UCICS 2026

2ND UNDERGRADUATE CONFERENCE ON INTELLIGENT COMPUTING AND SYSTEMS

EMPOWERING TOMORROW'S INNOVATORS

FULL PROCEEDINGS



Dept. of CSE
Varendra University

VARENDRA UNIVERSITY
বরেন্দ্র
বিশ্ববিদ্যালয়

Department of Computer Science and Engineering
Varendra University, Rajshahi, Bangladesh

ISBN: 978-984-37-0534-1

UCiCS 2026

2ND UNDERGRADUATE CONFERENCE ON INTELLIGENT COMPUTING AND SYSTEMS

17-18 JANUARY, 2026

Dept. of CSE
Varendra University

Department of Computer Science and Engineering
Varendra University, Rajshahi, Bangladesh

ISBN: 978-984-37-0534-1

2nd UNDERGRADUATE CONFERENCE ON INTELLIGENT COMPUTING AND SYSTEMS

EMPOWERING TOMORROW'S INNOVATORS

January 17-18, 2026
Varendra University
Rajshahi-6204, Bangladesh

- **Copyright©2026**
  Department of Computer Science and Engineering
  Varendra University

- **Chief Editor**
  M. Babul Islam

- **Editors**
  Md. Ariful Islam Nahid
  Md. Muktar Hossain

- **Contact Information**
  Department of Computer Science and Engineering
  Varendra University, Rajshahi-6204, Bangladesh
  Cell: +8801730406534
  Email: head.cse@vu.edu.bd

- **URL:** https://ucics.org

- **Cover Design**
  Md. Fatin Ilham

- **Print**
  The City Offset Printers
  Talaimari, Rajshahi

- **Published**
  17 January, 2026

- **ISBN:** 978-984-37-0534-1

Preface

The 2nd Undergraduate Conference on Intelligent Computing and Systems (UCICS), to be held on 17–18 January 2026, organized by the Department of Computer Science and Engineering, Varendra University, Rajshahi, Bangladesh provides a vibrant academic platform for undergraduate students to present, exchange, and discuss emerging ideas, innovations, and research findings in the broad domains of intelligent computing and modern systems.

Building on the success of the first edition, this conference aims to foster a research-oriented mindset among undergraduate students and encourage them to engage with contemporary challenges in areas such as Artificial Intelligence, Machine Learning, Data Science, Signal and Image Processing, Intelligent Systems, Internet of Things, Robotics, and related interdisciplinary fields. UCICS serves as a bridge between foundational undergraduate learning and advanced research, promoting critical thinking, creativity, and problem-solving skills.

The conference features technical paper presentations, poster sessions, invited talks, and keynote speeches by academics and industry professionals, offering participants valuable exposure to current research trends and real-world applications. It also provides an excellent opportunity for students to interact with peers, mentors, and experts, helping to build confidence, communication skills, and professional networks.

We extend our sincere appreciation to the authors for their enthusiastic participation and for contributing high-quality research work. We are also grateful to the reviewers, session chairs, organizing committee members, and volunteers for their dedication and efforts in making this conference possible. Special thanks go to the keynote speakers and invited guests for sharing their insights and experiences.

We hope that the 2nd Undergraduate Conference on Intelligent Computing and Systems will inspire young researchers, stimulate meaningful discussions, and contribute to the advancement of knowledge and innovation. We wish all participants a rewarding and enriching conference experience.

M. Babul Islam
Chief Editor, UCICS 2026
&
Professor
Department of Electrical and Electronic Engineering
University of Rajshahi, Bangladesh





বরেন্দ্র বিশ্ববিদ্যালয় ট্রাস্ট

Chairman
Varendra University Trust

Message

It is a great honor and privilege to extend my heartfelt greetings to the organizers, participants, and distinguished guests of the 2nd Undergraduate Conference on Intelligent Computing and Systems (UCICS-2026) at Varendra University. The outstanding success of UCICS-2025 has set a remarkable precedent, reflecting the university's steadfast commitment to promoting research, innovation, and academic excellence among undergraduate students. This year, UCICS-2026 builds upon that foundation, offering an even more engaging and intellectually stimulating platform for young scholars to showcase their talents.

Conferences such as UCICS play a vital role in shaping the academic and professional development of students. By presenting their research, interacting with peers, and engaging with experienced researchers, students are encouraged to think critically, develop creativity, and cultivate analytical and problem-solving skills. Events like UCICS-2026 are essential in fostering a culture of inquiry and innovation, preparing the next generation of professionals and leaders to contribute meaningfully to society and to address the complex challenges of the modern world.

I would like to take this opportunity to sincerely thank the faculty members of the Department of Computer Science and Engineering, Varendra University, for their tireless efforts in organizing this conference. Their dedication, mentorship, and commitment to guiding students have been central to the continued success and prestige of UCICS. Their hard work in ensuring the smooth conduct of the conference, maintaining high academic standards, and supporting students in their research endeavors is truly commendable.

I extend my best wishes to all participants, presenters, and contributors of UCICS-2026. May this conference inspire innovative ideas, promote collaboration across disciplines, and leave a lasting impact on the academic growth and professional journey of every attendee. I am confident that UCICS-2026 will continue the tradition of excellence established by its inaugural edition and further strengthen Varendra University's reputation as a hub of research and intellectual development.

Hafizur Rahman Khan



Vice Chancellor
Rajshahi University of Engineering & Technology

Message

It is with great enthusiasm and a profound sense of pride that I extend my heartfelt congratulations to the Department of Computer Science and Engineering at Varendra University, Rajshahi, for organizing the 2nd Undergraduate Conference on Intelligent Computing and Systems (UCICS-2026). The resounding success of UCICS 2025, held on February 26–27, has set a remarkable benchmark in promoting undergraduate research and fostering academic excellence. Building upon this strong foundation, UCICS 2026 stands as an inspiring continuation of that vision.

The previous edition of the conference not only provided a vibrant platform for students to showcase their innovative research but also demonstrated the immense potential and dedication of our young scholars. The overwhelming participation, high-quality submissions, and impactful keynote sessions collectively contributed to a memorable and truly grand event. Such achievements reaffirm the importance of nurturing curiosity, encouraging collaboration, and facilitating knowledge exchange among aspiring researchers.

In today's rapidly evolving world, advancements in science and technology are central to national development and global competitiveness. By cultivating a research-oriented mindset and empowering students to explore emerging areas of intelligent computing, UCICS 2026 promises to further strengthen the academic and professional growth of the next generation. The Department of Computer Science and Engineering at Varendra University has consistently upheld a tradition of excellence, organizing conferences that attract distinguished scholars, researchers, and industry experts from around the world. I am confident that this year's edition will once again serve as a catalyst for groundbreaking ideas and meaningful academic interactions.

UCICS 2026 embodies the spirit of innovation, perseverance, and intellectual curiosity. Through rigorous academic engagement, thoughtful discussions, and exposure to cutting-edge research, students will continue to develop analytical abilities and a forward-looking perspective essential for addressing real-world challenges.

I sincerely commend the organizing committee for their dedication, professionalism, and unwavering commitment to maintaining the highest academic standards. Their efforts in sustaining and elevating the quality of this prestigious conference are truly admirable.

I extend my best wishes for the continued success of UCICS 2026. May it build upon the achievements of the previous year, inspire transformative discoveries, foster lasting collaborations, and contribute meaningfully to the future of scientific research and technological advancement.

