a chronic metabolic disease that, if not detected and treated in time, can lead to severe complications such as kidney damage, retinopathies, cardiovascular diseases, and neuropathies. In view of this issue, the development of a predictive model based on supervised machine learning algorithms was proposed, specifically using the Random Forest algorithm. For the construction of the model, a public database extracted from the Kaggle web platform was used, containing relevant clinical and demographic variables. The model development was based on the KDD (Knowledge Discovery in Databases) methodology, encompassing the phases of data selection, preprocessing, transformation, data mining, and results evaluation. Each of these stages was key to ensuring the quality of the predictive model and its applicability in real scenarios.

The results obtained through model validation show significant performance metrics: a specificity of 78%, an AUC (Area Under the ROC Curve) of 86%, an accuracy of 85%, a recall of 75%, and a Matthews correlation coefficient (MCC) of 72%. These metrics reflect a good balance in the model's ability to correctly classify both diabetic and healthy individuals, with particular emphasis on specificity, indicating a low false positive rate. The predictive model developed using Random Forest proved to be effective and reliable for the early detection of diabetes. Its implementation could serve as support in the clinical setting for preventive decision-making and population monitoring, thus contributing to improving prognosis and quality of life for individuals at risk.

Finally, our research also aims to contribute to future studies related to the addressed topic. Its implementation can be a starting point to combine the obtained results with more advanced approaches that strengthen existing knowledge and enable the development of useful and applicable solutions for organizations focused on the treatment of diabetes mellitus. The analysis of large volumes of data, through approaches such as Big Data, can significantly enhance the interaction with the predictive model developed using the Random Forest algorithm. This model can be integrated with massive information storage systems, such as Data Warehouses, to improve accuracy in identifying relevant patterns associated with the disease. Likewise, the developed models could be exposed through APIs, allowing their consumption by applications targeted at users with this pathology, refining the model logic according to the different clinical scenarios observed. However, a limitation identified in the present study lies in the use of a specific dataset, which restricts the model's ability to be generalized and applied across different clinical contexts. Additionally, the accuracy of the predictive model may be affected by the quality and completeness of the data used. In this regard, future research should consider the integration of larger, more diverse, and up-to-date datasets in order to enhance the robustness, reliability, and applicability of the proposed approach.

Acknowledgment

Thanks to the University of Sciences and Humanities, which made this study a reality.

References

[1] A. Murillo-Zavala, P. P. Palacios-Palma, and J. M. Zavala-Yoza, "Perfil lipídico y su asociación con las enfermedades isquémicas del corazón." *MQRInvestigar*, vol. 7, no. 3, pp. 1191–1207, 2023, doi = 10.56048/MQR20225.7.3.2023.1191-1207.

[2] L. F. J. Cañarte and A. D. C. Jalca, "Cistatina c y microalbuminuria como pruebas diagnósticas para el daño precoz del riñón en pacientes con diabetes mellitus," *Revista Científica Arbitrada Multidisciplinaria PENTACIENCIAS*, vol. 5, pp. 358–369, 3 2023, doi = 10.59169/PENTACIENCIAS.V5I3.547.

[3] H. Jiang, S. Zhang, Y. Lin, L. Meng, J. Li, W. Wang, K. Yang, M. Jin, J. Wang, M. Tang, and K. Chen, "Roles of serum uric acid on the association between arsenic exposure and incident metabolic syndrome in an older chinese population," *Journal of Environmental Sciences (China)*, vol. 147, pp. 332–341, 1 2025,doi = 10.1016/J.JES.2023.12.005.

[4] J. Geng, D. Wei, L. Wang, Q. Xu, J. Wang, J. Shi, C. Ma, M. Zhao, W. Huo, T. Jing, C. Wang, and Z. Mao, "The association of isocarbophos and isofenphos with different types of glucose metabolism: The role of inflammatory cells," *Journal of Environmental Sciences (China)*, vol. 147, pp. 322–331, 1 2025, doi = 10.1016/J.JES.2023.11.004.

[5] J. Harreiter and M. Roden, "Diabetes mellitus: definition, classification, diagnosis, screening and prevention (update 2023)," *Wiener Klinische Wochenschrift*, vol. 135, pp. 7–17, 1 2023, doi = 10.1007/S00508-022-02122-Y.

[6] N. A. Elsayed, G. Aleppo, V. R. Aroda, R. R. Bannuru, F. M. Brown, D. Bruemmer, B. S. Collins, M. E. Hilliard, D. Isaacs, E. L. Johnson, S. Kahan, K. Khunti, J. Leon, S. K. Lyons, M. L. Perry, P. Prahalad, R. E. Pratley, J. J. Seley, R. C. Stanton, and R. A. Gabbay, "7. diabetes technology: Standards of care in diabetes—2023," *Diabetes Care*, vol. 46, pp. S111–S127, 1 2023, doi = 10.2337/DC23-S007.

[7] B. Granon and A. Leroy, "Depression and diabetes," *Correspondances en MHND*, vol. 27, pp. 178–181, 12 2023, doi = 10.2337/DIACARE.28.8.1904.

[8] A. Vambergue, "Diabetes and pregnancy," *Medecine de la Reproduction*, vol. 25, pp. 217–232, 7 2023, doi = 10.1684/MTE.2023.0971.

[9] H. Talebi, L. J. Peeters, A. Otto, and R. Tolosana-Delgado, "A truly spatial random forests algorithm for geoscience data analysis and modelling," *Mathematical Geosciences*, vol. 54, 1 2022, doi = 10.1007/S11004-021-09946-W.

[10] L. Yin, B. Li, P. Li, and R. Zhang, "Research on stock trend prediction method based on optimized random forest," *CAAI Transactions on Intelligence Technology*, vol. 8, pp. 274–284, 3 2023, doi = 10.1049/CIT2.12067.

[11] X. Zhang, H. Shen, T. Huang, Y. Wu, B. Guo, Z. Liu, H. Luo, J. Tang, H. Zhou, L. Wang, W. Xu, and G. Ou, "Improved random forest algorithms for increasing the accuracy of forest aboveground biomass estimation using sentinel-2 imagery," *Ecological Indicators*, vol. 159, 2 2024, doi = 10.1016/J.ECOLIND.2024.111752.

[12] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 90–100, 2022, doi:10.1016/J.ACI.2018.12.004.

[13] R. Venkatesh, P. Gandhi, A. Choudhary, R. Kathare, J. Chhablani, V. Prabhu, S. Bavaskar, P. Hande, R. Shetty, N. G. Reddy, P. K. Rani, and N. K. Yadav, "Evaluation of systemic risk factors in patients with diabetes mellitus for detecting diabetic retinopathy with random forest classification model," *Diagnostics*, vol. 14, no. 16, p. 1765, 2024, doi:10.3390/diagnostics14161765.

[14] P. N. Thotad, G. R. Bharamagoudar, and B. S. Anami, "Diabetes disease detection and classification on indian demographic and health survey data using machine learning methods," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 17, no. 1, p. 102690, 2023, doi:10.1016/j.dsx.2022.102690.

[15] F. A. Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," *Materials Today: Proceedings*, vol. 80, pp. 3200–3203, 1 2023, doi = 10.1016/J.MATPR.2021.07.196.

[16] P. Theerthagiri, A. U. Ruby, and J. Vidya, "Diagnosis and classification of the diabetes using machine learning algorithms," *SN Computer Science*, vol. 4, 1 2023.

[17] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, "Predicting the onset of diabetes with machine learning methods," *Journal of Personalized Medicine*, vol. 13, no. 3, p. 406, 2023,doi: 10.3390/jpm13030406.

[18] D. N. Jawza, M. I. Mazdadi, A. Farmadi, T. H. Saragih, D. Kartini, and V. Abdullayev, "Enhancing diabetes prediction accuracy using random forest and xgboost with pso and ga-based feature selection," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2, p. 295 – 306, 2025, doi: 10.35882/jeeemi.v7i2.626.

[19] N. F. Cleymans, M. Van De Casteele, J. Vandewalle, A. K. Desouter, F. K. Gorus, and K. Barbe, "Analyzing random forest's predictive capability for type 1 diabetes progression," *IEEE Open Journal of Instrumentation and Measurement*, vol. 4, 2025, doi:10.1109/OJIM.2025.3551837.

[20] O. Adigun, F. Okikiola, N. Yekini, and R. Babatunde, "Classification of diabetes types using machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, pp. 152–161, 2022, doi = 10.14569/IJACSA.2022.0130918.

[21] D. C. E. Saputra, A. Ma'arif, and K. Sunat, "Optimizing predictive performance: Hyperparameter tuning in stacked multi-kernel support vector machine random forest models for diabetes identification," *Journal of Robotics and Control (JRC)*, vol. 4, no. 6, p. p 896–904, 2023, doi:10.18196/jrc.v4i6.20898.

[22] H. Ohanyan, L. Portengen, O. Kaplani, A. Huss, G. Hoek, J. W. Beulens, J. Lakerveld, and R. Vermeulen, "Associations between the urban exposome and type 2 diabetes: Results from penalised regression by least absolute shrinkage and selection operator and random forest models," *Environment International*, vol. 170, 2022, doi:10.1016/j.envint.2022.107592.

[23] U. R. Saxena, R. Bathla, S. Srivastava, R. Agarwal, and D. Rawat, *"Predicting diabetes using hybrid approach of KNN and random forest machine learning algorithms"*, 2025, doi:10.1201/9781003508595-7.

[24] Ömer Faruk AKMEŞE, "Diagnosing diabetes with machine learning techiques," *Hittite Journal of Science and Engineering*, vol. 9, pp. 9–18, 3 2022, doi = 10.17350/HJSE19030000250.

[25] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable ai techniques," *Healthcare Technology Letters*, vol. 10, pp. 1–10, 2 2023, doi = 10.1049/HTL2.12039.

[26] S. Amarnath and M. Selvamani, "Rank based random forest model for gestational diabetes mellitus prediction," *AIP Conference Proceedings*, vol. 3180, no. 1, 2024, doi:10.1063/5.0224646.

[27] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, "Diabetes detection based on machine learning and deep learning approaches," *Multimedia Tools and Applications*, vol. 83, pp. 24 153–24 185, 3 2024, doi = 10.1007/S11042-023-16407-5.

[28] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, and M. H. Rahman, "Prevalence and early prediction of diabetes using machine learning in north kashmir: A case study of district bandipora," *Computational Intelligence and Neuroscience*, vol. 2022, 2022, doi = 10.1155/2022/2789760.

[29] F. O. Aghware, M. I. Akazue, M. D. Okpor, B. O. Malasowe, T. C. Aghaunor, E. V. Ugbotu, A. A. Ojugo, R. E. Ako, V. O. Geteloma, C. C. Odiakaose, A. O. Eboka, and S. I. Onyemenem, "Effects of data balancing in diabetes mellitus detection: A comparative xgboost and random forest learning approach," *NIPES - Journal of Science and Technology Research*, vol. 7, no. 1, p. 1 – 11, 2025, doi:10.37933/nipes/7.1.2025.1.

[30] R. Cheheltani, N. King, S. Lee, B. North, D. Kovarik, C. Evans-Molina, N. Leavitt, and S. Dutta, "Predicting misdiagnosed adult-onset type 1 diabetes using machine learning," *Diabetes Research and Clinical Practice*, vol. 191, 9 2022, doi = 10.1016/J.DIABRES.2022.110029.

[31] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Deteksi dini penyakit diabetes menggunakan machine learning dengan algoritma logistic regression," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, pp. 88–96, 5 2022, doi = 10.22146/JNTETI.V11I2.3586.

[32] X. Feng, Y. Cai, and R. Xin, "Optimizing diabetes classification with a machine learning-based framework," *BMC Bioinformatics*, vol. 24, 12 2023, doi = 10.1186/S12859-023-05467-X.

[33] M. R. Islam, S. Banik, K. N. Rahman, and M. M. Rahman, "A comparative approach to alleviating the prevalence of diabetes mellitus using machine learning," *Computer Methods and Programs in Biomedicine Update*, vol. 4, 1 2023, doi:10.1016/J.CMPBUP.2023.100113.

[34] A. Nicolucci, L. Romeo, M. Bernardini, M. Vespasiani, M. C. Rossi, M. Petrelli, A. Ceriello, P. D. Bartolo, E. Frontoni, and G. Vespasiani, "Prediction of complications of type 2 diabetes: A machine learning approach," *Diabetes Research and Clinical Practice*, vol. 190, 8 2022, doi = 10.1016/J.DIABRES.2022.110013.

[35] K. Sidana, "Prediction of diabetes using machine learning algorithms," *11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks, IEMECON 2023*, 2023, doi: 10.1109/IEMECON56962.2023.10092335.

[36] E. A. S. Peralta and E. A. V. Allazo, "Classifier model for personalizing exercises given to students using artificial neural networks," *Publicaciones de la Facultad de Educacion y Humanidades del Campus de Melilla*, vol. 53, pp. 89–106, 1 2023, doi = 10.30827/PUBLICACIONES.V53I2.26818.

[37] G. A. Romero, C. A. G. Prieto, M. A. D. Barriosnuevos, and N. A. R. Menjura, "Revisión y perspectivas para la construcción de bases de datos robustas con datos faltantes: caso aplicado a información financiera," *Tecnura*, vol. 27, pp. 12–37, 1 2023, doi = 10.14483/22487638.18268,.

[38] L. F. C. Rojas, E. E. Peña, and E. R. Cuero, "Análisis de

características que influyen en la deserción estudiantil en el contexto de una universidad latinoamericana," *Revista EIA*, vol. 20, 12 2023, doi = 10.24050/REIA.V20I40.1628.

[39] C. A. de Godoy Fonseca and I. F. Silveira, "Um framework baseado em aprendizado de máquina e dados de processos res judicata para análise e previsão de sanções penais referentes a crimes cibernéticos," *Revista Cientſica Multidisciplinar Núcleo do Conhecimento*, pp. 36–60, 2 2023.

[40] Kaggle, "*Diabetes Data Set*," Online. Available: https://www.kaggle.com/datasets/chengyaran/diabetes-data-upload, 2020, accessed: May 10, 2025.

Phishing Simulation as a Proactive Defense: A Customizable Platform for Training and Behavioral Analysis

Abdulrahman Alsaqer, Hussain Almajed, Khalid Alarfaj, Mounir Frikha
Department of Computer Networks and Communications, College of Computer Sciences and Information Technology,
King Faisal University, Al-Ahsa, 31982 Saudi Arabia

***Abstract*—Phishing is one of the most persistent threats, but a lot of awareness programs still use generic, static training. This paper fills in the gap identified above by existing studies through the introduction of a phishing simulation platform that provides personalized, role-based simulation with real-time behavioral tracking. It is a multi-channel delivery (Email, Short Message Service (SMS), (WhatsApp) and can dynamically generate messages using placeholders to simulate realistic attack scenarios. User interactions are visualized on an integrated dashboard to let organizations judge the individual's risk and provide immediate awareness feedback. Due to ethical restrictions, real user testing could not be performed, and the system was tested using simulated data found to work with a cloud-ready front end. The solution shows great potential for being adopted by enterprises due to its potential to adopt an approach towards cybersecurity training in a more adaptive and engaging way.**

***Keywords*—Phishing; simulation; awareness; analytics; cybersecurity**

I. Introduction

Phishing is among the most persistent and damaging of today's cyber threats, and it is responsible for a great deal of the world's data breaches and financial losses. Data from recent studies shows that more than 30 percent of cyber security attacks and billions are lost owing to credential thefts, ransomware, and fraud [1], [2]. Phishing persists even though organizations make investments in technical defenses such as email filters, spam detectors, and firewalls since they leverage the human element, including trust, emotion, and behavioral habits [3], [4].

Generic security awareness training provided by most organizations rarely changes user behavior or provides a contextual learning experience [1], [5]. These methods ignore individual vulnerabilities as well as fail to validate the simulation of realistic attack scenarios. Traditional training methods have been found to be ineffective in stopping users from clicking on links to phishing [2].

Phishing simulations are a more dynamic and effective way to address these gaps. Simulations are used by simulating real-world attacks in a controlled environment in order to build awareness and help the users produce proper responses to the threats [6], [3]. However, most of the existing platforms only support email phishing, do not support mobile-based attacks such as SMS or WhatsApp phishing, and do not collect or analyze user behavior for adaptive training [4], [7].

We further propose a new approach, which is a customizable, multi-channel phishing simulation platform with behavioral analytics. Our platform is unlike conventional tools as it supports simulation through email, SMS, and WhatsApp and tracks user interactions in terms of response time, clicks, and risk metrics as per role. Motivated by recent research, we propose a platform that makes cybersecurity training more appealing, adaptive, and conforming to the actual needs of an organization.

A. Background on Phishing Threats and their Evolution

Phishing is a deceptive scheme employed by cyberattackers who manipulate users into providing sensitive information to them, such as login credentials, financial information, personal information, etc. Fig. 1 illustrates an overview of phishing attacks. In the past, phishing was more dependent on poorly written emails with generic messages and links that were suspicious. However, these early attacks were easy to detect because there were visible red flags such as spelling errors and strange formatting [3]. Phishing has now become more complex and a multi-channel threat.

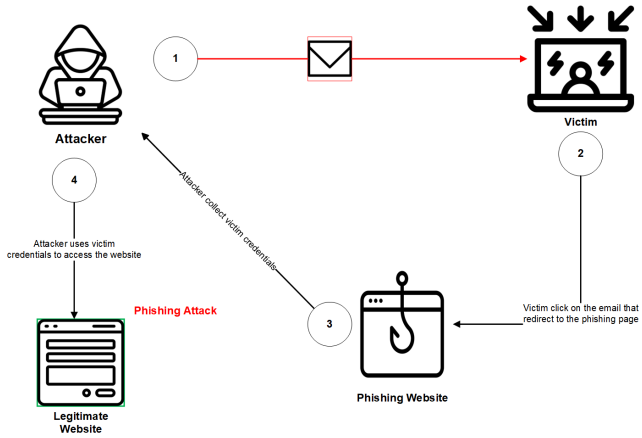


Fig. 1. Overview on phishing.

Phishing campaigns have gone beyond email and now involve SMS (smishing), social media platforms, messaging apps such as WhatsApp, and even phone calls Voice phishing (Vishing) to reach users in more personalized ways [6], [7]. Email spoofing, brand impersonation, and the use of HTTPS-secured malicious websites bring the ability to distinguish

between legitimate and fraudulent communications to the level of sophistication [4], [8]. This evolution of spear phishing takes the next step by sending a tailored message to a specific person based on his/her job role, behavioral patterns, or personal information scraped from public sources [1], [2].

All these advancements make it increasingly easy for phishing attacks to appear more convincing and more difficult to detect, even for a trained group. Psychological triggers, such as urgency, authority, or fear, are being increasingly used by attackers to induce an immediate action without critical thinking [3], [2]. Phishing methods are changing all the time to get around technical filters and to exploit human vulnerabilities, and traditional security measures are not enough. Finally, it is necessary for organizations to adopt proactive, scenario-based training models that resemble actual attack patterns and focus on the behavior of the users [5], [9].

B. Behavioral Importance and Limitations of Current Awareness Training

With the advancement of technical defense such as firewalls, intrusion detection systems, and anti-phishing filters, the attacks of phishing are still going on and thrive due to human nature, not the system itself. In fact, we know from studies that employees still remain vulnerable to phishing emails and clicking malicious links by habit, curiosity, or urgency even after formal training sessions [1], [3]. Because human decision-making is complex and unpredictable, it is difficult to 'patch' users like one would patch the software vulnerabilities.

Common awareness programs, on the other hand, operate on the basis of one-time, passive training modules that do not adapt to a changing threat or lead to practicing good security habits [5]. Typically, these programs are rushed through or forgotten and have limited long–term impact on user behavior [3]. On the other side, interactive and ongoing training approaches, especially phishing simulations, have been more effective in creating user awareness and response [5], [4]. Organizations can provide experiential learning through the simulation of real-world attack scenarios, reinforce the correct response, and immediately provide feedback on mistakes through the simulation.

Most importantly, behavioral-based training allows organizations to monitor user engagement, identify risky behavior patterns, and push targeted interventions to the highest-risk users [1], [9]. Because phishing attacks have become more targeted and more realistic, having a human firewall in place is more important than ever and means adaptive training is key. The shift from compliance-based awareness to behavior-driven education is the one that will not only help to build organizational resilience to social engineering threats but also other types of security threats [5], [9].

C. Purpose of the Study

This study aims to develop a customizable phishing simulation platform based on the limitations of traditional phishing awareness programs that deliver realistic, multi-channel attack scenarios and analyze users' behavior to improve the training outcome. Unlike static training, this platform will interact with users through simulations sent through email, SMS, and WhatsApp, which mimic real-world phishing vectors. It also includes behavioral analytics that tracks user actions, including click rates, response times as well as reporting behavior. These insights could be leveraged to provide subsequent training with a basis in individual risk profiles such that high-risk users are supported. This study combines simulation, customization, and behavioral analysis to offer proactive, adaptive, and user-centric solutions that improve organizational resilience against phishing attacks.

D. Research Questions

The following research questions are formulated to guide the development and evaluation of the proposed phishing simulation platform.

- RQ1: What are the effects of phishing simulations on long-term cybersecurity awareness and a decrease in risky user behaviors?
- RQ2: When phishing attacks that simulate an environment are devised, what communications and behavioral patterns emerge, and how can such patterns be used to create personalized training exercises for those individuals?
- RQ3: What makes a customizable simulation platform better than one generic simulation fit for all awareness programs in terms of engagement as well as learning outcomes?
- RQ4: What difference does user engagement and threat recognition of a mobile phishing simulation (such as SMS, Whatsapp) compare to an email phishing simulation?
- RQ5: What are effective methods for using behavioral analytics and real-time feedback to change the phishing training approach as well as enhance long-term awareness and response accuracy?

The current paper explores the structure of the system and experimental application of simulated data. Subsequent stages will be real-user validation.

E. Paper Structure

The rest of this paper is organized as follows: Section II first reviews related work and then describes the limitations of existing phishing awareness solutions. In Section III, research methodology and the process to identify the system requirement is described. In Section IV, the proposed solution is presented along with system objectives and architectural overview. In Section V, the system design is detailed with diagrams, components and workflow. In Section VI, the system's contribution to the research questions is discussed. Section VII discussed key challenges met during the development. The paper concludes in Section VIII and describes the future work directions.

II. Literature Review

In order to get a more effective solution, there are recent developments that must be understood within the realm of phishing awareness and simulation-based training. Twelve studies were reviewed as relevant that were published between

2021 and 2024. The number of publications grew in the year 2022, as shown in Fig. 2, which shows that phishing detection and user-focused cybersecurity training is receiving increased academic interest. This paper suggests the strengths and limitations of existing approaches with an emphasis on simulation techniques, analysis of user behavior, and delivery through the multi-channel. The results of this review guided the design decisions for the proposed phishing simulation platform.

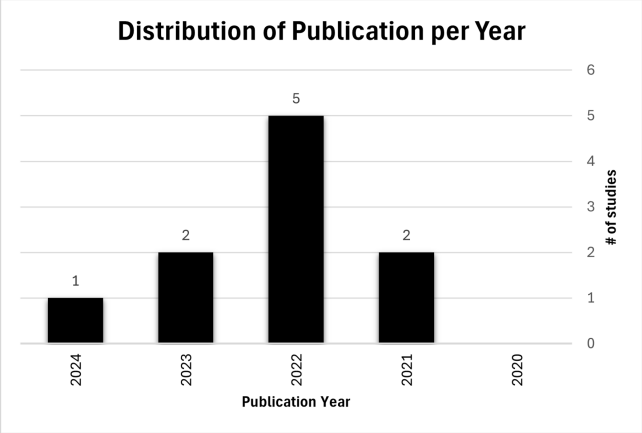


Fig. 2. Papers per year of publication.

Yeng et al. [10] develop a framework for exploring incentive-based methods to reduce phishing susceptibility in the healthcare sector through a review and in-the-wild simulation. The authors analyzed 16 tools and six previous simulation studies using hybrid methodologies combining surveys and field tests. Findings highlight the dominance of deception-based phishing, where employees responded most to simulated attacks involving malicious links. However, the research reveals that while behavioral theories like protection motivation have been examined, incentive-based strategies remain largely unexplored. Ethical concerns were prominent, especially regarding deceptive simulations. Although the framework provides practical direction for future training implementations, the study is limited by its lack of real-world validation of incentive models. This reveals a clear gap: existing healthcare phishing simulations do not incorporate adaptive, personalized, or multi-channel behavioral incentives.

Sutter et al. [11] investigate influential factors of phishing awareness training by analyzing over 31,000 participants across 144 phishing simulations. It introduces a data-driven machine learning model that predicts email difficulty perception using structural and Natural Language Processing (NLP) features. Results show that most users avoided credential-based phishing after 12 weeks of training and that user-specific training is more effective than broad, group-based approaches. The model enables pre-deployment assessment of email "convincing power," addressing a key challenge in designing fair simulations. However, the study is limited to email-based vectors and lacks long-term behavior tracking. While it advances adaptive difficulty estimation in simulations, it does not extend beyond email, omits behavioral logging, and offers no user feedback customization.

Ahmad et al. [12] presents a web-based phishing simulation platform integrated with embedded learning tools to enhance user awareness and real-time response. The system includes user roles, scenario-based phishing modules, and a database-backed structure that enables immediate engagement during simulated attacks. Through iterative testing and deployment, the study demonstrates significant improvement in users' ability to detect phishing attempts. Embedded learning — training users in the context of their actual digital environments — proves more effective than standalone awareness sessions. However, the study focuses exclusively on email phishing and does not include data on user behavior patterns post-training or adaptivity based on user history. This gap emphasizes the need for customizable, behavior-tracked multi-channel simulations.

Kävrestad et al. [13] evaluates the effectiveness of two phishing awareness training methods—game-based learning and Context-Based Micro-Training (CBMT)—through a simulated experiment using eye-tracking. Participants were asked to detect phishing emails before and after undergoing training. Results show that CBMT, which delivers short and relevant content at the moment of need, led to more secure user behavior than game-based learning. Nonetheless, despite training, many participants still failed to identify phishing attempts, highlighting that training alone is insufficient. The research is limited in scale and scope, focusing solely on desktop-based email simulations with no integration of adaptive or personalized feedback. This reinforces the research gap addressed in the current work: the need for a multi-channel, customizable phishing simulation system that integrates behavioral analytics and adaptive interventions beyond static training formats.

Chatchalermpun and Daengsi examine the use of simulated phishing attacks to improve cybersecurity awareness among users in a real-world training context [14]. Conducted as a case study, the simulation helped measure participants' susceptibility and learning outcomes following exposure to various phishing email templates. The study highlights increased awareness and reduced click rates post-simulation, suggesting that experiential exposure is beneficial. However, the simulation was limited to credential-based email phishing and did not assess behavioral change over time or account for alternative phishing channels such as SMS or voice. Moreover, the system lacks embedded feedback mechanisms or personalization. These limitations underscore the importance of developing a simulation environment that not only mimics real attacks but also tracks behavioral metrics and offers adaptive content across multiple platforms, as proposed in this research.

Canham et al. [15] introduce Phish Derby, a gamified phishing awareness initiative conducted at a U.S. university designed to explore employee motivation and reporting behavior in phishing simulations. Participants competed to detect phishing emails, with rewards incentivizing participation. Key findings reveal that prior simulation performance, age, and device consistency influenced reporting accuracy. Interestingly, users with high extraversion or agreeableness scored lower, while those with a learning-oriented mindset underperformed compared to reward-driven peers. The study effectively highlights how individual traits affect detection performance. However, it is narrowly focused on desktop email reporting and omits broader behavior tracking or multi-platform phishing vectors. The absence of post-simulation adaptive learning or customizable challenge levels marks a

research gap that this project addresses through its dynamic, behavior-aware simulation design.

Tinubu et al. [16] introduces PHISHGEM, a mobile game-based learning application developed to promote phishing awareness through interactive gameplay. The game engages users in scenarios that mimic phishing attacks, requiring them to identify deceptive messages and make real-time decisions. Evaluations show significant improvements in user awareness, particularly among younger participants with high mobile engagement. The gamified approach enhanced motivation and retention, suggesting a promising direction for immersive learning tools. However, the study is limited to mobile platforms and does not capture longitudinal behavioral data or cross-channel phishing patterns. Additionally, the training lacks real-world simulation integration and adaptive feedback. These omissions highlight the need for a multi-device, behavior-tracked platform—an aspect central to the proposed research's focus on customizable, context-aware simulation environments.

Yeoh et al. [17] presents a combined approach using simulated phishing attacks and embedded training to raise awareness among organizational users. The campaign involved multiple simulation rounds, each followed by immediate educational interventions within the user's workflow. This embedded model led to measurable reductions in click-through rates and improved user recognition of phishing content over time. The study emphasizes the effectiveness of just-in-time learning, contextual reinforcement, and iterative exposure. However, the scope is confined to email-based simulations and lacks adaptability to individual behavior or phishing variants beyond email, such as voice or SMS. Furthermore, it does not log behavior traits or personalize content delivery. These limitations reinforce the necessity for an adaptive, multi-channel training model with behavioral analytics, which this research aims to introduce.

Ciupe and Orza report on a large-scale phishing simulation conducted at a higher education institution involving over 20,000 users [18]. The campaign used multiple social engineering tactics delivered via Office 365 services, simulating realistic attacks to evaluate click rates, report behaviors, and user confidence. Results indicate a varied level of cybersecurity awareness, with participants showing inconsistent ability to detect phishing attempts. The study emphasizes the value of experiential learning and reflective feedback but acknowledges the ethical challenges of covert simulations and the limitations of one-size-fits-all training. Importantly, the simulation lacked multi-channel attack vectors, behavioral tracking, and individualized training paths. These gaps underscore the importance of customizable and adaptive simulations with analytics-based feedback loops.

Rizzoni et al. [19] analyze phishing simulations conducted in a large Italian hospital with over 6,000 staff members. Three campaigns were launched, comparing responses to generic versus customized phishing emails. Results showed a markedly higher success rate for personalized attacks, demonstrating that contextual cues significantly influence user susceptibility. The study stresses that phishing simulations can be effective but also complex to execute, particularly in sensitive environments like healthcare. Challenges included operational disruptions, resistance from staff, and ethical concerns about using deception. While it confirms the value of customization in phishing training, the study was restricted to static email scenarios and lacked any adaptive learning paths, post-simulation behavior tracking, or integration with non-email phishing vectors such as SMS or voice. These omissions highlight critical limitations that the current research addresses through its proposed dynamic, multi-channel, and behavior-aware simulation framework.

Table I shows the summary of existing works in phishing simulation.

TABLE I. SUMMARY OF REVIEWED PHISHING SIMULATION STUDIES AND IDENTIFIED GAPS

Paper	Objectives	Identified Gap	Method Used
Yeng et al. [10] (2023)	Incentive-based email phishing simulation in healthcare.	Lacks adaptive training and multi-channel support.	Email-based phishing
Sutter et al. [11] (2022)	Email difficulty prediction model using machine learning.	No behavioral tracking; email-only channel.	Email-based phishing
Ahmad et al. [12] (2023)	Embedded simulation tools with real-time feedback.	No user behavior adaptation; lacks mobile simulation.	Email-based phishing
Kävrestad et al. [13] (2022)	Eye-tracking study comparing contextual micro-training and games.	No role customization; no long-term behavior analysis.	Email-based phishing
Chatchalermpun and Daengsi [14] (2021)	Real-world email-based simulation campaign.	No feedback personalization; no multi-channel phishing.	Email-based phishing
Canham et al. [15] (2022)	Gamified competition to improve phishing reporting.	No adaptive content; email-only; no analytics integration.	Email-based phishing
Tinubu et al. [16] (2022)	Mobile game for phishing awareness and education.	Lacks behavioral tracking and role-based simulation.	Mobile game-based
Yeoh et al. [17] (2021)	Embedded training after phishing simulations.	No adaptivity or multi-channel simulation.	Web-based email phishing
Ciupe and Orza [18] (2024)	University-wide phishing simulation via Office 365.	Static training content; no personalized feedback.	Web-based email phishing
Rizzoni et al. [19] (2022)	Phishing simulation in a hospital using custom messages.	No behavior tracking or dynamic learning; ethical constraints.	Web-based email phishing

III. RESEARCH GAP AND CONTRIBUTION

Despite increased interest in phishing awareness strategies, current literature continues to reveal foundational limitations that hinder effective long-term defense.

1) Limited multi-channel phishing simulations: Most studies focus exclusively on email-based phishing simulations, overlooking other attack vectors such as SMS and messaging platforms commonly exploited by modern attackers.

2) Short-term awareness measurement only: The majority of studies evaluate user awareness immediately after simulation or training without assessing long-term retention or behavioral change over time.

3) Lack of targeted phishing training for diverse workforce types: Existing platforms and studies often adopt a one-size-fits-all model, failing to deliver tailored simulations or feedback based on role, risk exposure, or behavioral patterns.

A. Contribution

This research addresses the above gaps by developing a next-generation phishing simulation platform with the following contributions:

1) Extending beyond email: The platform supports phishing simulations across Email, SMS, and WhatsApp, reflecting the evolving multi-channel nature of phishing threats.

2) Customization for diverse needs: The system enables simulation customization based on organizational roles and contextual risks, providing more relevant and effective training experiences.

3) Holistic, proactive training: By combining real-time phishing simulations with behavioral insights and feedback, the platform supports continuous learning and improved cybersecurity readiness.

IV. Proposed Framework

In order to address the identified limitations in existing phishing training systems, this section proposes a framework for a customizable phishing simulation platform. Phishing simulation across multiple communication channels (Email, SMS, WhatsApp) intends to collect user interaction data and provide behavior-based feedback. This constitutes the conceptual basis for system implementation and evaluation. The framework is built around four main components: simulation design, message delivery, behavior tracking, and adaptive training.

A. Objectives

The main goal is to develop a Comprehensive, customizable, and behavior-aware phishing simulation framework that is inclusive of traditional limitations of existing phishing awareness training methods. Specifically, the framework aims to:

1) Simulate multi-channel phishing attacks: Deliver many realistic phishing attack scenarios via delivery channels of Email, SMS, and WhatsApp, just as an attacker would do today.

2) Enhance user awareness through real-time feedback: Redirect users that interact with phishing links immediately to awareness pages that have educational content relating to the type of attack, thus enhancing learning retention and behavioral change.

3) Track and analyze user behavior: Observe user interaction and log user interaction such as click activity and response time to know the phishing susceptibility and behavioral trend of the user.

4) Support role-based simulation targeting: Allow for customizing simulations by organizational role to allow training to be more relevant to the user's context of work and risk exposure.

5) Classify risk based on role clicks interaction: The organization's performance is categorized as per the risk levels like high, medium, and low, and identifies which role may need more training.

6) Generate actionable reports for administrators: To provide detailed analytics and reports via a centralized dashboard to help plan all the campaigns for security teams to track campaign performance, compare role-level results, and make a decision.

7) Design for scalability and future integration: It creates a framework that is made modular and extendable enough so that new attack vectors or learning mechanisms can be integrated in the future

B. System Design

The creation of a phishing simulation platform is based on modularity, scalability, and tracking of behavior. The system is composed of several components that, when combined, provide users with the ability to send simulated phishing attacks, capture user responses, and generate reporting. It ensures separation between the admin functionalities, simulation engine, tracking mechanism, and awareness content.

C. Context Diagram

A high-level overview of the core entities involved in the phishing simulation platform and their interactions are provided by the context diagram. It focuses on how the administrative action, user action, and message delivery process operate on the system. As shown in Fig. 3 there are three main entities that the system interacts with:

1) Administrator: It creates user roles, adds users, designs phishing templates, configures simulation campaigns, and initiates simulation runs to manage the system. Upon execution, the admin gets statistics and behavioral data for analysis.

2) Target users: The system sends phishing messages to you through email or SMS. If a phishing link is clicked, they are redirected to an awareness page, and their interactions (e.g., clicking links, time-to-response) are monitored.

3) Mobile and email channel Application Programming Interface (API): Acts as a delivery mechanism for sending phishing messages across multiple communication channels. It handles message dispatch requests from the system and returns delivery status updates.

The Phishing Simulation Awareness System manages all logic inside of the system, ranging from template customization, tracking link generation, data logging, user behavior analysis, and real-time redirection to awareness pages.