Prof. S M Abdur Razzak



Vice Chancellor
Rajshahi Medical University

Message

I am delighted and honored to address the organizers, participants, and esteemed guests of the 2nd Undergraduate Conference on Intelligent Computing and Systems (UCICS-2026) at Varendra University, Rajshahi, Bangladesh. Following the exceptional success of UCICS-2025, this year's event continues the university's dedication to fostering research, innovation, and scholarly achievement among undergraduate students. UCICS-2026 serves as a vital forum for emerging scholars to showcase their work, exchange ideas, and delve into new frontiers in computing and technology.

In our fast-changing world, the quest for knowledge and research is essential to both academic and societal advancement. Conferences such as UCICS equip students with crucial opportunities to hone their analytical, creative, and problem-solving abilities while staying abreast of the latest developments in their disciplines. Through dialogue with peers and established researchers, students build confidence, widen their outlook, and nurture a research-focused approach beneficial to their future careers.

I also wish to express my sincere gratitude to the faculty of the Department of Computer Science and Engineering at Varendra University for their steadfast commitment, mentorship, and hard work. Their dedication in planning this conference, guiding students, and upholding rigorous academic standards has been instrumental to UCICS's accomplishments.

UCICS-2026 is more than a research showcase, it is a tribute to curiosity, determination, and innovation. By encouraging a spirit of inquiry and teamwork, the conference motivates students to think boldly, challenge limits, and make significant contributions to academic and practical problems.

My warmest wishes go to all involved in UCICS-2026 participants, keynote speakers, and contributors alike. May this gathering continue to spark innovation, deepen collaboration, and positively influence each attendee's academic journey and professional path.

Prof. Dr. Md. Jawadul Haque



Vice Chancellor
University of Rajshahi

Message

It is a pleasure to extend my warm greetings to the Department of Computer Science and Engineering at Varendra University on the occasion of the 2nd Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026). The remarkable achievement and enthusiastic response witnessed during UCICS 2025 have clearly demonstrated the commitment of the department toward nurturing a research-driven academic culture among undergraduate students.

The previous edition of this conference stood out not only for the originality of the student-led research but also for the exceptional organization, scholarly engagement, and collaborative spirit it fostered. Its grand success has set a strong precedent and created high expectations for this year's event. I am confident that UCICS 2026 will continue this positive momentum and provide an even more impactful environment for young researchers to explore emerging domains in intelligent computing and related technologies.

Higher education today demands a balance of knowledge creation, analytical skill, and practical insight. Events such as UCICS 2026 play a significant role in guiding students toward meaningful research endeavors while helping them develop confidence and academic maturity. By connecting aspiring scholars with experienced researchers, the conference contributes to building a vibrant academic ecosystem that benefits not only the participants but also the broader scientific community.

I appreciate the vision and dedication of the organizing committee for their tireless efforts in arranging this program for the second consecutive year. Their perseverance reflects the commitment of Varendra University to academic excellence and research promotion.

I convey my best wishes for the successful hosting of UCICS-2026. May this conference continue to inspire creativity, strengthen scholarly dialogue, and motivate students to pursue innovative research that contributes to societal advancement.

S.H. Naqib

Prof. Dr. Saleh Hasan Naqib



VARENDRA UNIVERSITY
বরেন্দ্র বিশ্ববিদ্যালয়

Vice Chancellor &
Conference Chair

Message

It is my pleasure to welcome the participants, researchers, faculty members, and distinguished guests to the 2nd Undergraduate Conference on Intelligent Computing and Systems (UCICS-2026) at Varendra University, Bangladesh. The resounding success of UCICS-2025 has set a high benchmark, reflecting the dedication of the respective committees to promote undergraduate research, innovation, and academic excellence. As a continuation, UCICS-2026 aims to provide an even more engaging and dynamic platform for young scholars to present and share their research contributions.

Conferences like UCICS play a vital role in fostering analytical thinking, critical reasoning, and problem-solving skills among students and researchers. Beyond technical knowledge, they encourage collaboration, interdisciplinary learning, and exposure to emerging trends in the current trend of computing and technology. Such experiences prepare young researchers to face real-world challenges confidently leading to contribute to the society.

I sincerely thank the faculty members of the Department of Computer Science and Engineering, Varendra University for their tireless efforts, guidance, and mentorship in organizing this conference. Their commitment to maintain high academic standards, supporting students, and ensuring the smooth conduct of the event is truly commendable for the success of UCICS.

UCICS-2026 is not only a platform for presenting research but also a celebration of curiosity, creativity, and perseverance. It provides the young researchers with the confidence to think beyond conventional boundaries, approach problems innovatively, and engage meaningfully with the academic community. By building on the achievements of the previous edition, this year's conference promises to bridge the gap between industry and academia, inspire new ideas, foster collaboration, and strengthen the ecosystem of knowledge and research.

I extend my heartfelt best wishes to young researchers, participants, speakers, and contributors involved in UCICS-2026. May this conference be a significant and enriching a forum that not only shares knowledge but also sparks creativity, drives meaningful innovation, and leaves an enduring, positive mark on the academic growth of prospective academicians.

Prof. Dr. Md. Khademul Islam Molla



Conference Co-Chair
UCICS 2026

Message

It gives me an immense pleasure to be a part of the 2nd Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026), organized by the Department of Computer Science and Engineering, Varendra University, Bangladesh. It's also a great privilege and honour for me to greet and welcome all participants of this conference at our magnificent campus as the Co-Chair of this conference. The idea of organizing conferences for undergraduate students is unique and it is not much prevalent even in the first world countries. Nevertheless, it has a huge impact among young scholars in terms of participating in cutting-edge research, creating a platform to foster the research-driven mindsets, building connections, gaining knowledge, exchanging ideas, and demonstrating excellence.

I must congratulate the CSE Department of Varendra University for coming up with this thought-provoking idea and organizing such a conference in two consecutive years. The grand success of UCICS 2025 has motivated the organizers to repeat it on a regular basis. This conference received a good number of quality papers from the student participants that led to the publication of a Special Volume of Varendra International Journal for Interdisciplinary Research (VIJIR), which is scheduled to be published on the inaugural day of UCICS 2026. As the Editor-in-Chief of this journal, I congratulate all young authors who have inspired us with their outstanding research in the ICT area. I sincerely hope and believe UCICS 2026 would outperform its predecessor in every respect.

I seize this opportunity to congratulate the organizing committee, keynote speakers, plenary speakers, invited distinguished speakers, all committees and sub-committees, authors, reviewers, participants, volunteers, sponsors, and last but not the least the VU authority in advance for their time and efforts to make this event a grand success. I have a dream. I want this conference to be upgraded to an International Conference soon. This team possesses all the potential, so it will be just a matter of time. Let us put our hands together to make this dream a reality.

I wish UCICS 2026 a grand success. May this initiative serve as a catalyst for undergraduate research and innovation in Bangladesh and across the world.

A. H. M. Rahmatullah Imon

A. H. M. Rahmatullah Imon

Contents

Program Schedule	15
Technical Session Schedule	17
Organizing Committee	23
International Advisory Committee	23
National Advisory Committee	23
Technical Program Committee	24
Program Committee	25
Publication Committee	25
Promotional Committee	26
Finance and Sponsorship Committee	27
Registration and Kit Distribution Committee	28
Event Management and Decoration Committee	28
Food and Refreshment Committee	29
Invitation and Hospitality Management Committee	30
Banner Committee	31
Certificate and Crest Committee	31
Technical Support Committee	32
Disciplinary Committee	32
Idea Contest Committee	33

PID	Title	
Keynote-1	Speech prosody and its modeling	34
Keynote-2	Reservoir computing using superconductor	35
Keynote-3	Research in undergraduate years	36
Keynote-4	Speech Brain–Computer Interfaces Using Intracranial EEG: Advances, Challenges, and Prospects	38
Keynote-5	Eye-gaze-based Human-Intention Detection	39
Keynote-6	Trustworthy AI for Cyber Defense in the Age of Large Language Models	40
Keynote-7	AI-Driven Wearables and Decision Support Systems for Digital Healthcare	41
Keynote-8	Rethinking Engineering Education in the Generative AI Era	42
Invited Talk-1	Controlling Malaria with the help of agent-based models	43
Invited Talk-2	Embedded Systems–Driven Industrial Transformation in Bangladesh: Challenges and Opportunities	44
Invited Talk-3	Effectiveness of the Frontal Pole EEG Signals in Diagnosing MDD	45
Invited Talk-4	Intelligent Machine Perception: Computer Vision, Techniques for Human–Robot Interaction and Autonomous Vehicle	46
Invited Talk-5	Distributed Database Management Systems: Powering Modern Data	47
Industrial Talk-1	Designing RoCE Networks for AI Workloads	48

PID	Title	
Industrial Talk-2	IT Project Delivery framework	49
Industrial Talk-3	Think Smart, Build Bold: Innovate, Fail, and Soar	49
Industrial Talk-4	Leveraging AI in AI Chip Manufacturing	50
Industrial Talk-5	From Wheel to AI — Humanity's Journey of Innovation and the Future of the Job Market	51
Industrial Talk-6	Incident Management at Google	52
Industrial Talk-7	The Architecture of Scalable Leadership from Amazon's Engineering Culture	52
5	Hybrid Feature Fusion with Lightweight ANN for Efficient Diabetic Retinopathy Screening Across Datasets	53
6	Assessing Mental Health in Technology Professionals Using Explainable Machine Learning	58
7	AgriCall: An IoT-Integrated GSM-Based Irrigation Control System for Sustainable Agriculture	64
11	Beyond Text: A Review of Emoji-Enhanced Notifications and Their Impact on User Attention in HCI	69
17	Exploring Multi-Model Machine Learning Approaches for Pathogenicity Classification of Somatic Gene Mutations	75
18	Swin Transformer-Based Multi-Class Classification of Gallbladder Diseases: A Comparative Study with Lightweight CNNs	81
23	Enhancing Community Group Recognition Using CNN Architectures.	87
24	A Hybrid Prophet–LSTM Framework for Enhanced Monthly Rainfall Forecasting in Climate-Sensitive Regions: A Case Study of Cox's Bazar, Bangladesh	93
25	Cardiovascular Disease Prediction Using Modern Machine Learning Pipelines on the UCI Heart Disease Dataset	99
29	PneuXNet: An Interpretable CNN Framework for High-Accuracy Pneumonia Detection Using Multi-Scale and Adaptive Attention	103
32	Design and Analysis of an Ultra-Low-Power CMOS Front-End with Enhanced CMRR for Wearable Electrocardiogram Monitoring	109
36	Angular Dependence of RF Absorption in Brain and Muscle Tissues at 2.45 GHz: A Comparative SAR Analysis	115
37	OMR-Fallback: A Robust Hybrid Pipeline for Multi-Format Educational Assessment Integrating Computer Vision, OCR, and Deep Learning	121
44	Rainfall Decomposition and Forecasting Across Six Climatic Regions of Bangladesh: A Statistical and Deep Learning Approach	127
47	A Random Forest–Based Predictive Model for Assessing Children's Educational Opportunities in Fishermen Communities Using Socioeconomic and Family Factors	133
48	Synergistic Clinical-Genomic Feature Fusion with Optimized Machine Learning Models for Enhanced Breast Cancer Classification	139
49	A Systems-Oriented Review of the Duan et al. O(m log2/3 n) SSSP Algorithm	144

PID	Title	
54	Towards Robust Word-Level Bangla Sign Language Recognition: A Comparison of Spatio-Temporal Graph and Hybrid Sequence-Based Deep Models	148
58	Privacy Preserving and Secure Cataract Detection from Fundus Images Using an Edge-Enhanced Federated Deep Learning Framework	154
59	Autism Spectrum Disorder Detection Using a Fine-Tuned VGG16 Model	160
62	IoT-Enabled Multisensory Smart Cane for the Visually Impaired: Design, Implementation, and Performance Evaluation	166
66	A Transfer Learning- Based Ensemble Model for Automated Tea Leaf Disease Detection	172
69	Rice Leaf Disease Detection Using Vision Transformer and Transfer Learning-Based Architectures	178
70	Optimized ACS-BTF: A Steganographic Method for Secure Text Generation via Large Language Models Using Contrastive Sampling and Token Filtering	184
72	A Comprehensive Review of Artificial Intelligence Possibilities in Photovoltaic Systems: The Bangladesh Perspective	190
73	AI as an Algorithmic Crutch: The Impact of AI Use on Cognitive Skills and Academic Psychology of Bangladeshi University Students	196
75	Elevated View Gait Biometrics: Efficient Skeleton-Based Recognition Using UAV Captured Videos	202
76	The Role of Family and Institutional Support in Student Well-being: A Structural Equation Modeling Study	208
80	Intelligent Multi-Agent Coordination with Adaptive Knowledge Structures for Complex Web Workflows	213
82	Gait Recognition with EfficientNet-B0: Transfer Learning for Enhanced Biometric Performance	219
85	Improving Openness in the Identification of Heart Disease via SHAP and LIME Explainability	225
87	Hybrid Ensemble Machine Learning Modeling for Tumor Prediction Using COSMIC Differential Methylation Data	231
91	TeaScan: A Comprehensive Survey of Tea Leaf Disease Detection Using Machine Learning and Deep Learning Techniques	237
99	The Role of Agile Methodology in Enhancing Product Management Efficiency: A Review	243
100	Skeleton-Based Medical Action Recognition Using Spatio-Temporal Graph Neural Networks	249
105	Impact of Online Learning Environments on Attention and Retention among University Students	255
108	Hybrid Ensemble Learning for Coding vs. Non-coding Somatic Variant Classification	261
109	EpidermNet: A Task-Specific Deep Learning System for Scabies and Eczema Detection	267
110	Deep Learning-Based Detection and Classification of Organic, General, and Recyclable Waste in Aquatic and Terrestrial Environments	273