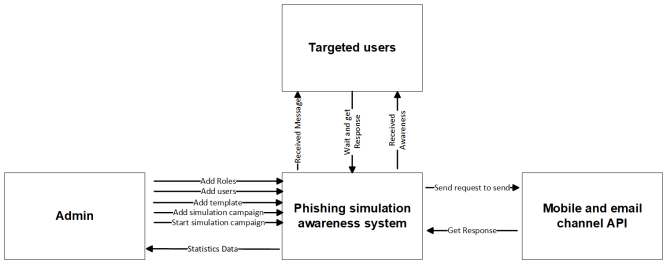


Fig. 3. Context diagram of the phishing simulation awareness system.

D. Use Case Diagram

The functional overview of the use case diagram answers the question of how the primary actors, which are the administrator and the targeted user, interact with the phishing simulation platform. The system is described from campaign configuration through to user interaction all the way to delivery of user feedback.

Fig. 4 shows the use case diagram, and the administrator responsibilities are listed in Table II.

TABLE II. ADMINISTRATOR USE CASES

Functionality	Description
Manage Roles	Decide what different user groups (e.g., Sales, Marketing) should be able to simulate.
Manage Users	Set and assign users to pre-defined roles for campaign targeting.
Manage Templates	Build Email, SMS, and WhatsApp phishing templates for delivery along with placing holder to substitute with a link generated by the system and a targeted user name.
Create Campaigns	Create phishing simulations and link them to roles.
Launch Simulation	Initiate an immediate phishing simulation.
View the Campaign List and Details	Check on all current and past campaigns with status updates.
View Tracking Data	Look into link clicks and response times of users.

The targeted user responsibilities are listed in Table III.

TABLE III. TARGETED USE CASES

Functionality	Description
Receive Message	The assigned communication channel is used to message users with phishing messages.
Click the Link	on the interaction with the phishing link, the user behavior is tracked.
View Awareness Page	Upon clicking a phishing link, the user is automatically directed to an awareness page for immediate feedback.

From this diagram, the system will cover all functional requirements of the system as well as support interaction between administrative users and end-user participants.

E. System Architecture

The proposed phishing simulation platform uses a layered and modular architecture to make sure that the system is scalable, maintainable, and secure. As depicted in Fig. 5, the system consists of many interconnected components that are partitioned into six key layers.

1) Frontend layer – admin dashboard: Administrators have a secure, web-based interface to this layer. It enables:

- User and role management.
- Template design.
- Campaign creation and tracking.
- Real-time navigation across simulation statistics.

The dashboard is constructed by modern web technologies and hence responds and provides intuitive interaction.

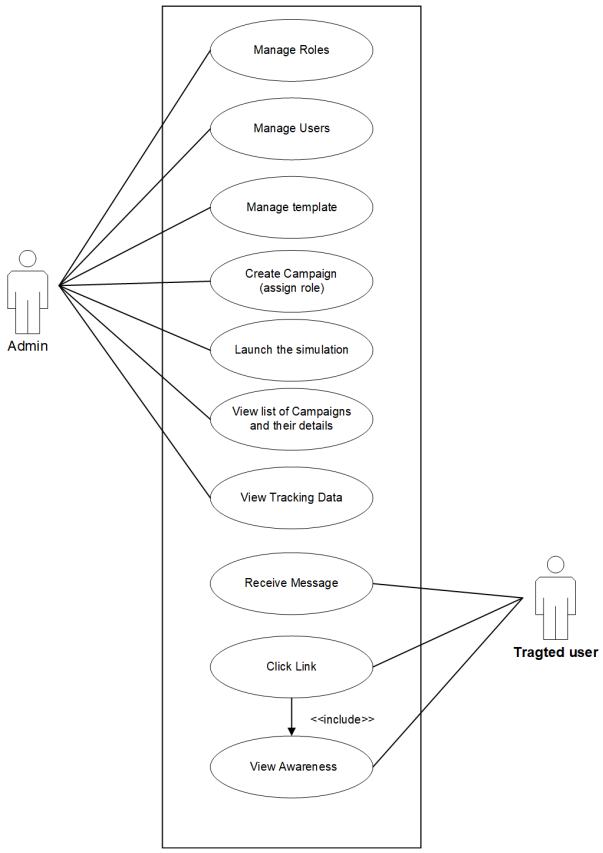


Fig. 4. Use case diagram of the phishing simulation platform.

2) Backend logic layer: Implemented using React.js, this layer:

- Handles campaign logic.
- Creates a unique phishing link for each user template pair.
- Logs user activity (clicks, timestamps).
- Assigns behavior-based risk scores.

3) Communication API layer: It is responsible for sending phishing messages via using resend API for sending email and looking for SMS and Whatsapp.

This layer communicates with the backend, acting as a middle layer to dispatch phishing content and log delivery status using a secure request-response mechanism.

4) Tracking engine:

- It creates individualized URLs for every user.
- To capture the users' actions, such as response time and frequency of clicks.

5) Awareness module: In case of a user interaction with a phishing simulation:

- This redirects them to a tailored awareness page.
- The page provides microlearning content or training tips delivered just in time to reinforce security behavior.

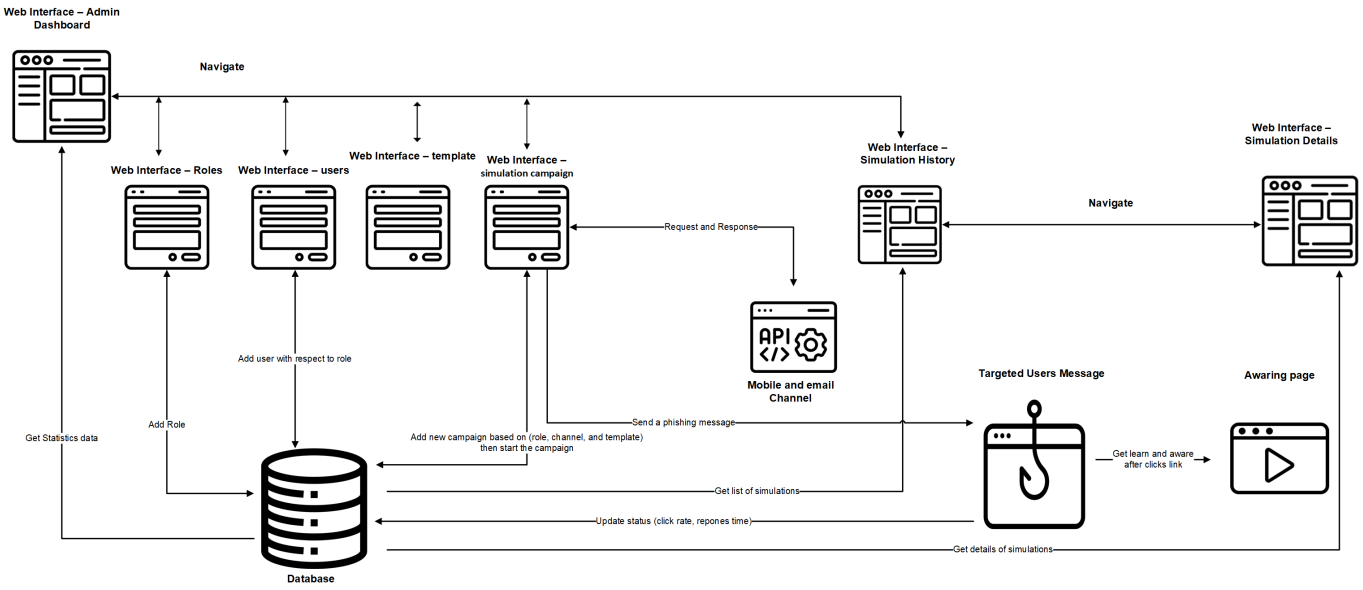


Fig. 5. System architecture of the phishing simulation platform.

6) Database layer: All simulation data, including is stored and secured on the system.

- User roles and profiles.
- Campaign settings and templates.
- Click activity and behavioral logs.

F. Database Design

Designing the database for the phishing simulation platform is based on the normalized relational model to ensure data integrity, efficient querying, and scalability. It is capable of catching main entities like users, roles, simulations, templates, and behavior tracking. In this section, we present an Entity Relationship (ER) model at a high level, as well as the corresponding relational schema to be used for the implementation.

1) ER Diagram: As shown in Fig. 6, the logical structure of the data used in the simulation system is modeled by the ER diagram.

Table IV captures the key relationship between the entities involved.

TABLE IV. THE RELATIONSHIPS BETWEEN KEY ENTITIES.

Entity	Goal
Users	those who are phished in phishing simulations. A user can be assigned to one of the roles.
Roles	Determine user categories (e.g., HR, IT) to target for simulation.
Phishing Templates	Using pre-defined messages for each campaign based on their type and delivery channel.
Simulations	The admin will create a campaign using templates and targeting specific roles.
Simulation Tracking	Tracking individual users' interactions with phishing simulations (delivery status, response time, and click behavior).

Key relationships include:

TABLE V. PHYSICAL IMPLEMENTATION OF THE PHISHING SIMULATION DATA MODEL WITH PK AND FK RELATIONSHIPS

Table	Role
Users	Contains basic user details along with the details of role association.
Roles	The definition of roles used to group users.
Phishing Templates	Stores the message content, channel type, and creation date.
Simulations	Metadata about the campaign, including channel, name, and the link to templates.
Simulation Tracking	Records user responses in the form of click_time (initial NULL value), track_id, and delivery status.
Simulations_Targets	Connects simulations to targeted roles.

- One-to-many relationship between Roles and Users.
- Many-to-one between Simulations and Templates.
- Users and Simulations, they are many-to-many relationships, and we will resolve this through Simulations Tracking.

2) Database schema diagram: Fig. 7 illustrates the way the ER model is implemented using the Structured Query Language (SQL) relational tables. It describes what the data types are, the Primary Keys (PK) and Foreign Keys (FK) throughout the system.

The schema adheres to the normalization principles of the Third Normal Form (3NF), aiming at eliminating redundancy and ensuring relational consistency. It makes possible user performance tracking, report generation, and behavior analysis at individual and group levels. Table V represents the tables and role of each for the system.

G. Workflow Diagram

The phishing simulation platform is a streamlined, direct execution workflow that keeps administrators able to initiate

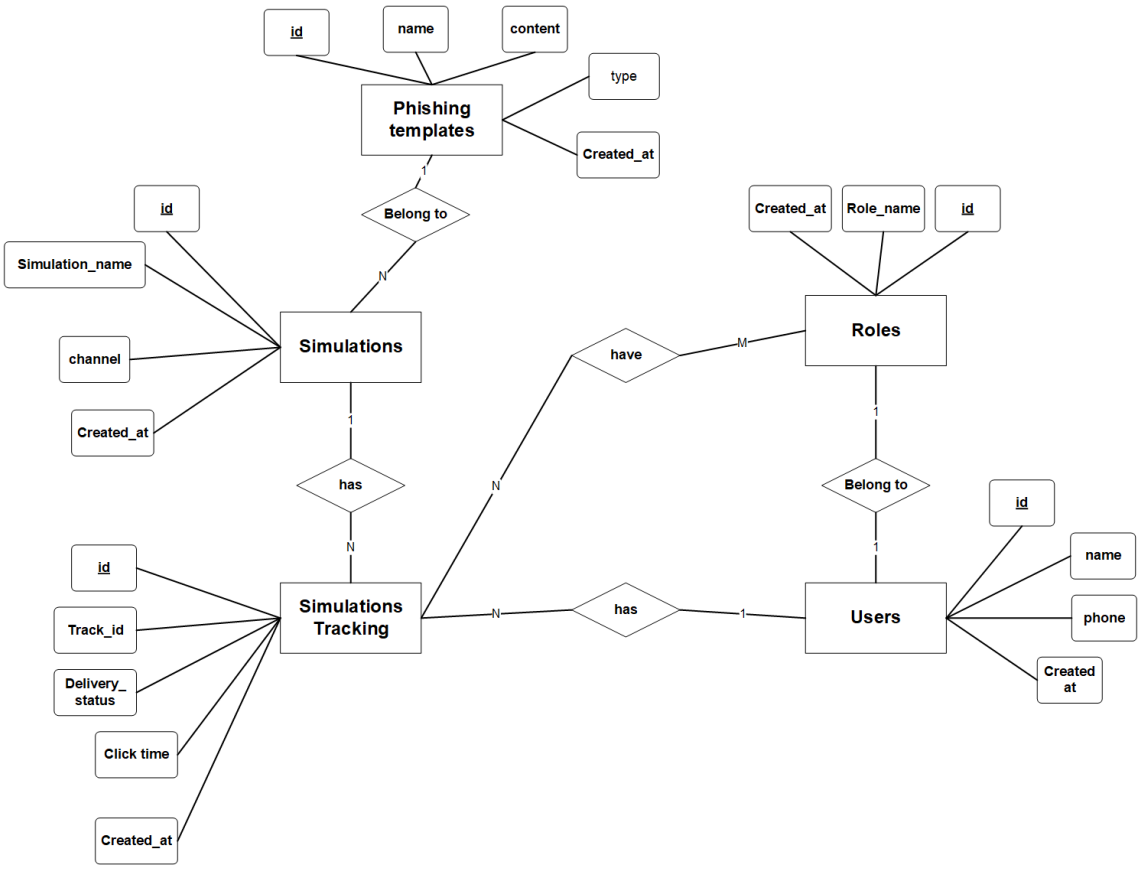


Fig. 6. ER Diagram of the phishing simulation system.

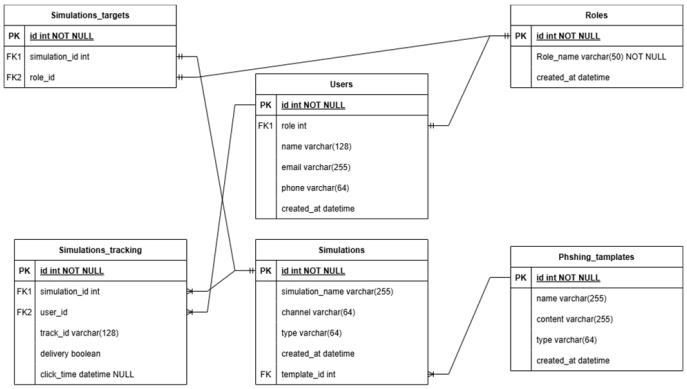


Fig. 7. Database schema diagram of the phishing simulation system.

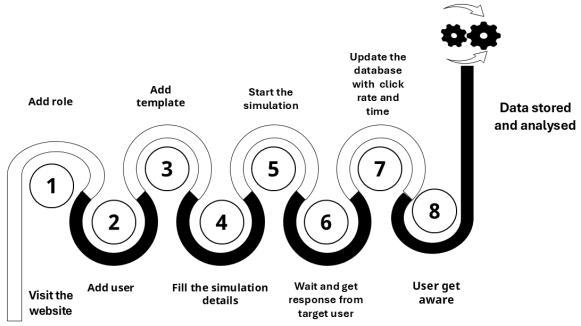


Fig. 8. System workflow of the phishing simulation platform.

and monitor phishing simulations without much delay in configuration. The workflow shown in Fig. 8 details the workflow as eight steps in succession: administrative setup, campaign launch, user interaction, and behavior tracking.

1) Step-by-step workflow:

a) System access: The process is initiated by logging into the platform via a secure web interface by the administrator.

b) User registration: User profiles are set up and assigned to pre-defined (organizational) roles like HR and Finance so that specific simulations can be run against them.

c) Template configuration with placeholders: Phishing templates with dynamic placeholders are created by the administrator. The backend system automatically replaces these placeholders at runtime.

- {{user_name}}: Replaced with the targeted user's name.
- {{tracking_url}}: Replaced with a unique tracking link generated for the user-template pair and simulation campaign.

- {{random_number}}: Generate a random number (Fake invoice or verification code etc).

Using this mechanism allows a simulation to be more realistic and more personalized, thus making it more effective in emulating real phishing attempts.

d) Simulation setup: By selecting a template, delivery channel, and target user groups (roles), it is the admin who configures the simulation. It is saved and immediately ready for launch.

e) Immediate campaign execution: Once the backend is launched, it dynamically fills in the placeholders with actual values and dispatches the phishing messages through Email, SMS, or WhatsApp API. A customized message with a user's personal data and a unique tracking link is sent to each user.

f) User interaction logging: The system records user actions including: delivery status, link click events, and time to response.

g) Behavioral data collection and risk scoring: The data of user interaction is stored in the database. For instance, the system calculates risk scores and classifies an organization's risk level according to predefined thresholds.

h) Awareness feedback delivery: Upon interacting with phishing links, users are redirected to a training page with immediate awareness content. The reinforcement of learning through real-time feedback and behavioral reinforcement is reiterated by this step.

This is a behavior-focused and fully dynamic workflow that allows for the multi-channel delivery of phishing simulations. It makes use of placeholders to increase the realism of messages and its supportable and scalable deployment.

V. System Implementation

Thereafter, the phishing simulation platform was implemented by translating the proposed architecture and system design into a working web-based system. The development took a modular approach to allow for each component to be developed, tested, and improved independently: user management, template creation, simulation launch, behavior tracking, and reporting.

A. Technology Stack

Table VI shows the tools and technologies used for building the platform.

TABLE VI. List of Tools and Technology

Component	Technology / Tool
Backend Development	React.js
Frontend Development	HTML and CSS
Database	PostgreSQL
Message Delivery APIs	resend API for email
Development Environment	Visual Studio Code
Diagram Modeling	Microsoft Visio and draw.io.

B. Development Modules

Several modular components were used to implement the platform, and each module plays a particular role in the phishing simulation workflow.

- Admin Dashboard: Fig. 9 shows the web interface for administrators to navigate users, roles, other interfaces, and review reports.

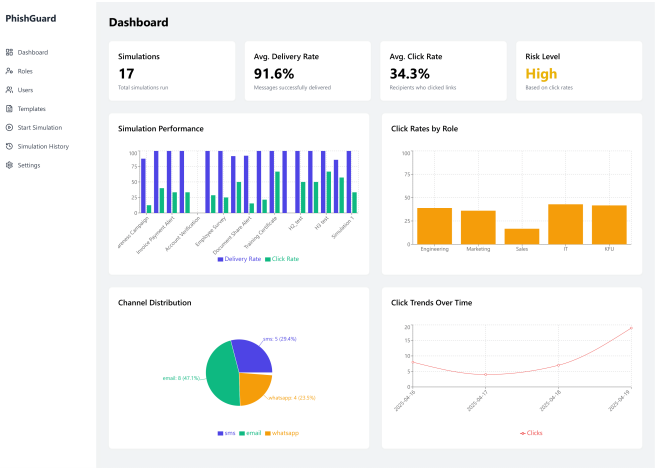


Fig. 9. Phishing simulation admin dashboard.

- Template Engine: Enables dynamic placeholder injection such as {{user_name}}, {{tracking_url}} and {{random_number}} which leaves room for maximizing personalization. Fig. 10 shows an example from the system.

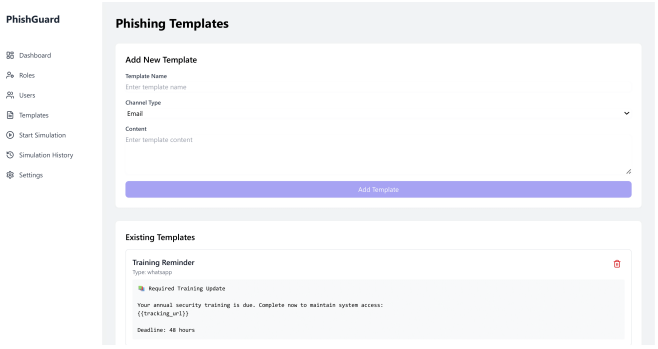


Fig. 10. Phishing simulation adding template page.

- Simulation Dispatcher: It sends out phishing messages through integrated Email, SMS, and WhatsApp APIs. Fig. 11 shows the interface of creating a new simulation, and this gets triggered after clicking the button.
- Tracking Mechanism: Unique tracking links record the user clicks, the timestamps, and response metrics. Fig. 12 shows the email received from the system after the launching of the simulation.
- Awareness Module: Immediately trains users who fall for simulations by redirecting them to an awareness page as shown in Fig. 13.

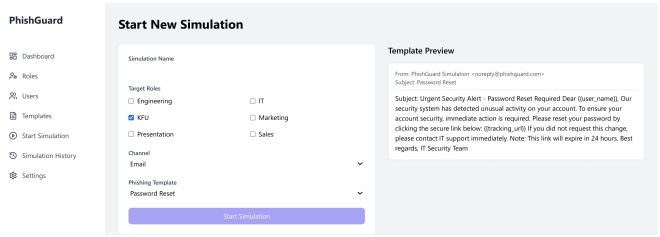


Fig. 11. Configuration setting of the simulation campaign.

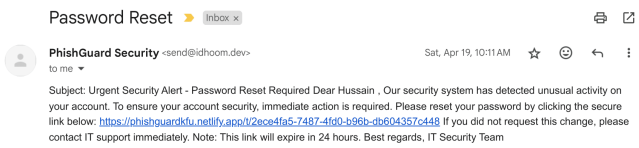


Fig. 12. Phishing simulation email received with a unique tracking link.

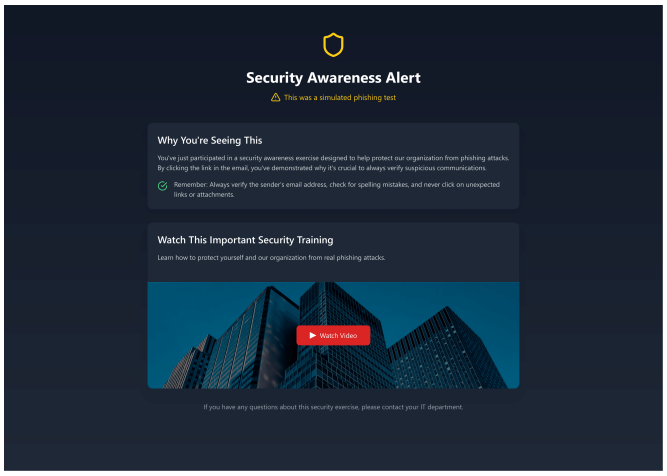


Fig. 13. The awareness page from the system.

- Database Integration: It normalizes the storage of all simulation data, logs, and user profiles.

C. Testing and Verification

To verify, individual core modules were tested.

- Correct rendering and substitution of template placeholders.
- A reliable way of interacting with an API for message delivery.
- Make sure accurate tracking of click events and response time.
- User interface that is responsive and compatible with browsers Chrome, Firefox, and Edge.

D. Deployment

The platform was deployed using Netlify, a cloud service platform that is known for being simple and powerful when deploying web applications. Finally, Netlify was used to host frontend components (admin dashboard and awareness pages), as they allowed for really fast, continuous deployment. This approach made for easy updates and a reliable way to host the web interface.

This section contains data and visuals generated from sample values to show the capabilities of the system. No real user interaction was recorded.

E. Behavioral Analytics and Reporting

The phishing simulation platform also includes a set of in-depth analytics that allows administrators to measure user behavior, campaign performance, and phishing awareness in terms of users, roles, and communication channels. The collection of data happens in real-time and creates an intuitive dashboard to visualize the data.

1) Admin dashboard reports: The key visual and statistical insights are available on the admin dashboard, which are:

- Total number of simulations conducted.
- The average delivery and click-through rates for each Simulation campaign.
- Low, Medium, High risk classification levels.
- Email, SMS, WhatsApp channel usage summary.
- Click trends over time.
- Click rate by role.

These visualizations that are shown before in Fig. 9 enable organizations to discover vulnerable users, assess the efficiency of training campaigns, and decide on the follow-up intervention.

2) Simulation history and detailed reports: All of the simulations executed within the platform are maintained on a complete history, which administrators can view:

- Simulation campaign names and associated templates.
- The targeted roles.
- Delivery rate and click rate performance.
- Channel of communication type used.
- Date and time of execution.

Fig. 14 shows lists of past simulation campaigns with delivery and engagement metrics across targeted roles.

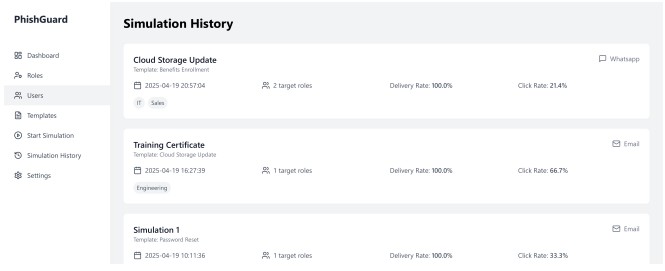


Fig. 14. Simulation history log – sample data.

By selecting a simulation campaign, you can discover detailed analytics and view these analytics:

- Date and time of execution of the simulation.
- Number of targeted roles.
- Delivery rate and click rate performance of the simulation campaign.
- The click rate per role as a bar chart.
- Overall response to the phishing message by click.
- List of users tragedy respect to user role along with delivery Status, click Status and the response date and time.

Fig. 15 shows an individual simulation campaign with its information and analytics.

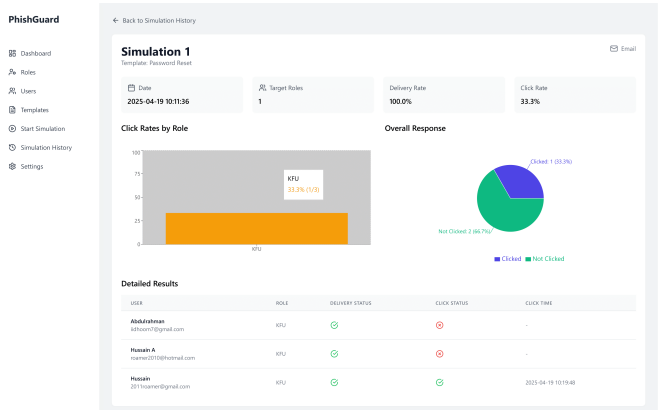


Fig. 15. Specific simulation campaign - analysis – sample data.

VI. Discussion

As a result, in this study, a customizable phishing simulation platform was presented for the improvement of user-specific cybersecurity awareness by means of targeted phishing and real-time behavioral analytics through multi-channel delivery. The proposed approach improves the traditional awareness methods by providing personalized simulations of the attacks through Email, SMS, and WhatsApp, which are a real attack vector. It generates phishing content dynamically using placeholders such as {{user_name}} and {{tracking_url}}, increases the user's engagement and offers an authentic environment to recognize phishing in familiar scenarios. Also, the behavioral tracking engine assists organizations in identifying click patterns and response times and identifies high-risk individuals with whom training intervention data can be linked. Furthermore, this also provides immediately awareness feedback to support just in time learning to reinforce safe practices when the users are the most vulnerable. Being adaptable, realistic and insightful at the user level, this directly fills the gaps in current generic, one-size-fits-all, one size fits all training programmes. Moreover, the system is deployed using a cloud-ready frontend with Netlify and has a modularity in the architecture that would make it a good candidate for real-world product adoption in enterprise environments where ongoing, customized phishing awareness is needed.

RQ1-RQ5 were created to serve as the system development framework and future evaluation. This paper is focused on the design, architecture, and demonstration of the platform through computational simulation. The above specifies the way each of the questions is answered at this step:

- RQ1 (long-term awareness and behavioral change): The proposed platform has modules of behavioral-tracking (e.g., click behavior, response time), which are proposed as a precondition to assess the long-term awareness and the following behavioral change of users.
- RQ2 (emerging behavioral patterns): The behavioral data produced on the platform enables a researcher to record the user response patterns based on the role and channel. These patterns provide a basis where emergent trends of behavior are identified.
- RQ3 (advantages of customization): The current platform enables the customization of simulations by the role and behavior of the user, unlike the standard, non customized training environment.
- RQ4 (engagement across platforms): Support multi-channel, but the study has not explored the differences between engagement across the channels.
- R5: (real-time feedback and analytics): The current platform delivers real-time awareness pages upon phishing link, but the study has not address the effectiveness and consider as future work.

The discussion in the current paper provides that the platform in question is intentionally designed in such a way that it can address the technical concerns identified, but its general effectiveness has not been tested and consider as future work.

VII. Challenges to the Study

Several challenges were faced in developing the phishing simulation platform. However, because of ethical restrictions, the system was not tested with real users, and all behavioral data were simulated inputs that restricted the ability to measure training effectiveness on real users. The platform was developed to support multi-channel delivery through Email, SMS, and WhatsApp, however, it was not possible to see how users might react differently on these platforms. Furthermore, the awareness page that is opened when a phishing link is clicked is also created with basic informational content and does not yet have interactive features like quizzes or scenario training. These improvements are slated for future versions in order to drive warmer engagement and better learning outcomes. In addition, there is no public dataset found that contains phishing simulation analytics or user behavioral responses with which to benchmark the system outputs against real-world baselines.

VIII. Conclusions

In this study, we design and implement a customizable phishing simulation platform to tackle challenges associated with traditional cybersecurity awareness training. From a review of the existing literature and tools, the research gap was identified that most of the awareness programs are

generic, limited to email simulation, not personalized, and do not provide real-time behavioral insights. The proposed platform then comes up with the multi-channel simulation delivery via Email, SMS, and WhatsApp, role-specific targeting, and dynamic template generation using placeholders such as {{user_name}} and {{tracking_url}}. All of these features make the simulation more realistic and engaging to the user. A behavioral tracking engine is also integrated within the system to track user interactions, assess risk levels, and provide immediate feedback via an awareness page, thus supporting just-in-time learning. The platform provides an administrator dashboard to analyze in detail campaign performance as well as user behavior for a given campaign and help to make the platform adaptable to a particular organization's needs. Due to ethical restrictions, the platform was tested with simulated data. Nevertheless, a cloud-ready frontend was deployed, and the platform shows potential for real-world use. Future work will be to conduct live evaluations, to enhance the awareness content with interactivity learning, and to complete integration of all communication channels for full multi-platform simulation capability.

FUNDING

This work was funded by King Faisal University, Saudi Arabia. [Project No. GRANT KFU252300].

ACKNOWLEDGMENT

This work was supported through the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Project No. GRANT KFU252300].

CONFLICTS OF INTEREST

All authors declare no conflict of interest.

REFERENCES

[1] H. Shahbaznezhad, F. Kolini, and M. Rashidirad, "Employees' behavior in phishing attacks: what individual, organizational, and technological factors matter?" *Journal of Computer Information Systems*, vol. 61, no. 6, pp. 539–550, 2021.

[2] U. Divakarla and K. Chandrasekaran, "Predicting phishing emails and websites to fight cybersecurity threats using machine learning algorithms," in *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*. IEEE, 2023, pp. 1–10.

[3] M. Rader and S. Rahman, "Exploring historical and emerging phishing techniques and mitigating the associated security risks," *arXiv preprint arXiv:1512.00082*, 2015.

[4] B. Naqvi, K. Perova, A. Farooq, I. Makhdoom, S. Oyedeji, and J. Porras, "Mitigation strategies against the phishing attacks: A systematic literature review," *Computers & Security*, vol. 132, p. 103387, 2023.

[5] M. L. Jensen, R. T. Wright, A. Durcikova, and S. Karumbaiah, "Improving phishing reporting using security gamification," *Journal of Management Information Systems*, vol. 39, no. 3, pp. 793–823, 2022.

[6] A. K. Jain and B. B. Gupta, "A survey of phishing attack techniques, defence mechanisms and open research challenges," *Enterprise Information Systems*, vol. 16, no. 4, pp. 527–565, 2022.

[7] S. G. Abbas, I. Vaccari, F. Hussain, S. Zahid, U. U. Fayyaz, G. A. Shah, T. Bakhshi, and E. Cambiaso, "Identifying and mitigating phishing attack threats in iot use cases using a threat modelling approach," *Sensors*, vol. 21, no. 14, p. 4816, 2021.

[8] C. K. Olivo, A. O. Santin, and L. S. Oliveira, "Obtaining the threat model for e-mail phishing," *Applied soft computing*, vol. 13, no. 12, pp. 4841–4848, 2013.

[9] M. A. Bawazir, M. Mahmud, N. N. A. Molok, and J. Ibrahim, "Persuasive technology for improving information security awareness and behavior: literature review," in *2016 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*. IEEE, 2016, pp. 228–233.

[10] P. K. Yeng, B. Yang, M. A. Fauzi, and P. Nimbe, "A framework for exploring incentive methods towards reducing phishing susceptibility in healthcare: Based on a review and in-the-wild-field study approach," *2023 Intelligent Methods, Systems, and Applications (IMSA)*, pp. 228–234, 2023.

[11] T. Sutter, A. S. Bozkir, B. Gehring, and P. Berlich, "Avoiding the hook: influential factors of phishing awareness training on click-rates and a data-driven approach to predict email difficulty perception," *Ieee Access*, vol. 10, pp. 100 540–100 565, 2022.

[12] B. M. Ahmad, S. M. Ahmed, and D. E. Sylvanus, "Enhancing phishing awareness strategy through embedded learning tools: a simulation approach," *Archives of Advanced Engineering Science*, pp. 1–14, 2023.

[13] J. Kävrestad, A. Hagberg, M. Nohlberg, J. Rambusch, R. Roos, and S. Furnell, "Evaluation of contextual and game-based training for phishing detection," *Future Internet*, vol. 14, no. 4, p. 104, 2022.

[14] S. Chatchalermpun and T. Daengsi, "Improving cybersecurity awareness using phishing attack simulation," in *IOP Conference Series: Materials Science and Engineering*, vol. 1088, no. 1. IOP Publishing, 2021, p. 012015.

[15] M. Canham, C. Posey, and M. Constantino, "Phish derby: Shoring the human shield through gamified phishing attacks," in *Frontiers in Education*, vol. 6. Frontiers Media SA, 2022, p. 807277.

[16] C. O. Tinubu, O. J. Falana, E. O. Oluwumi, A. S. Sodiya, and S. A. Rufai, "Phishgem: a mobile game-based learning for phishing awareness," *Journal of Cyber Security Technology*, vol. 7, no. 3, pp. 134–153, 2023.

[17] W. Yeoh, H. Huang, W.-S. Lee, F. Al Jafari, and R. Mansson, "Simulated phishing attack and embedded training campaign," *Journal of Computer Information Systems*, vol. 62, no. 4, pp. 802–821, 2022.

[18] A. Ciupe and B. Orza, "Reinforcing cybersecurity awareness through simulated phishing attacks: Findings from an hei case study," in *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2024, pp. 1–4.

[19] F. Rizzoni, S. Magalini, A. Casaroli, P. Mari, M. Dixon, and L. Coventry, "Phishing simulation exercise in a large hospital: A case study," *Digital Health*, vol. 8, p. 20552076221081716, 2022.

Reducing Computational Complexity in CNNs: A Focus on VGG19 Pruning and Quantization

Md. Mijanur Rahman[1], Anik Datta[2], Md. Sabiruzzaman[3], Md Samim Ahmed Bin Hossain[4]
Assistant Professor, Dept. of CSE, Southeast University, Dhaka, Bangladesh[1]
Dept. of CSE, Southeast University, Dhaka, Bangladesh[2,3,4]

***Abstract*—The Convolutional Neural Network (CNN) models are effective in computer vision strategies and have gained popularity due to their strong performance in visual tasks. Nevertheless, models with architectures such as VGG19 are expensive in terms of computational resources and require huge memory, which limits their usage on low-end devices. The study examines how efficiency can be increased in the model VGG19 by using model compression techniques, like pruning (structured and unstructured) and quantization (8-bit and 4-bit Quantization-Aware Training - QAT). The efficiency of the individual compression approaches was tested by thoroughly exploring the VGG19 with the MNIST, CIFAR-10, and Oxford-IIIT Pet datasets. Each model was evaluated against the baseline based on measures of accuracy, model size, inference time, and complexities of the model, CPU usage, and memory usage. The applied QAT approach reduced the model size by 75% with a drop in computational cost across all methods. In addition, the 8-bit quantitative assessment allowed for substantial system compression alongside increased speed delivery with minimal impact on accuracy. The highest compression and sparsity achieved by 4-bit QAT was 48%, which was not effective as it reduced accuracy on complex datasets, with additional computational overhead on T4 GPU. Structured pruning resulted in faster inference, but unstructured pruning also demonstrated a good result in retaining accuracy and even improving it. To simplify the VGG19 structure, pruning and quantization mechanisms are suggested in order to simplify the architecture to implement the model on edge devices sufficiently, without compromising prediction performance.**

Keywords*—*VGG19 Model optimization; model compression; pruning, quantization; structured pruning; unstructured pruning; memory management; quantization-aware training; 8-bit; 4-bit

I. Introduction

CNNs have experienced significant advancements during the last few years to solve a wide range of computer vision [1] problems that include simple image classification [2] tasks alongside complex object detection [3] problems. The VGG19 CNN model stands among multiple models as a significant impactor because it significantly contributes to deep learning by accurately addressing challenging image classification tasks [4]. Although efficient, direct deployment of VGG19 on resource-limited devices, i.e., smartphones and edge computing setups, is infeasible because of its enormous parameter size and excessive computational demands in terms of floating-point operations per second (FLOPs) [5]. Model compression methods, which include pruning and quantization, have addressed these challenges effectively. Quantization achieves maximum savings in memory usage plus computation requirements through its usage of weight and activation bit widths, which makes it ideal for devices with minimal resources [6], [7]. Model performance enhancement through pruning strategies involves connection reduction for both size reduction and computational complexity reduction [8].