PID	Title	
111	A Standardized and Interpretable Deep Learning Pipeline for 17-Class Brain Tumor MRI Classification	279
113	Predicting Mental Health Risk from Social Media Usage Patterns Using Machine Learning	285
115	Clinical Utility of Machine Learning Algorithms in Predicting the Presence of Heart Disease: A Cross-Validated Analysis	291
118	Machine Learning-Driven Early Detection of PCOS: Addressing Data Imbalance and Model Optimization	297
119	Stroke Risk Prediction Using Machine Learning: A Comprehensive Evaluation	302
121	Deep Transfer Learning-Based Multi-class Classification of Bone Tumors Using Medical Imaging	308
122	SkillMap-ML: A Data-Driven Skill Mapping and Machine Learning Framework for Classifying, Predicting, and Visualizing Digital Competency Gaps in University Students	314
124	Two-Stage Bengali Sign Language Recognition Using YOLOv8 Hand Detection and Lightweight CNN Backbones	320
126	A Machine Learning–Integrated Endocrine Biomarker Framework for Sensitive Detection of Polycystic Ovary Syndrome	326
128	An Explainable AI Framework for PCOS Diagnosis Integrating Oversampling, Ensemble Learning, and Kernel-Based Classification	332
129	Predicting Climate Factors in Rajshahi using Hybrid Deep Learning and Machine Learning Models	338
131	Effective Feature Extraction Based for Epileptic Seizure Detection Using Classical Machine Learning	344
132	AI Based medical Transparent and Non Transparent waste classification	349
133	Experimental Validation of Classical LSB Substitution Algorithms on Modern High Resolution Datasets	355
138	Plant Leaf Disease Detection using CNN	359
141	Privacy-Preserving and Secured Multi-Site Cataract Detection from Fundus Images with EVA-02 Vision Transformers through Federated Deep Learning (FDL)	364
142	Early Diabetes Prediction Using Multi-Level Ensemble Machine Learning with Clinical Feature Engineering	370
143	Deep Learning based CNN Framework for Automated Glaucoma Detection	376
145	Multiclass MRI Brain Tumor Classification using Hybrid CNN–Transformer Architecture	381
148	Analysis of Lung Cancer Risk Associated with Long-Term Smoking Using Machine Learning Approaches	387
149	A Comparative Methodology Study on Cyberbullying Detection Using Machine Learning and Natural Language Processing	393
150	Assessing the Brain and Kidney Tissues for Their Radiation Absorption Attributes	399
151	Subject Independent Epileptic Seizure Recognition Using 1D-EEGNet Model	405
152	Spectral Regularization: Controlling Latent Variance	409
153	Activation Patching for Whisper Interpretability: Mapping the Acoustic-to-Linguistic Transition in Whisper	415

Program Schedule

16 January, 2026

Conference Kit Distribution	January 16, 2026	15:00 – 17:00	Information & Kit Collection Booth

Day – 1: 17 January, 2026

Breakfast (08:45 – 09:15)

Conference Kit Distribution	January 17, 2026	08:30 – 09:30	Information & Kit Collection Booth
Inaugural Session	January 17, 2026	09:30 – 10:30	Auditorium Hall
Keynote Session (Brain–Computer Interface)	January 17, 2026	10:30 – 10:55	
Keynote Session (Cyber Security)	January 17, 2026	10:55 – 11:20	

Tea Break-01 (11:20 – 11:40)

Keynote Session (Human–Computer Interaction)	January 17, 2026	11:40 –12:05	Auditorium Hall
Keynote Session (Digital Healthcare)	January 17, 2026	12:05 –12:30	
Keynote Session (Generative AI)	January 17, 2026	12:30 –12:55	

Prayer and Lunch Break (13:00 – 15:00)

Technical Session-1			
Invited Talk (Bioinformatics)	January 17, 2026	15:00 – 17:30	Auditorium Hall
Technical Session-1.1 (Signal, Image, and Video Processing)			
Invited Talk (Bioinformatics and Embedded Systems & IOT)	January 17, 2026	15:00 – 17:30	311
Technical Session-1.2 (Bioinformatics and Embedded Systems & IOT)			
Technical Session-1.3 (AI & Machine Learning)	January 17, 2026	15:00 – 17:30	313
Technical Session-1.4 (Signal, Image and Video Processing)	January 17, 2026	15:00 – 17:30	314

Prayer and Tea Break-02 (17:30 – 18:00)

Cultural Program	January 17, 2026	18:00 – 20:00	Auditorium Hall
Gala Dinner			Varendra University

Day – 2: 18 January, 2026

Breakfast (08:30 – 09:00)

Technical Session-2			
Invited Talk (Biomedical Engineering and Data Science & Big Data Analysis)	January 18, 2026	09:00 – 10:30	311
Technical Session-2.1 (Biomedical Engineering and Data Science & Big Data Analysis)			
Technical Session-2.2 (AI & Machine Learning)	January 18, 2026	09:00 – 10:30	313
Technical Session-2.3 (Biomedical Engineering)	January 18, 2026	09:00 – 10:30	314
Technical Session-2.4 (AI & Machine Learning)	January 18, 2026	09:00 – 10:30	414
Industrial Talk (IT Project Management)	January 18, 2026	10:30 – 10:50	Auditorium Hall
Industrial Talk (Innovation & Entrepreneurship)	January 18, 2026	10:50 – 11:10	
Industrial Talk (Computer Networks and Distributed Systems)	January 18, 2026	11:10 – 11:30	

Tea Break-01 (11:30 – 11:40)

Keynote Session (Speech Prosody)	January 18, 2026	11:40 – 12:05	Auditorium Hall
Keynote Session (Neural Computing)	January 18, 2026	12:05 – 12:30	
Industrial Talk (Engineering Culture, Ownership and Career Growth)	January 18, 2026	12:30 – 12:50	
Industrial Talk (Technology Management & Strategy)	January 18, 2026	12:50 – 13:10	
Industrial Talk (AI Chip Manufacturing)	January 18, 2026	13:10 – 13:30	

Prayer and Lunch Break (13:30 – 15:00)

Industrial Talk (Digital Transformation & Jobs)	January 18, 2026	15:00 – 15:20	Auditorium Hall

Technical Session-3			
Invited Talk (AI & Machine Learning)	January 18, 2026	15:20 – 17:00	311
Technical Session-3.1 (AI & Machine Learning)			
Technical Session-3.2 (Biomedical Engineering)	January 18, 2026	15:20 – 17:00	313
Invited Talk (AI & Machine Learning, Signal, Image & Video Processing)	January 18, 2026	15:20 – 17:00	314
Technical Session-3.3 (AI & Machine Learning, Signal, Image & Video Processing)			
Technical Session-3.4 (Computer Networks & Cyber security, ICT Applications and Software Engineering)	January 18, 2026	15:20 – 17:00	414

Prayer and Tea Break-02 (17:00 – 18:00)

Award Ceremony and Closing	January 18, 2026	18:00 – 19:00	Auditorium Hall

Technical Session Schedule

Day – 1: 17 January, 2026

Session ID	Track	Speaker	Chair	Time	Venue
Keynote Session	Brain–Computer Interface	Toshihisa Tanaka, PhD	Dr. Md. Shahid Uz Zaman	10:30 – 10:55	Auditorium Hall
	Cyber Security	Dr. Kaushik Roy		10:55 – 11:20	

Tea Break-01 (11:20 – 11:40)

Session ID	Track	Speaker	Chair	Time	Venue
Keynote Session	Human–Computer Interaction	Dr. Kosin Chamnongthai	A.H.M. Rahmatullah Imon, Ph.D.	11:40 –12:05	Auditorium Hall
	Digital Healthcare	Dr.Mohammad Ali Moni		12:05 –12:30	
	Generative AI	Dr. Faisal Tariq		12:30 –12:55	

Prayer and Lunch Break (13:00 – 15:00)

Technical Session (TS) - 01					
Invited Talk-1 by Dr. M. Sohel Rahman					
Invited Talk-2 by Dr. Mosaddek Hossain Kamal					
Session ID	**Track**	**Paper ID**	**Chair**	**Time**	**Venue**
Invited Talk Session	Bioinformatics	Invited Talk-1	Dr. Tanvir Ahmed	15:00 – 15:15	Auditorium Hall
TS-1.1	Signal, Image, and Video Processing	5, 37, 54, 58, 59, 66	Dr. Md. Sohel Rana	15:15 – 17:30	
Invited Talk Session	Bioinformatics and Embedded Systems & IOT	Invited Talk-2	Dr. Boshir Ahmed	15:00 – 17:30	311
TS-1.2		87, 142, 7, 32, 62			
TS-1.3	AI & Machine Learning	6, 11, 17, 23, 24, 91	Dr. Mohammod Abdul Motin	15:00 – 17:30	313
TS-1.4	Signal, Image and Video Processing	75, 82, 109, 121, 124, 131	Dr. Md. Abdur Rahim	15:00 – 17:30	314

Prayer and Tea Break-02 (17:30 – 18:00)

Event	Time	Venue
Cultural Program	18:00 – 20:00	Auditorium Hall
Gala Dinner		Varendra University

Day – 2: 18 January, 2026

Technical Session (TS) - 02					
Invited Talk-5 by Dr. Md. Anisuzzaman Siddique					
Session ID	**Track**	**Paper ID**	**Chair**	**Time**	**Venue**
Invited Talk Session	Biomedical Engineering and Data Science &Big Data Analysis	Invited Talk-5	Dr. Md. Johirul Islam	09:00 – 10:30	311
TS-2.1		44, 49, 105, 129, 150			
TS-2.2	AI & Machine Learning	113, 115, 119, 122, 132, 138	Dr. Md. Golam Rashed	09:00 – 10:30	313
TS-2.3	Biomedical Engineering	111, 118, 126, 128, 148	Dr. Sinthia Shabnam Mou	09:00 – 10:30	314
TS-2.4	AI & Machine Learning	47, 48, 69, 72, 143	Dr. Foez Ahmed	09:00 – 10:30	414

Session ID	Track	Speaker	Chair	Time	Venue
Industrial Talk	IT Project Management	Mohammed Golam Mostofa	Dr. M. Babul Islam	10:30 – 10:50	Auditorium Hall
	Innovation & Entrepreneurship	Nazme Mohsina		10:50 – 11:10	
	Computer Networks and Distributed Systems.	Dr. Abdul Alim		11:10 – 11:30	

Tea Break-01 (11:30 – 11:40)

Session ID	Track	Speaker	Chair	Time	Venue
Keynote Session	Speech Prosody	Keikichi Hirose, PhD	Prof. Muhammad Sajjadur Rahim	11:40 – 12:05	Auditorium Hall
	Neural Computing	Edmund Soji Otabe		12:05 – 12:30	
Industrial Talk	Engineering Culture, Ownership and Career Growth	Rafsan Jany	Dr. Md. Iqbal Aziz Khan	12:30 – 12:50	
	Technology Management & Strategy	Shakil Ahmed		12:50 – 13:10	
	AI Chip Manufacturing	Dr. Gazi Saurav Islam		13:10 – 13:30	

Prayer and Lunch Break (13:30 – 15:00)

Session ID	Track	Speaker	Chair	Time	Venue
Industrial Talk	Digital Transformation & Jobs	Nizam Uddin	Dr. Mahboob Qaosar	15:00 – 15:20	Auditorium Hall

Technical Session (TS) - 03					
Invited Talk-4 by Dr. Dipankar Das					
Invited Talk-3 by Dr. M. Babul Islam					
Session ID	**Track**	**Paper ID**	**Chair**	**Time**	**Venue**
Invited Talk Session	AI & Machine Learning	Invited Talk-4	Dr. Dipankar Das	15:20 – 17:00	311
TS-3.1	AI & Machine Learning	80, 100, 108, 110			
TS-3.2	Biomedical Engineering	18, 25, 29, 36, 85	Dr. Asif Zaman	15:20 – 17:00	313
Invited Talk Session	AI & Machine Learning, Signal, Image & Video Processing	Invited Talk-3	Dr. Md. Ekramul Hamid	15:20 – 17:00	314
TS-3.3		141, 149, 145, 151, 153			
TS-3.4	Computer Networks & Cyber security, ICT Applications and Software Engineering	70, 73, 76, 99, 133, 152	Dr. Rezwana Sultana	15:20 – 17:00	414

Prayer and Tea Break-02 (17:00 – 18:00)

Event	Time	Venue
Award Ceremony and Closing	18:00 – 19:00	Auditorium Hall

Keynote Speakers	Industrial Speakers
1. Toshihisa Tanaka, PhD Professor, Tokyo University of Agriculture and Technology, Japan	1. Mohammed Golam Mostofa IT Delivery Lead (Informatics), Blue Cross Blue Shield of Michigan
2. Dr. Kaushik Roy Professor and Chair, Jefferson-Pilot/Ron McNair Endowed Chair, Department of Computer Science North Carolina A&T State University, NC, USA	2. Nazme Mohsina Global Technical Director, AMCA International
3. Dr. Kosin Chamnongthai Professor at King Mongkut's University of Technology Thonburi (KMUTT)	3. Dr. Abdul Alim Senior Network Engineer, IBM
4. Dr. Mohammad Ali Moni Program Lead, AI and Digital Health Technology. Artificial Intelligence and Cyber Futures Institute, Charles Sturt University, Australia Professor (Research), Washington University of Science & Technology, USA Adjunct Professor, Daffodil International University, Bangladesh Senior Fellow (Hon), The University of Queensland, Australia	4. Md. Nizam Uddin Co-Founder & Chief Executive Officer, weDevs
5. Dr. Faisal Tariq Associate Professor, James Watt School of Engineering, University of Glasgow	5. Rafsan Jany Software Development Engineer, Amazon
6. Keikichi Hirose, PhD Professor Emeritus, University of Tokyo, Japan	6. Shakil Ahmed Senior Software Engineer (SRE), Google
7. Edmund Soji Otabe Professor, Kyushu Institute of Technology, Japan	7. Dr. Gazi Saurav Islam R&D Engineering Manager at Intel Corporation

Invited Speakers
1. Dr. M. Sohel Rahman Member, Public Service Commission, Bangladesh. Professor, Department of Computer Science & Engineering, Bangladesh University of Engineering and Technology
2. Dr. Mosaddek Hossain Kamal Professor, Department of Computer Science & Engineering, University of Dhaka, Bangladesh
3. Dr. Md. Anisuzzaman Siddique Professor, Department of Computer Science & Engineering, University of Rajshahi, Bangladesh
4. Dr. Dipankar Das Professor, Department of Information & Communication Engineering, University of Rajshahi, Bangladesh
5. Dr. M. Babul Islam Professor, Department of Electrical and Electronic Engineering, University of Rajshahi, Bangladesh