Although the advancement in model compression has been remarkable, most of the current research mainly concentrates on large-scale data and high-resource computing settings. However, their performance is hardly well-tested in environments where datasets are comparatively small and computer resources are modest. The testing of their performance needs more investigation in situations that use small datasets alongside minimal computer resources. Different pruning techniques like layer-wise pruning [9], structured pruning [10], activation-based pruning [11], and different quantization techniques like mixed-precision and post-training quantization [12], [13], have been studied separately on different models. There is limited modeling understanding of where the efficiency of these techniques is jointly implemented on the VGG19 model with constrained availability of resources. This study attempts to solve this gap by making a thorough evaluation of pruning and quantization techniques performed on VGG19.

To address this gap, the following research questions are asked:

- RQ1: What level of impact do various pruning methods (structured versus unstructured) have on the performance considering trade-offs of VGG19 on resource-limited datasets?
- RQ2: Which bit-width (8-bit QAT versus 4-bit QAT) of the encoded quantization is the best to keep the accuracy and simultaneously maximize the compression?
- RQ3: What are the performance of these optimization methods on data with different complexities (MNIST, CIFAR-10, Oxford-IIIT Pet)?

The gap is to be filled by comparing the accuracy, size, and processing costs of models, then evaluating the weight distribution and parameter sparsity norms [14]. To the best of our knowledge, such an in-depth, visual comparison of VGG19 with different reporting conditions, namely structured and unstructured pruning and 8-bit and 4-bit QAT.

This study compares and contrasts these approaches and studies their behavior under both isolated and combined settings and resource-limited data. The key contributions of this study are: i) an empirical evaluation of many pruning and quantization techniques on VGG19 and ii) their analysis of utility-by-constraints of usefulness under resource-constrained conditions, hence generalizing beyond the large-scale studies

done typically in the literature on model compression techniques.

The structural framework of the study is as follows: the literature review is given in Section II, methodology in Section III, the experimental results in Section IV, detailed discussion of the results in Section V, the future directions in Section VI, and conclusion with final remarks in Section VII.

II. Related Work

Research into deep learning model improvement seeks to enhance VGG19 and its contemporaries. Different research studies investigated pruning alongside quantization and transfer learning methods to minimize complex model sizes with no negative impact on accuracy levels. Resource-restricted devices such as mobile devices and embedded systems benefit most from optimization techniques because optimizing performance takes priority on such devices.

Researchers adopted VGG19 for detecting defects on strip steel surfaces through successive solutions of restricted datasets, according to study [15]. The team obtained 97.8% accomplishment in accuracy through the implementation of new preprocessing approaches and model enhancement strategies. Until it conducted another training process, the detection model failed to detect recently introduced defect types and missed apparent defects in its evaluation. The authors conducted a comprehensive research study of diverse pruning and quantization methods in [16]. According to the research findings, model pruning results in an accurate yet significantly smaller model, approximately 13 times smaller compared to the original version, while quantization offers efficiency benefits, although this comes at a cost to precision levels [17].

The fusion filter and channel pruning approach from reference number [18] eliminated 83.8% of model parameters and 71% of computations but obtained slightly better accuracy results. The reinforcement learning-based pruning techniques [19] use automatic management for pruning speeds to achieve 73% maximum computation reduction before accuracy starts to degrade beyond traditional methods of kernel pruning along with guided pruning and soft filter pruning. Scientists across the board have conducted substantial research that led to the development of various quantization techniques. Many scientists have extensively researched and developed various techniques for quantization. The study in [20] analyzed VGG19 along with MobileNet when their parameters were quantized to 8 bits. VGG19 demonstrated higher quantization resilience since it maintained only an 8.58% decline in accuracy, whereas MobileNet experienced substantial accuracy degradation. Research indicates that combined quantization and pruning methods create effective solutions for the objective. Using a 4-bit quantization on the models, it has shown more than 30% variable reduction and as much as 50% memory reduction in CNN models, with minimal damage to accuracy on real-world edge devices [21].

The improvement of VGG19 and ResNet34 model accuracy and complexity comes from Minimax Concave Penalty (MCP)-based pruning techniques [22]. The technique adopted DeepPruningES by [23] along with its evolutionary strategy-based method to achieve up to 80.89% FLOPs reduction on ResNet56 through the generation of multiple model variants with various trade-off selections. A technique that merges weight and feature-level pruning with quantization methodology benefits MobileNetV2 by sustaining a 0.37% accuracy decrease after pruning 50% of the weights while improving operational performance [24].

VGG19 demonstrates its deployment capabilities in real-world applications apart from optimization purposes, thus showcasing its usable nature and versatility. Several improved forms of VGG19 demonstrate high performance levels when used for medical diagnosis applications. The proposed E-VGG19 system by Kandhro et al. detected skin cancer in real-time through pooling/dense layers and the combination of CNN features and SVM and KNN classifiers [25]. The collaboration between Faghihi and colleagues led to 98.18% skin lesion classification accuracy when using VGG19/VGG16 together with feature-modified AlexNet through transfer learning and dropout [26]. The combination of VGG19 along with morphological features and Otsu thresholding and watershed segmentation resulted in more than a 99% success rate for parasitic organism detection per Kumar et al. [27]. Tuned VGG19 models showed better performance than pipeline workflow combinations of previous architectural models for the detection of COVID-19 and pneumonia on X-ray dataset analysis [28], [1]. Research findings proved VGG19's superiority in medical imaging applications over newer structures, DenseNet and ResNet [29]. Enhancing the performance on imbalanced COVID-19 datasets greatly relied upon data augmentation along with hyperparameter optimization techniques [30].

The study by Ahmad et al. demonstrated how VGG19 successfully detected tomato leaf diseases by fine-tuning its last layers when trained on combined field and laboratory samples [31]. Remote sensing technology reached peak performance in small aerial target detection under class imbalance through the hybrid VGG19 and ResNet50 architecture combined with the Spiral Search Grasshopper Algorithm [32]. Accounting for road infrastructure monitoring involved the optimization of VGG19 through Q-learning applied on a two-layer optimizer that managed redundancy in filters and reached a 96.4% F1-score in drone-based road damage detection [33]. The VGG19 network underwent minor modifications to create a face anti-spoofing system, which achieved a 100% success rate in detecting spoofed attacks [34]. The Table Extraction Model (TEM) developed by Iqbal et al. used VGG19 as its base structure to exceed heuristic scanning methods in extracting document tables [35]. The VGG19 model functions as part of content-based image retrieval (CBIR) systems for image retrieval purposes. By adding texture and color features to its design, the model delivered better accuracy for images with semantic similarities [36].

Despite its dated design, VGG19 remains a flexible and effective baseline for optimization research. The model's basic design makes it an optimal choice for optimization research through pruning and quantization and transfer learning and hybrid approaches. The analyzed studies possess various areas that still need improvement. However, some of the biggest VGG19 struggles are in the sense that it has not worked on small datasets or subsets of datasets as appropriate, imbalances, compromises with precision and compression, and not being able to generalize outside of the domain without transfer learning, and there do not exist efficient compression frameworks.

Constructing accurate and highly deployable VGG19 versions requires the resolution of particular challenges to service data-scarce and resource-limited systems.

III. Methodology

The research applied pruning and quantization processes to VGG19 architecture when assessing three benchmark datasets consisting of MNIST, CIFAR-10, and Oxford-IIIT Pet. The datasets contain visual data categories with original dimensions of 28×28 and 32×32 pixels, thus providing efficient computation while supporting complex image identification. Data augmentation procedures and preprocessing methods allowed all images to be uniformly resized to 224×224 pixels, even though VGG19 maintained its fixed input requirements. Among the available datasets, the MNIST provides 70,000 grayscale digits of handwriting distributed across 10 classes, and CIFAR-10 shows 60,000 color object pictures divided into 10 categories. A total of 7,349 images derived from the Oxford-IIIT Pet dataset with 2 classes were utilized, but training and evaluation occurred using 5,000 strategically chosen illustration samples from each collection. The data was split into training and experimental samples, where the data partitioned itself into training and experimental parts, keeping an 8:2 ratio, which ensured model training with diverse datasets while maintaining testing reliability on new data samples. The decision to split the data 80:20 was made by evaluating how training robustness relates to testing strength.

The torchvision library in PyTorch provides comprehensive image processing capabilities through data loading and preprocessing features that contain built-in functionality to control datasets alongside transform functionalities and augmentation options.

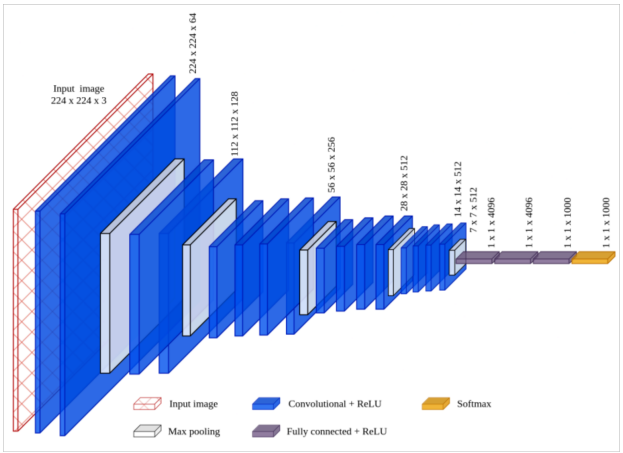


Fig. 1. VGG19 Architecture.

The VGG19 convolutional neural network developed by the Oxford Visual Geometry Group introduces a total of 19 layers composed of 16 convolutional layers together with 3 fully connected layers and 5 max-pooling layers (see Fig. 1) [37].

Through its 3x3 kernel with 1 stride, the model can master small-scale spatial features within each layer. After each convolutional layer, the model passes through ReLU non-linearity for introducing complexity. The max-pooling layers progressively shrink the spatial dimensions but keep the most important of the feature representations. To begin the multi-class classification process, the last fully connected (FC) layer will include the final softmax activation function, and binary classification will be applied to the Oxford-IIIT Pet dataset.

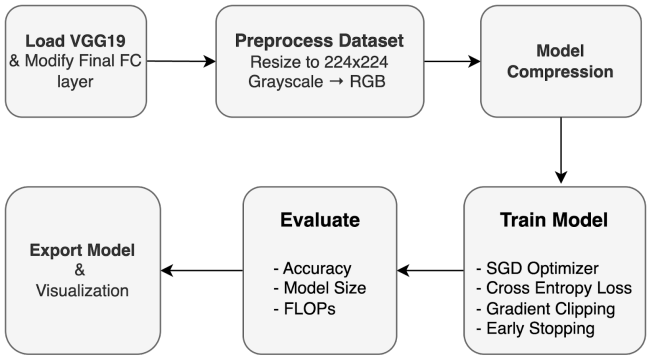


Fig. 2. VGG19 Model optimization and evaluation pipeline.

Fig. 2 represents the operational sequence of modifying, compressing, training, and evaluating the VGG19 model. Testing and training using optimization methods serve as intermediate steps before exporting the model using compression methods.

A. Pruning

The model simplification technique of pruning functions by removing insignificant weights and layers or neurons to maintain performance quality. This process substantially decreases both memory usage and computational expenses for the model. Deploying models on low-end devices makes for efficient performance and function use.

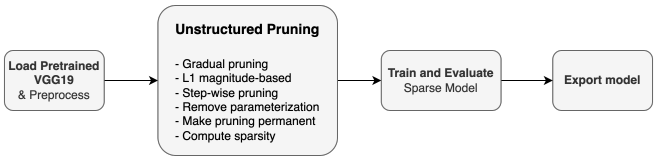


Fig. 3. Unstructured pruning training.

1) Unstructured pruning: In unstructured pruning, the importance of the weights is analyzed, and less important weights are removed. This is usually determined by a gradient-based method during backpropagation. Although this technique significantly reduces the size of the model, the sparse weight matrix has a limited impact on improving the efficiency of standard hardware implementations.

The removal of unimportant weights through unstructured pruning produces reduced models that preserve most of their operational accuracy. Fine-tuning post-pruning is used to recover any possible accuracy loss.

The methodology for unstructured pruning implementation on a pre-trained VGG19 model is illustrated in Fig. 3. L1 magnitude-based pruning uses step-wise methods to remove parameterization before enforcing permanent sparsity, followed by training and evaluation to export the model.

a) Mathematical representation: L1-norm Based Unstructured Pruning: For each weight tensor w in a module, the importance of each weight is computed as the L1-norm, as shown in Eq. (1).

$$\text{importance} = |W| \tag{1}$$

b) Gradual pruning strategy: In the implementation of this strategy, the pruning is applied in several iterations and does not remove a fixed percentage of the weights in one pass, as shown in Eq. (2). The pruning amount is gradually increased over a series of steps, starting from an initial amount and progressing to a target amount, as shown in Eq. (3).

$$\text{Step Increment: } A_{\text{steps}} = \frac{A_{\text{final}} - A_{\text{init}}}{S} \tag{2}$$

$$\text{Pruning Amount Update: } A_i = A_{\text{init}} + i \times A_{\text{steps}} \tag{3}$$

c) Sparsity calculation: After applying pruning, the sparsity of each module is calculated as Eq. (4) as the percentage of weights that have been set to zero.

$$\text{sparsity} = \left(\frac{\text{number of zero weights}}{\text{total number of weights}} \right) \times 100 \tag{4}$$

d) Overall pruning summary: Individual layers are then pruned, a summary is calculated overall, and parameters are aggregated across all layers, as given by Eq. (5).

$$\text{overall_sparsity} = \left(\frac{T - R}{T} \right) \times 100 \tag{5}$$

where, T = total parameters and R = remaining parameters

2) Structured pruning: Structured pruning functions differently from unstructured pruning because it removes entire neurons, channels, or filters, which results in hardware-friendly and deployment-ready implementation. The implementation of this technique becomes more feasible in practical applications because hardware compatibility surpasses unstructured pruning methods. The method shows outstanding success in minimizing accuracy losses while reducing the convolutional layers of VGG19.

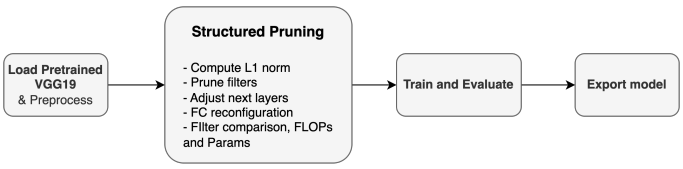


Fig. 4. Structured pruning training.

Fig 4 illustrates structured pruning on VGG19 pretraining. The system first calculates L1 norm values to locate insignificant filter elements for pruning while it readjusts subsequent layers, reorganizes the fully connected layers, measures FLOPs and parameter reduction, and then retrains and generates an exported model.

Structured pruning creates compatibility with hardware accelerators because it allows stable memory patterns that enhance inference speed significantly. The process cuts down the model size with computational complexity because it omits unnecessary channels and neurons to maintain both model performance levels and feature extraction capacity.

3) Mathematical formulation: Filter Importance Computation: For each convolutional filter, its importance is measured by the L1-norm of its weights, Eq. (6).

$$I_i = \sum_j |w_{ij}| \tag{6}$$

a) Pruning decision: For a convolutional layer with n_{filters} filters and a pruning ratio r, the number of filters to prune is given by Eq. (7).

$$n_{\text{prune}} = \lfloor r \times n_{\text{filters}} \rfloor \tag{7}$$

Filters are ranked by their L1-norm importance $I_i = \sum_j |w_{ij}|$, and the n_{prune} filters with the smallest I_i are removed.

b) Adjusting subsequent layers: The input channels for the next convolutional layer are modified according to the output channels that were previously trimmed, according to Eq. (8). This is done by re-indexing the weight tensor along the channel dimension:

$$\text{new input channels} = \text{previously kept indices} \tag{8}$$

B. Quantization

Quantization is a model compression technique that takes high-precision 32-bit floating-point weights and activations to low bit-value such as 8-bit or 4-bit integer. This heavily cuts down the memory usage and computational cost at the price of a little loss of accuracy in edge devices. So it is possible to guarantee good performance along with model functionality.

1) Precision-based quantization: Precision-based quantization, especially fixed-point quantization, is widely used to reduce the bit-width of weights and activations. When applied to the convolutional, activation, and fully connected layers of VGG19 [38]. The model implements 8-bit and 4-bit quantizations for all its convolutional layers in addition to ReLU activations and fully connected layers.

2) Quantization-aware training: QAT is a training procedure in which the quantization operation is applied during the forward operation of the model, i.e., the low-bit representation used for inference[39]. But the full precision of the parameters is saved in the backward pass, that is, the back-propagation, so that the model can learn and correct the quantization errors. This allows the model to be aware of quantization effects during the training phase, and it retains high performance and efficiency at the inference period. Mathematically, the quantized forward pass is represented as Eq. (9).

$$x_{\text{quant}} = q(x) \tag{9}$$

$q(x)$ is the quantization function. This way, inference-time quantization's accuracy is preserved, and QAT is one of the best methods for compressing deep learning models.

3) Module fusion: Efficient quantization, together with faster inference speed, needs module fusion to serve as a critical preparation step. The model gains advantages from reduced overhead, lower quantization error, and improved runtime efficiency through single fused modules retaining the same computed output resulting from combining convolution and ReLU layers.

4) Quantization and dequantization: These equations describe the dictionary defining the mapping of floating-point numbers to their integer representation: The quantized value was calculated with Eq. (10), and the dequantized float value was calculated with Eq. (11).

$$\text{Quantization: } q = \text{round}\left(\frac{x}{s}\right) + z \tag{10}$$

$$\text{Dequantization: } x = s \times (q - z) \tag{11}$$

5) 8-Bit QAT: The forward pass of quantized VGG19 models is composed of a sequence of quantized convolutions and fully connected layers, with in-between ReLUs and average pooling.

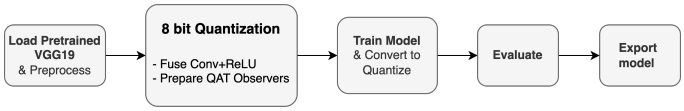


Fig. 5. 8-bit Quantization aware training.

Fig. 5 shows the pipeline for the quantization of 8-bit VGG19 models during preprocessing. The quantification method starts with convolutions and ReLUs before preparing QAT observers, then conducts training followed by quantized model conversion and evaluation and final export processing.

6) Kaiming normal initialization for weights: This method makes the weights have the appropriate value such that the variance of the activations remains roughly the same across layers. It is also good at signalling in deep nets using ReLU activations, helping to deal with issues with the vanishing or exploding gradients. The weights w are sampled from a normal distribution defined as Eq. (12).

$$w \sim \mathcal{N}\left(0, \frac{2}{n_{\text{in}}}\right) \tag{12}$$

7) 4-Bit QAT: A structure similar to the 8-bit configuration was followed, but all quantization functions and weight representations were replaced with their 4-bit counterparts.

The workflow for 4-bit QAT to apply on the VGG19 model appears in Fig. 6. The procedure starts with importing the model while performing the required input preparation.

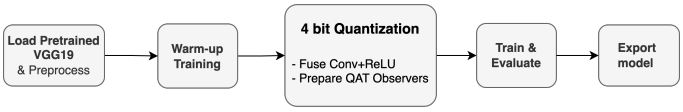


Fig. 6. 4-bit Quantization Aware Training

After weight stabilization, the short warm-up training begins. The implementation of QAT begins when 4-bit quantization combines convolution and ReLU layers and then adds QAT observers for all the layers. The model proceeds to fine-tuning through a training and evaluation stage, after which it becomes exportable for deployment.

The choice of pruning rates and quantization bits was done under the premise of preliminary testing that optimized the trade-offs. It was found that a pruning rate of 30% yields the best trade-off between compressing and maintaining the accuracy gained from the preliminary experiments. Training the hyperparameters was performed based on known model compression methods. In the process, it was found that cosine annealing learning rate is an absolute requirement for achieving stable QAT convergence.

8) Cosine annealing learning rate equation: Eq. (13) gives the dynamical learning rate during the period of training epochs by implementing a cosine that provides a progressive decay within the process of learning.

$$n_t = n_{\min} + \frac{1}{2}(n_{\max} - n_{\min})\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_{\max}} \times \pi\right)\right) \tag{13}$$

C. Combined Metrics for Performance Evaluation

1) Model size: Eq. (14) was used in calculating the model size.

$$\text{Model Size (MB)} = \sum_{p \in \text{parameters}} \frac{\text{numel}(p) \times \text{element_size}(p)}{1024^2} \tag{14}$$

2) Inference Time (T): Eq.(15) defines the inference time.

$$T_{\text{batch,avg}} = \frac{\sum_{i=1}^{N} T_{\text{batch}}^{(i)}}{N} \tag{15}$$

3) Parameter Reduction (R): The percentage decrease in model size due to compression [see Eq. (16)]:

$$R = \left(1 - \frac{S_{\text{compressed}}}{S_{\text{original}}}\right) \times 100 \tag{16}$$

4) Incremental memory usage: Eq. (17) defines incremental memory usage.

$$\Delta M = M_{\text{current}} - M_{\text{baseline}} \tag{17}$$

5) FLOPs counting: Eq. (18) defines FLOPs counting.

$$\text{FLOPs} = 2 \times \sum_{p \in \text{parameters}} \prod_{i=1}^{\text{ndim}(p)} \dim(p) \tag{18}$$

6) Early stopping: The early stopping mechanism continuously monitors the validation loss $L_{\text{val},t}$. If, for a given epoch t, the loss difference

$$L_{\text{val},t} - L_{\text{val},t+\Delta t} \tag{19}$$

Eq. (19) is less than min_delta for patience consecutive epochs, the algorithm concludes that further training is unlikely to yield significant improvements and stops the training process.

IV. Experimental Results

The evaluation in this part utilizes experimental tests on VGG19 optimizations to study structured and unstructured pruning together with quantization effects on performance outcomes. Three popular benchmark datasets—MNIST, CIFAR-10, and Oxford-IIIT Pet Dataset—are used for the evaluation to measure the actual performance of these optimizations. The tests are mainly conducted based on several important indicators, including model size, FLOPs, classification accuracy, inference latency, CPU utilization, and memory consumption.

In this study, the training and evaluations are run on Kaggle P100 GPUs to maximize computational efficiency. After the training phase, the developed models were exported and subsequently subjected to evaluation. The experimental benchmarking of the models was conducted using CPUs on Google Colab. The 4-bit model inference overhead was specifically measured on Google Colab CPUs during this evaluation phase. Although 4-bit quantization significantly reduces model size and memory consumption, the overhead arises because standard CPUs are optimized for 8-bit or higher precision operations. In the absence of native 4-bit hardware support, additional computational cost is incurred due to packing, unpacking, and emulating low-bit arithmetic during inference.

The optimization process of VGG19 baseline models starts through standard training interventions on each dataset. The model then undergoes both structured and unstructured pruning procedures, which enhance resource utilization. The efficiency of parameter utilization in the model has improved since unnecessary parameters are now being removed. Tests of both 4-bit and 8-bit quantization techniques seek to enhance the model compression efficiency.

The collected findings help to determine the impact of these optimization methods on system performance. Model efficiency faces an opposing force against inferential performance when implementing these optimization techniques. The findings from the research demonstrate valuable information regarding the realistic implementation prospects of the optimized VGG19 model.

1) Model efficiency summary: Examination of the advantages of pruning and quantization on the VGG19 architecture was conducted via inspection of performance metrics, including decreased parameters, increased FLOPs, and overall sparse network. Eq. (14) was used in calculating the model sizes of each configuration. The performance metrics of the baseline model and its structured and unstructured pruning implementations appear in Table I, Eq. (16), and Eq. (17).

TABLE I. Model Efficiency Comparison

Metric	Original	Structured	Unstructured	8-bit	4-bit
Parameters (M)	139.58	98.70	80.27	139.58	139.58
Pruning Rate (%)	–	30	30	–	–
FLOPs (G)	0.28	9.77	0.28	0.28	0.28
Sparsity (%)	0	0	42.50	1.59	48.10
Model Size (MB)	532.57	376.64	532.57	133.42	133.20
Size Reduction (%)	–	29.28	0	74.95	74.99

The VGG19 system requires 139.58 million model parameters and computation of 0.28 GFLOPs Eq. (18). A structured pruning technique with a 30% rate allows our model to maintain 98.70 million parameters while lowering FLOPs to 9.77 GFLOPs. In contrast, the unstructured pruning presents a higher pruning ratio of 42.50%; thus, the total parameter in the compressed model is reduced to 80.27 million. Overall sparsification of the pruned network is 42.50%, greatly reducing computational complexity. For 8-bit QAT and 4-bit QAT, the observed model size reduction is almost 75% [38].

For training efficiency, the pruned model was trained over 30-45 epochs with early stopping with a batch size of 32. The optimization method decreases algorithm processing complexity without affecting accuracy performance[40].

2) Best weight distribution after model quantization: These experiments show that this approach effectively can compress the model, reduce the number of FLOPs, and keep a balance between efficiency and performance, potentially demonstrating the most effective technique to compress deep neural networks with a slight loss in accuracy.

3) Best layer comparison after model compression: In summary, the results from Fig. 14 and Fig. 13 show that, although it offers compression benefits, structured pruning keeps accuracy. Excessive quantization methods, like using 4-bit, need exact calibration precision when working with challenging dataset materials.

4) Inference performance and norm-based metrics: Table III shows the effect of the comparison depending on inference time [Eq. (15)], CPU consumption, memory consumption, norm-based weight measures (L1, L2), and sparsity.

The best way for real-time inference with a minimal loss of model goodness was 8-bit quantization, which reliably delivered the fastest inference time and less memory usage across all datasets. 4-bit quantization provided the highest sparsity (up to 48%) and considerable model compression along with competitive inference speed with overhead.

The structured pruning method decreased inference time and memory consumption, though it did not introduce weight sparsity because it primarily modified network architecture instead of parameter numbers.

TABLE II. CLASSIFICATION PERFORMANCE

Dataset	Approach	Epochs	Accuracy	Top-1 Acc.	Top-5 Acc.
MNIST	Baseline Model	13	98.00%	98.00%	100.00%
	Structured Pruning	30	97.60%	97.60%	100.00%
	Unstructured Pruning	16	98.10%	98.10%	100.00%
	8-bit Quantization	17	97.50%	97.50%	100.00%
	4-bit Quantization	9	98.30%	98.30%	100.00%
CIFAR-10	Baseline Model	11	86.30%	86.30%	99.80%
	Structured Pruning	30	71.80%	71.80%	98.10%
	Unstructured Pruning	19	86.50%	86.50%	99.70%
	8-bit Quantization	21	85.70%	85.70%	99.60%
	4-bit Quantization	11	71.80%	71.80%	84.80%
Oxford-IIIT Pet	Baseline Model	9	96.74%	96.74%	–
	Structured Pruning	18	94.02%	94.02%	–
	Unstructured Pruning	15	97.96%	97.96%	–
	8-bit Quantization	13	97.42%	97.42%	–
	4-bit Quantization	10	89.67%	89.67%	–

Note: Top-5 accuracy was not computed for the Oxford-IIIT Pet dataset due to its binary classification nature.

TABLE III. INFERENCE PERFORMANCE AND NORM-BASED METRICS

Dataset	Approach	Inference Time (ms)	CPU Usage (%)	Peak Memory (MB)	L1 Norm	L2 Norm	Sparsity
MNIST	Baseline Model	849.286	98.530	41.762	772058	100.57	0%
	Structured Pruning	483.952	99.700	35.641	536877	83.23	0%
	Unstructured Pruning	829.155	99.547	24.812	665095	98.75	7.93%
	8-bit Quantization	472.216	99.827	0	771630	100.09	1.59%
	4-bit Quantization	489.674	99.352	0	622668	96.62	48.10%
CIFAR-10	Baseline Model	842.537	99.675	30.719	772053	100.57	0%
	Structured Pruning	473.458	99.819	23.977	536882	83.23	0%
	Unstructured Pruning	837.766	99.562	30.699	666228	98.75	0.12%
	8-bit Quantization	480.096	99.626	0	771622	100.09	1.59%
	4-bit Quantization	485.351	99.577	0	616684	95.94	48.42%
Oxford-IIIT Pet	Baseline Model	837.163	99.610	30.594	771776	100.55	0%
	Structured Pruning	475.043	99.717	36.094	536555	83.20	0%
	Unstructured Pruning	836.408	99.533	36.816	664427	98.67	0.66%
	8-bit Quantization	471.534	99.866	0.258	771024	100.01	1.59%
	4-bit Quantization	488.039	99.521	0	648955	100.61	48.11%

The results from unstructured pruning methods were inconsistent because memory consumption improved in certain use cases, yet inference speed boosts, along with sparsity degrees remained dependent on the dataset and generally low. Considering smaller and less redundant models, all optimization methods lowered L1 and L2 norms from the baseline.

V. DISCUSSION

This research demonstrates that both pruning and quantization methods are effective in improving the VGG19 model efficiency for deep learning functions. The experimental figures make visible the different trade-offs and benefits produced by each optimization method implemented in this study.

Fig. 7 demonstrates. Both quantization techniques reduced the model size extensively, although they left the actual parameter count unchanged because quantization modifies computational bit representation instead of parameter number. Structured pruning maintained a higher parameter count at

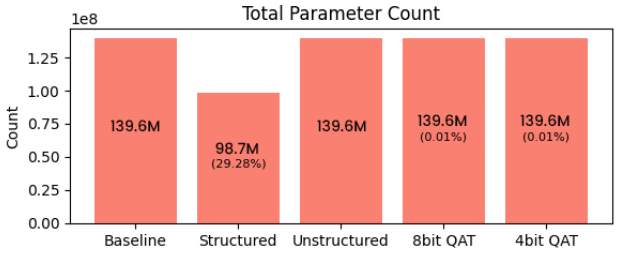


Fig. 7. Total parameter count of VGG19 variants.

98.7 million, but the parameter total in unstructured pruning dropped to 80.27 million.

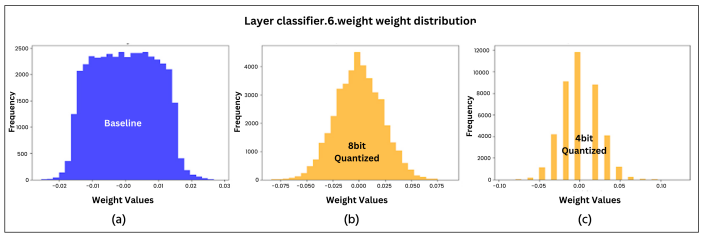


Fig. 8. Best QAT weight distribution.

The effectiveness of QAT was depicted in Fig. 8 by its weight distribution. The baseline distribution is plotted as shown in Fig. 8(a), and the 8-bit quantization, which is plotted in Fig. 8(b), still has a relatively smooth distribution. Fig. 8(c), however, shows that the use of 4-bit quantization leads to the doubling of weights around the zero value, which means that the model weights itself to the larger quantization error and reduced precision.

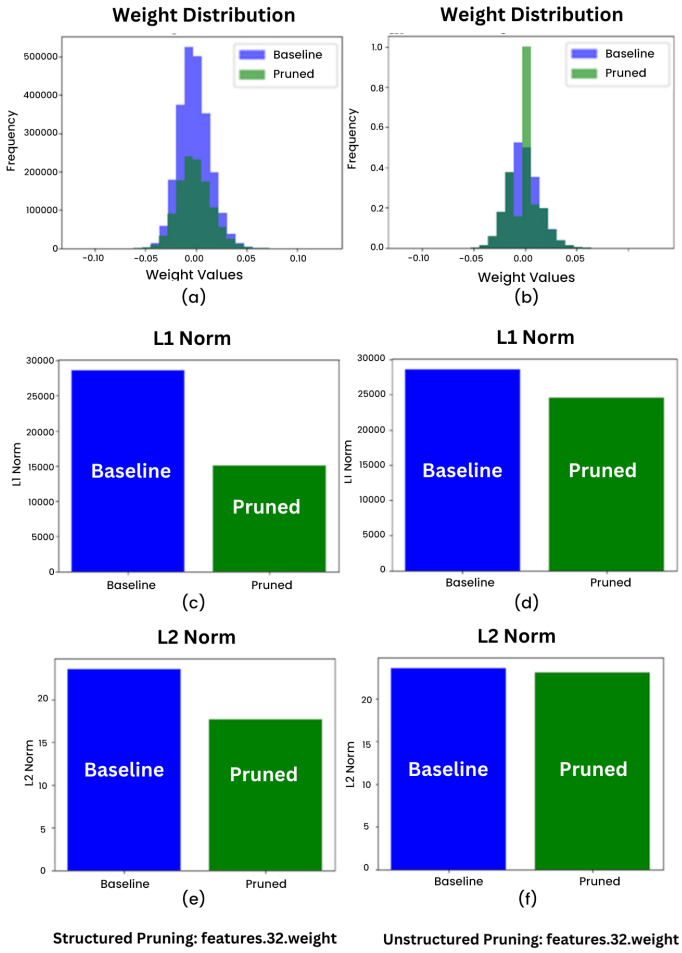


Fig. 9. Best pruned layer compression.

Fig. 9 displays particular impacts of pruning, since this model compression comparison acts at the level of a layer. Structured pruning [Fig. 9(a),Fig. 9(c),Fig. 9(e)] preserved the functionality of convolutional layers by removing the most redundant parts of deeper fully connected layers. In particular, Fig. 9(a) is the distribution of weight before and after pruning, Fig. 9(e) and Fig. 9(c) show the decreasing L2 and L1 norms, which means sparsity of parameters. The effect of unstructured pruning is shown in Fig. 9(b),Fig. 9(d) and Fig. 9(f), with the more focused post-pruning weight distribution shown in Fig. 9(b) and the L2 and L1 norm comparison seen in Fig. 9(d) and Fig. 9(e), respectively. The pruning was used to maintain the feature extraction capability of the model and to trim its structure.

As it is presented in Table II, five different architectures of VGG19 were assessed on three different datasets: MNIST, CIFAR-10, and Oxford-IIIT Pet Dataset. The models will be denoted as Baseline, Structured Pruning, Unstructured Pruning, 8-bit Quantization, and 4-bit Quantization. All tested models in the MNIST evaluation exceeded 97.5% accuracy levels. With a 4-bit quantized model, the accuracy level reached 98.30% while exceeding baseline standards at 98.00%. The 9-epoch training duration marked the minimum needed for this model. Unstructured pruning added slightly to the baseline with the accuracy of 98.10%, indicating sparsity-based pruning also works for lesser complexity of datasets. Despite being marginally reduced (97.60% and 97.50% in structured trimming and 8-bit quantization, respectively), they maintained complete top 5 accuracy. Using unstructured pruning proved to be a slightly better option when it was applied to CIFAR-10 as multi-class (86.50% vs. 86.30%). This was because it could keep the accuracy high for the more complicated dataset. 8-bit quantization hit a good accuracy (85.70%) with just small losses. The performance of the models deteriorates extensively to 71.80% when applying structured pruning and 4-bit quantization techniques to the dataset. The Oxford-IIIT Pet dataset by binary classification proved the substantial performance improvement brought up through unstructured pruning, achieving 97.96%, better than the basic 96.74%. Both 8-bit quantization and structured pruning resulted in comparable results to the baseline, with performances of 97.42% and 94.02%, respectively. However, 4-bit quantization decreased the most.

1) Results: Training figures using MNIST, Oxford-IIIT Pet, and CIFAR-10 datasets appear in Fig. 10(a), Fig. 11(a), Fig. 12(a) to show training behavior and validation and corresponding training and validation accuracies are presented in Fig. 10(b), Fig. 11(b), Fig. 12(b) to show training and validation accuracies. The training curves show that unstructured pruning techniques, together with other pruning methods, brought stability to training operations while achieving accurate test results throughout all epochs. The results in Fig. 12 show that structured pruning and 4-bit QAT [41] struggle on the complex CIFAR-10 dataset because they produce slow convergence and an enlarged generalization gap during aggressive compression efforts.