Session Chairs
1. Dr. Md. Shahid Uz Zaman Head & Professor, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology
2. A.H.M. Rahmatullah Imon, Ph.D. Professor, Department of Computer Science & Engineering, Director, Center for Interdisciplinary Research, Varendra University
3. Dr. Tanvir Ahmed Coordinator, Department of Electrical & Electronic Engineering, Varendra University
4. Dr. Md. Sohel Rana Professor, Department of Electrical & Electronic Engineering, Rajshahi University of Engineering & Technology
5. Dr. Boshir Ahmed Professor, Dean, Faculty of Electrical & Computer Engineering, Rajshahi University of Engineering & Technology
6. Dr. Mohammod Abdul Motin Associate Professor, Department of Electrical & Electronic Engineering, Rajshahi University of Engineering & Technology
7. Dr. Md. Abdur Rahim Professor, Department of Computer Science and Engineering, Pabna University of Science and Technology
8. Dr. Md. Johirul Islam Professor, Physics, Rajshahi University of Engineering & Technology
9. Dr. Md. Golam Rashed Associate Professor, Department of Information and Communication Engineering, University of Rajshahi
10. Dr. Sinthia Shabnam Mou Professor, Department of Electrical & Electronic Engineering, University of Rajshahi
11. Dr. Foez Ahmed Associate Professor, Department of Information & Communication Engineering, University of Rajshahi
12. Dr. M. Babul Islam Professor, Department of Electrical and Electronic Engineering, University of Rajshahi
13. Dr. Muhammad Sajjadur Rahim Professor, Department of Information & Communication Engineering, University of Rajshahi
14. Dr. Md. Iqbal Aziz Khan Professor, Department of Computer Science & Engineering, University of Rajshahi
15. Dr. Mahboob Qaosar Associate Professor, Department of Computer Science & Engineering, University of Rajshahi
16. Dr. Dipankar Das Professor, Department of Information & Communication Engineering, University of Rajshahi
17. Dr. Asif Zaman Professor, Department of Computer Science & Engineering, University of Rajshahi
18. Dr. Md. Ekramul Hamid Professor, Department of Computer Science & Engineering, University of Rajshahi
19. Dr. Rezwana Sultana Associate Professor, Department of Electrical and Electronic Engineering, Varendra University

Organizing Committee

Chair

- **Prof. Dr. Md. Khademul Islam Molla**
  Vice Chancellor
  Varendra University

Co-chair

- **Dr. A.H.M. Rahmatullah Imon**
  Professor Emeritus, Ball State University, USA
  Director, Center for Interdisciplinary Research (CIR)
  Professor, Dept. of CSE.

Secretary

- **Muhammad Sajjadur Rahim**
  Professor
  Dept. of ICE
  Rajshahi University

Joint-secretary

- **Sabina Yasmin**
  Coordinator
  Dept. of CSE
  Varendra University

Members

Prof. Dr. Kazi Khairul Islam
Professor, Dept. of EEE, Varendra University
Shamim Ahsan Parvez
Executive Director, Varendra University
All Faculty Members
Dept. of CSE, Varendra University

International Advisory Committee

Dr. Keikichi Hirose, University of Tokyo, Japan
Dr. Toshihisa Tanaka, Tokyo University of Agriculture and Tech., Tokyo, Japan
Dr. Fahim Hasan Khan, Dept. of CSE, Polytechnique State University, California, USA

National Advisory Committee

Prof. Dr. M. Kaykobad, Dept. of CSE, BRAC University
Prof. Dr. Mamun Bin Ibne Reaz, School of Engineering, Technology and Sciences, Independent University, Bangladesh
Prof. Dr. Md. Shahid Uz Zaman, Dept. of CSE, Rajshahi University of Engineering & Technology
Prof. Dr. Md. Rabiul Islam, Dept. of CSE, Rajshahi University of Engineering & Technology
Prof. Dr. Md. Ekramul Hamid, Dept. of CSE, Rajshahi University
Prof. Abu Bakar Md. Ismail, Dept. of EEE, Rajshahi University
Prof. Dr. M. Mamun-Ur-Rashid Khandker, Dept. of EEE, Rajshahi University

Technical Program Committee

Chair

- **Dr. Shamim Ahmad**
  Professor
  Dept. of CSE
  Varendra University

Co-chair

- **Dr. Md. Iqbal Aziz Khan**
  Professor
  Dept. of CSE
  University of Rajshahi

Secretary

- **Dr. Md. Golam Rashed**
  Professor
  Dept. of ICE
  University of Rajshahi

Joint-Secretary

- **Arifa Ferdousi**
  Lecturer
  Dept. of ICE
  Rajshahi University

Members

Dr. Yoshikazu Washizawa, University of Electro-Communications, Japan
Prof. Dr. Bimal Kumar Pramanik, Rajshahi University, Bangladesh
Prof. Dr. Dipankar Das, Rajshahi University, Bangladesh
Prof. Dr. Mirza A. F. M. Rashidul Hasan, Rajshahi University, Bangladesh
Dr. Md. Nazrul Islam Mondal, Rajshahi University of Engineering & Technology, Bangladesh
Prof. Dr. Md. Rabiul Islam, Rajshahi University of Engineering & Technology, Bangladesh
Prof. Dr. Boshir Ahmed, Rajshahi University of Engineering & Technology, Bangladesh
Prof. Dr. Halida Homyara, Rajshahi University, Bangladesh
Prof. Muhammad Sajjadur Rahim, Rajshahi University, Bangladesh
Prof. Dr. Md. Al Mamun, Rajshahi University of Engineering & Technology, Bangladesh
Prof. Dr. Md. Ali Hossain, Rajshahi University of Engineering & Technology, Bangladesh
Dr. Md. Rabiul Islam, Neurological Institute, Cleveland Clinic, USA
Dr. Md. Shahriar Hossain, University of Queensland, Australia
Dr. Shamsad Parvin, Buffalo State University, USA
Dr. Shaikh Arifuzzaman, University of Nevada, Las Vegas, USA
Dr. Sania Zahan, University of Western Australia, Australia
Dr. Muhammad Yeasir Arafat, Chosun University, South Korea
Dr. Md. Johirul Islam, Rajshahi University of Engineering & Technology, Bangladesh
Dr. Md Nakib Hayat Chowdhury, Associate Professor and Head, CSE Department, BAUST
Dr. Mahmudul Hasan, Head, Dept. of CSE, Comilla University

Foez Ahmed, Associate Professor, Dept. of ICE, Rajshahi University
Dr. Md. Farukuzzaman Khan, Dept. of CSE, Islamic University
Dr. Asif Zaman, Professor, Dept. of CSE, Rajshahi University
D.M. Asadujjaman, Lecturer, Dept. of CSE, Varendra University

Program Committee

Chair

- **Umme Rumman**
  Assistant Professor
  Dept. of CSE
  Varendra University

Co-chair

- **Mst. Jannatul Ferdous**
  Assistant Professor
  Dept. of CSE
  Varendra University

Secretary

- **Ayesha Akter Lima**
  Lecturer
  Dept. of CSE
  Varendra University

Members

Sumaiya Tasnim, Dept. of CSE, Varendra University
Tanver Ahmed, Dept. of CSE, Varendra University
Afroza Islam, Dept. of CSE, Varendra University
Afifa Tasneem Quanita, Dept. of CSE, Varendra University
Rokaiya Tasnim, Dept. of CSE, Varendra University
Mst. Mazeda Noor Tasnim, Dept. of CSE, Varendra University
Mst. Nafia Islam Shishir, Dept. of CSE, Varendra University
Humayra Tasnim, Dept. of CSE, Varendra University
Mohsiul Mumit Alik, Dept. of CSE, Varendra University

Publication Committee

Chair

- **Dr. M. Babul Islam**
  Professor
  Dept. of EEE
  University of Rajshahi

Co-chair

- **Dr. Md. Ariful Islam Nahid**
  Professor
  Dept. of EEE
  University of Rajshahi

Secretary

- **Mohammad Kasedullah**
  Lecturer
  Dept. of CSE
  Varendra University

Joint-secretary

- **Md. Muktar Hossain**
  Lecturer
  Dept. of CSE
  Varendra University

Members

Tanver Ahmed, Dept. of CSE, Varendra University
Sumaiya Tasnim, Dept. of CSE, Varendra University
Md. Taufiq Khan, Dept. of CSE, Varendra University
Md. Rakibul Islam, Dept. of CSE, Varendra University
Md. Alamin Hossain Pappu, Dept. of CSE, Varendra University