Devices that operate at edge points need maximum performance throughout inference procedures to qualify for deployment. The optimization algorithms in Fig. 14 reveal lower norm values than basic model protocols for L1 and L2 norm metrics. Their functionality depends on parameter sets, which must be minimal and simple. The structured pruning approach

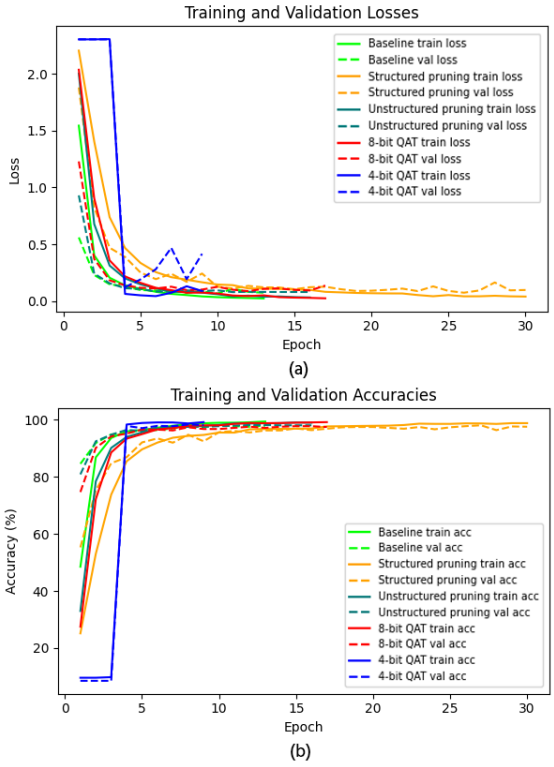


Fig. 10. MNIST training and validation.

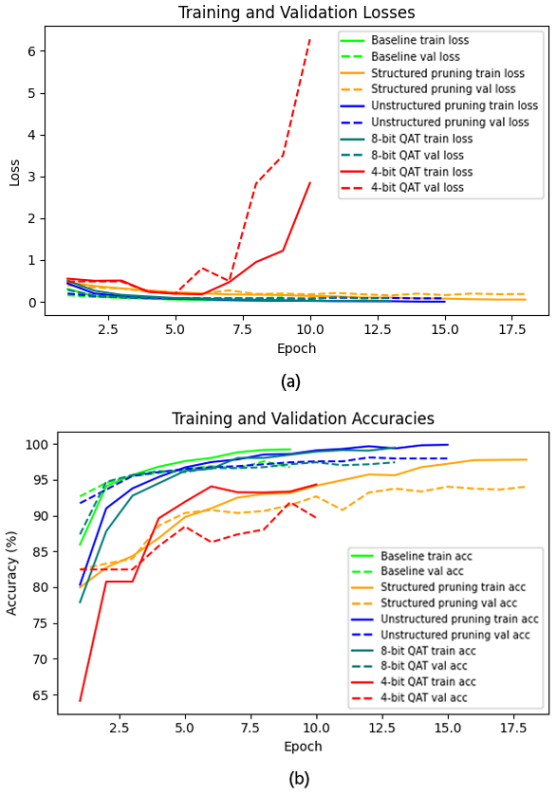


Fig. 11. The Oxford-IIIT pet training and validation.

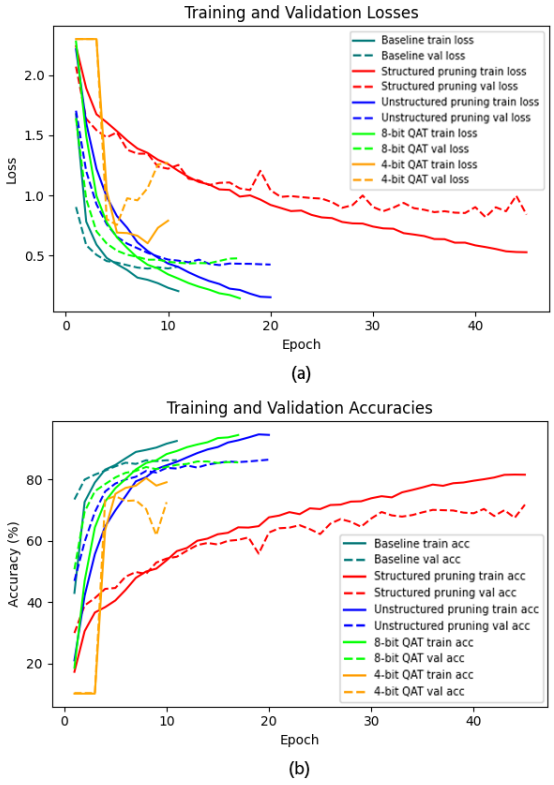


Fig. 12. CIFAR-10 training and validation.

effectively reduces both norms while preserving parameter density because network reduction outperforms weight removal in performance improvement.

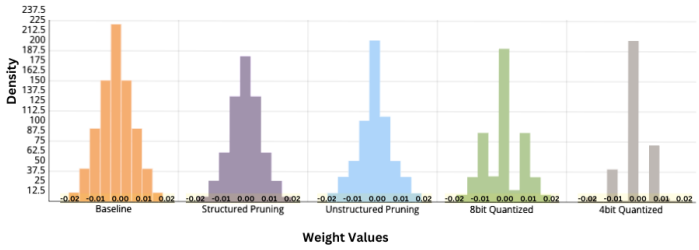


Fig. 13. Average weight distribution of VGG19 variants.

Fig. 13 shows the typical weight distribution statistics across all variants, which validates this observation. The precise and uniform weight distribution that occurs when pruning methods are structured leads to better hardware efficiency and more foreseeable inference behavior. The weight bins of the 4-bit quantized model became discrete from heavy quantization because of this pattern, which allowed it to achieve sparsity levels exceeding 48%, according to performance data.

The numerical data found in Table III proves the findings presented previously. The deployment times for 8-bit quantization kept its approximate duration at ≈472 milliseconds while simultaneously demonstrating zero memory consumption and minimal deviations in L1 and L2 norms when compared to the baseline model. 4-bit quantization outpaced 8-bit QAT

regarding inferential speed, even though it achieved peak sparsity levels together with minimum memory usage. This makes 8-bit QAT the most effective option for achieving performance-quality equilibrium.

Pruning methods that do not follow structured guidelines prove to be the most suitable techniques for preserving exact accuracy on complex datasets. The sparse retraining technique achieved superior results to baseline measurements across the Oxford-IIIT Pet data through its 97.96% accuracy, which surpassed the 96.74% baseline score, thus indicating that the method enhances generalization effectiveness. The model limitation was reduced through data augmentation and early stopping techniques, although no regularization methods such as dropout or L2 penalties were implemented. A minor degree of model bias may result from this implementation when applying it to complex datasets.

The use of four bits during quantization resulted in substantial accuracy deterioration when implementing it for CIFAR-10 and Oxford-IIIT datasets. Researchers need to develop complex quantization methods that choose the best quantization parameters when sensitivity levels change between different network parts.

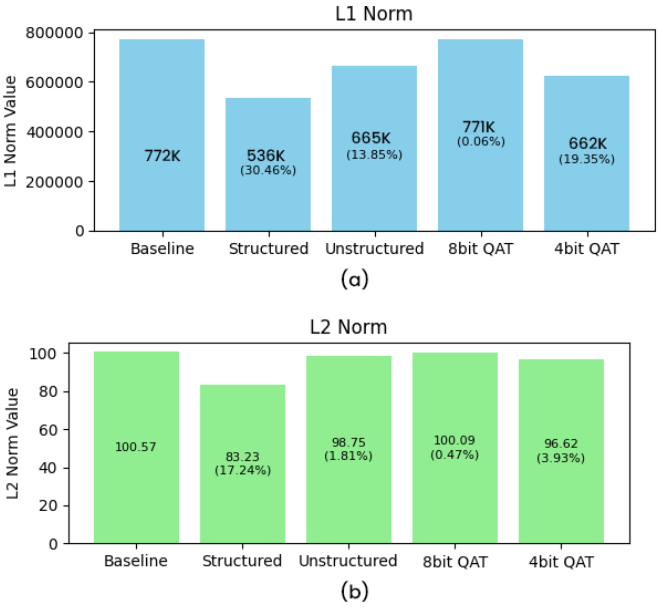


Fig. 14. L1, L2 norm of VGG19 variants.

The computational efficiency of quantized models is undeniable, according to Fig. 14 and Table III, thus making them optimal picks for mobile devices that require energy-efficient approaches, research shows.

A. Parameter Influence Analysis

According to our findings in the three datasets:

1) Pruning rate (30%): The one chosen when unsuccessful experimental tests indicated optimum accuracy-compression balance. Both the larger rates (40%) resulted in a significant loss in accuracy on the CIFAR-10, and the lower ones (20%) did not provide enough beneficial compression.

2) Quantization precision: Inference speed remained high on 8-bit with the same baseline accuracy as the more digitally prioritized architectures. The 4-bit had the best compression (75%), but it experienced brittle accuracy loss (10 to 15%) and overhead, as non-native compute code had to run on the T4 GPU.

3) Learning rate schedule: Cosine annealing was crucial to quantized models, and a fixed learning rate was not working well during QAT fine-tuning runs.

4) Batch size (32): Trade-off memory limitations and stability of training under all compression methods and data sets.

VI. Future Work

Future investigations should explore mixed-precision quantization methods, a scenario in which various layers of the network are quantized (i.e., encoded to reduce precision) at different bit depths in order to achieve the best out of precision and out of the computational budget. The search for hardware acceleration optimization methods, like Tensor Processing Units (TPUs) or Application-Specific Integrated Circuits (ASICs), might enable the adoption of quantum models, especially in resource-constrained environments. An additional and more promising region of research is the development of adaptive rephrase precision in real-time in accordance with the behavior of the input or application-defined constraints, thus providing robust and flexible models.

Moreover, combining structured pruning with QAT is an attractive research area that can achieve the highest computational efficiency with little loss of model accuracy. Benchmarking those optimization techniques against huge sets of various types of datasets, including realistic large-scale ones, would bring valuable knowledge to their generic acceptability. Finally, energy usage trend analysis among available quantized models in the deployment environment will be done to identify which models are energy efficient. This would be a major step in making greener artificial intelligence (AI) solutions without sacrificing AI performance and with less environmental impact of deep learning models.

VII. Conclusion

In this study, we optimized the model VGG19 by model compression techniques, structured pruning, unstructured pruning, and QAT 8-bit and 4-bit precision. We aimed at speeding up deep convolutional networks through limited resource platforms without compromising the accuracy of classification.

Experimental tests performed on the three benchmarks MNIST, CIFAR-10, and Oxford-IIIT Pet have shown that unstructured pruning and 8-bit QAT provided the best tradeoff between model sizes and predictive accuracy. The inference time and the model size were well-traded, and the same produced an acceptable accuracy by structured pruning. At the same time, the 4-bit QAT achieved the best compression rate (approximately 75%) and sparsity (approximately 48%), but degradation in accuracy occurred on more complex datasets, such as CIFAR-10 and Oxford-IIIT Pet.

These findings indicate that structured connectivity and lower-bit quantization-based pruning have the potential to greatly improve deploying VGG19 to edge devices and that

they can be further improved with the help of training strategies such as early stopping and cosine annealing. Nonetheless, the active quantization, such as 4-bit QAT, has to be tuned a little further to eliminate the inaccuracy, particularly with complex visual information.

Future work will look at the hybrid approaches to the compression of neural networks, especially at the intersection of mixed-precision quantization and structured pruning, towards pursuing an ideal balance between performance and intellectual accuracy of the inference. In addition, evaluating energy usage and optimization in comparison to edge-concentrated hardware acceleration will also be ideal in making these techniques practical in real-world low-power applications.

REFERENCES

[1] J. Ma, G. Wang, L. Zhang, and Q. Zhang, "Restoration and enhancement on low exposure raw images by joint demosaicing and denoising," *Neural Networks*, vol. 162, pp. 557–570, 2023.

[2] J. Sun, W. Yao, T. Jiang, D. Wang, and X. Chen, "Differential evolution based dual adversarial camouflage: Fooling human eyes and object detectors," *Neural Networks*, vol. 163, pp. 256–271, 2023.

[3] S. Mascarenhas and M. Agarwal, "A comparison between vgg16, vgg19, and resnet50 architecture frameworks for image classification," in *Proc. Int. Conf. Disrupt. Technol. Multidiscip. Res. Appl. (CENTCON)*, 2021, pp. 96–99.

[4] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher flops for faster neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 12 021–12 031.

[5] J.-H. Luo and J. Wu, "Autopruner: An end-to-end trainable filter pruning method for efficient deep model inference," *Pattern Recognition*, vol. 107, p. 107461, 2020.

[6] A. Kuzmin, M. Nagel, M. Van Baalen, A. Behboodi, and T. Blankevoort, "Pruning vs quantization: Which is better?" *Advances in neural information processing systems*, vol. 36, pp. 62 414–62 427, 2023.

[7] X. Qu, D. Aponte, C. Banbury, D. P. Robinson, T. Ding, K. Koishida, I. Zharkov, and T. Chen, "Automatic joint structured pruning and quantization for efficient neural network training and compression," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15 234–15 244.

[8] Y. Ding and D.-R. Chen, "Optimization based layer-wise pruning threshold method for accelerating convolutional neural networks," *Mathematics*, vol. 11, no. 15, p. 3311, 2023.

[9] Y. He and L. Xiao, "Structured pruning for deep convolutional neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12–34, 2023.

[10] T. Ganguli and E. K. P. Chong, "Activation-based pruning of neural networks," *Algorithms*, vol. 17, no. 1, p. 48, 2024.

[11] X. Liu, M. Ye, D. Zhou, and Q. Liu, "Post-training quantization with multiple points: Mixed precision without mixed precision," *arXiv preprint arXiv:2002.09049*, 2020. [Online]. Available: https://arxiv.org/abs/2002.09049

[12] X. Wu, E. Hanson, N. Wang, Q. Zheng, X. Yang, H. Yang, S. Li, F. Cheng, P. P. Pande, J. R. Doppa, K. Chakrabarty, and H. Li, "Block-wise mixed-precision quantization: Enabling high efficiency for practical reram-based dnn accelerators," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 43, no. 1, pp. 1–14, 2024.

[13] H. Zhang, L. Wang, and T. Chen, "An improved vgg19 transfer learning strip steel surface defect recognition deep neural network based on few samples and imbalanced datasets," *Applied Sciences*, vol. 11, no. 6, pp. 1–15, 2021.

[14] J. Kim, P. Luo, and F. Wu, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 512, pp. 23–45, 2021.

[15] R. Das, A. Kumar, and S. Mehta, "A transfer learning with structured filter pruning approach for improved breast cancer classification on point-of-care devices," *Computers in Biology and Medicine*, vol. 125, pp. 55–67, 2021.

[16] Y. Huang and B. Singh, "A compact parallel pruning scheme for deep learning model and its mobile instrument deployment," *Mathematics*, vol. 10, no. 12, pp. 1–17, 2022.

[17] F. Wang, J. Li, and T. Xu, "Automatic group-based structured pruning for deep convolutional networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 5, pp. 2345–2356, 2022.

[18] J. Brown and M. White, "Do all mobilenets quantize poorly? gaining insights into the effect of quantization on depthwise separable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 77–88.

[19] V. Kulkarni, A. Sharma, and P. Dubey, "Methodologies of compressing a stable performance convolutional neural network in image classification," *Neural Processing Letters*, vol. 51, no. 3, pp. 567–580, 2020.

[20] B. Shah and A. Mehra, "Pruning deep neural network models via minimax concave penalty regression," *Applied Sciences*, vol. 14, no. 9, pp. 1–18, 2024.

[21] H.-H. Chin, R.-S. Tsay, and H.-I. Wu, "A high-performance adaptive quantization approach for edge cnn applications," *arXiv preprint arXiv:2107.08382*, 2021.

[22] K. Ghosh, R. Patel, and D. Narang, "Training deep neural networks with joint quantization and pruning of features and weights," in *International Conference on Learning Representations (ICLR)*, 2022, pp. 1–12.

[23] I. A. Kandhro *et al.*, "Performance evaluation of e-vgg19 model: Enhancing real-time skin cancer detection and classification," *Heliyon*, vol. 10, p. e31488, 2024.

[24] I. Ahmad *et al.*, "Optimizing pretrained convolutional neural networks for tomato leaf disease detection," *Complexity*, vol. 2020, p. Article ID 8812019, 2020.

[25] B. K. Nyaupane, "Pneumonia of chest x-ray images detection using vgg architectures," *LEC Journal*, vol. 4, no. 1, pp. 43–48, 2022.

[26] R. Siddiqi and S. Javaid, "Deep learning for pneumonia detection in chest x-ray images: A comprehensive survey," *Journal of Imaging*, vol. 10, no. 8, p. 176, 2024.

[27] M. M. A. Monshi *et al.*, "Covidxraynet: Optimizing data augmentation and cnn hyperparameters for improved covid-19 detection from cxr," *Computers in Biology and Medicine*, vol. 133, p. 104375, 2021.

[28] H. Samma *et al.*, "Evolving pre-trained cnn using two-layers optimizer for road damage detection from drone images," *IEEE Access*, 2021, early access.

[29] S. D. Thepade *et al.*, "Face presentation attack identification optimization with adjusting convolution blocks in vgg networks," *Intelligent Systems with Applications*, vol. 16, p. 200107, 2022.

[30] M. Z. Iqbal, N. Garg, and S. B. Ahmed, "Table extraction with table data using vgg-19 deep learning model," *Sensors*, vol. 25, no. 1, p. 203, 2025.

[31] A. Stateczny, G. U. Kiran, G. Bindu, K. R. Chythanya, and K. A. Swamy, "Spiral search grasshopper features selection with vgg19-resnet50 for remote sensing object detection," *Remote Sensing*, vol. 14, no. 21, p. 5398, 2022.

[32] K. T. Ahmed *et al.*, "Maximum response deep learning using markov, retinal & primitive patch binding with googlenet & vgg-19 for large image retrieval," *IEEE Access*, 2021, early access.

[33] A. Faghihi, M. Fathollahi, and R. Rajabi, "Diagnosis of skin cancer using vgg16 and vgg19 based transfer learning models," arXiv preprint arXiv:2404.01160, 2024.

[34] Y. Kumar *et al.*, "Enhancing parasitic organism detection in microscopy images through deep learning and fine-tuned optimizer," *Scientific Reports*, vol. 14, no. 5753, 2024.

[35] M. Fu, H. Yu, J. Shao, J. Zhou, K. Zhu, and J. Wu, "Quantization without tears," arXiv preprint arXiv:2411.13918, 2024.

[36] F. E. Fernandes and G. G. Yen, "Pruning deep convolutional neural networks architectures with evolution strategy," *Information Sciences*, vol. 541, pp. 345–361, 2020.

[37] A. Bouguettaya and H. Zarzour, "Cnn-based hot-rolled steel strip surface defects classification," *The International Journal of Advanced Manufacturing Technology*, 2024.

[38] X. Li, Y. Chen, and J. Wang, "A mutual learning framework for pruned and quantized networks," *Journal of Computer Science & Technology*, vol. 38, no. 23, 2023.

[39] C. Gernigon, S.-I. Filip, O. Sentieys, C. Coggiola, and M. Bruno, "Adaqat: Adaptive bit-width quantization-aware training," in *2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS)*. IEEE, 2024, pp. 442–446.

[40] Q. Lu, "Tinyml computer vision using coarsely-quantized log-gradient input images," Stanford University, 2023. [Online]. Available: https://purl.stanford.edu/fb372vt6975

[41] P. P. F. Delgado, "Real-time implementation of 3d lidar point cloud semantic segmentation in an fpga," University of Minho, 2022.

Fake News Detection on Kashmir Issue Using Machine Learning Techniques

Misbah Kazmi[1], Sadia Nauman[2], Sadaf Abdul Rauf[3],
S. Ali[4], Ali Daud[5], Bader Alshemaimri[6]
Department of Computer Science, Fatima Jinnah Women University, Rawalpindi, Pakistan[1,3]
Department of Communication and Media Studies, Fatima Jinnah Women University, Rawalpindi, Pakistan[2]
Department of Computer Science, National University of Science and Technology, Islamabad, Pakistan[4]
Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates[5]
Software Engineering Department, King Saud University, Riyadh, Saudi Arabia[6]

***Abstract*—Focusing events are sudden, impactful occurrences that spark widespread discussions. Analyzing fake news during such events is challenging due to limited and short-lived datasets. Online fact checkers are slow in identifying fake news, and internet communities and forums become the primary source of news, allowing unchecked dissemination. This study proposes a machine learning approach to predict fake news during the revocation of Article 370 in Kashmir as a focusing event. Small dataset from 20th August till 2nd September is collected and user profile parameters are utilized for effective classification. Five classifiers were employed, with Random-Forest and Logistic-Regression achieving the highest F1 scores of 74 per cent. Results identifies prevalent words in true and false news tweets, aiding in fake news detection. This approach mitigates misinformation during events with limited data, contributing to a reliable online environment. The research is valuable for major geopolitical shifts, natural disasters, and social movements.**

***Keywords*—Classification algorithm; fake news; Kashmir issue; machine learning techniques**

I. Introduction

Focusing events are short lived events that capture public attention and direct it towards a specific aspect of a problem [1]. They can sway politics by mobilizing public support for one side of a debate, potentially influencing political outcomes [6]. Internet communities and forums have become a powerful and cost-effective platform for sharing news, leading to the dissemination of both real and fake information [32], [11]. Fake news detection involves predicting misleading articles intended to deceive readers [40], [62]. During Hurricane Sandy, a study found that 86 per cent of tweets sharing fake images were re-tweets, illustrating the alarming speed at which misinformation spreads [44].

Fake news aims to mislead readers by presenting false information, posing a high risk as it is driven by motives such as social popularity, political agendas, or manipulation of public opinion [43] [39] [50]. Detecting fake news on social media is an emerging research field of great importance to mitigate its adverse impact on the public and the news ecosystem [42].

However, existing research on fake news detection [59], [60], [26] has primarily focused on general-purpose datasets or misinformation in stable, long-term contexts (e.g., elections or health crises). Very limited studies explore detection during focusing events — short, high-impact political developments where data is sparse and rapidly evolving. This creates a significant gap in developing real-time fake news detection methods tailored for such scenarios [19]. This leaves an important gap in the development of real-time fake news detection techniques that are specific to these situations. Our research study tries to address this gap by applying conventional machine learning methods to short-term time series data on Twitter, focusing on the revocation of Article 370 in Kashmir.

In the field of natural language processing (NLP) and computational linguistics, Machine learning(ML) and Deep learning(DL) techniques are being used as powerful analysis tools for identification, classification and detection of language related issues [51] [41]. Our research focuses on Article 370 revocation by India in the state of Jammu and Kashmir as a significant event [29]. In August 2019, the Indian government abolished Article 370, bringing the region under direct central government control [8]. The event generated extensive discussion on Twitter, with a bombardment of visual and textual data [18]. For scenarios like focusing events, where limited data size is available. Traditional ML techniques are effective in classification problems, as DL models require a large amount of training data to produce quality results [40], [59]. Recently, these techniques have been explored as the starting point towards the automation of task selection mechanism of Crowd-Sourced software development (CSSD) platforms [45].

The research question in focus in this study is: What are the performance levels of traditional machine learning models in identifying fake news in high-impact short-duration political events when applied to event-specific social media data?

This study utilizes User profiles on the Twitter social network, for measuring fake news identification using explicit and implicit features [41] [53]. For news tweet dataset, data of twitter users responses and news sources on the revocation of Article 370 in Kashmir was collected. Motivation behind the data collection is identification of fake and real news spread across the platform in the backdrop of the event. Data collection period ranged from 20th August to 02 September, 2019 [52]. The credibility of collected tweet news was checked using Google News search engine, through valid sources and parameters, such as the presence of a "Verified" symbol on twitter account sharing the news [47], friends and following counts of the twitter account [16], re-tweet counts, tweet

Algorithm 1 Fake News Tweet Detection using Logistic Regression

for $Eachtweet = 1, 2, \ldots$ **do**
  for $tCount \leftarrow \sum(numberofwords)$ **do**
    Tokenize tweet by word
  end for
  Apply pre-processing techniques
  $Tweet \Leftarrow words - (stopword$ U $punctuationword)$
  Tweet = Tweet.wordNet Lemmatizer(words)
  for $EachTweet$ **do**
    Tweet classify.logisticRegression(TweetHyperParams)
    if $Tweet$ = True **then**
      $TweetLabel \Leftarrow real$
    else if $Tweet$ = False **then**
      $TweetLabel \Leftarrow fake$
    end if
  end for
end for

creation date, and user reviews about the given tweet [10] [36] [13].

This study introduces a system that automatically detects and evaluates fake news on Twitter, specifically focusing on the revocation of Article 370 in Kashmir. The system captures tweets from various sources and utilizes profile parameters to classify them as fake or real. It trains a model by extracting relevant features and employs selected classifiers. F1 score is used to evaluate model performance.

Our research primarily contributed in:

- Affirming the credibility of traditional ML techniques to analyse "focusing events". The framework identifies fake tweets from different sources and enhance the ability to discern the credibility of news.
- Classification of the most prevalent words in true and false news tweets, highlighting the linguistic patterns that can aid in detecting fake news.

Our study integrates various ML techniques to effectively detect and evaluate fake news on Twitter. We compared five different ML algorithms based on their F1 scores and accuracy. The study presents research outcomes, system limitations, and utilizes word clouds and statistics to classify prevalent words in news tweets, highlighting linguistic patterns for fake news detection, the process is shown in Algorithm 1 on the task of fake news detection.

The rest of the study is organized as follows. The background information about the Kashmir conflict, fake news, and machine learning methods are provided in Section II. The proposed methodology such as data collection, feature extraction and classification algorithms are explained Section III. The results are discussed in Section IV and compared in terms of classifier performance. The study ends with the limitations and future directions in Section V.

II. Background

The related work can be divided into three sub-sections; Conflict of Jammu and Kashmir and revocation of Article 370 of Indian constitution as its focusing event, Fake news and its widespread impact in the backdrop of such events and lastly ML techniques as effective approaches for mitigating the spread and impact of fake news.

A. Conflict of Jammu and Kashmir

Territory of Jammu and Kashmir has been a source of contention between Pakistan and India for over 70 years, leading to four wars in 1948, 1965, 1971, and 1999[56]. Additionally, various groups within Kashmir have their own agenda, which span from seeking affirmative discrimination to advocating for a separate political status or even outright secession [38]. This intricate mix of factors contributes to the intricate nature of the Kashmir conflict.

In August 2019, the Indian government made a significant move by revoking Article 370 of the Indian Constitution. This decision eliminated region's separate constitution and flag, bringing it under the direct control of the central government of India. The decision received both support and opposition, sparking debates about its impact on the region's identity, security, and relations with the rest of India [8]. The event triggered a widespread response across the world and on social media, particularly on Twitter, with an influx of visual and textual data [18].

B. Fake News and its Impact

Cognitive theories [15], [20] suggest that due to our inherent gullibility, people are prone to believing information that aligns with their pre-existing views (confirmation bias). Consequently, individuals may readily share such information without proper verification, further distorting facts that do not align with their beliefs. In the past, foreign actors primarily used fake news for political manipulation. However, it is now increasingly being leveraged across various domains [2] [25] to influence public opinion. During COVID-19 pandemic, misleading and fake information spread all across social media which resulted in false claims and a surge in pseudo scientific health treatments related to the virus and potential remedies [7]. The landscape of fake news has undergone a significant transformation in recent years[55], [35]. Traditionally, the dissemination of misinformation was primarily through conventional media channels. However, the rise of online news platforms and social media has drastically altered the ecosystem of fake news [4] [33] [27]. Hence, the detection of fake news has garnered significant attention in recent years.

C. Machine Learning Techniques

Previously, ML models are being used for prediction problems [30] [37] to eventually help organizations with decision making and identifying future trends. [10] proposed a system utilizing 45 features to predict news accuracy, capturing tweet author specifications, tweet stream properties, contextual aspects, and trends. In [22], the authors proposed a modest approach to fake news detection using the naive Bayes classifier, incorporating user accounts, reverse image searching, and cross-verification of sources. In [12], the authors discussed DL algorithms such as RNN and LSTMs. In Twitter account and bot detection, [14] introduced a system using Support Vector Machine Neural Network (SVM-NN) and other ML

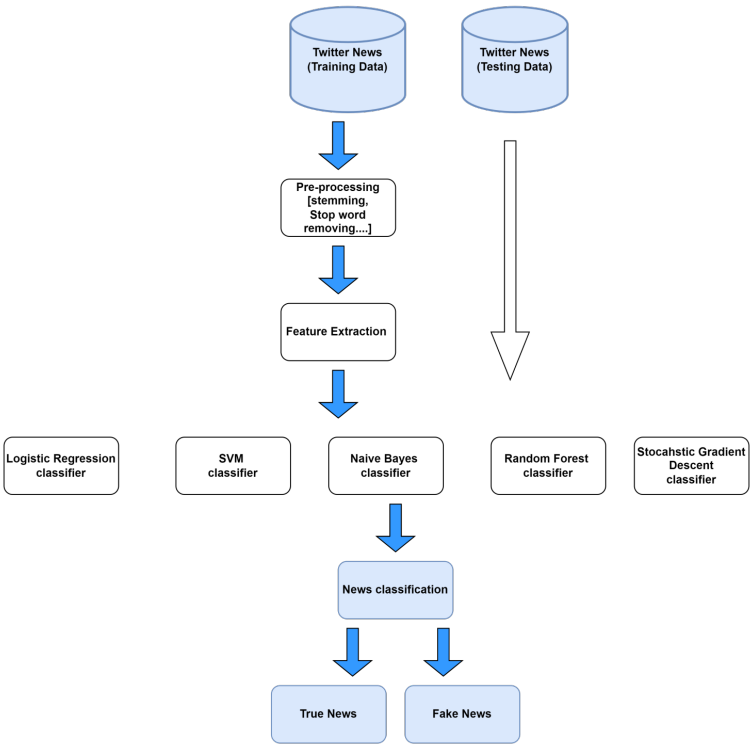


Fig. 1. Machine learning based process of fake news classification.

algorithms. In [17], [46], and [21], the authors explored similar techniques for detecting fake accounts. In [41], the authors measured trust levels of users to identify fake news, employing explicit profile features. In [49], the authors proposes an improved decision tree by changing threshold for their bi-criteria optimization. These improved decision trees are useful for fake news detection.Recently, [45] [54] proposes multi-modal fake news detection using ensemble learning. Modal exploits publicly available dataset and uses NLP techniques for textual information pre-proccessing and gauging the sentiment from each news.

III. Proposed System

An initial set of 6570 general tweets on the revocation of article 370 was collected. During the pre-processing step, tweet redundancy was removed and relevant tweet news content was kept. Graphical data was excluded from dataset due to the challenges in determining the authenticity of images and the complexity involved in assessing hidden objects within them [51], [24]. As a result, 402 news tweets are obtained for model training. Tweet news dataset was filtered to include English news content and remove any unwanted columns and empty rows. Before training, some hyper parameters are fixed for classification and filtered dataset was distributed into train and test data. Table I provides initial distribution of our real and fake news tweets dataset. Fig. 1 shows a high-level overview of the machine learning-based process used to classify fake news.

A. Data Pre-Processing

Fig. 1 illustrates the classification process of the proposed system. After tweet news dataset collection, data pre-processing techniques are applied to remove URLs, extra spaces, hashtags, and other intricacies. The tweets are further refined using stop word and punctuation removal, sentence segmentation, lower casing, and tokenization. These techniques removed all data redundancies and transformed tweets into a unified form for ML algorithms [23]. Additionally, N-gram features were extracted and using TF (Term Frequency) and TF-IDF (Term Frequency-Inverted Document Frequency) a feature matrix is formed. At the end, results of five different ML algorithms - Stochastic Gradient Descent (SGD), Logistic Regression (LR), Random Forest (RF) classifiers, Linear SVM and Naive-Bayes (NB) were compared based on F1 score and accuracy using the Python Natural Language Toolkit (NLTK).

Main steps of data pre-processing techniques being used are:

1) Splitting: This step involves breaking a text document into individual words or tokens [9].

2) Stop words removing: Stop words are used words with no significant meaning in given context. Removing stop words reduces noise and improve the efficiency of text analysis [17].

3) Stemming: Stemming reduce the words to the word base form called “Stem”. For classification method speed and accuracy, we used the commonly used Porter stemmer [28].

4) N-gram model: N-gram model represents a sequence of two words or 2-grams. N-grams store spatial information and serve as an ordered document representation [10]. In

proposed methodology, Bag of words and N-grams are used to represent the documents and do the classification [48]. An N-gram classifier is built and used to differentiate between fake and real news tweets.

5) Features extraction method: However, high dimensionality is a challenge for text categorization since documents which often contain large number of words and phrases. To address this issue, two primary feature selection methods are investigated: Term Frequency (TF) and Term Frequency-Inverted Document Frequency (TF-IDF) [31]. They can mitigate the impact of redundant and irrelevant features on classifier accuracy and performance [34].

B. Prediction Algorithm

Prediction model for fake tweet news identification is built by implementation of five ML classifiers from scratch. These algorithms include LR model, NB classifier model, SVM model, RF model, and Stochastic Gradient Descent. Classification and prediction of fake news was checked across each classifier. To achieve consistency and reproducibility of experiments, we applied standardized hyperparameter configurations to all machine learning models, which are the default parameters in scikit-learn.

C. Tweet Analysis

To understand the distribution of real and fake news feature over the dataset, the study on each feature has been made using tools and packages such as seaborn, wordcloud and NLTK. Fig. 2 shows the total distribution of extracted tweets from 20th August to 30th August, 2019. A higher frequency trend of tweets is visible between 23rd August and 30th August as well as the frequency of true and false news tweets during this time period are visualized in Fig. 3. Table I illustrates the distribution of followers and re-tweet counts of the twitter user and created tweet respectively in true and false news.

TABLE I. FAKE AND REAL TWEETS DISTRIBUTION

Label	Count
True	175
False	125

It is evident from the chart that false news tends to have more followers compared to true news, indicating a higher level of dissemination for false information. Overall, these analysis help to characterize (and patterns) real and fake news in the included tweet news dataset. The analysis also represent a baseline for sentiment prediction of twitter users when a focusing event occurs [3].

D. Most Repeated Words in Extracted Tweets

Using word clouds, display the most frequent words related to the Kashmir conflict. Additionally, separate word clouds are created for true and false news categories in training data, as depicted in Fig. 4, respectively.

E. Tweet with more Re-tweets and Favourite Counts

Within the corpus, tweet with most number of favourite counts and re-tweets was selected and visualised using numpy max function as shown in Table II.

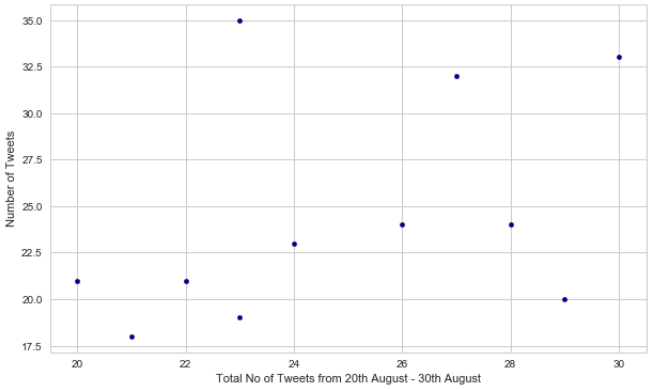


Fig. 2. Total number of tweets from 20th to 30th August.

TABLE II. TWEET HAVING MORE FAVOURITE COUNTS AND RE-TWEETS

The tweet with more likes is:
"Foreign Journals need permit to work in Kashmir. So all the hostile anti-india coverage in Western liberal media has been feed to them by their minions in India". Number of likes: 168
The tweet with more re-tweets is:
"Pakistani Envoy to U.S says world should take notice to avoid new conflict with India over Kashmir". Number of re-tweets: 488

F. Most Engagement of Twitter Users in a Given Time

Level of user engagement is identified using python 2D plotting library Matplotlib to plot a time series of tweet between given date and number of likes and re-tweets within news dataset. As per Fig. 5 , the most favourite tweet counted is between 23rd and 30th August.

G. Sources

To identify the geolocation of the tweeter users creating the tweet, different sources of extracted tweets are plotted in Fig. 6. Pakistan with the ratio of 24.52 is the The highest generator of relevant tweets with India being second highest generator of relevant tweets i.e, 23.56. Fig. 7 provides break down of true and false tweets generated from these location.

H. Results

We conducted experimentation and analysis on a dataset of 6,570 tweets relating to the revocation of Article 370 of Indian constitution by India in disputed territory of Jammu and Kashmir. After data cleaning, removing redundancy and non English tweets, five classifier models were trained on a refined dataset of 402 tweets. The overall efficiency of our classifiers is presented in Table III. We compared the F1 scores achieved by our classifiers and obtained the highest score of 74 per cent by RF classifier and LR classifier. In terms of accuracy, the RF classifier and LR classifier also achieved the highest accuracy of 82 per cent.

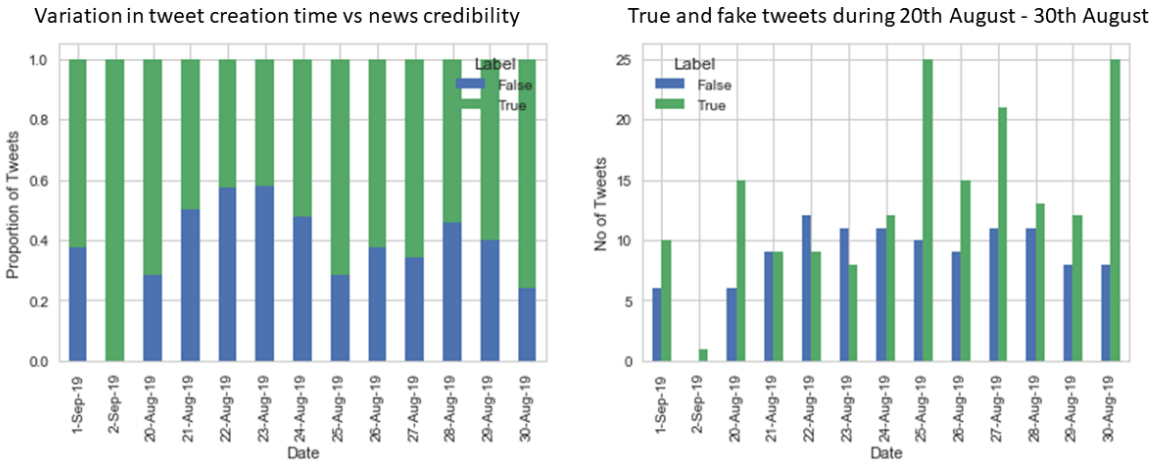


Fig. 3. Analyzing temporal variations in twitter time creation: News credibility and spread of True versus Fake tweets (Aug 20 to Aug 30).

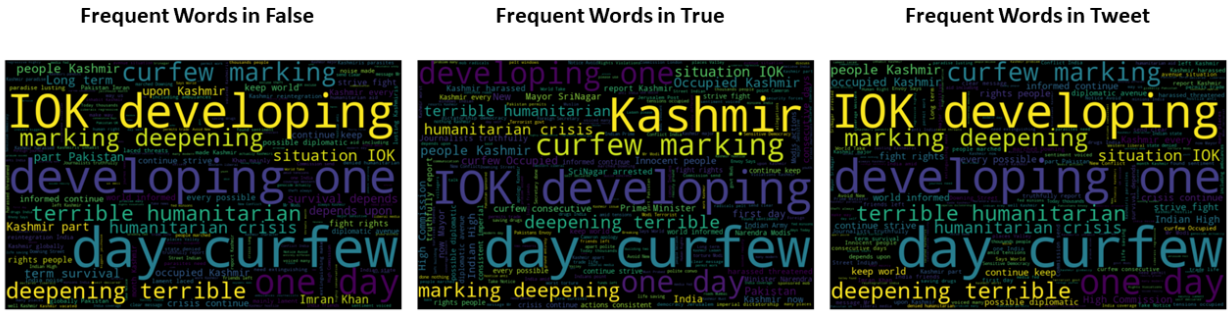


Fig. 4. Comparison of frequent words in fake and real tweets.

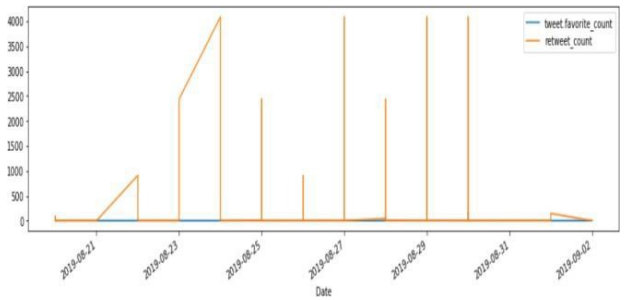


Fig. 5. Most engagement of Twitter user within the selected time span.

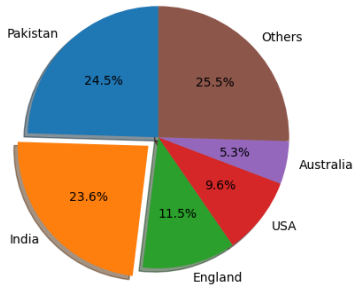


Fig. 6. Tweets percentage based on active Twitter user location.

IV. Discussion

The F1 scores achieved by our best models is 74 per cent which can be attributed primarily to the limited size of our training dataset. Since the tweets on Kashmir Conflict span over few weeks, obtaining a large dataset was not possible, and it was not sufficient to train a DL model. However, our traditional ML models have demonstrated promising results.

Furthermore, our dataset specifically focused on textual

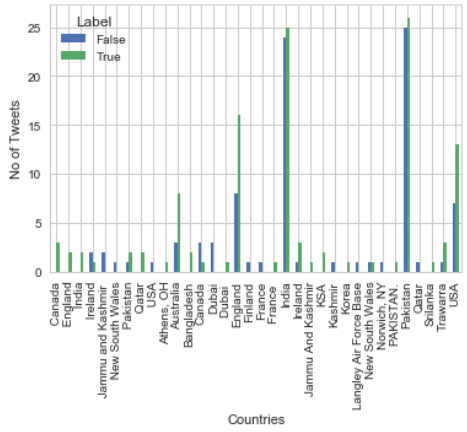


Fig. 7. Comparison of true and fake no. of tweet based on country.

TABLE III. COMPARISON OF DIFFERENT MACHINE LEARNING MODEL BASED ON PERFORMANCE METRICS ON FAKE NEWS DETECTION TASK

Model	F1-Score	Precision	Recall	Accuracy
Random forest	0.74	0.68	0.82	0.82
Naïve Bayes	0.71	0.67	0.76	0.62
SVM	0.73	0.67	0.79	0.79
Logistic regression	0.74	0.68	0.82	0.82
stochastic gradient decent	0.64	0.64	0.63	0.69

data. It is worth noting that existing online fact-checkers, primarily concentrate on fact-checking statements pertaining to American politics [5]. Hence, verifying credibility of real-time news content presented a challenge. Consequently, the task of maintaining an available dataset for our research proved to be challenging, resulting in a decrease in performance when using the "Kashmir Dataset" with a relatively smaller number of instances. Despite these limitations, our research demonstrates the potential of ML techniques in detecting fake news, especially given the constraints of the dataset. We believe that with larger and diverse datasets, our models could achieve higher accuracy and F1 scores.

When data is small, traditional ML has several advantages over deep learning approaches, which typically require large annotated datasets and extensive computational resources. First, it performs well with an F1-score of 74% and accuracy of 82% on a small, event-specific Twitter dataset that represents focusing events. Secondly, our system achieves interpretability and robustness, which cannot be obtained by many end-to-end neural models given that we combine textual features with user profile attributes (e.g., follower ratio, re-tweet count, and verification status). These advantages are reflected in recent comparative studies [61], that compared classical ML algorithms (RF, LR, etc.) to transformer-based models on several real-world datasets and concluded that the traditional ones are still competitive. Likewise, the survey [58] of ML and DL architectures found that Logistic Regression exceeded transformer models in limited environments. Our results confirm these findings because it has been proved that traditional ML is frequently more feasible, more resource-friendly, and interpretable in low-resource, high-impact settings.

V. CONCLUSION

We have developed a system that uses traditional ML techniques to automatically detect and assess the credibility of fake news on Twitter using a comparatively smaller dataset. Our system achieved the highest accuracy of 82 per cent, with RF and LR classifiers. Our study is a contribution to the combating of fake news because we offer a strong system of identification and evaluation of news credibility that gives impressive outcomes using a small dataset. Our experimental results affirm the credibility of traditional ML techniques to analyse "focusing events". Moreover, classification of the most prevalent words in true and false news tweets is useful in highlighting the linguistic patterns that can aid in detecting fake news.

These results indicate the success of our method in real-time, event-specific detection of misinformation, particularly, in the situation when data is scarce. The accuracy and interpretability were achieved by the combination of user profile features and linguistic patterns. Our evaluation does not only provide powerful predictive performance but also adds practical knowledge of how to design lightweight detection systems that can be applicable to political crises and other rapidly changing environments. The framework can be incorporated into early warning systems by journalists, fact-checking organizations, or policy response teams to track misinformation in real-time during political or crisis events. Theoretically, it helps to fill the gap in the under-researched field of fake news detection in focusing events, proving that even in the conditions of limited data, the standard ML models can be effective.

Although the results are promising, this research has limitations. The dataset was small and narrow, only concerned with English tweets during a certain time frame. Such limitations can influence the applicability of the model to more extensive or multilingual fake news detection procedures. Future work will focus on expanding the dataset, incorporating additional feature selection methods, and diversifying language coverage. Additionally, employing different DL techniques, such as fine-tuning a pretrained model can potentially improve the results.

REFERENCES

[1] Seffen, B. and Patt, A., 'A historical turning point? Early evidence on how the Russia-Ukraine war changes public support for clean energy policies." Energy Research & Social Science. vol. 91, pp. 102758, 2022.

[2] Alper Darendeli ,Aixin Sun ,Wee Peng Tay . The geography of corporate fake news. Plos One, 2024.

[3] Li, Hui Ru and Hu, Yan Rong and Liu, Hong Jiu. A novel LASSO-ATT-LSTM model of stock price prediction based on multi-source heterogeneous data. Journal of Intelligent & Fuzzy Systems 1–11, 2023.

[4] Simko, Jakub and Racsko, Patrik and Tomlein, Matus and Hanakova, Martina and Moro, Robert and Bielikova, Maria. (2021). A study of fake news reading and annotating in social media context. New Review of Hypermedia and Multimedia (pp. 97–127), 2021.

[5] Zhou, Xinyi and Zafarani, Reza. A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR) (pp. 1–40), 2020.

[6] Sabatier, Paul A. An advocacy coalition framework of policy change and the role of policy-oriented learning therein, 129–168, 1988.

[7] Naeem, Salman Bin and Bhatti, Rubina and Khan, Aqsa. An exploration of how fake news is taking over social media and putting public health at risk. Health Information & Libraries Journal (pp. 143–149), 2021.

[8] Ali, Rizwan and Mustafa, Usman. Kashmir dispute: Emerging complexities after abrogation of Article 370. Journal of Humanities, Social and Management Sciences (JHSMS) (pp. 13–25), 2021.

[9] Conroy, Nadia K and Rubin, Victoria L and Chen, Yimin. Automatic deception detection: Methods for finding fake news. Proceedings of the association for information science and technology (pp. 1–4), 2015.

[10] Buntain, Cody and Golbeck, Jennifer. Automatically identifying fake news in popular twitter threads. 017 IEEE international conference on smart cloud (smartCloud), 208–215, 2017.

[11] Torabi Asr, Fatemeh and Taboada, Maite. Big Data and quality data for fake news and misinformation detection. Big Data and Society. (pp. 127–132). SAGE Publications Sage UK: London, England, 2019.

[12] Girgis, Sherry and Amer, Eslam and Gadallah, Mahmoud. Deep learning algorithms for detecting fake news in online text. In 2018 13th international conference on computer engineering and systems (ICCES) (pp. 93–97), 2018.

[13] Xiao, Cao and Freeman, David Mandell and Hwa, Theodore. Detecting clusters of fake accounts in online social networks. In Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security (pp.91–101), 2014.

[14] Khaled, Sarah and El-Tazi, Neamat and Mokhtar, Hoda MO. Detecting fake accounts on social media. 2018 IEEE international conference on big data (big data), (3672-3681), 2018.

[15] Shu, Kai and Dumais, Susan and Awadallah, Ahmed Hassan and Liu, Huan. Detecting fake news with weak social supervision. IEEE Intelligent Systems, 96–103, 2020.

[16] Benevenuto, Fabricio and Magno, Gabriel and Rodrigues, Tiago and Almeida, Virgilio. Detecting spammers on twitter. Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), 2010.

[17] Ahmed, Hadeer and Traore, Issa and Saad, Sherif. Detection of online fake news using n-gram analysis and machine learning techniques. Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1, 2017.

[18] Gabel, Sabrina and Reichert, Lilian and Reuter, Christian. Discussing conflict in social media: The use of Twitter in the Jammu and Kashmir conflict. Media, War and Conflict (pp. 504–529), 2022.

[19] Kumar, Srijan and West, Robert and Leskovec, Jure. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In Proceedings of the 25th international conference on World Wide Web, 2016.

[20] Jain, Akshay and Kasbe, Amey. Fake news detection. 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2018.

[21] Manzoor, Syed Ishfaq and Singla, Jimmy. Fake news detection using machine learning approaches: A systematic review. 2019 3rd international conference on trends in electronics and informatics (ICOEI), 230–234, 2019.

[22] Granik, Mykhailo and Mesyura, Volodymyr. Fake news detection using naive Bayes classifier. In 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON), 2017.

[23] Dey, Amitabha and Rafi, Rafsan Zani and Parash, Shahriar Hasan and Arko, Sauvik Kundu and Chakrabarty, Amitabha. Fake news pattern recognition using linguistic analysis. In 2018 Joint 7th International Conference on Informatics, Electronics and Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR) (pp. 305–309), 2018.

[24] Di Domenico, Giandomenico and Sit, Jason and Ishizaka, Alessio and Nunan, Daniel. Fake news, social media and marketing: A systematic review. Journal of Business Research (pp. 329–341), 2021.

[25] Juan R. Martinez-Rico ,Lourdes Araujo,Juan Martinez-Romo. Building a framework for fake news detection in the health domain. Plos one, 2024.

[26] Shu, Kai and Mahudeswaran, Deepak and Wang, Suhang and Lee, Dongwon and Liu, Huan. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data,Mary Ann Liebert, Inc., publishers (pp. 171–188), 2020.

[27] Kumar, Srijan and Shah, Neil. False information on web and social media: A survey. arXiv preprint arXiv:1804.08559, 2018.

[28] Seo, Youngkyung and Seo, Deokjin and Jeong, Chang-Sung. FaNDeR: fake news detection model using media reliability. TENCON 2018-2018 IEEE Region 10 Conference (pp. 1834-1838), 2018.

[29] Khan, Ajmal and Khan, Azmat. From Burhan Wani to Abhi Nandan: A Comparative Analysis of India and Pakistan's Newspaper Coverage of Kashmir Conflict. NUST Journal of International Peace and Stability, 2020.

[30] Mahmood, Z., Daud, A., & Abbasi, R. A. (2021). Using machine learning techniques for rising star prediction in basketball. Knowledge-Based Systems, 211, 106506.

[31] Parikh, Shivam B and Atrey, Pradeep K. Media-rich fake news detection: A survey. 2018 IEEE conference on multimedia information processing and retrieval (MIPR) (pp. 436–441), 2018.

[32] Vieweg, Sarah and Hughes, Amanda L and Starbird, Kate and Palen, Leysia. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1079–1088), 2010.

[33] Fernandez, Miriam and Alani, Harith. Online misinformation: Challenges and future directions. Companion Proceedings of the The Web Conference 2018 (pp. 595–602), 2018.

[34] Samonte, Mary Jane C. Polarity analysis of editorial articles towards fake news detection. In Proceedings of the 2018 1st International Conference on Internet and e-Business (pp. 108–112), 2018.

[35] Castillo, Carlos and Mendoza, Marcelo and Poblete, Barbara. Predicting information credibility in time-sensitive social media. Internet Research, 2013.

[36] Clark, Eric M and Williams, Jake Ryland and Jones, Chris A and Galbraith, Richard A and Danforth, Christopher M and Dodds, Peter Sheridan. Sifting robotic from organic text: a natural language approach for detecting automation on Twitter. Journal of computational science, 2016.

[37] A. Daud, N. u. Islam, X. Li, I. Razzak and M. K. Hayat, "Identifying Rising Stars via Supervised Machine Learning," in IEEE Transactions on Computational Social Systems, vol. 10, no. 5, pp. 2374-2384, Oct. 2023, doi: 10.1109/TCSS.2022.3178070.

[38] Behera, Navnita Chadha. The Kashmir conflict: Multiple fault lines. Journal of Asian Security and International Affairs (pp. 41–63), 2016.

[39] Ahmed, Naveed Ur Rehman, Badshah, Afzal, Adeel, Hanan, Tajammul, Ayesha, Duad, Ali, and Alsahfi, Tariq. "Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects," *IEEE Access*, 2024.

[40] Rubin, Victoria L and Conroy, Niall J and Chen, Yimin. Towards news verification: Deception detection methods for news discourse. Hawaii International Conference on System Sciences (pp. 5–8), 2015.

[41] Hussain, Kamal, Saeed, Zafar, Abbasi, Rabeeh, Sindhu, Muddassar, Khattak, Akmal, Arafat, Sachi, Daud, Ali, and Mushtaq, Mubashar. "Towards understanding the role of content-based and contextualized features in detecting abuse on Twitter," *Heliyon*, vol. 10, no. 8, 2024, Elsevier.

[42] Martinez-Gil, Jorge and Chaves-Gonzalez, Jose Manuel. Transfer learning for semantic similarity measures based on symbolic regression, Journal of Intelligent and Fuzzy Systems(pp. 1–13), 2023.

[43] Shu, Kai and Wang, Suhang and Liu, Huan. Understanding user profiles on social media for fake news detection. 2018 IEEE conference on multimedia information processing and retrieval (MIPR) (pp. 430–435), 2018.

[44] Saeed, Zafar, Abbasi, Rabeeh Ayaz, Maqbool, Onaiza, Sadaf, Abida, Razzak, Imran, Daud, Ali, Aljohani, Naif Radi, and Xu, Guandong. "What's happening around the world? A survey and framework on event detection techniques on Twitter," *Journal of Grid Computing*, vol. 17, pp. 279–312, 2019, Springer.

[45] Helmstetter, Stefan and Paulheim, Heiko. Weakly supervised learning for fake news detection on Twitter. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 274–277), 2018.

[46] Abbasi, Rabeeh Ayaz, Maqbool, Onaiza, Mushtaq, Mubashar, Aljohani, Naif R., Daud, Ali, Alowibdi, Jalal S., and Shahzad, Basit. "Saving lives using social media: Analysis of the role of Twitter for personal blood

donation requests and dissemination," *Telematics and Informatics*, vol. 35, no. 4, pp. 892–912, 2018, Elsevier.

[47] Gupta, Aditi and Lamba, Hemank and Kumaraguru, Ponnurangam and Joshi, Anupam. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. Proceedings of the 22nd international conference on World Wide Web (pp. 729–736), 2013.

[48] Banjar, Ameen, Ahmed, Zohair, Daud, Ali, Abbasi, Rabeeh Ayaz, and Dawood, Hussain. "Aspect-Based Sentiment Analysis for Polarity Estimation of Customer Reviews on Twitter," *Computers, Materials & Continua*, vol. 67, no. 2, 2021.

[49] Yasmin, A., Haider, W., Daud, A., & Banjar, A. Machine learning based software effort estimation using development-centric features for crowdsourcing platform. Intelligent Data Analysis, (Preprint), 1-31.

[50] Alowibdi, Jalal S., Alshdadi, Abdulrahman A., Daud, Ali, Dessouky, Mohamed M., and Alhazmi, Essa Ali. "Coronavirus pandemic (COVID-19): Emotional toll analysis on Twitter," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 17, no. 2, pp. 1–21, 2021, IGI Global.

[51] M. Luqman, M. Faheem, W. Y. Ramay, M. K. Saeed and M. B. Ahmad, "Utilizing Ensemble Learning for Detecting Multi-Modal Fake News," in IEEE Access, vol. 12, pp. 15037-15049, 2024, doi: 10.1109/ACCESS.2024.3357661.

[52] Atodiresei, Costel-Sergiu and selea, Alexandru and Iftene, Adrian. Identifying fake news and fake users on Twitter. Procedia Computer Science, Elsevier. (pp. 451–461), 2018.

[53] Gurajala, Supraja and White, Joshua S and Hudson, Brian and Matthews, Jeanna N. Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach. Proceedings of the 2015 international conference on social media and society (pp. 1–7), 2015.

[54] Figueira, Alvaro and Oliveira, Luciana. The current state of fake news: challenges and opportunities. Procedia computer science (pp. 817–825), 2017

[55] Kumar, Mahesh and Kumar, Devender. An efficient gravitational search decision forest approach for fingerprint recognition. Kuwait Journal of Science, 2023.

[56] Saeed, R., Afzal, H., Rauf, S. A., & Iltaf, N. (2023). Detection of offensive language and its severity for low resource language. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(6), 1-27.

[57] Alghamdi, J., Lin, Y., & Luo, S. (2022). A comparative study of machine learning and deep learning techniques for fake news detection. *Information*, 13(12), 576. MDPI.

[58] Binay, A., Binay, A., & Register, J. (2024). Fake news detection: Traditional vs. contemporary machine learning approaches. *Journal of Information & Knowledge Management*, 23(05), 2450075. World Scientific.

[59] Nasser, M., Arshad, N. I., Ali, A., Alhussian, H., Saeed, F., Da'u, A., & Nafea, I. (2025). A systematic review of multimodal fake news detection on social media using deep learning models. *Results in Engineering*, 26, 104752. Elsevier.

[60] Yi, J., Xu, Z., Huang, T., & Yu, P. (2025). Challenges and innovations in LLM-powered fake news detection: A synthesis of approaches and future directions. In *Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security* (pp. 87–93).

[61] Alghamdi, J., Lin, Y., & Luo, S. (2022). A comparative study of machine learning and deep learning techniques for fake news detection. *Information*, 13(12), 576. MDPI.

[62] Hu, B., Mao, Z., & Zhang, Y. (2025). An overview of fake news detection: From a new perspective. *Fundamental Research*, 5(1), 332–346. Elsevier.

Text Classification Using Enhanced Binary Wind Driven Optimization Algorithm

Jaffar Atwan[1], Mohammad Wedyan[2], Ahmad Hamadeen[3], Qusay Bsoul[4], Ayat Alrosan[5], Ryan Alturki[6]
Prince Abdullah Bin Ghazi Faculty, AlBalqa Applied University, Jordan[1,3]
Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid, Jordan[2]
College of Information Technology, Amman Arab University, Amman, 11953, Jordan[4]
School of Information Technology, Skyline University College, Sharjah, United Arab Emirates[5]
Department of Software Engineering, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia[6]

***Abstract*—Document classification using supervised machine learning is now widely used on the internet and in digital libraries. Several studies have focused on English-language document classification. However, Arabic text includes high variation in its morphology, which leads to high extracted features and increases the dimensionality of the classification task. Towards reducing the curse of dimension in Arabic text classification, a wrapper feature selection method is proposed in this study. In more detail, a hybrid metaheuristic model based on the Wind Driven and Simulated Annealing is designed to solve FS task in Arabic text, known as WDFS. The Wind Driven method is initially introduced to optimize the Fs task in the exploration phase. Then, WD is hybridized with simulated annealing as a local search in the exploitation phase to enhance the solutions located by the WD. Three classifiers are utilized to evaluate the selected features using the proposed WDFS: K-nearest Neighbor, Naïve Bayesian, and Decision Tree. The proposed WDFS method was assessed on selected four groups of files from a benchmark TREC Arabic text newswire dataset. Comparative results showed that the WDFS method outperforms other existing Arabic text classification methods in term of the accuracy. The obtained results reveal the high potentiality of WDFS in reliably searching the feature space to obtain the optimal combination of features.**

***Keywords*—Text classification; Arabic documents; wind driven optimization algorithm; simulating annealing; feature selection**

I. Introduction

There is a large and growing amount of machine-readable information out there, which makes it more difficult to use and understand. It provides resources that make it possible to automatically recognize and organize huge amounts of text [1]. Technologies concerning document classification that are automatically applied to huge amounts of diverse document information have grown due to the early progress on the Internet and the escalation of electronic information. Among document classification tasks, Arabic document classification is the most recent challenge in the field of document classification because of the Arabic language's complexity [2]. Additionally, there are a lot of documents conducted almost every day which has made it difficult to accurately detect Arabic documents. Thus, many studies have focused on identifying the difficulties facing compilations of standard Arabic documents classifier [3, 4, 5], encouraging more studies for to improve the performance of the Arabic document classifiers. Most Arabic document classifiers are notoriously difficult to handle huge amounts of documents and to properly identify Arabic documents. As a result, this issue is considered the primary Arabic classification problem. There are two main aspects to the obstacles facing Arabic classifiers. The first is the huge diversity of features resulting from different Arabic word variants, and the second aspect is another operation of the document classification system may be affected by the failure of classifiers to handle stemming. Thus, the algorithm will be used to create a stemming rule that relies on the grammatical rules of speech to address the difficulty of processing the Arabic text while bypassing the morphological and grammatical rules [6].

Arabic document classification task is the process of classifying documents based on their content into predefined category classifications [7]. In this task, large, enormous features are extracted from the text, and they may amount to hundreds or thousands. Due to the high dimensions of the feature space, classifier performance may be affected and more time may be required to replace the term with alternative concepts. By sifting through the text and identifying key semantics, the number of characteristics can be reduced; however, this reduction is not necessarily optimal as some of the omitted terms may be significant characteristics [8]. Choosing the right features is the most critical application in machine learning. Substantially, the most important features identify the learning objective where the attention of the learning algorithm highlights the most useful data for analysis and of the future prediction [1, 9, 10, 11]. Concepts from testing theory were used to create a correlation-based feature selection method and evaluate a set of machine learning algorithms that have been trained on various combinations of real-world and compositional issues. Feature selection (hereinafter referred to as FS) is simple and quick to implement because it removes redundant and irrelevant data while also iteratively improving the performance of learning algorithms. Comparison of the results of this technology with a contemporary feature identified from the literature [12] requires much less calculation.

Although several FS methods have been introduced to assist in Arabic text classification, most of these studies focus on filter FS methods, such as X, which introduced the Chi-square filter for Arabic text classification, and Y, which introduced the cosine similarity for this task. These filter methods were utilized to rank the features but not to select a suitable subset. Using wrapper FS methods with an optimization algorithm assists in choosing an appropriate subset, which increases the performance of text classification.

In designing an FS method, the main goal is to find the most appropriate subset of features; hence, FS problems

are considered NP-hard problems [7]. Using exact search algorithms with the NP-hard problem will generate all possible feature subsets. For instance, assuming we have N text words (i.e., features), then using FS methods such as wrapper methods will generate 2^N subsets of features using a specific learning algorithm (i.e., KNN), which is computationally costly. Whereas FS aims to keep the number of features selected to a minimum while maximizing classification performance, it can be considered an optimization problem. To choose wisely, it is usually necessary to find a middle ground between these two goals. It is best to think of the feature selection process for text classification as a multi-objective task rather than a single-objective task. Numerous meta-heuristic algorithms based on multi-objective approaches have recently been employed to solve text FS issues in an encapsulated framework [3].

The following contributions were made to the classification of documents in this study:

- The stemming process is introduced as a preprocessing step for Arabic text classification, which groups words of the same root into a single word.
- The Bag of Words method is used as a preprocessing technique to extract features from Arabic documents.
- The BWDO-SA wrapper-based FS method is investigated to select the most relevant features from the extracted features. SA is introduced into the BWDO to improve the original BWDO's capacity for exploitation.
- Different learning algorithms, including Decision Tree, Naïve Bayesian, and K-nearest Neighbor, were investigated with BWDO-SA for classifying Arabic documents.

The rest of this study are organized as following; Section II discusses an overview of the related work for text classification. The research methodology considered in this study and the proposed framework of Arabic text classification is provided in Section III. Section IV introduces the experiment dataset and parameter settings and the performance measures. Section V describes the experiment results and discussion. Finally, Section VI contains the conclusion and the future work.

II. Related Works

Several previous studies on classification in Arabic language are reviewed to uncover new classes of classifiers, as well as to address issues related to them, with the aim of bridging the gap between feature selection and the impact of using classifier algorithms, such as optimization feature selection and popular classifiers. An actual evaluation is conducted on studies related to Arabic classifiers to determine the potential for high-performance Arabic classifiers with a small number of selected features. Additionally, the problems and obstacles related to feature selection are identified and discussed, along with their impact on optimization feature selection, by reviewing text classifiers and classifier algorithms. Text categorization (TC) is a type of supervised learning activity that entails classifying documents according to labels already assigned to a set of training documents. Knowledge engineering, in which a manually generated set of rules is used to classify documents into predefined categories, was the most widely used strategy until the development of machine learning approaches in TC.

In [13], the authors explained that the incorporation of machine learning (ML) in TC offers several advantages, such as reducing costs and time by relying on expert workforce only, without compromising precision. K-Near Neighbor Learning (KNN), Decision Tree (DT), Neural Networks, least Squares Fitting (LLSF) and Naive Bayesian (NB) have all been developed and applied to TC [14, 15, 16]. All of these learning strategies perform equally well when there are more than 300 documents in each category, as measured by their comparison. However, when there are less than ten positive training papers in each category, DT, KNN, and LLSF outperform Neural Networks and NB.

The volume of textual data currently available has increased, leading to initiatives such as Arabic language filtering [17] that aim to filter and highlight incoming papers for unwanted or unwelcome content. This has led to the increasing importance of transcript taxonomy (TC). Sentiment analysis [18] which aims to ascertain the overall sentiment expressed in a document, is another use. By providing classification models with examples labeled for the problem at hand, such as appropriate classifications (labels), supervised learning methods can be used to solve problems associated with text classification. Therefore, labels of unlabeled documents can be predicted using these model [3, 19, 17, 20, 21, 16, 10]. The task still requires more work of pre-defining classes and assigning class labels to training set documents, despite the fact that several supervised machine learning algorithms have been applied to TC.

Researchers in [22] compared different Arabic classifiers using BOW as extraction without FS. The comparison showed that NB performs better than DT and tree-based J48. Another study conducted by [23] compared NB, tree based J48, and IB1 using BOW as an extraction. The study reported NB to be better in classification tasks compared to FS methods. In [24], the authors compared DT, AdaBoost, and RF using BOW as extraction and found that DT is the best classifier. In addition, they contrasted IG with Chi-Square and deduced that both methods are comparable when using DT as a classifier and Chi-Square as an FS. However, RF was the best classifier, compared to DT with FS. A study done by [25] utilized six FS methods and two classifiers (DT and NB) on six datasets. The study merged ODFFS and TFFSs to generate a hybrid method (HBM) to propose the FSEPO for parameter optimization. The report showed that NB is the best classifier compared to DT and HBM in enhancing optimal FS. In [26], the authors conducted a comparison among various classifiers for Arabic text. The text preprocessing steps included removing digits, dashes, punctuation marks, and other non-Arabic characters, as well as filtering out stop words without using FS. The results of the comparison revealed that the decision tree (D.T) classifier performed better than the Naïve Bayesian (NB) method, K-Nearest Neighbors (KNN), and NaïveBayes Multinomial (NBM) classifiers. In [27], the authors compared Decision tree (C5.0) SVM using stop words as extraction with Chi-Squared statistic FS. The study reported a Decision tree (C5.0) to be better in classification tasks compared to SVM.

The selection of feature subsets is a common problem for

Arabic classifiers due to the large dimensionality of textual Arabic types. Therefore, the feature selection (FS) process is important to reduce these dimensions and to choose a large number of high-quality attributes without significantly influencing the performance. The investigation of the problem can be in two areas: the first is to focus on the selection of significant features and eliminate redundant ones.

The problem of selecting the best subset of features from the original feature space without affecting the accuracy of the classifiers is known as FS. The second perspective involves enhancing the prediction performance of the classifiers by considering their specific problems. Existing literature has shown that several approaches have been investigated for FS with different techniques of Arabic classifiers. Most importantly, FS methods should consider the properties of the learning algorithm and the problem domain. The effects of different FS methods on different learning frameworks and fields may vary, and therefore, ongoing studies are being conducted to develop new methods for handling FS and addressing dimensionality problems, particularly for Arabic classifiers. Hence, the aim of using FS approaches may include performance enhancement, data simplification, visualization, as well as feature dimensionality reduction [28, 29].

Various approaches have been applied for dimensionality reduction, including IG, Chi-Square Statistic, MI, TS, and DFT. Chi-square scale the level of dependence between a particular term and a particular category, i.e., A measure of how strongly a term is believed to indicate whether a document belongs to a particular category [30]. Several researchers have reformulated the statistic and deployed it for document categorization [30, 31]. In [32], the authors designed a classifier for Arabic document filters by employing Chi-square and IG on ML techniques, namely, BA tree-based J48 and DT. From their report, the Chi-square and IG showed similar performance as the BA and tree-based J48 classifiers, while DT without FS performed better than Chi-square and IG with FS. IG determines the number of bits of information acquired for each prediction class when the presence or absence of a term is determined in a document.

In [33], the authors suggested an improved Feature selection method for Arabic text classification. To enhance the effectiveness of classification, a number of experiments were performed, comparing improved chi-square with three known types of FS, namely chi-square, information gain, and mutual information, using the SVM classifier on a dataset of 5070 Arabic documents. The researchers evaluated these methods using classification measures, The results showed that the Arabic text classification model performs significantly better when the improved chi-square is combined with the SVM classifier. In [34], the authors suggested an innovative and new technique of feature selection that relies on practical swarm optimization to enhance the work of text classification. The proposed method was compared with several other methods, including genetic algorithm, chi-square, and the information gain, using Matlab and the Reuters_21578 dataset. The proposed method efficiency and effectiveness were evaluated using the nearest neighbor classifier. In conclusion, the findings indicated that the proposed method surpassed other methods. Marie-Sainte and Alalyani [35], introduced a new firefly algorithm that relies on FS for Arabic Text Classification (ATC). In this study, various experiments and comparisons were conducted using six feature selection techniques in combination with one state-of-the-art techniques that used the SVM classifier. These techniques were applied to the OSAC dataset, and three different techniques were used to measure execution accuracy. The outcomes showed that the suggested technique (ATC_FA) produced superior performance.

In [36], the authors presented a hybrid method combining the Flower Pollination Algorithm (FPA) and the Adaboost algorithm. Each algorithm served a different purpose, with FPA focusing on feature selection and Adaboost aiming to categorize text contents. The experiments have been made on Reuters-21578, WEBKB, and CADE 12 datasets, and the outcomes demonstrated that the suggested approach outperforms NB-K-Means, KNN-K-Means, and other learning models in terms of FS and performance. In [37], the authors proposed a new hybrid algorithm that combines bee colony optimization (BCO) and ant colony optimization (ACO) to improve FS in 2017. This approach was applied to 13 datasets from the University of California, using Decision Tree (DT) classifiers. The experiments demonstrated that the hybrid model achieved higher categorization accuracy and better feature selection compared to other methods, including ACO Hybrid ACO, ACO-NN Hybrid, Based FS, IQR Bee, CatFish Binary PSO, PSO-SVM, and ABC-FS Swarm based Hybrids such as ACO-PSO Hybrid, and ABC-DE Hybrid, as evidenced by the Prediction Accuracy (PA) measure. In [29], the authors used the improved feature selection technique of whale optimization with SA. The outcomes were contrasted with those of other population-based techniques were outperformed them in terms of data classification performance.

For feature extraction and selection [38], the authors exploited the Aravec embedding model. Aravec is a large dataset trained using an Arabic twitter data domain. The extracted features were exploited to classify twittes using a multilayer bidirectional long short term memory (BiLSTM). The model outperforms the state of the art models.

The authors of Targio and Tubishat [39] proposed an approach that used a hybrid fillter-wrapper method based on Principal Component Analysis (PCA) to filter the features and select relevant ones. These features will be exploited for categorizing Arabic documents. To select more relevant features GreyWolf Optimizer (GWO) as wrapper approach is performed. To evaluate the classification task the Logistic Regression (LR) is used. Three Arabic datasets were exploited to perform the experimentations. These laters show that the approach-based PCA-GWO outperforms the baseline classifiers.