Promotional Committee

Chair

- **Sabina Yasmin**
  Associate Professor
  Dept. of CSE
  Varendra University

Co-chair

- **Mst. Jannatul Ferdous**
  Assistant Professor
  Dept. of CSE
  Varendra University

Secretary

- **Susmit Jahan Rose**
  Lecturer
  Dept. of CSE
  Varendra University

Joint-secretary

- **D.M. Asadujjaman**
  Lecturer
  Dept. of CSE
  Varendra University

Members

Ayesha Akter Lima, Dept. of CSE, Varendra University
Md. Fatin Ilham, Dept. of CSE, Varendra University
Akib Ikbal, Dept. of CSE, Varendra University
Zannatul Mifta, Dept. of CSE, Varendra University
Md. Musfiqur Rahman Mridha, Dept. of CSE, Varendra University
Mohammad Faisal Al-Naser, Dept. of CSE, Varendra University
Khalid Sakib, Dept. of CSE, Varendra University
Md. Mahfujur Rahman, Dept. of CSE, Varendra University

External Members

Dr. Md Nakib Hayat Chowdhury
Associate Prof. and Head, Dept. of CSE, Bangladesh Army University of Science and Technology
Ananya Sarker
Assistant Professor & Head, Dept. of CSE, Bangladesh Army University of Engineering & Technology
Md. Ashraful Islam
Assistant Professor, Dept. of CSE, Bangladesh University of Business and Technology
Nakib Aman
Lecturer, Dept. of CSE, Pabna University of Science and Technology
Ananna Hoque Shathi
Lecturer, Dept. of CSE, Bangladesh Army University of Science and Technology

Finance and Sponsorship Committee

Chair

- **Shamim Ahsan Parvez**
  Executive Director,
  Varendra University

Co-chair

- **Umme Rumman**
  Assistant Professor
  Dept. of CSE
  Varendra University

Secretary

- **Shamim Reza**
  Lecturer
  Dept. of CSE
  Varendra University

Members

Abdus Salam, Deputy Director, Finance and Accounts, Varendra University
Md. Khalid Sakib, Dept. of CSE, Varendra University
Md. Mahfujur Rahman Arnob, Dept. of CSE, Varendra University
Ipshita Tasnim Raha, Dept. of CSE, Varendra University
D.M. Asadujjaman, Dept. of CSE, Varendra University
Ahmed-Al-Azmain, Dept. of CSE, Varendra University
Abdullah Tamim, CIR, Varendra University

Registration & Kit Distribution Committee

Chair

- **Md. Mizanur Rahman**
  Assistant Professor
  Dept. of CSE
  Varendra University

Co-chair

- **Dr. Ahammad Hossain**
  Associate Professor
  Dept. of CSE
  Varendra University

Secretary

- **A.S.M Delwar Hossain**
  Lecturer
  Dept. of CSE
  Varendra University

Members

Md. Ruhul Amin, Dept. of CSE, Varendra University
Md. Shahid Ahammed Shakil, Dept. of CSE, Varendra University
Md. Adnan Sami, Dept. of CSE, Varendra University
Md. Rakibul Islam, Dept. of CSE, Varendra University
Md. Khalid Sakib, Dept. of CSE, Varendra University
Tanzila Zinat, Assistant Registrar, Dept. of CSE, Varendra University
Mahmud Murshed, Senior Executive, Dept. of CSE, Varendra University
Abdullah Tamim, CIR, Varendra University
Md. Wadud Jahan, Executive, Dept. of CSE, Varendra University
Md. Arif Hossen, Executive, Dept. of CSE, Varendra University

Event Management and Decoration Committee

Chair

- **Sumaia Rahman**
  Lecturer
  Dept. of CSE
  Varendra University

Co-chair

- **Md. Toufikul Islam**
  Lecturer
  Dept. of CSE
  Varendra University

Secretary

- **Ipsita Tasnim Raha**
  Lecturer
  Dept. of CSE
  Varendra University

Joint-secretary

- **Salma Akter Lima**
  Lecturer
  Dept. of CSE
  Varendra University

Members

Samira Tareque, Dept. of CSE, Varendra University
Israt Jahan Rinky, Dept. of CSE, Varendra University
Protik Chakroborty, Dept. of CSE, Varendra University
Sumaya Hannan Shova, Dept. of CSE, Varendra University
Afroza Islam, Dept. of CSE, Varendra University
Md. Musfiqur Rahman Mridha, Dept. of CSE, Varendra University
Zannatul Mifta, Dept. of CSE, Varendra University
Mohammad Faisal Al – Naser, Dept. of CSE, Varendra University
Rokaiya Tasnim, Dept. of CSE, Varendra University
Asim Moin Saad, Dept. of CSE, Varendra University
Anupoma Barman Shetu, Dept. of CSE, Varendra University
Md. Alamin Hossain Pappu, Dept. of CSE, Varendra University

Food and Refreshment Committee

Chair

- **Dr. Ahammad Hossain**
  Associate Professor
  Dept. of CSE
  Varendra University

Co-chair

- **Md. Mizanur Rahman**
  Assistant Professor
  Dept. of CSE
  Varendra University

Secretary

- **A.S.M Delwar Hossain**
  Lecturer
  Dept. of CSE
  Varendra University

Members

Shorav Paul, Dept. of CSE, Varendra University
Md. Taufiq Khan, Dept. of CSE, Varendra University
Arun Kumar Sikder, Dept. of CSE, Varendra University
Md. Fayzul Islam, Dept. of CSE, Varendra University
Susmita Paul, Dept. of CSE, Varendra University
Mazeda Noor Tasnim, Dept. of CSE, Varendra University
Mahmud Murshed, Senior Executive, Dept. of CSE, Varendra University
Md. Arif Hossen, Executive, Dept. of CSE, Varendra University
Monjur Rahman, Executive, Dept. of CSE, Varendra University

Invitation and Hospitality Management Committee

Chair

- **Sushmit Jahan Rose**
  Lecturer
  Dept. of CSE
  Varendra University

Co-chair

- **Md. Jamil Chaudhary**
  Lecturer
  Dept. of CSE
  Varendra University

Secretary

- **Md. Shahid Ahammed Shakil**
  Lecturer
  Dept. of CSE
  Varendra University

Joint-secretary

- **Md. Sirajur Rahaman**
  Deputy Director
  Planning and Development
  Varendra University

Members

Arshad Wasif, Dept. of CSE, Varendra University
Md. Mahfujur Rahman, Dept. of CSE, Varendra University
Md. Taufiq Khan, Dept. of CSE, Varendra University
Asim Moin Saad, Dept. of CSE, Varendra University
Zuairia Raisa Bintay Makin, Dept. of CSE, Varendra University

Banner Committee

Chair

- **Sumaia Rahman**
  Assistant Professor
  Dept. of CSE
  Varendra University

Co-chair

- **Md. Fatin Ilham**
  Lecturer
  Dept. of CSE
  Varendra University

Secretary

- **Mohammad Faisal Al – Naser**
  Lecturer
  Dept. of CSE
  Varendra University

Members

Md. Shahid Ahammed Shakil, Dept. of CSE, Varendra University
Md. Jamil Chaudhary, Dept. of CSE, Varendra University
Sushmit Jahan Rose, Dept. of CSE, Varendra University
Ipshita Tasnim Raha, Dept. of CSE, Varendra University