In the work [40], the authors present a feature section approach that is composed of two stages. the first one exploited the filter model. The second one used and the multi-objective wrapper model, an upgraded version of the Whale Optimization Algorithm (WOA) with Particle Swarm Optimization (PSO). Four benchmark text corpora were used to evaluate the proposed approach which shows encouraging results. The study of [41] presents an approach that uses NLP and machine learning methods for text classification. To tune the hyperprameter of machine learning algorithms two techniques were exploited: Grid Search and Random Search. Experimentations were performed using CNN Arabic dataset. The results showed

that random search is more efficient than grid search in terms of accuracy, precision, recall, and F1-score and execution times for some of the algorithms.

In this research we have analyzed a number of previous studies that help us better understand this study. Various concepts, including TC, is one of the most important applications that are in high demand today, and the most used classifiers like k-Nearest Neighbors (KNN), Naive Bayes (NB), and Decision Trees (DT), have been discussed. Given the huge quantity of online information that is constantly increasing, the findings of these studies are valuable in terms of saving effort and reducing costs for users and developers. Additionally, feature selection depends on optimization techniques which result in reducing redundant and unnecessary features in texts.

III. Methodology

The proposed methodology embraces a wrapper-FS method to select the most suitable subset of features from the original feature set to improve the performance of Arabic document classification. Fig. 1 below displays the principal methodology used to conduct this study, which consists of several phases: data pre-processing, feature extraction, feature selection, and finally, the classification phase. The proposed BWDO-SA is introduced within the feature selection phase.

A. Data Pre-Processing

The TC system begins with the pre-processing stage. This step is important for document representation because it pre-processes the collection's text data to determine the indexed features or terms. Several procedures are performed during this phase to clean and clarify the text data [42, 43]: First, a tokenization and normalization procedure is used to remove unnecessary terms such as punctuation marks, non-Arabic letters, diacritics, and digits [44]. Second stop-word removal, which focuses on eliminating unimportant words for machine-learning tasks. The third phase is the stemming phase, which is used in text classification to find the stem word for several related terms. The following steps provide additional details regarding these processes for Arabic document classification.

1) Tokenizing and normalizing of data: When text is divided into separate units by using either space or special symbols. As such, every word in a text is represented in a unit. Such a method is called tokenization. For instance, (خير جليس في الزمان كتاب) it explains the method tokenizing works in a text .(خير_جليس_في_الزمان_كتاب) Furthermore, normalization is beneficial for carrying out before the stemming task, particularly in the Arabic text is being performed due to the text normalization that decreases several types of characters in Arabic languages to enable one uniform character to act to those characters. The following steps followed in the pre-processing.

- Remove diacritics, punctuation marks, hyphens, numbers, digits, and non-Arabic letters.
- Substitute ,آ ,إ as well as أ by ا
- Substitute the last ة by ه.
- Substitute the last ى by ي

2) Stop word removal: The removal of these words will not have any major impact on the document's meanings, but rather will be clearer to smooth the process of the meaning interpretation of the document [6, 44, 45, 43, 46]. Fig. 2 shows the document sample words after the removal of the stop-word.

3) Stemming: A stemming algorithm reduces all the inflectional derivational variants of words into a common form called the stem [45]. For example, the words, 'work', 'works', 'working', 'worked', and 'worker' are taken from the root (stem) 'work'. The root of a word is obtained by removing all or some of the affixes attached to the word. The Light10 stemmer was used in the stemming process following the same steps in [47]. This process applies to words containing additional characters whether suffixes or prefixes to get the root of these words [48]. There are four types of affixes shown in Table I.

B. Feature Extraction

Text pre-processing is a necessary step in TC using ML technologies. This entails converting the text into forms that learning algorithms can use. The document representation vector space model, which is widely used, was first described in [49]. Each dimension in the "d" vector that this model uses to describe each text corresponds to a different term in the document set space term. The text document is often converted into a term frequency vector to implement the document processing process [34, 85]. Each document was represented by a vector, where each term acts as an attribute and the value of the attribute is the document's TF*IDF weight [50] a statistical method for measuring the importance of a term for a document in a group. The most common terminology weighing method (TF*IDF) takes this property into account. According to this method, the term weight in the document d is inversely related to the number of documents in the group where the term appears (IDF) and proportional to the number of times the word appears in the document (TF). When using the TF-IDF weighting strategy, terms are given weight based on their frequency in a document using a factor that ignores their importance if they are present in most documents, especially when the term is thought to have poor discriminatory power [see Eq. (1)].

$$w_i = tf_i . \log(\frac{N}{n_i}) \quad (1)$$

where, N denotes the total amount of documents in the collection, w_i denotes the weight of the term in the document, tf denotes its frequency, and n_i denotes the number of documents in which the term appears. Since each token in each document represents a dimension in the feature space, this transformation raises the problem of high dimensionality of the feature space. As a result, large computational costs significantly reduce classification effectiveness [51].

C. Feature Selection

Choosing a subset of features for an Arabic text classification system is crucial. Generally, in ML, three methods for feature selection: embedded methods, wrapper, and filter. By using these techniques, a cost-effective and accurate text

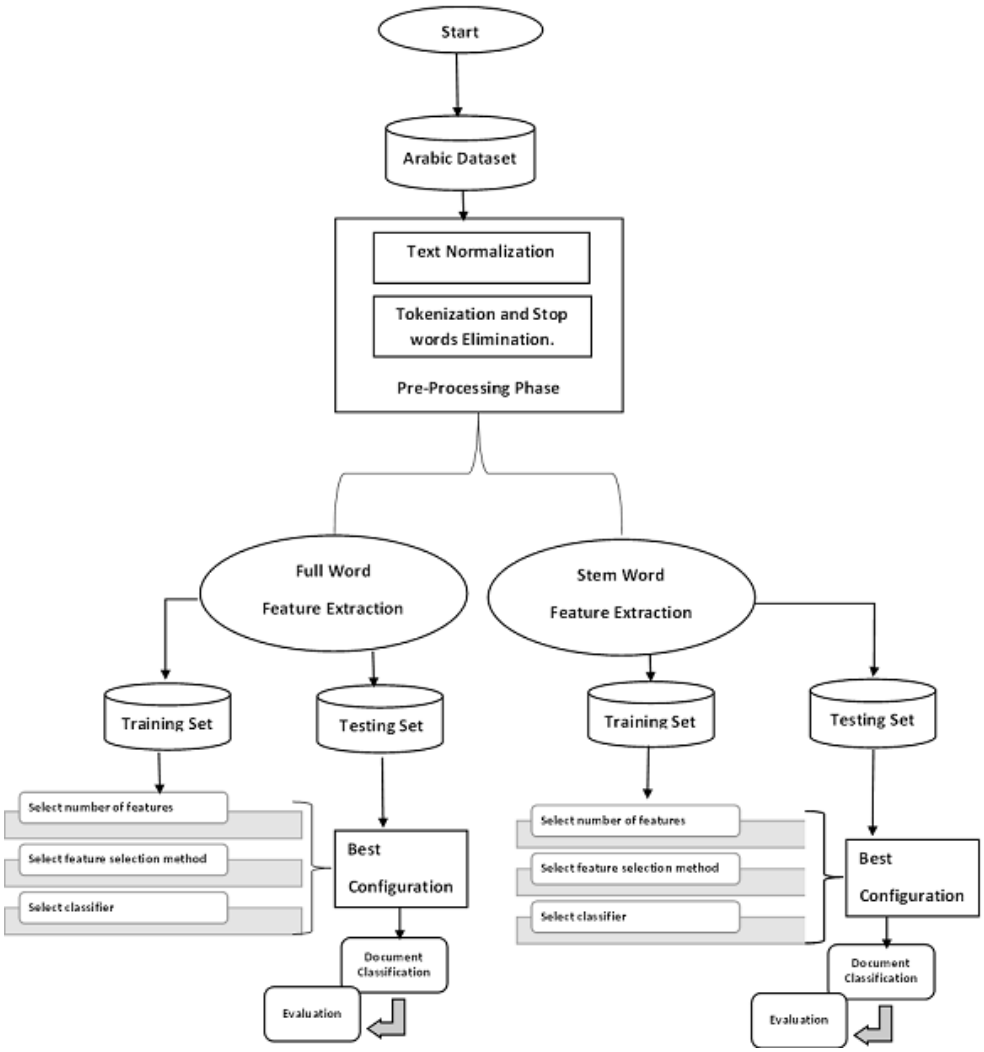


Fig. 1. The primary methodology of this research.



Fig. 2. Document sample words after the removal of the stop-words.

TABLE I. UTTERANCES OF AFFIXES INWORD(مهنوريلي, (مهنوعدخيلي .

Postfix	Suffix	Root	prefix	Antefix
مه	ون	يرى	ي	ل
مه	ون	خدع	ي	ل
Pronoun meaning (they)	Termination of conjugation	يرى meaning (See) and خدع Meaning (Deceive)	A letter meaning the tense and the person of conjugation	Preposition meaning (to)

classification task can be achieved [52]. In this study, wrapper methods were examined.Fig. 3 illustrates the three primary factors used in wrapper methods for identifying the optimal feature set [53]: First, the ideal feature combination from the initial dataset is found using a search approach (for instance, an optimization technique). Second, a machine learning model (such a classifier) is used to assess the produced subset of features by comparing them to the validation dataset. Third, the dependability of the chosen feature subset is evaluated using an objective function criterion.

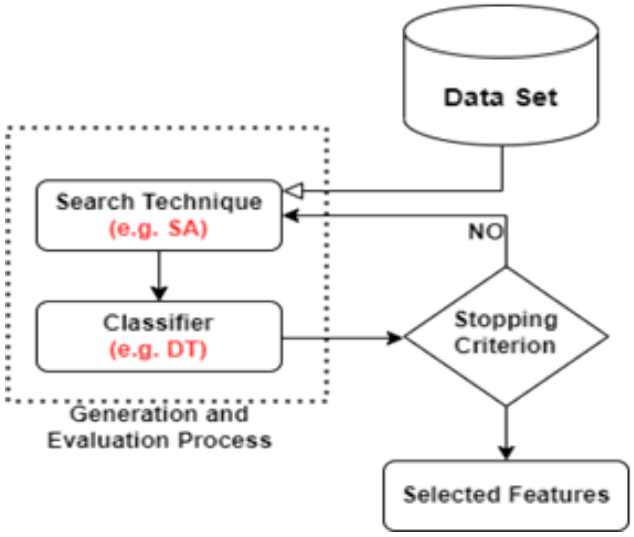


Fig. 3. Wrapper-Based feature selection method using SA and DT classifier.

As the increases in the amount of Arabic language text documents in classification, it is becoming increasingly difficult. Thus, FS techniques play a crucial part in deciding what features are the most relevant for the classification model while ignoring the redundant and the irrelevant ones. This study proposes a hybrid wrapper-based FS method for Arabic document classification, based on a combination of binary wind-driven optimization (WDO) and simulated annealing (SA) algorithms, namely BWDO-SA.

D. The Proposed BWDO-SA for Feature Selection

In this section, the proposed BAWDO-SA method is investigated, and its early stages of standard WDO, solution initialization, binary WDO, simulated annealing, and the proposed BAWDO-SA are discussed in the following:

1) Wind Driven Optimization (WDO): The Earth's atmosphere, where winds blow in an attempt to balance horizontal differences in air pressure, served as the inspiration for WDO [54]. The term "wind" refers to the horizontal movement of air, particularly in the troposphere, which is the lowest layer of Earth's atmosphere. The thickness of the troposphere varies with latitude and extends from the earth's crust to a height of about 18 km [55].

This phase aims to employ supervised WDO for feature selection. In this research, a supervised WDO approach was applied, using the accuracy of classifiers as the objective measure of performance for feature selection. The study suggested using three classifiers by combining feature selections to identify, detect document categories and assess the three classifiers' performance using WDO's supervised feature selection. In other words, this proposed method was designed to detect the most important features that would improve the classifier's performance while reducing the number of features selected. Fig. 4 shows the pseudo-code of WDO as feature selection [55].

The major strengths of the WDO high priority are that it [56]:

- Contains many transactions that facilitate the process of restricting the search field.
- Uses additional terms such as (gravitational and Coriolis force) in the equation of speed update.
- Classifies the population based on their pressure.
- Checks the boundaries to prevent any air expulsion from the search area.
- Affects problems related to both discrete and continuous value parameters [55].

2) Simulated annealing: The escape probability of a local optimum problem served as the basis for the development of the SA employing Hill Climbing-based (HC) techniques [56]. Using the simulated annealing method [57], is a heuristic solution based on the HC technique. Fig. 5 illustrates the algorithm. SA begins by establishing a random number from the available solutions and then generates neighboring solutions based on the same random number.

After generating the surrounding solutions, the algorithm verifies each one's credentials before attempting to select the most acceptable number for the next stage. With the probability defined by the Boltzmann probability value, P=e^(-θ/T), the method also accepts minimum solutions, where θ is the difference between the Solbest and Soltrial evaluations of the objective function. T is the temperature, which, in accordance with a cooling plan, drops periodically during the search phase. In [58], the authors used SA in order to solve numerous attribute reduction issues in SimRSAR. The solution took into account that each case corresponds to a certain set of attributes in SimRSAR, and a nearby arithmetic attribute can be found using the chosen attribute. $T(t+1) = 0.93 * T(t)$.

Set user parameter of the WDO: $P_{opt}, \alpha, g, RT, and\ c\ and\ Maximum_Iteration$.

Create randomly initial population of feature selection using Eq (2, 3)

$$AirParcel\ feature = [X_1, X_2, X_3 \ldots] \quad (2)$$

$$\text{AirParcels of feature} = \begin{bmatrix} AirParcel_1 \\ AirParcel_2 \\ AirParcel_3 \\ \vdots \\ AirParcel_{Popt} \end{bmatrix} \begin{bmatrix} x_1^1 x_2^1 x_3^1 & \cdots & x_{Nfeatures}^1 \\ \vdots & \ddots & \vdots \\ x_1^{Popt} x_2^{Popt} x_3^{Popt} & \cdots & x_{Nfeatures}^{Popt} \end{bmatrix} \quad (3)$$

Randomize the velocity and the position of air parcels

Set the fitness function using Eq (4)

$$fitness = \beta \gamma_R(D) + \alpha \frac{|R|}{|N|} \quad (4)$$

Evaluate the population and find the minimum pressure value for air parcels ($P_0 = min(f(x_i))$)

Set global $P = P_0$

for all i do

for all j do

Update the velocity by Eq (5)

$$u_{new} = (1-\alpha)u_{cur} - gx_{cur} + \left(RT \left|\frac{1}{i} - 1\right| (x_{opt} - x_{cur})\right) + \left(\frac{cu_{cur}^{otherdim}}{i}\right) \quad (5)$$

Choose random dimension

Choose velocity based on random dimension

Check velocity by Eq (6)

$$u_{new}^* = \begin{cases} u_{max} & if\ u_{new} > u_{max} \\ -u_{max} & if\ u_{new} < -u_{max} \end{cases} \quad (6)$$

Update air parcel positions by Eq (4.27)

$$x_{new} = x_{cur} + (u_{new}\Delta t) \quad (7)$$

Call CMAES

Return the new set of WDO coefficients for $i+1$

Evaluate the new solution and update $P(i,j)$ using Eq (4)

if $P(i,j) < P0$ then

global $P = P(i,j)$

end if

end for

Rank air parcels and find the local optimum ($xlocal$)

end for

Rank air parcels and find the global optimum ($global$)

Return the values of $xglobal$ and $fmin = min(f(xbest))$

End of the algorithm

Fig. 4. Pseudo-code of WDO as a feature selection.

ize (temperature T, random starting features selected)
While cool_iteration <= max_iterations
Cool_iteration = cool_iteration + 1
Temp_iteration = 0
While temp_iteration <- nrep
Choose a new point from he tions.
Compute the objective function of produced c
$cost_j = f\left(x^j_{1e}, x^j_{featuatures@}\right), j = 1,2,3,\ldots,P_{opt}$
δ= current_cost-previous_cost
if $\delta < 0$, approve neighbour
else, approve with possibility exp($-\delta/T$)
end while
$T = \alpha \cdot T\ (0 < \alpha < 1)$
end while
end while

Fig. 5. Pseudo-code simulating annealing.

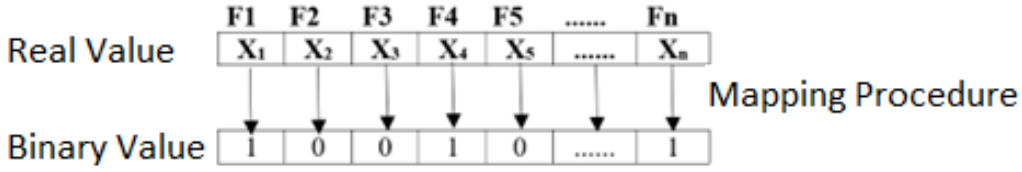


Fig. 6. Representation of the selected features.

3) Binary WDO for feature selection: The task of FS is a binary discrete problem in which features can be represented by a binary solution or vector. The values of this vector are either one for the selected features or zero for the omitted features, which are not selected, as illustrated in Fig. 6.

The basis of the WDO was initially implemented for continuous optimization tasks. Thus, to implement the WDO for the FS task, a transform function is required to convert the WDO to a binary WDO (BWDO). This study uses a sigmoid transform function, as shown in Fig. 7, to convert the solution to binary format (i.e., 0 or 1).

Eq. (2) illustrates the formula for the sigmoid function for a solution x.

$$T(X_i^j(k)) = \frac{1}{1 + exp(-X_i^j(k))} \quad (2)$$

where, $T(X^j{}_i(k))$represents the transform of the dimension j in the solution x_i (i.e., air particle) at iteration k.

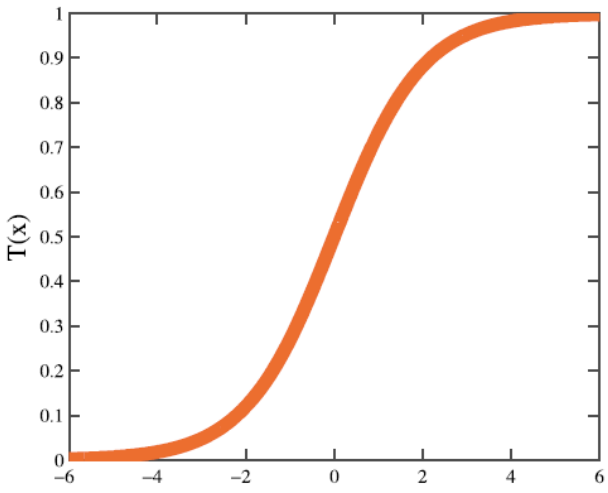


Fig. 7. Sigmoid function.

In the iteration, k, the location of each air particle xik of each pair particle i is converted into binary format using the following equation:

$$X_i^j(k) = \begin{cases} 1 \; rand \; \geq \; T(X_i^j(k) \\ 0 \; rand \; < \; T(X_i^j(k) \end{cases} \quad (3)$$

The above Eq. (3) is utilized in the FS approach in this study to determine the probability of changing the location of the elements in the solution.

The BWSO begins the optimization process with a randomly constructed initial population of N air particles. Each particle designates one solution with a binary vector of dimension D, which represents the candidate features, i.e., a subset of features. In this vector, dimension D characterizes the entire set of features in the original dataset. Fig. 4 shows the pseudocode of WDO as a feature selection.

4) Evaluation using fitness function: The major objective of using the feature selection optimization approach is Reducing the number of selected features while getting the best classification accuracy. In other words, fewer selected features provide better opportunities to improve the solution classification performance.

All possible available solutions are evaluated according to the proposed witness selection procedure and the final publisher, which depends on three classifiers, namely KNN, SVM, and NB [59] to classify the answer correctly based on the amount of features chosen. To achieve the highest level of precision while maintaining an appropriate balance of all features chosen for each of the minimum solutions for feature selection, the fitness function (objective function) in Eq. (4) is utilized for each WDO as well as SA assessment techniques for agent searches.

$$fitness = \beta_{\gamma_R}(D) + \alpha \frac{|R|}{|N|} \quad (4)$$

The BWDO is a population-based metaheuristic method with high exploration performance. However, the main limitation of the BWDO is the degradation of its exploitation capabilities. This is because it employs a blind operator to enhance the local area surrounding the best solutions discovered thus far without regard for the quality of the present solutions. This study introduces an innovative FS method based on the hybridization of BWDO and SA, known as BWDO-SA. This method is applied to the FS task in the domain of Arabic text classification. The SA method, which is a local search algorithm, is utilized within the BWDO to replace the blind exploitation phase. Specifically, the SA seeks the best solution in the vicinity of the best-known and randomly selected solutions.Fig. 8 shows the flowchart of the proposed Hybrid of BWDO-SA algorithm.

E. Classification

We investigated NB, DT and KNN, which are frequently used for text classification and proven to give satisfactory results, to test the effectiveness of the proposed methodology [60]. Additionally, these classifiers meet nearly all of the difficulties encountered in this paper thanks to the following qualities: In supervised algorithms, the categorical structure of a given database is taken to be known beforehand. For the supervised algorithms to map documents into predefined labels, they need a set of labeled documents. As noted earlier, it can be difficult to determine the exact class and naming of training sets, especially in large databases. Therefore, the most popular supervised algorithms, including DT, KNN, and NB, will be the focus of this section.

1) KNN: KNN has been established as a well-liked instance-based learning method and has shown efficacy in a number of text classification tasks [60]. The algorithm flow can be summarized as follows: From the given training documents, KNN first determines the k-nearest neighbors. The flow of the algorithm can be summarized as follows: First, the KNN identifies the k-nearest neighbors from the given training documents. Second, based on the category labels of these neighbors, the category of the test document is established. The proven k-nearest method usually sets the test document with the name of the most common class. Each neighbor's contribution to the weighted KNN is determined by how close it is to the test document. This is an extension of the regular KNN. To determine the class score of a document, the similarities between neighbors in each class are added. Or to put it another way, the category score for document x can be represented as in Eq. (5):

$$score(c_j, x) = \sum_{d_i \subset N(x)} \cos(x_i, d_j), y(d_i, c_j) \quad (5)$$

The training document is =d_i , a collection of the k training documents that are closest to x=N(x), similarity between x and d_i =cos(x,d_i) , value of a=1 in function y(di,c_j) if belongs to category, and otherwise 0. The results of both evaluation methods are presented in the results part of this study, which examined the original KNN as well as its weighted version by varying the values of the parameter k.

2) Naïve Bayesian (NB): NB was constructed as a probability-based model that uses the joint probability of terms and categories in a test document in order to calculate the probability of categories assigned to that document [60]. It is the classifier's assumption that all words in each class are conditionally independent of all words in the other class that gives rise to the "naive" feature of the classifier. In contrast to non-NB classifiers, computational procedures can be made simpler by learning parameter values for each word separately based on the idea of independence.

As per [61], Polynomial modeling and multivariate Bernoulli model were both major models used for text classification using NB. Bayes' rule is used by the two models to accomplish the categorization of documents [60] [see Eq. (6)]:

$$P(c_j|d_j) = \frac{P(c_j) \; P(d_j|c_j)}{d_i} \quad (6)$$

where, d_i= test document, C_j= class. Given d_i is the test document, c_j posterior probability for each category's, i.e., , $P(c_j|d_j)$and d_i assigned to the category with the highest probability. The calculation of $P(c_j|d_j)$ calls for an estimation

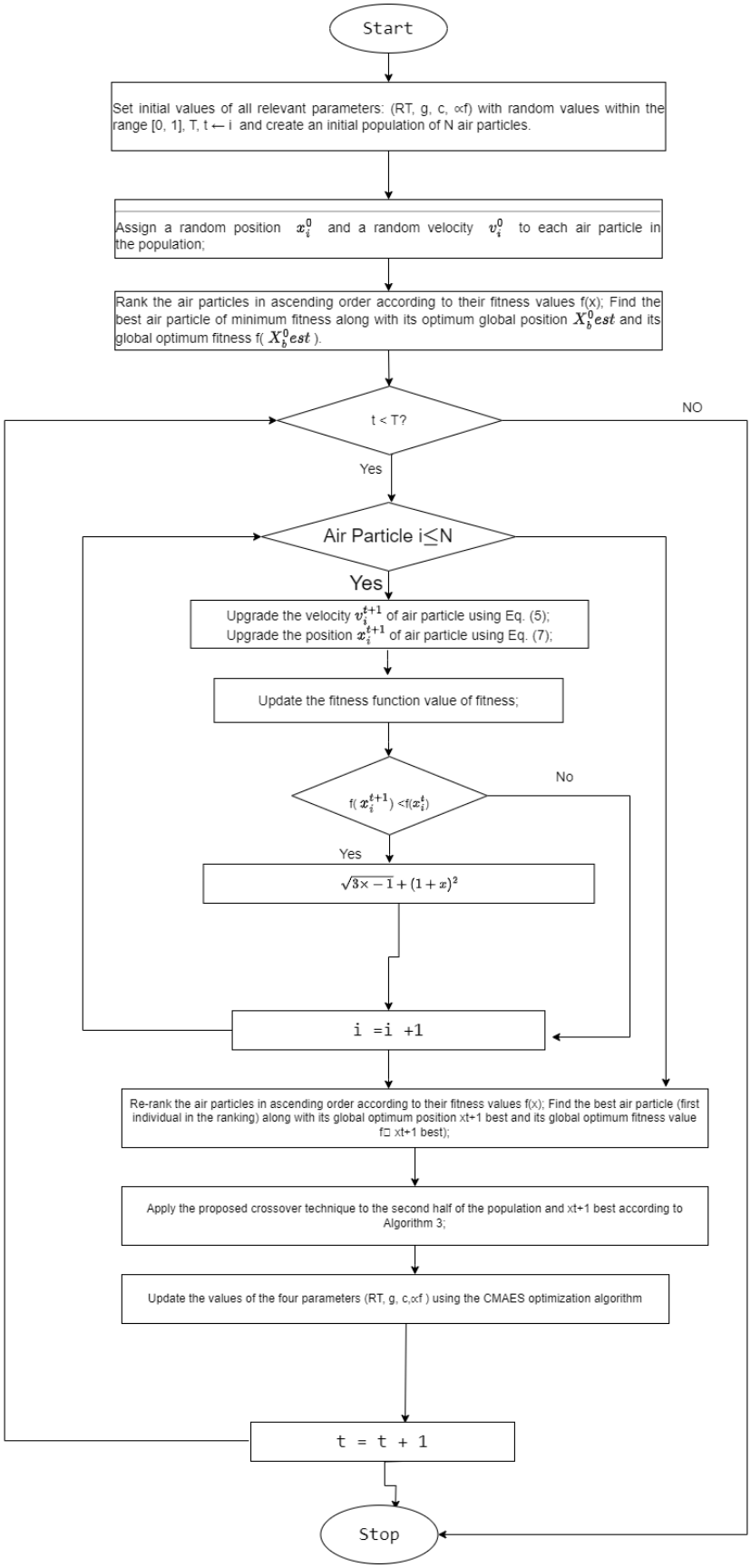


Fig. 8. Flowchart of the proposed hybrid of BWDO-SA algorithm.

of the P(Cj) and $P(c_j|d_j)$based on training set. It should be noted that P(Cj) is the same for each category. Hence, it can be eliminated from the calculation. In doing so, the prior probability of the category P(Cj) will be thus estimated as Eq. (7):

$$pc_j^{\wedge} = \frac{\sum_{i=1}^{N} y(d_i,\ c_j)}{N} \tag{7}$$

where, the amount of training set =N and thus y(di, $c_{j)}$estimated as Eq. (8):

$$y(d_i, c_j) = \begin{cases} d_i \in c_j \\ 0\ otherwise \end{cases} \tag{8}$$

Consequently, the percentage of training set which belongs to Cj is employed to determine the a priori probability of a class Cj. The multinomial model and multivariate Bernoulli model estimate $P(c_j|d_j)$ the parameters in different ways.

3) Decision Tree (DT): A Decision Tree text classifier is a non-parametric supervised learning method, represented as a tree in which the internal nodes are named terms and the leaves are tagged with classes [62]. The decision tree is one of the most popular and successful learning algorithms due to its many features, including lack of parameters, simplicity, ease of understanding, and ability to deal with different types of data. During the learning phase, a decision tree is built by recursively splitting the data based on attribute values and class labels, resulting in a large search space.

A decision tree is a greedy, top-down algorithm and iterative process from the training data entered until it reaches an empty tree. The feature that contains more information about the content is better defined to divide and is considered as the root division feature, Subsequently, subsets satisfying the partitioning features values are generated from the training data. This algorithm works on every subset repeatedly until all instances of a subset belong to the same category [63]. There are many applications in which the decision tree is configured, including its ability to deal with the problem of classification of incomplete data very well . In addition, it can replace statistical procedures to obtain data, extract text, find missing data, enable search engines, and use in many medical fields.

IV. Experiments

A. DataSet

The collection of data contains a set that was created to evaluate the creation of Arabic classifiers by extracting Arabic text, for which 800 documents have been chosen as a portion of TREC 2001 [64]. Realism and Criteria For such a set, TREC queries were developed: TREC 2001 included 25 queries, and relevant appropriate judgments were created using the grouping technique for the set of queries. Accordingly, Table II definition of the classifier's component of TREC includes four categories and the number of documents.

TABLE II. Summary Description of Arabic Data Set

Classes	# documents
1	200
2	200
3	200
4	200

B. Parameter Setting

Regarding the available optimal approach, it could be employed based on the three classification methods of the algorithm, considering this special method. For cross-validation, entire datasets are split. In conclusion, the validations are also categorized, using the evaluation strategy from [65]. While K-fold cross-validation and K-5 folds are utilized for both training and validation, the other folds are employed for testing.

The overall iterations of M could be applied to this process. After that, each optimizer unit could assess the times of K*M per dataset. To obtain the data used for training, the size of all parameters should be determined. Optimum results can be achieved when the number of repetitions reaches 50. The population size in this case should be 28. Each algorithm can be used for five total iterations. For feature selection, random feature selection should be used.

This section explores the algorithmic solution under several settings using 5 vital parameters: P_opt (population size), g (gravitational constant), RT coefficient, C (coriolis effect), and α (constants alpha). The vertical force is the gravitational force (g) directed from highest pressure to lowest pressure. The constant universal gas is represented by RT, where both R and T represent temperature. The constants C and α are utilized in velocity, while P_opt represents the number of Air Parcels. The gravitational force coefficient, g, and the friction coefficient, α, should only vary in the range [0, 1]. However, C and RT might include a wider range of values, such as [0, 5]. This section focused on the effects of variations in one of the parameters. This section focuses on the effects of variations in one of the parameters. Specifically, the three scenarios presented in Table III are evaluated. Furthermore, empirical studies have shown that the best results are achieved when there is a linear relation between P_opt and the number of features. Each scenario is evaluated for 50 iterations, with the cost value of the solution remaining as the fitness function value. The lowest fitness function value [Equation (8)] is considered the best fitness. Among the nine scenarios, the lowest fitness function value was related to scenario S2.3. The evaluation system used for wind-driven feature selection, as indicated in Section 3.3.3 (dmax = 1E03). Accordingly, case S2.3 was selected for conducting tests in this section, while other scenarios had higher fitness function scores. The parameters were set to P_{opt}=6, g=0.2, RT =3, C=0.4 and alph=0.4.

C. Performance Evaluation

The labeled test of the document component often determines the external quality measure, as part of its methodology, the classifiers that are produced are compared to the classes that have been labeled, and the extent to which documents

TABLE III. Parameter Settings of Arabic Data Sets

Scenario	Popt	g	RT	C	A
S1.1	2	0.1	1	0.2	0.1
S1.2	4	0.1	1	0.2	0.1
S1.3	6	0.1	1	0.2	0.1
S2.1	2	0.2	3	0.4	0.4
S2.2	4	0.2	3	0.4	0.4
S2.3	**6**	**0.2**	**3**	**0.4**	**0.4**
S3.1	2	0.3	5	0.6	0.8
S3.2	4	0.3	5	0.6	0.8
S3.3	6	0.3	5	0.6	0.8

from the same class are assigned to the same class is measured. In this work, the employed external quality measure is accuracy, as it is a commonly used quality measure in text mining, along with the number of features. The confusion matrix, also known as the evaluation measures in classification issues, is often defined from a matrix utilizing the number of instances that are properly and incorrectly identified for each class. The confusion matrix for a binary classification issue with just positive and negative classes was displayed in Table IV.

TABLE IV. Confusion Matrix

	The expected Class	
Real class	positive	real class
P	TP	TN
N	FP	FN

There, it is shown how FP, FN, TP, and TN is:

- False Positives (FP): Negative events that are incorrectly seen as positive.
- False Negatives (FN): Positive occurrences that are incorrectly interpreted as bad.
- True Positives (TP): The instances of positivity that are accurately anticipated to be positive.
- True Negatives (TN): The negative occurrences that are accurately predicted as negative.

Percentage Accuracy (ACC) is an evaluation measure widely used in practice that evaluates the effectiveness of a classifier based on the percentage of accurate predictions it makes. Thus, the ACC of the classifier is calculated as Eq. (9):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (9)$$

V. Results and Discussion

The planned contribution of this section is to respond to the research queries raised by the study. The purpose of this section is to achieve the study objective of evaluating the performance of Arabic classifiers NB, DT, and KNN in terms of classification performance, whether with or without selecting features, as well as with and without stemming of words. Additionally, the proposed BWDO-SA algorithm is employed for feature selection to enhance the performance of Arabic classifiers by reducing the dimensions of their features. Finally, the proposed framework of the Arabic classifier evaluated by accuracy and the amount of features. The results of the proposed framework focus on the reduction of the number of features using the optimization FS approach and their enhancement. The results presented in this section highlight the significant implementations using a number of features and the accuracy as evaluation metrics. Evaluation methodology used is also explained, followed by a discussion of the obtained outcomes. In conclusion, The outcomes of the combined WDO and SA techniques are shown in this section. Table V and Table VI present the results of the three classifiers' accuracy with and without the stemmer as a pre-processing step. The results show that DT achieved higher accuracies compared to NB and K-NN in both cases. Without the stemmer, DT achieved an accuracy of 90%, which was higher than NB (33.83%) and K-NN (26.11%). With the stemmer, DT achieved an accuracy of 93%, higher than NB (35%) and K-NN (26.36%). When comparing between the use of stemmer and without stemmer, it can be observed that the stemmer resulted in better accuracy for the classifiers. In terms of feature reduction, without the stemming, the feature count evaluation showed that reducing the feature count was not effective, whereas with the stemmer, better reduction of the number of features was achieved, as shown in Fig. 9 and Fig. 10.

Fig. 11 and Fig. 12 displays the results of integrating various metaheuristic algorithms as feature selection methods with the DT classifier, in terms of precision and the total amount of features, both when a stemmer is used as a preprocessing step and when it is not. The outcomes show that our suggested algorithm, WDO with DT, and SA with DT using the stemmer achieved higher precision compared to other feature selection algorithms. Both WDO with DT and SA with DT obtained the same precision value of 97.01%, which was higher than WOA with DT (94.7%), GA with DT (94.5%), PSO with DT (91.79%), HS with DT (91.5%), and SA with DT (87.2%). Without a stemmer as a preprocessing step with feature selection, WDO with DT and GA with DT achieved the same precision value of 95.02%, higher than WDO with DT (92.2%). In addition, WOA with DT and HS with DT obtained the same precision value of 90.2%, PSO with DT (90%), and SA with DT (85.46%).

Regarding the number of features, using DT classifier, WDO as a feature selection technique method, and using a stemmer as a pre-processing step, achieved outstanding results by significantly reducing the number of features. Specifically, WDO with DT reduced the features to 4231 out of 4616, which was lower compared to our proposed algorithm WDO and SA with DT, where it reduced the features to 8909. HS with DT had 14175 features, SA with DT had 14937 features, PSO with DT had 34787 features, and GA with DT and WOA with DT had the same value of 46167 features without any change in the number of features. On the other hand, the findings revealed that our proposed algorithm WDO and SA with DT, without using a stemmer as a preprocessing step, succeeded in reducing the amount of features and achieved good results. When compared to other algorithms, it produced 7236 out of 91756 features, which was the best result. The distribution of results was as follows: WDO with DT reduced to 9289

TABLE V. THREE CLASSIFIER WITHOUT STEMMER

Classifier	#of features before selected	# of features before selected	Acc.
DT	**91756**	**91756**	**90.298**
KNN	91756	91756	26.119
NB	91756	91756	33.830

TABLE VI. THREE CLASSIFIER WITH STEMMER

Classifier	# of features after selected	# of features before selected	Acc.
DT	46167	46167	**93.781**
KNN	46167	46167	26.368
NB	46167	46167	35.074

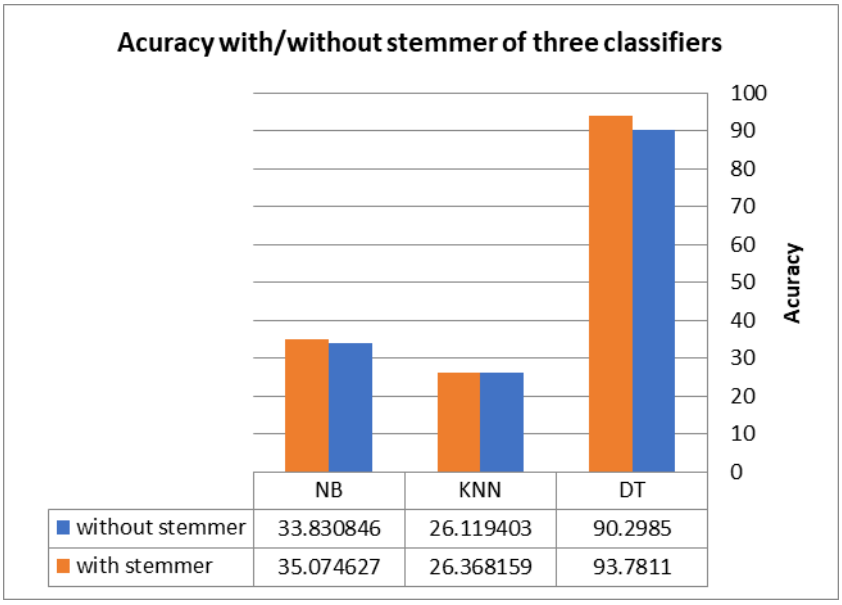


Fig. 9. Accuracy with/without stemmer of three classifiers.

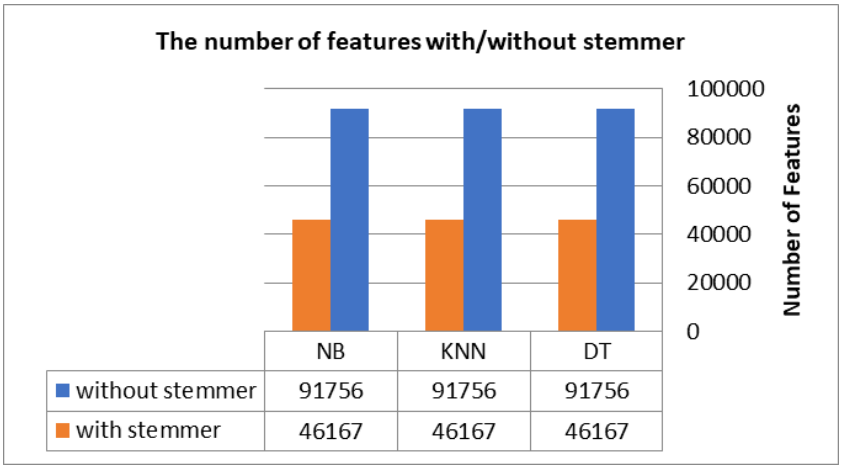


Fig. 10. Number of features with/without stemmer.

features, SA with DT to 19834 features, HS with DT to 28230 features, GA with DT to 45559 features, PSO with DT to 45711 features, and WOA with TD to 45989 features (see Table VII and Table VIII).

When comparing the integration of many metaheuristic algorithms as feature selection methods with DT, with and without stemmer, in terms of accuracy, the algorithms with DT classifier and stemmer achieved better accuracy than those without stemmer as show in Fig. 11.

However, the number of features is another measure for evaluation, It showed that the performance of DT classifier algorithms in feature count reduction was worse when they did not include a stemmer tool, while the algorithms with DT classifier and stemmer performed better, as shown in Fig. 12.

Fig. 13 and Fig. 14 present the results of hybridizing metaheuristic algorithms as feature selection methods with the KNN classifier, in terms of features size and accuracy, with using stemmer and without it as preprocessing. When stemmer was used, it is clear that our suggested algorithm, WDO and SA with KNN classifier, achieved better results compared to other algorithms. Specifically, WDO and SA with KNN obtained an accuracy that was 70.6% higher than WDO and SA with KNN, and HS with KNN, which had the same accuracy of 67.6%. GA with KNN and PSO with KNN had the same accuracy of 25.12%, WOA with KNN had an accuracy of 24.37%, and SA with KNN had an accuracy of 21.64%. However, when without stemmer was used as preprocessing with some optimization algorithms as feature selection, the results revealed that WDO with KNN classifier achieved an accuracy that was 67.6% higher than WDO and SA with KNN (61.9%), HS with KNN (41.5%), GA with KNN (24.6%), PSO with KNN (24.3%), WOA with KNN (24.1%), and SA with KNN (21.6%).

Regarding the size of features, the suggested algorithm WDO and SA with KNN classifier, when using stemmer as preprocessing, demonstrated superior performance as the results showed a significant reduction in the number of features. Specifically, SA with KNN and WDO reduced the features size to 519 out of 46167, which was lower than WDO with KNN (582 features), HS with KNN (633 features), SA with KNN (19528 features), PSO with KNN (23083 features), GA with KNN (23099 features), and WOA with KNN (23143 features). On the other hand, when stemmer was not used as preprocessing, the results showed that WDO and SA with KNN still provided good results in reducing the number of features, with some minor differences. Specifically, WDO and SA with KNN reduced the number of features to 472 out of 91756, while HS with KNN reduced to 243 out of 91756 features, WDO with KNN to 320, SA with KNN to 43859, PSO with KNN to 45454, WOA with KNN to 45939, and GA with KNN to 45994 (see Table IX and Table X).

When comparing different metaheuristic algorithms as feature selection with KNN classifier, using stemmer and without stemmer as pre-processing, it is observed that the algorithms with KNN classifier and stemmer achieved better accuracy, as shown in Fig. 13. However, the number of features is another evaluation measure, which indicates that the results without stemmer were worse in terms of reducing features size, while using stemming is more effective, as results in Fig. 14.

Fig. 15 and Fig. 16 present the results of hybridizing metaheuristic algorithms as feature selection with NB classifier for accuracy and number of features, both with and without the use of a stemmer for preprocessing. With stemmer, it is evident that the proposed algorithm WDO and SA with NB achieved better outcomes in comparison to other algorithms in terms of accuracy. Specifically, WDO and SA with NB showed 38.3% higher accuracy than GA with NB (37.56%), PSO with NB (36.3%), WDO with NB and SA with NB (34.3%), HS with NB (34.07%), and WOA with NB (24.37%).

On the other hand, without stemmer as preprocessing, our proposed algorithm WDO and SA with NB outperformed other algorithms, achieving an accuracy of 37.8%. The results for other algorithms were as follows: GA with NB (33.58%), PSO with NB (33.3%), WDO with NB and HS with NB (33.08%), SA with NB (31.34%), and WOA with NB (31.09%).

Regarding the number of features, the WDO and SA feature selection with NB classifier using stemmer performed well, as evidenced by the results, and reduced the number of features to 5354 out of 46167, which was lower than other algorithms. The results for other algorithms were as follows: HS with NB (15119), WDO with NB (17730), GA with NB (22965), SA with NB (23124), PSO with NB (23162), and WOA with DT (23162). On the other hand, the results without using stemmer showed that the WDO and SA with NB outperformed other algorithms, achieving 6899 features out of 91756, which was less than each of WDO with NB (25589), SA with NB (44358), PSO with NB (45666), GA with NB (45821), WOA with NB (46143), and HS with NB (91756) [see Table XI and Table XII].

In general, when comparing different metaheuristic algorithms as feature selection with NB classifier with and without stemmer, the algorithms with NB classifier with stemmer showed better accuracy, as depicted in Fig. 15. Moreover, the features size is another evaluation measure, which revealed that the results without stemmer were less effective in reducing the features size in comparison to those with stemmer, as shown in Fig. 15 and Fig. 16.

The three classification algorithms, namely NB, KNN, and DT, used in this study were discussed in detail. Additionally, the proposed algorithm, Wind Driven Feature Selection (WDFS), and simulated annealing (SA) were explained in full, including the modifications made to use them as feature selection techniques and the parameters employed in our proposed algorithm. Finally, the results of the research experiments were given, showcasing the impact of using stemmer and feature selection with the three classifiers on our dataset. Furthermore, a comparison was made between the results of WDFS and SA with the three classifiers separately, with/without using stemmer. The findings revealed that using WDFS and SA with stemmer positively affected the accuracy and number of features in classification, leading to improved classifier accuracy and reduced feature count.

VI. Conclusion and Future Work

Arabic text classification is a difficult computational task due to the high content of natural Arabic language text on the Internet. FS is a critical step in the Arabic TC process, as there are thousands of possible features to consider. This study

TABLE VII. DO AND SA WITH DT ALGORITHMS WITHOUT USING A STEMMER

DT	# of features after selected	# of features before selected	Acc.
GA+ DT	91756	45559	0.9502
HS+ DT	91756	28230	0.9029
PSO+ DT	91756	45711	0.9004
WDO+ DT	91756	9289	0.9228
SA+ DT	91756	19834	0.8546
WOA+ DT	91756	45989	0.9029
WDO+SA+DT	91756	**7236**	**0.9502**

TABLE VIII. WDO AND SA WITH DT ALGORITHMS WITH USING A STEMMER

Classifier	# of features after selected	# of features before selected	Acc.
GA+DT	46167	46167	0.945274
HS+DT	46167	14175	0.915423
PSO+DT	46167	34787	0.917910
WDO+DT	46167	**4231**	**0.9701**
SA+DT	46167	14937	0.872698
WOA+DT	46167	46167	0.947761
WDO+SA+DT	46167	**8909**	**0.9701**

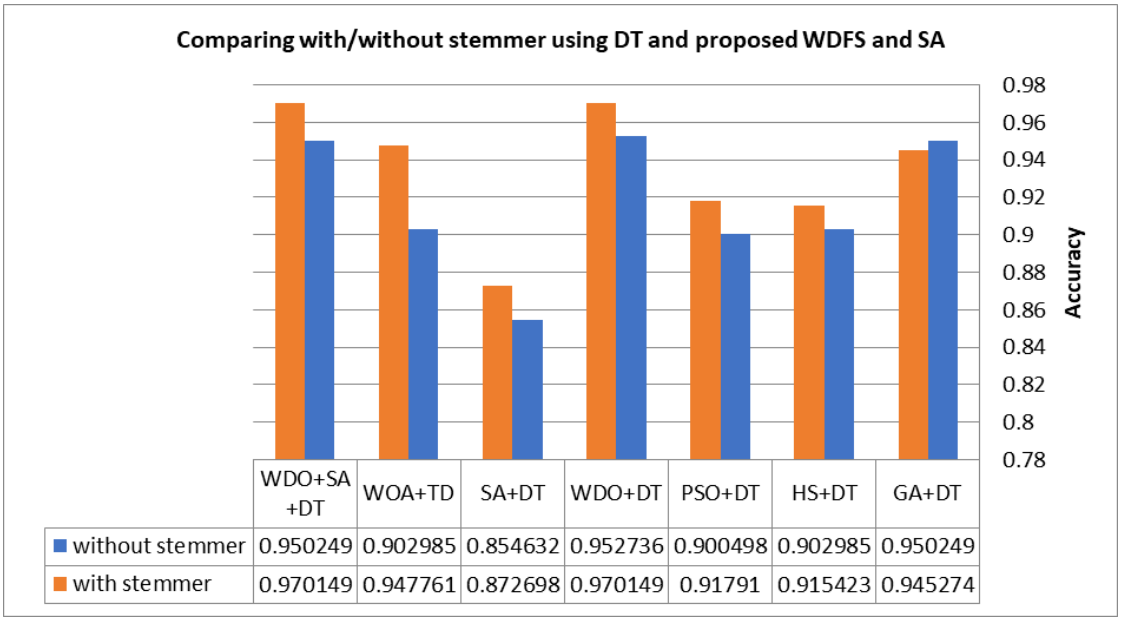


Fig. 11. Comparing with/without stemmer using DT and proposed WDFS and SA.

TABLE IX. WDO AND SA WITH KNN ALGORITHMS WITHOUT USING STEMMER

KNN	# of features after selected	# of features before selected	Acc.
GA+KNN	91756	45994	0.24626
HS+KNN	91756	**243**	**0.41542**
PSO+KNN	91756	45454	0.24378
WDO+KNN	91756	**320**	**0.67661**
SA+KNN	91756	43859	0.21641
WOA+KNN	91756	45939	0.24129
WDO+SA+KNN	91756	**472**	**0.61940**

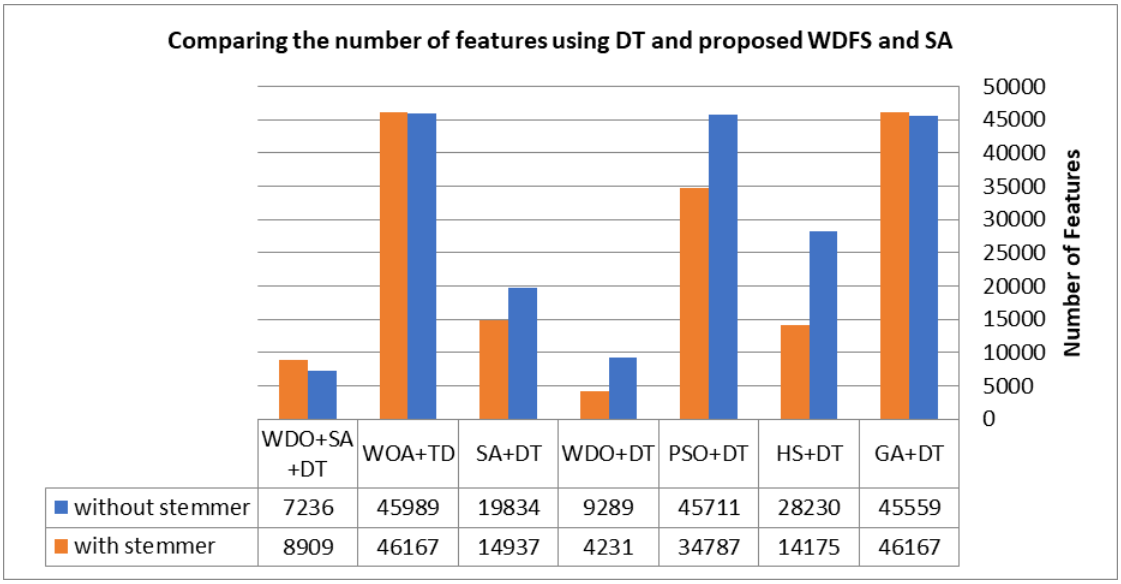


Fig. 12. Comparing the number of features using DT and proposed WDFS and SA.

TABLE X. WDO AND SA WITH KNN ALGORITHMS WITH USING STEMMER

KNN	# of features after selected	# of features before selected	Acc.
GA+KNN	46167	23099	0.2512
HS+KNN	46167	**633**	**0.6766**
PSO+KNN	46167	23083	0.2512
WDO+KNN	46167	**582**	**0.6766**
SA+KNN	46167	19528	0.2164
WOA+KNN	46167	23143	0.2437
WDO+SA+KNN	46167	**519**	**0.7064**

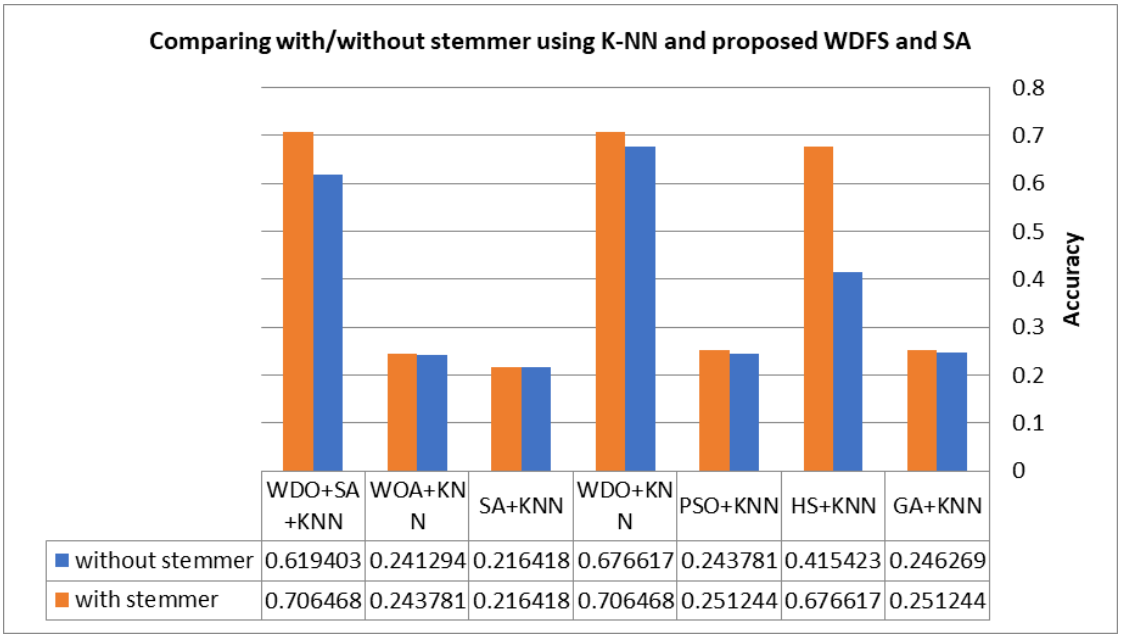


Fig. 13. Comparing with/without stemmer using K-NN and proposed WDFS and SA.

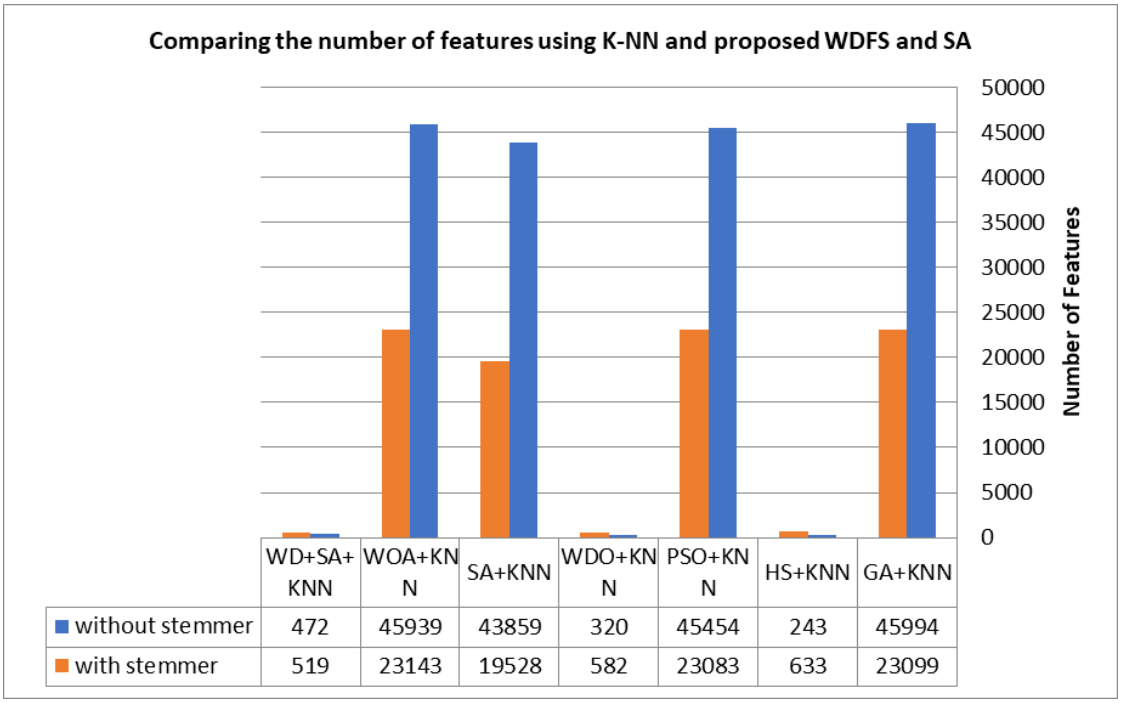


Fig. 14. Comparing the features size using proposed WDFS and SA and K-NN.

TABLE XI. WDO AND SA WITH NB ALGORITHM WITHOUT USING STEMMER

NB	Number of features after selected is	Number of features before selected is	Acc.
GA+NB	91756	45821	0.335821
HS+NB	91756	91756	0.33085
PSO+NB	91756	45666	0.333333
WDO+NB	91756	25589	0.330846
SA+NB	91756	44358	0.313433
WOA+NB	91756	46143	0.310945
WDO+SA+NB	91756	**6899**	**0.378109**

TABLE XII. WDO AND SA WITH NB ALGORITHM WITH USING STEMMER

NB	Number of features after selected is	Number of features before selected is	Acc.
GA+NB	46167	22965	0.375622
HS+NB	46167	15119	0.340796
PSO+NB	46167	23124	0.363184
WDO+NB	46167	17730	0.343284
SA+NB	46167	23124	0.343284
WOA+NB	46167	23162	0.243781
WDO+SA+NB	46167	**5354**	**0.383085**

introduces two wrapper FS methods based on the WDO algorithm for Arabic text classification. The first method uses the binary version of WDO, while the second method improves the binary WDO by hybridizing it with Simulated Annealing (SA) to create WDO-SA. The primary goal of this hybridization is to optimize the best found binary WDO solution after each major iteration. Three famed classifiers, K-NN, SVM, and Naive Bayes, it was used to assess the applicability of each subset of features to Arabic text classification. The proposed methods were validated on a complex Arabic benchmark dataset. The suggested WDO-SA with SVM classifier outperformed other approaches for the FS task in Arabic text categorization, according to the empirical study's findings. Furthermore, investigation results show that incorporating SA into binary WDO methods enhances the search ability to find feasible regions and exploit them. In the future, exploring ensemble machine learning with individual classifiers and investigating different hybridization optimization methods for Arabic text classification could be worthwhile research directions.

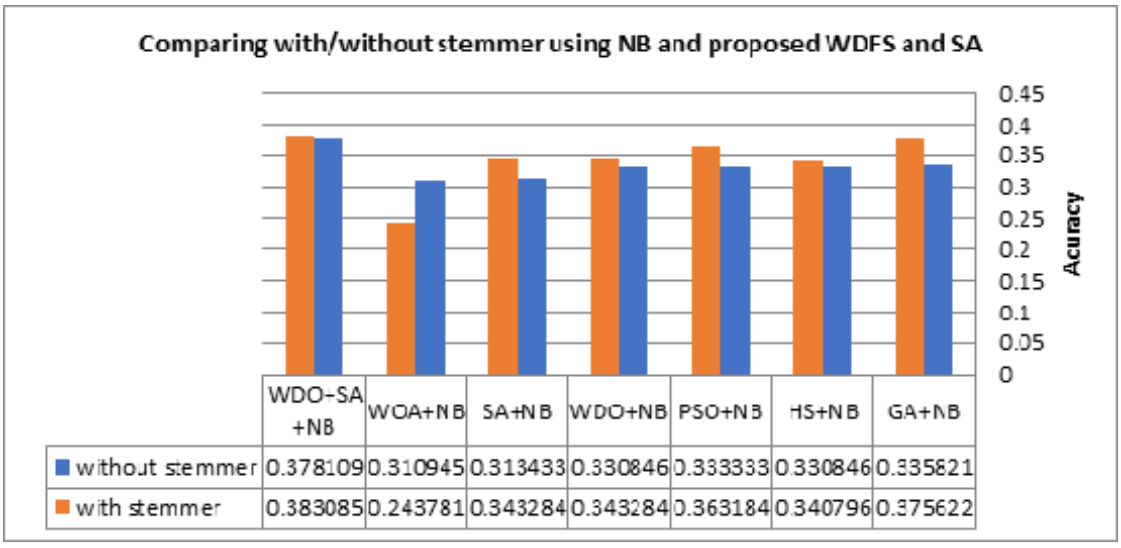


Fig. 15. Comparing with/without stemmer using NB and proposed WDFS and SA.

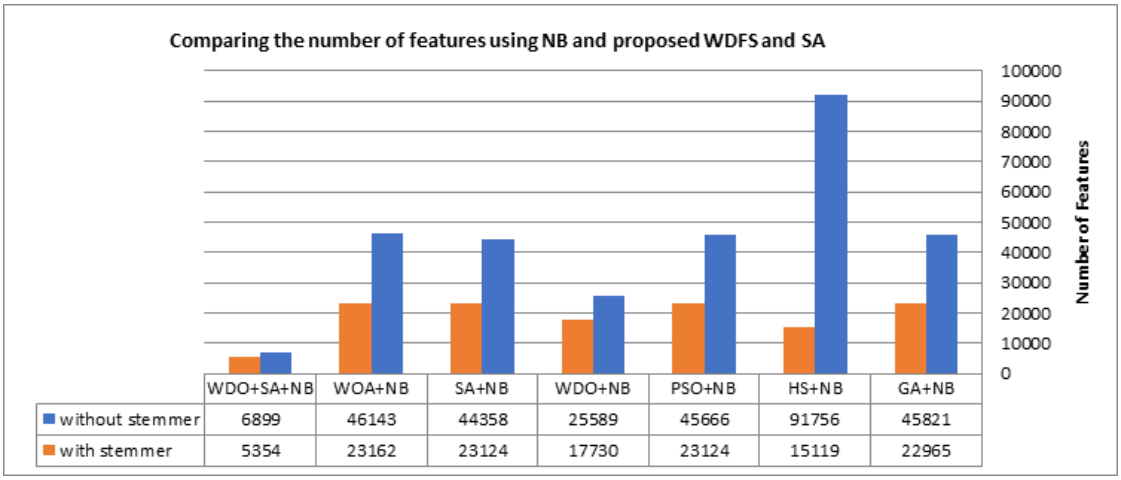


Fig. 16. Comparing the number of features using NB and proposed WDFS and SA.

References

[1] E. G. Dada, J. S. Bassi, H. Chiroma, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, pp. 1802–1802, 2019.

[2] A. Wahdan, S. Hantoobi, S. A. Salloum, and K. Shaalan, "A systematic review of text classification research based on deep learning models in Arabic language," *Int. J. Electr. Comput. Eng*, vol. 10, no. 6, pp. 6629–6643, 2020.

[3] A. Qadi, L. E. Rifai, H. Obaid, S. Elnagar, and A, "Arabic text classification of news articles using classical supervised classifiers," *2nd International Conference on New Trends in Computing Sciences (ICTCS*, pp. 1–6, 2019.

[4] D. Alsaleh and S. Larabi-Marie-Sainte, "Arabic text classification using convolutional neural network and genetic algorithms," *IEEE Access*, vol. 9, pp. 91 670–91 685, 2021.

[5] A. Omar, T. M. Mahmoud, T. Abd-El-Hafeez, and A. Mahfouz, "Multi-label arabic text classification in online social networks," *Information Systems*, vol. 100, pp. 101 785–101 785, 2021.

[6] E. Al-Shammari and J. Lin, "A novel Arabic lemmatization algorithm," *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, pp. 113–118, 2008.

[7] H. Chantar, M. Mafarja, H. Alsawalqah, A. A. Heidari, I. Aljarah, and H. Faris, "Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification," *Neural Computing and Applications*, vol. 32, pp. 12 201–12 220, 2020.

[8] S. Fodeh, B. Punch, and P. N. Tan, "On ontology-driven document clustering using core semantic features," *Knowledge and Information Systems*, vol. 28, pp. 395–421, 2011.

[9] G. Jain, M. Sharma, and B. Agarwal, "Spam detection on social media using semantic convolutional neural network," *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, vol. 8, no. 1, pp. 12–26, 2018.

[10] S. Saha, S. Dasgupta, and S. K. Das, "Spam mail detection using data mining: A comparative analysis," *Smart Intelligent Computing and Applications: Proceedings of the Second International Conference on SCI 2018*, vol. 1, pp. 571–580, 2019.

[11] F. Wang, T. Xu, T. Tang, M. Zhou, and H. Wang, "Bilevel feature extraction-based text mining for fault diagnosis of railway systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 49–58, 2016.

[12] A. El-Mageed, A. A. Gad, A. G. Sallam, K. M. Munasinghe, K. Abohany, and A. A, "Improved binary adaptive wind driven optimization algorithm-based dimensionality reduction for supervised classification," *Computers & Industrial Engineering*, vol. 167, pp. 107 904–107 904, 2022.

[13] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[14] Y. A. Alhaj, J. Xiang, D. Zhao, M. A. A. Al-Qaness, M. A. Elaziz, and A. Dahou, "A study of the effects of stemming strategies on arabic document classification," *IEEE Access*, vol. 7, pp. 32 664–32 671, 2019.

[15] B. Masand, G. Linoff, and D. Waltz, "Classifying news stories using memory based reasoning," *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–65, 1992.

[16] K. Nigam, J. Lafferty, and A. Mccallum, "Using maximum entropy for text classification," *IJCAI-99 Workshop on Machine Learning for Information Filtering*, vol. 1, pp. 61–67, 1999.

[17] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1048–1054, 1999.

[18] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv preprint cs/0205070*, 2002.

[19] Y. A. Alhaj, W. U. Wickramaarachchi, A. Hussain, M. A. A. Al-Qaness, and H. M. Abdelaal, "Efficient feature representation based on the effect of words frequency for Arabic documents classification," *Proceedings of the 2nd International Conference on Telecommunications and Communication Engineering*, pp. 397–401, 2018.

[20] H. Gupta, M. S. Jamal, S. Madisetty, and M. S. Desarkar, "A framework for real-time spam detection in Twitter," *10th International Conference on Communication Systems & Networks (COMSNETS)*, pp. 380–383, 2018.

[21] W. Chebil, M. Wedyan, M. Alazab, R. Alturki, and O. Elshaweesh, "Improving semantic information retrieval using multinomial naive Bayes classifier and Bayesian networks," *Information*, vol. 14, no. 5, p. 272, 2023.

[22] P. Parveen and P. Halse, "Spam mail detection using classification," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, pp. 347–349, 2016.

[23] S. Deepalakshmi and T. Velmurugan, "Empirical study of feature selection methods for high dimensional data," *Indian Journal of Science and Technology*, vol. 9, no. 39, pp. 1–6, 2016.

[24] M. Diale, C. V. D. Walt, T. Celik, and A. Modupe, "Feature selection and support vector machine hyper-parameter optimisation for spam detection," *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference*, pp. 1–7, 2016.

[25] Y. Liu, Y. Wang, L. Feng, and X. Zhu, "Term frequency combined hybrid feature selection method for spam filtering," *Pattern Analysis and Applications*, vol. 19, pp. 369–383, 2016.

[26] G. Raho, R. Al-Shalabi, G. Kanaan, and A. Nassar, "Different classification algorithms based on Arabic text classification: feature selection comparative study," *Int. J. Adv. Comput. Sci. Appl*, vol. 6, no. 2, pp. 23–28, 2015.

[27] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. Khorsheed, and A. Al-Rajeh, "Automatic Arabic text classification," *Proceedings of the international conference on the statistical analysis of textual data*, pp. 77–83, 2008.

[28] M. Mafarja and S. Abdullah, "A fuzzy record-to-record travel algorithm for solving rough set attribute reduction," *International Journal of Systems Science*, vol. 46, no. 3, pp. 503–512, 2015.

[29] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, 2017.

[30] M. Spitters, "Comparing feature sets for learning text categorization," *Content-Based Multimedia Information Access*, vol. 2, pp. 1124–1134, 2000.

[31] F. Yang, T. Sun, and C. Zhang, "An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization," *Expert Systems with Applications*, vol. 36, no. 6, pp. 9847–9852, 2009.

[32] A. Sharaff, N. K. Nagwani, and K. Swami, "Impact of feature selection technique on email classification," *Int. J. Knowl. Eng*, vol. 1, no. 1, pp. 59–63, 2015.

[33] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 2, pp. 225–231, 2020.

[34] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Text feature selection using ant colony optimization," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6843–6853, 2009.

[35] S. L. Marie-Sainte and N. Alalyani, "Firefly algorithm based feature selection for Arabic text classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 3, pp. 320–328, 2020.

[36] H. Majidpour and F. S. Gharehchopogh, "An improved flower pollination algorithm with AdaBoost algorithm for feature selection in text documents classification," *Journal of Advances in Computer Research*, vol. 9, no. 1, pp. 29–40, 2018.

[37] P. Shunmugapriya and S. Kanmani, "A hybrid algorithm using ant and bee colony optimization for feature selection and classification (AC-ABC Hybrid)," *Swarm and evolutionary computation*, vol. 36, pp. 27–36, 2017.

[38] E. A. H. Khalil, E. H. F. Houby, and H. K. Mohamed, "Deep learning for emotion analysis in Arabic tweets," *J Big Data*, vol. 8, pp. 136–136, 2021.

[39] I. Targio, M. Hashem, and Tubishat, "Hybrid Feature Selection Based on Principal Component Analysis and Grey Wolf Optimizer Algorithm for Arabic News Article Classification," *IEEE Access*, vol. 10, pp. 121 816–121 830, 2022.

[40] P. Dhal and C. Azad, "A fine-tuning deep learning with multi-objective-based feature selection approach for the classification of text," *Neural Comput & Applic*, 2023.

[41] I. Jamaleddyn, R. E. Ayachi, and M. Biniz, "An improved approach to Arabic news classification based on hyperparameter tuning of machine learning algorithms," *Journal of Engineering Research*, vol. 11, no. 2, pp. 2023–2023.

[42] A. Ayedh, G. Tan, K. Alwesabi, and H. Rajeh, "The effect of preprocessing on arabic document categorization," *Algorithms*, vol. 9, no. 2, pp. 27–27, 2016.

[43] B. Alhadidi and M. Alwedyan, "Hybrid Stop-Word Removal Technique for Arabic Language." *Egypt. Comput. Sci. J.*, vol. 30, no. 1, pp. 35–38, 2008.

[44] A. Jaffar, M. Wedyan, and H. Al-Zoubi, "Arabic text light stemmer," *Int. J. Comput. Acad. Res*, vol. 8, no. 2, pp. 17–23, 2019.

[45] J. Atwan, M. Wedyan, and H. Al-Zoubi, "Arabic text light stemmer," *Int. J. Comput. Acad. Res*, vol. 8, no. 2, pp. 17–23, 2019.

[46] J. Atwan, M. Wedyan, Q. Bsoul, A. Hamadeen, R. Alturki, and M. Ikram, "The effect of using light stemming for Arabic text classification," *International Journal of Advanced Computer Science and Applications*, no. 5, pp. 12–12, 2021.

[47] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Light stemming for Arabic information retrieval," in *Arabic computational morphology: knowledge-based and empirical methods*, and others, Ed. Springer, 2007, pp. 221–243.

[48] M. A. H. Omer and S. Ma, "Stemming algorithm to classify Arabic documents," *Journal of Communication and Computer*, vol. 7, no. 9, pp. 1–5, 2010.

[49] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[50] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[51] W. G. Indriyani and A. Rakhmadi, "Filter-wrapper approach to feature selection using PSO-GA for Arabic document classification with Naive Bayes multinomial," *IOSR Journal of Computer Engineering*, vol. 17, no. 6, pp. 45–51, 2015.

[52] K. C. Lin, K. Y. Zhang, Y. H. Huang, J. C. Hung, and N. Yen, "Feature selection based on an improved cat swarm optimization algorithm for big data classification," *The Journal of Supercomputing*, vol. 72, pp. 3210–3221, 2016.

[53] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[54] R. D. Thompson, "Atmospheric processes and systems," and others, Ed. Psychology Press, 2002.

[55] Z. Bayraktar, M. Komurcu, and D. H. Werner, "Wind Driven Optimization (WDO): A novel nature-inspired optimization algorithm and its application to electromagnetics," *IEEE Antennas and Propagation Society International Symposium*, pp. 1–4, 2010.

[56] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.
[57] F. Y. Vincent, S. W. Lin, W. Lee, and C. J. Ting, "A simulated annealing heuristic for the capacitated location routing problem," *Computers & Industrial Engineering*, vol. 58, no. 2, pp. 288–299, 2010.
[58] R. Jensen and Q. Shen, "Fuzzy-rough attribute reduction with application to web categorization," *Fuzzy Sets and Systems*, vol. 141, no. 3, pp. 469–485, 2004.
[59] N. S. Altman, "An introduction to kernel and nearest-neighbor non-parametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.
[60] T. M. Mitchell, "Machine Learning," in *Machine Learning*, vol. 1, 2007.
[61] A. Mccallum and K. Nigam, "A comparison of event models for naive bayes text classification," *AAAI-98 Workshop on Learning for Text Categorization*, vol. 752, pp. 41–48, 1998.
[62] P. Y. Pawar and S. H. Gawande, "A comparative study on different types of approaches to text categorization," *International Journal of Machine Learning and Computing*, vol. 2, no. 4, pp. 423–423, 2012.
[63] J. Su and H. Zhang, "A fast decision tree learning algorithm," in *Aaai*, and others, Ed., vol. 6. McGraw-Hill Science, 2006, pp. 500–505.
[64] D. Graff and K. Walker, "Linguistic Data Consortium (LDC)," 2001. [Online]. Available: Https://Catalog.Ldc.Upenn.Edu/LDC2001T55
[65] T. Hastie, R. Tibshirani, and J. H. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," *Biometrics*, vol. 66, no. 4, p. 746, 2010.

Comparison of Conventional Techniques for House Electricity Consumption Forecasting

Sandra Pajares Centeno, Hugo Alatrista-Salas
Escuela De Posgrado Newman, Tacna, Peru

***Abstract*—Electricity consumption monitoring is the automated process of recording, processing, and analyzing electricity usage in real time to make informed decisions. This research aims to implement an artificial intelligence- and deep learning-based methodology to forecast monthly electricity consumption in Tacna, Peru, and generate decision-making indicators. To this end, we used electricity consumption records from Electrosur S.A., the company responsible for electricity distribution and marketing in the departments of Tacna and Moquegua, from February 2015 to December 2022 (a total of 95 months). We compared three artificial intelligence models in this context: (i) eXtreme Gradient Boosting (XGBoost), (ii) Light Gradient Boosting (LGBM), and (iii) Prophet. While all models effectively forecasted electricity consumption, the Prophet model demonstrated superior performance, achieving a mean absolute percentage error (MAPE) of 0.7% compared to actual consumption values. Additionally, the study discusses the potential of recurrent neural networks to further enhance predictive accuracy.**

***Keywords*—Electricity consumption; forecasting; recurrent neural networks; deep learning**

I. Introduction

According to a report by the United Nations Department of Economic and Social Affairs, cities account for up to 80% of global energy consumption and 75% of carbon emissions, despite covering only 3% of the Earth's surface. This problem is prevalent in various countries, including in Peru, which emitted over 91.4 million tons of greenhouse gases (GHGs) in 2018. Peru also has one of the highest per capita GHG emissions in South America, at 2.8 tons. The indiscriminate use of electricity has resulted in our cities having one of the highest carbon footprints in Latin America. Additionally, international organisations such as the UN and the OAS are prioritising issues related to carbon emissions and climate change mitigation.

An increasing number of companies and government agencies are using artificial intelligence to improve the decision-making process [1]. In this sense, extracting information from data representing complex dynamics has been used to predict future events and make informed decisions in several areas, as discussed in [2], [3], [4], [5].

Electricity distribution and marketing companies, in particular, have processes that require continuous monitoring because the results of these processes directly impact the environment [6]. Therefore, it is essential to forecast energy consumption in the short, medium, and long term for environmental impact studies and capacity expansion. Power or energy consumption monitoring involves the automated recording, processing, and analysis of power network data in real time. The purpose of monitoring is to enable informed decision-making. Several algorithms have been implemented to predict electricity consumption [7], [8], [9], [10].

Energy consumption analysis has been applied to various fields, including the analysis of traffic conditions [11], the energy consumption of IoT devices [12], [13], the estimation of electric vehicle consumption [14], [15], and the use of aerial imagery to predict energy consumption [16].

This paper focuses on forecasting time series data. Apart from the classification proposed in [17], machine learning and deep learning techniques have become very relevant in recent years for various prediction tasks [18], [19]. Many works have been proposed for time series prediction tasks, particularly for energy consumption estimation For example, the authors in [20] used deep neural networks, specifically long short-term memory (LSTM) neural networks. In their paper, the authors compare two ways of representing data: classical LSTMs with time series data and sequences using the sequence-to-sequence (S2S) method. Similarly, LSTM models have improved thanks to algorithms based on the bidirectional long-short-term memory (Bi-LSTM) model [21]. In this context, this paper compares three well-known techniques for predicting energy consumption and generating indicators to facilitate better energy management and decision-making in Tacna, Peru.

After the introduction, Section II summarizes the existing literature on the topic. Section III presents the definitions and conceptual framework. Section IV describes the methodology used in this study, and Section V details the dataset and results. The work concludes with the conclusions and suggestions for future work.

II. Related Works

Various authors have used machine learning and deep learning techniques in their literature to predict electricity usage in residential, commercial, and public buildings. For instance, [8] use algorithms such as Support Vector Machine (SVM) and decision trees to predict household-level electricity consumption in the US. Similarly, [10] use support vector regression to estimate electricity consumption in a Turkish city. In [22], the authors use the Random Forests algorithm to predict short-term energy consumption based on five datasets collected over a year from various buildings. Similar studies have been conducted on houses in various countries, including China [23], Malaysia [24], Brazil and Spain [25], France [26], Albania [27] and several countries simultaneously [28], [29].

Additionally, some authors have used a variety of techniques to estimate electrical energy consumption. For example,

in [7], the authors proposed a new fuzzy wavelet neural network model to analyse hourly electricity consumption. The model was validated using a database of electricity consumption from the Greek island of Crete. Similarly, [8] propose a method based on neural network back propagation and the improved particle swarm optimization (IPSO) algorithm for predicting electricity consumption and measuring the impact of changes and imbalances in energy consumption caused by the pandemic.

Other papers focus on more advanced techniques, such as deep learning. For instance, [30] the authors use *Long Short-Term Memory* In this text, the authors discuss the use of LSTM neural networks to predict household electricity consumption and their potential to reduce carbon emissions. The authors also discuss the potential of these networks to reduce carbon emissions. Additionally, they reference a study by [31] that proposes a model for predicting electric vehicle energy consumption using transfer learning. The proposed method utilizes data from the target domain as well as data or a model from a similar domain. The authors used data on electric vehicle trips, including the battery charge state, departure and arrival times at rest areas, and GPS data similar to the predicted training data.

In [32], three energy consumption prediction algorithms were compared: Asynchronous Advantage Actor-Critic (A3C), Deep Deterministic Policy Gradient (DDPG), and Recurrent Deterministic Policy Gradient (RDPG). The algorithms were tested on data collected from an office building. In this context, [33] predict monthly consumption using three deep learning models: Deep Fully Connected, Convolutional Neural Network, and Long Short-Term Memory Neural Networks. The authors also employed pre-processing techniques, such as normalization, to homogenize the input domain. Finally, they applied the algorithms to over 9 million sample datasets. In [34] study, data from 100 civil public buildings in China was used to predict short and medium-term electrical energy consumption. The authors proposed an algorithm based on a deep neural network and compared it with other algorithms such as the back-propagation algorithm, the Elman neural network, and the fuzzy neural network. In another study, [35] proposed integrating two deep learning models, including a long short-term memory (LSTM) model. They claim that combining LSTM and reinforcement learning agents (reinforced learning) is more efficient for predicting energy consumption at the University of Wollongong in Australia.

Additionally, several efforts have been made to propose efficient algorithms in terms of both prediction quality and ease of interpretation, particularly when the results are used to make informed decisions. For example, in [36], the authors use three interpretable models based on long short-term memory (LSTM) with self-attention to predict electrical energy consumption in a dataset related to an office building. Similarly, in [37], the authors use the SHAP tool to analyze the electrical energy consumption of an electric arc furnace used to produce stainless steel after an artificial neural network (ANN) makes a prediction. Through this study, the authors aim to provide practical examples of how SHAP can be used to analyse the contributions of each input variable to the consumption predictions generated by black box machine learning models when the data are associated with steel processes.

Furthermore, the authors of [38] propose a four-component methodology to predict and explain electrical energy consumption over approximately two million minutes collected from 2006 to 2010. These components include a power encoder, an auxiliary encoder for a two-dimensional, interpretable latent space, a deep-learning-based prediction model, and an explainer that identifies the most significant input features for predicting future demand. Finally, in [39], the authors focus on improving the efficiency of existing algorithms - implemented in the Darts library - for predicting the daily, weekly and monthly electricity consumption of 43 customers using the Neural Basis Expansion Analysis for Interpretable Time Series (N-BEATS) technique.

This research underscores the practical importance of adopting automated forecasting techniques for predicting electricity consumption, especially in contexts where current practices still depend on manual projections and basic trend extrapolations. Unlike complex models such as ARIMA, recurrent neural networks (RNNs), or Transformer-based architectures, which often require extensive parameter tuning and substantial computational resources, our findings demonstrate that simpler, well-established models like Prophet, XGBoost, and LGBM can deliver highly accurate results when applied to robust, well-structured datasets. This suggests that advanced forecasting performance does not always require sophisticated models, especially when the data is consistent in quality and time-based patterns. Our objective is not to conduct an extensive comparison of forecasting techniques, propose novel methods, or introduce new evaluation metrics. Instead, we aim to present a practical, evidence-based report reviewing existing algorithms and demonstrating the effectiveness of three simple models with promising results[1].

III. Basic Definitions

A. Electric Energy Consumption: Electrosur Tacna, Peru

The electricity balance illustrates the flow of electricity from generation to final consumption. Electricity travels through transmission and distribution networks. In 2022, Peru generated $59,713$ GWh of electricity and imported 32 GWh from Ecuador. The electrical energy balance, including production, consumption, and losses, is listed below:

- Of the total energy produced in Peru, $57,814$ GWh (97%) came from companies in the electricity market, while $1,898$ GWh (3%) came from industrial companies that produce electricity for their own use.
- Of the total electrical energy available, which includes national production plus imports, 1.4% is used for auxiliary services in power stations. Meanwhile, 11.1% is lost in transmission and distribution, leaving 87.4% to reach the final consumer.
- Of the energy produced by enterprises in the electricity market, 1.2% was used for auxiliary services in power stations, 11.3% was lost and 87.5% reached final consumers.
- Of the energy produced by industrial enterprises for their own use, 7.2% was used for auxiliary services

[1]The code and data are available at https://github.com/huvaso/electricityForecasting

of their plants, 5.78% was lost and 87.7% was used for their operations.

- Of the $48,092$ GWh of electrical energy sold by generation and distribution companies to final consumers, 38% ($18,410$ GWh) was sold on the regulated market, while 62% ($29,682$ GWh) was sold on the free market.

Electrosur S.A.'s main activity is distributing and marketing electrical energy in its concession area, which encompasses the Tacna and Moquegua departments. The company purchases the energy it distributes and markets from nine generation companies and the National Interconnected System of Peru.

In 2022, energy sales amounted to $399,786$ MWh, an increase of 2.3% compared to 2021. This increase is due to the progressive and sustained rise in consumption by both regulated and free customers, resulting from the economic reactivation of commercial activities in Tacna, Moquegua, and Ilo. ENEL Generación Perú was the company's main energy supplier in 2022.

In Peru, the forecast for electrical energy purchases (demand) is currently based on historical data projections, considering new user demand and population growth rates. The electricity distribution company's energy efficiency can be estimated by dividing its energy sales and purchases into periods. Therefore, its energy efficiency is estimated to be 89.97% in the 2022 financial year. One of the main activities undertaken in 2022 to reduce technical losses was the installation of two reactive compensation banks, one in Tacna with a total capacity of 300 kWh and one in Ilo with a total capacity of 300 kWh, for a total capacity of 600 kWh. This reduced technical energy losses to $36,500$ kWh per year.

Another important activity in reducing technical losses has been installing LED lights. A total of $14,084$ units were installed in Tacna ($7,452$), Moquegua ($2,863$), and Ilo ($3,769$). Due to the replacement of LED lights and the program to control losses, change drivers, recover energy, prevent illegal connections, and provide exceptional services, Electrosur S.A. is now among the electricity distribution companies in Peru with the lowest losses, below the average of 87.46%. Additionally, the loss of electrical energy was less than 10%, corresponding to $44,536$ MWh, which equates to approximately 15 million soles in economic terms.

Therefore, using historical data from a specific period and machine learning algorithms will enable more accurate purchase forecasts for annual planning and more confident economic and operational decisions. In terms of energy efficiency, correctly forecasting electricity demand will enable households to reduce their electricity bills.

B. Conventional Classifiers

1) XGBoost (eXtreme Gradient Boosting): This is an open-source library for performing machine learning tasks, especially regression and classification problems. It is based on the boosting algorithm, which combines several weaker models to form a stronger one. XGBoost uses decision trees and employs advanced techniques such as regularisation and gradient optimisation to improve model accuracy and performance. It is also widely used in data science competitions due to its effectiveness.

2) LGBM (LightGBM): This is a decision tree-based machine learning library designed to be efficient and scalable. Unlike XGBoost, LightGBM uses a leaf-wise rather than level-wise growth approach, which significantly improves training speed. It is designed to handle large datasets and high dimensionality.

3) Prophet: This is an open-source library developed by Facebook for time series forecasting. It is designed to handle time series data that may have trends, seasonality, holidays and also allows the inclusion of special events. Prophet can handle missing data and outliers, and it is suitable for short- and medium-term time series forecasting.

C. Evaluation Metrics

After training the model, it is essential to evaluate its performance on unseen data using evaluation metrics. The choice of metrics depends on the problem's specific context and preferences regarding error interpretation. The objective is to minimize these metrics' values to enhance the prediction model's accuracy.

1) MAE (Mean Absolute Error): measures the average size of absolute errors between predictions and actuals in a time series. It is useful for understanding the average size of errors regardless of their direction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

Where: n is the number of observations; y_i is the actual value at time i; and $\hat{y}_i$ is the prediction at time i.

2) MSE (Mean Squared Error): measures the average squared error between predictions and actual values in a time series. The squared term emphasizes larger errors, making it more sensitive to outliers.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

Where: n is the number of observations; y_i is the actual value at time i; and $\hat{y}_i$ is the prediction at time i.

3) MAPE (Mean Absolute Percentage Error): measures the average percentage of absolute errors with respect to the actual values. It is useful to evaluate the relative error in percentage terms.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{|y_i|}\right) \times 100 \tag{3}$$

Where: n is the number of observations; y_i is the actual value at time i; and $\hat{y}_i$ is the prediction at time i.

IV. Methodology

Predictive techniques based on temporal data are now fundamental to data science and artificial intelligence because they can model and predict patterns, trends, and behaviors in time series data.

In this context, and after analysing the work presented in the literature [40], the methodology described in Figure 1 was proposed in this paper. The dataset contains the periodic electricity consumption of each subscriber in kWh. The dataset was preprocessed before building a time series. This process can include various steps, such as filling in missing data, detecting outliers, and normalizing. This step depends on the dataset. Next, the time feature is preprocessed. Depending on the algorithm, the dataset must meet certain format specifications. Then, the time series is built according to the desired temporal granularity (daily, monthly, etc.). After building the time series, three forecasting algorithms were compared using the quality metrics described in the previous section. Finally, the predictions were visualized to better understand consumption behavior patterns.

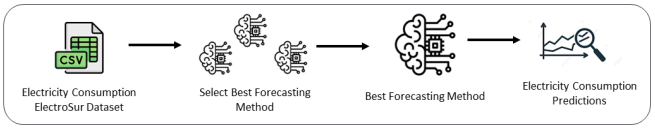


Fig. 1. Proposed methodology.

V. Experiments and Results

This section describes the available dataset. Additionally, the main findigs of our approach were detailed and discussed.

A. Dataset

The required data for the study was provided by Electrosur S.A., Tacna's electricity distribution company in Peru. The data, which corresponded to the electricity consumption data for the district, province, and department of Tacna, was obtained from the commercial system database. It is worth noting that the district's users are subject to a residential tax.

The research will examine data from February 2015 to December 2022. This includes the supply number and active energy consumption in kilowatt-hours (kWh) per user for each month and year within this period. The supply number is a code that identifies the service user. It is also known as the customer number that appears on the electricity consumption voucher. Active energy consumption (AEC) is measured in kWh and corresponds to a user's energy consumption as recorded by an electricity meter, also known as a wattmeter, which is calibrated in billing units. AEC is calculated by subtracting the electric meter reading from the previous month from the current month's reading, resulting in the billed consumption for the month.

It is important to note that there are $8,171$ users, each with 88 months of billing information. These records correspond to active energy consumption from February 2015 to December 2022, resulting in 719,048 total records of electrical energy consumption.

Table I presents the main statistical characteristics of the data by year.

As expected, the number of services increases each year due to the progressive growth of urban centers. However, contrary to urban growth, the average electricity consumption expenditure in Tacna is decreasing. Figure 2 shows a slight increase in 2019; however, the average consumption decreased during the period of the 2020-2021 (Covid-19) pandemic. Furthermore, the maximum annual electricity consumption value in Tacna increases in 2020 and 2021, during the period of the pandemic. Note that the quarantine period began in March 2020.

B. Results

First, any missing, null, or duplicate data in the dataset was addressed. This cleaning process is essential to ensure that the models are based on high-quality data. Next, a data transformation process was performed. This process includes converting data into appropriate formats, creating new features from existing ones, and rescaling variables to facilitate modeling. Since the number of houses has increased over the years, our particular dataset focuses on only those houses with records for each of the 95 months.

Additionally, to explore which model works best with our dataset, the average monthly consumption of all houses was calculated. This produced one time series representing the average consumption of houses in the Tacna districts over the 95 months (see Figure 3).

The training set is used to train the model, while the validation set is used to evaluate its performance (see figure 4). When working with time series, the data set must be divided according to chronological order, not randomly. In this sense, our data sets correspond to:

- Training set: Consumption corresponding to the dates from February 2012 to May 2021 (76 months).
- Validation set: Consumption corresponding to the period from June 2021 to December 2022 (19 months).

Identifying the features (independent variables) that will be used to predict the target variable (dependent variable) is essential for training the model. Then a machine learning algorithm is used to fit the model to the training data. The model learns the patterns and relationships in the data that can be used for predictions.

Moreover, three learning algorithms were chosen: eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting (LGBM), and Prophet. The Mean Absolute Error, Mean Squared Error, and Mean Absolute Percentage Error were used to evaluate performance.

To train the model with XGBoost and LightGBM, new features were generated from the data: Year, Month and Quarter. An example of these new features is shown in Table II.

The Prophet algorithm generates features related to trends and seasonality. These features relate to different components of the model and their prediction intervals: (i) The trend component captures the general direction of the time series.

TABLE I. STATISTICAL FEATURES OF THE DATA SET BY YEAR

Features	2015	2016	2017	2018	2019	2020	2021	2022
Count	6004.00	6156.00	6346.00	6676.00	7437.00	7506.00	8013.00	8247.00
Mean	123.49	123.43	122.18	116.98	117.91	117.23	110.68	106.30
Std.Deviation	175.54	185.99	185.04	172.37	190.08	183.41	197.29	181.57
Minumum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	38.63	38.58	39.16	36.39	34.33	34.75	28.75	27.33
50%	91.13	89.00	89.33	84.83	82.41	84.75	77.83	73.33
75%	153.13	150.83	149.02	144.33	142.08	144.33	136.83	129.66
Maximum	4156.45	4312.58	4483.83	3706.41	4052.41	4327.09	9000.00	5075.00

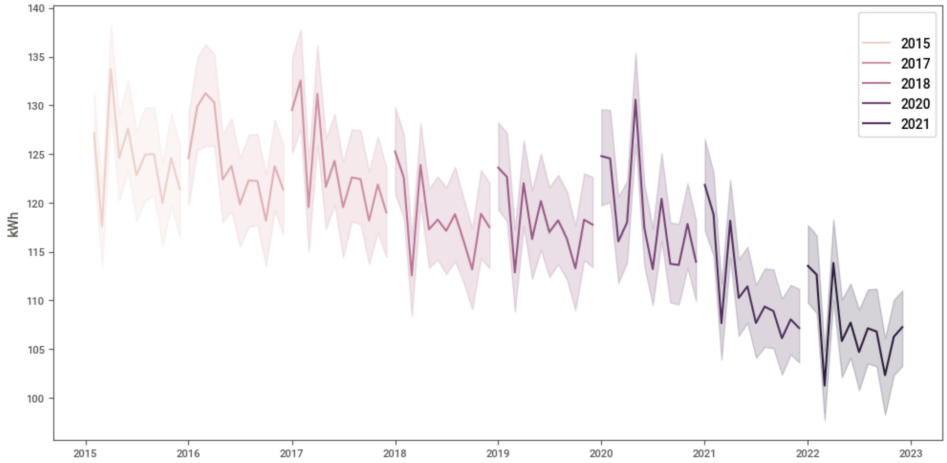


Fig. 2. Electricity consumption per year.

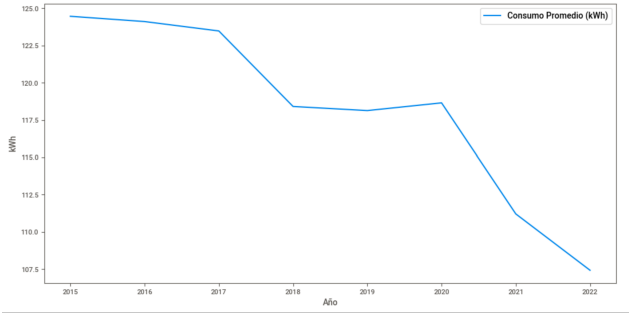


Fig. 3. Time series representing the mean consumption in Tacna.

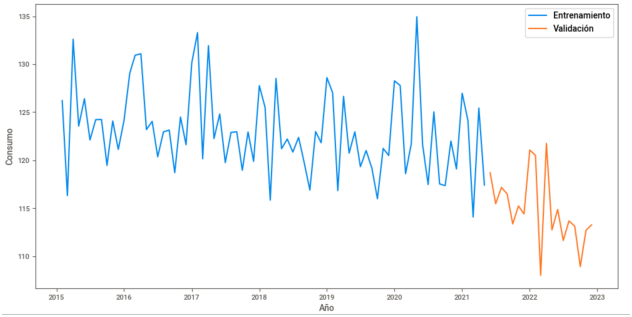


Fig. 4. Splitting the dataset.

It represents the long-term growth or decline trajectory in the data. (ii) Variables representing the additive components of the model, including seasonality and holiday effects. (iii) Variables representing the annual component of the model, which captures seasonal patterns over the years; and (iv) variables representing the multiplicative components of the model.

Table III shows the values obtained for each method according to the evaluated metrics. Based on these results, we can conclude that the Prophet model is the best. Furthermore, Figure 5 shows that the Prophet model's predictions are closer to the actual consumption values than those of the other methods.

TABLE II. NEW ADDITIONAL FEATURES

Period	Consumptions	**Four-month Period**	**Month**	**Year**
2015-02-01	126.208909	1	2	2015
2015-03-01	116.323349	1	3	2015
2015-04-01	132.613204	2	4	2015
2015-05-01	123.545468	2	5	2015
2015-06-01	126.405208	2	6	2015

C. Discussion

The current methodology for forecasting electricity demand in Peru relies primarily on extrapolating historical data while

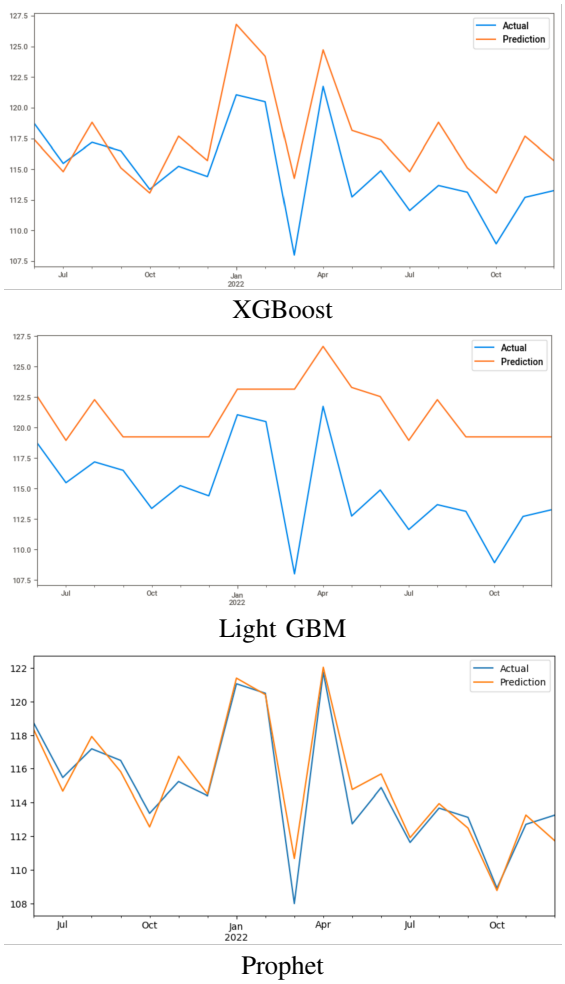


Fig. 5. Forecasting of electricity consumption in validation dataset.

TABLE III. METRICS EVALUATED FOR THE THREE ALGORITHMS

Algorithm	MAE	MSE	MAPE
XGBoost	3.026	12.264	0.027
LGBM	6.196	48.351	0.055
Prophet	0.771	1.056	0.007

considering demographic growth and the addition of new users. While this traditional approach provides a baseline estimation, it often lacks the granularity and adaptability required for dynamic energy consumption environments. In this context, this study research proposes a data-driven methodology that uses machine learning and statistical forecasting models, namely eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting (LGBM), and Prophet, to improve the accuracy of electricity consumption forecasts in Tacna, Peru.

The results indicate that the Prophet model outperforms the other two algorithms across all evaluated metrics (MAPE, MAE, and MSE), with a particularly low MAPE (see Table III), demonstrating its strong suitability for capturing seasonal and trend-based patterns in time series data. The proposed methodology benefits from several strengths. First, the dataset is extensive and consists of monthly active energy consumption records from $8,171$ users over an 8-year period, resulting in over 700,000 observations. This rich dataset allows for robust model training and evaluation. Second, the inclusion of real operational data from Electrosur S.A., a key electricity provider, ensures practical relevance and reflects actual consumption patterns in the region.

However, some limitations must be acknowledged. While the proposed models are effective in forecasting, they do not yet incorporate exogenous factors, such as weather patterns, economic activity, or special events, that could affect consumption behavior. Additionally, while Prophet shows excellent performance, its black-box nature may limit its usefulness to decision-makers who prefer transparent models. From a policy and operational perspective, improving forecasting accuracy could support Electrosur S.A.'s ongoing efforts to reduce technical losses, such as installing reactive compensation banks and LED lighting. Enhanced forecasting could also support targeted interventions for high-consumption users, energy-saving initiatives, and more efficient demand planning.

VI. CONCLUSION

This study presented a data-driven approach to forecasting monthly electricity consumption in Tacna, Peru. The study used nearly eight years of historical data from the Peruvian Electricity Administration Office. Implementing and comparing three machine learning models demonstrated the efficacy of AI-based forecasting techniques in the energy sector. Of the evaluated models, Prophet demonstrated the highest predictive accuracy, achieving a mean absolute percentage error (MAPE) of 0.7%, highlighting its robustness for time series forecasting in this domain.

Future work will focus on exploring deep learning approaches, particularly recurrent neural networks (RNNs) and transformers, to better model temporal consumption patterns, as discussed in [41], [42]. Comparative analyses will be conducted to assess the performance of these models against traditional ensemble methods such as XGBoost and LGBM. Finally, a real-time alert system to identify users with potentially high electricity consumption. This system will allow us to provide targeted energy-saving recommendations and support demand management efforts.

REFERENCES

[1] M. Havey, *Essential business process modeling*. " O'Reilly Media, Inc.", 2005.

[2] J. Bascompte and R. V. Solé, "Rethinking complexity: modelling spatiotemporal dynamics in ecology," *Trends in Ecology & Evolution*, vol. 10, no. 9, pp. 361–366, 1995.

[3] A. McGovern, D. John Gagne, N. Troutman, R. A. Brown, J. Basara, and J. K. Williams, "Using spatiotemporal relational random forests to improve our understanding of severe weather processes," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 4, no. 4, pp. 407–429, 2011.

[4] K. S. Oberoi and G. Del Mondo, "Graph-based pattern detection in spatio-temporal phenomena," in *16th Spatial analysis and geomatics conference (SAGEO 2021)*, 2021.

[5] K. V. Rao, A. Govardhan, and K. C. Rao, "Spatiotemporal data mining: Issues, tasks and applications," *International Journal of Computer Science and Engineering Survey*, vol. 3, no. 1, p. 39, 2012.

[6] J. W. Lee, "The contribution of foreign direct investment to clean energy use, carbon emissions and economic growth," *Energy policy*, vol. 55, pp. 483–489, 2013.

[7] M. Amina, V. Kodogiannis, I. Petrounias, and D. Tomtsis, "A hybrid intelligent approach for the prediction of electricity consumption," *International Journal of Electrical Power & Energy Systems*, vol. 43, no. 1, pp. 99–108, 2012.

[8] S. Nallathambi and K. Ramasamy, "Prediction of electricity consumption based on dt and rf: An application on usa country power consumption," in *2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE)*. IEEE, 2017, pp. 1–7.

[9] X. Li, Y. Wang, G. Ma, X. Chen, J. Fan, and B. Yang, "Prediction of electricity consumption during epidemic period based on improved particle swarm optimization algorithm," *Energy Reports*, vol. 8, pp. 437–446, 2022.

[10] K. Kavaklioglu, "Modeling and prediction of turkey's electricity consumption using support vector regression," *Applied Energy*, vol. 88, no. 1, pp. 368–375, 2011.

[11] P. Li, Y. Zhang, K. Zhang, and M. Jiang, "The effects of dynamic traffic conditions, route characteristics and environmental conditions on trip-based electricity consumption prediction of electric bus," *Energy*, vol. 218, p. 119437, 2021.

[12] G. Bedi, G. K. Venayagamoorthy, and R. Singh, "Development of an iot-driven building environment for prediction of electric energy consumption," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 4912–4921, 2020.

[13] A. Chaudhari, B. Rodrigues, and S. More, "Automated iot based system for home automation and prediction of electricity usage and comparative analysis of various electricity providers: Smartplug," in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, 2016, pp. 390–392.

[14] C. Beckers, I. Besselink, J. Frints, and H. Nijmeijer, "Energy consumption prediction for electric city buses," in *Proceedings of the 13th ITS European Congress, Brainport, The Netherlands*, 2019, pp. 3–6.

[15] Z. Yi and P. H. Bauer, "Adaptive multiresolution energy consumption prediction for electric vehicles," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 515–10 525, 2017.

[16] M. Rosenfelder, M. Wussow, G. Gust, R. Cremades, and D. Neumann, "Predicting residential electricity consumption using aerial and street view images," *Applied Energy*, vol. 301, p. 117407, 2021.

[17] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, "Modeling and forecasting building energy consumption: A review of data-driven techniques," *Sustainable Cities and Society*, vol. 48, p. 101533, 2019.

[18] A. G. Salman, B. Kanigoro, and Y. Heryadi, "Weather forecasting using deep learning techniques," in *2015 international conference on advanced computer science and information systems (ICACSIS)*. Ieee, 2015, pp. 281–285.

[19] Y. Lin, N. Mago, Y. Gao, Y. Li, Y.-Y. Chiang, C. Shahabi, and J. L. Ambite, "Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning," in *Proceedings of the 26th ACM SIGSPATIAL international conference on advances in geographic information systems*, 2018, pp. 359–368.

[20] D. L. Marino, K. Amarasinghe, and M. Manic, "Building energy load forecasting using deep neural networks," in *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2016, pp. 7046–7051.

[21] J. Huang, C. Pang, W. Yang, X. Zeng, J. Zhang, and C. Huang, "A deep learning neural network for the residential energy consumption prediction," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 17, no. 4, pp. 575–582, 2022.

[22] A.-D. Pham, N.-T. Ngo, T. T. H. Truong, N.-T. Huynh, and N.-S. Truong, "Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability," *Journal of Cleaner Production*, vol. 260, p. 121082, 2020.

[23] Y. Liu, H. Chen, L. Zhang, X. Wu, and X.-j. Wang, "Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in china," *Journal of Cleaner Production*, vol. 272, p. 122542, 2020.

[24] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: Case study in malaysia," *Developments in the Built Environment*, vol. 5, p. 100037, 2021.

[25] T. C. Carneiro, P. A. Rocha, P. C. Carvalho, and L. M. Fernández-Ramírez, "Ridge regression ensemble of machine learning models applied to solar and wind forecasting in brazil and spain," *Applied Energy*, vol. 314, p. 118936, 2022.

[26] A. J. del Real, F. Dorado, and J. Durán, "Energy demand forecasting using deep learning: Applications for the french grid," *Energies*, vol. 13, no. 9, p. 2242, 2020.

[27] A. Qosja, E. G. Dhamo, D. Georges, L. Nikolla, and A. Cela, "Comparative analysis of neural network and radial basis functions approaches for electricity consumption forecasting in five different regions in albania," in *2024 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, 2024, pp. 1–6.

[28] D. Sen, K. M. Tunç, and M. E. Günay, "Forecasting electricity consumption of oecd countries: A global machine learning modeling approach," *Utilities Policy*, vol. 70, p. 101222, 2021.

[29] U. Şahin, S. Ballı, and Y. Chen, "Forecasting seasonal electricity generation in european countries under covid-19-induced lockdown using fractional grey prediction models and machine learning methods," *Applied Energy*, vol. 302, p. 117540, 2021.

[30] Z. Zheng, H. Chen, and X. Luo, "Spatial granularity analysis on electricity consumption prediction using lstm recurrent neural network," *Energy Procedia*, vol. 158, pp. 2713–2718, 2019.

[31] A. Fukushima, T. Yano, S. Imahara, H. Aisu, Y. Shimokawa, and Y. Shibata, "Prediction of energy consumption for new electric vehicle models by machine learning," *IET Intelligent Transport Systems*, vol. 12, no. 9, pp. 1174–1180, 2018.

[32] T. Liu, Z. Tan, C. Xu, H. Chen, and Z. Li, "Study on deep reinforcement learning techniques for building energy consumption forecasting," *Energy and Buildings*, vol. 208, p. 109675, 2020.

[33] R. F. Berriel, A. T. Lopes, A. Rodrigues, F. M. Varejao, and T. Oliveira-Santos, "Monthly energy consumption forecast: A deep learning approach," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 4283–4290.

[34] L. Lei, W. Chen, B. Wu, C. Chen, and W. Liu, "A building energy consumption prediction model based on rough set theory and deep learning algorithms," *Energy and Buildings*, vol. 240, p. 110886, 2021.

[35] X. Zhou, W. Lin, R. Kumar, P. Cui, and Z. Ma, "A data-driven strategy using long short term memory models and reinforcement learning to predict building electricity consumption," *Applied Energy*, vol. 306, p. 118078, 2022.

[36] Y. Gao and Y. Ruan, "Interpretable deep learning model for building energy consumption prediction based on attention mechanism," *Energy and Buildings*, vol. 252, p. 111379, 2021.

[37] L. S. Carlsson, P. B. Samuelsson, and P. G. Jönsson, "Interpretable machine learning—tools to interpret the predictions of a machine learning model predicting the electrical energy consumption of an electric arc furnace," *Steel research international*, vol. 91, no. 11, p. 2000053, 2020.

[38] J.-Y. Kim and S.-B. Cho, "Explainable prediction of electric energy demand using a deep autoencoder with interpretable latent space," *Expert Systems with Applications*, vol. 186, p. 115842, 2021.

[39] A. K. Shaikh, A. Nazir, I. Khan, and A. S. Shah, "Short term energy consumption forecasting using neural basis expansion analysis for interpretable time series," *Scientific reports*, vol. 12, no. 1, p. 22562, 2022.

[40] D. Mariano-Hernández, L. Hernández-Callejo, M. Solís, A. Zorita-Lamadrid, O. Duque-Pérez, L. Gonzalez-Morales, F. S. García, A. Jaramillo-Duque, A. Ospino-Castro, V. Alonso-Gómez *et al.*, "Analysis of the integration of drift detection methods in learning algorithms for electrical consumption forecasting in smart buildings," *Sustainability*, vol. 14, no. 10, p. 5857, 2022.

[41] R. Kumar, B. Shakila, and M. Prakash, "A novel hybrid deterministic temporal fusion transformer and dynamic recurrent neural network architecture for multi-horizon multi-variate electric vehicle energy consumption forecasting: mitigation pathway," *International Journal of Information Technology*, pp. 1–19, 2025.

[42] J. He, Y. Li, X. Xu, and D. Wu, "Energy consumption forecasting for oil and coal in china based on hybrid deep learning," *PloS one*, vol. 20, no. 1, p. e0313856, 2025.