Certificate and Crest Committee

Chair

- **Shamim Reza**
  Lecturer
  Dept. of CSE
  Varendra University

Co-chair

- **Ahmed-Al-Azmain**
  Lecturer
  Dept. of CSE
  Varendra University

Secretary

- **Akib Ikbal**
  Lecturer
  Dept. of CSE
  Varendra University

Members

Md. Fatin Ilham, Dept. of CSE, Varendra University
Md. Adnan Sami, Dept. of CSE, Varendra University
Adrita Alam, Dept. of CSE, Varendra University
Humayra Tasnim, Dept. of CSE, Varendra University

Technical Support Committee

Chair

- **Mohammad Kasedullah**
  Lecturer
  Dept. of CSE
  Varendra University

Co-chair

- **Faridujjaman**
  Deputy Director
  ICT Cell
  Varendra University

Secretary

- **Md. Ashraf Ali Joy**
  Executive
  Dept. of CSE
  Varendra University

Members

Md. Wadud Jahan, Executive, Dept. of CSE, Varendra University
Md. Tareq Ben Musha, Executive, Dept. of CSE, Varendra University
Mahmud Hasan, Executive, Dept. of CSE, Varendra University
Monjur Rahman, Executive, Dept. of CSE, Varendra University
Md. Shahidul Islam, Executive, ICT Cell, Varendra University

Disciplinary Committee

Chair

- **Dr. Md. Abdul Awal**
  Associate Professor and Head
  Dept. of Public Health
  Varendra University

Co-chair

- **Md. Mizanur Rahman**
  Assistant Professor
  Dept. of CSE
  Varendra University

Secretary

- **Md. Toufikul Islam**
  Lecturer
  Dept. of CSE
  Varendra University

Joint-secretary

- **Md. Nour Noby**
  Lecturer
  Dept. of CSE
  Varendra University

Members

A.S.M Delwar Hossain, Dept. of CSE, Varendra University
Akib Ikbal, Dept. of CSE, Varendra University
Md. Ruhul Amin, Dept. of CSE, Varendra University
Md. Shahid Ahammed Shakil, Dept. of CSE, Varendra University

Idea Contest Committee

Chair

- **Dr. Shamim Ahmad**
  Professor & Head
  Dept. of CSE
  Varendra University

Co-chair

- **Md. Musfiqur Rahman Mridha**
  Lecturer
  Dept. of CSE
  Varendra University

Members

Md. Rakibul Islam, Dept. of CSE, Varendra University
Md. Farhan Tanvir Nasim, Dept. of CSE, Varendra University
Mohsiul Mumit Alik, Dept. of CSE, Varendra University
Adrita Alam, Dept. of CSE, Varendra University
Shahara Laila, Dept. of CSE, Varendra University



Keikichi Hirose, PhD
Professor Emeritus
University of Tokyo, Japan

Title: Speech prosody and its modeling

Abstract

In recent years, great progress has been made in artificial intelligence, and now machines can hear and speak like humans. However, this technology relies on huge amounts of data collected in situations similar to those in which the applications will be used. As for speech output, while high-quality has been realized for "well-formed" reading style speech, researchers are struggling to increase expressiveness and add variances (i.e. various utterance styles) in generated speech without relying on a huge speech corpus. Generally speaking, speech has two features: segmental and supra-segmental. The former plays a dominant role in conveying the linguistic content of utterances, while the latter, which is often called as prosody, plays an important role in highlighting important words, expressing speaker's intentions and emotions, regulating turns in conversation, and so on. Researches on prosody is important in realizing expressing/affective speech like humans.

In the talk, I will first explain speech prosody including how it contributes in human communications, and then introduce a quantitative model for fundamental frequency contours, which are the most important physical features of prosody. The talk will present some analysis results, and demonstrate how the model can be applied in language/pronunciation learning and speech synthesis.

Biography

Keikichi Hirose received the Ph. D. degree in electronic engineering in 1977 from the University of Tokyo. He was a professor of the University of Tokyo from 1994. He retired in 2015, and received Professor of Emeritus title. From March 1987 to January 1988, he was Visiting Scientist at the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, U.S.A. He served as a project professor at National Institute of Informatics during June 2019 - March 2023. He has been engaged in a wide range of research on spoken language processing, including analysis, synthesis, recognition, dialogue systems, and computer-assisted language learning. From 2000 to 2004, he was Principal Investigator of the national project "Realization of advanced spoken language information processing utilizing prosodic features." He served as the general chair for INTERSPEECH 2010, Makuhari, Japan. Since 2010, he served as the Chair of ISCA SProSIG until 2018. Also, he served as a board member of ISCA during 2009-2017. He is a member of International Advisory Council from January 2021 (to 2024). He received ISCA fellow grade in 2018. He became an honorary member, Polish Phonetic Association, in 2013. For his long-term and remarkable contribution to spoken language processing, he received Achievement Award from Acoustical Society of Japan in 2020. In 2015, he was honored as a Named Person of Merit in Science and Technology by the Mayor of Tokyo.



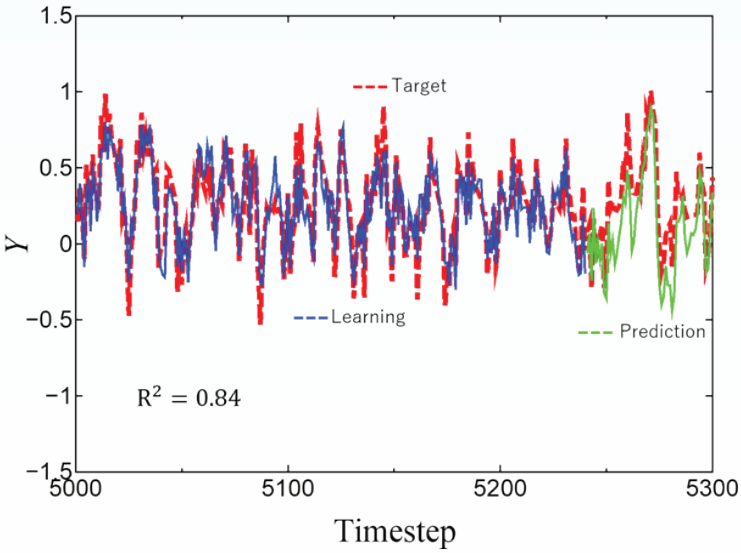
Edmund Soji Otabe
Professor
Kyushu Institute of Technology, Japan

Title: Reservoir computing using superconductor

Abstract

In recent years, machine learning-based classification has gained significant attention; however, the high computational cost associated with training remains a major challenge. As a potential solution, research into physical reservoir computing, which leverages nonlinear physical phenomena, has been actively pursued. Despite this, the range of physical phenomena currently utilized as reservoirs is still quite limited.

Superconductors are materials that exhibit nonlinear relationships between the applied current and the resulting electric field. In this study, we investigated the potential of superconducting phenomena as physical reservoirs by visualizing these phenomena through simulations and using the resulting electric field data to perform various time-series prediction tasks.



First, we solved the Time-Dependent Ginzburg-Landau (TDGL) equations using the Affine Integrator (AFI) method to visualize two-dimensional superconducting phenomena. Initial and boundary conditions were applied to the order parameter Ψ, and its values Ψij at coordinates (i, j) were updated at each time step τ. The magnitude and phase of Ψ were visualized, along with the magnetic flux density, current density, and electric field within the superconductor.

Next, we applied a sinusoidally varying current density u(t) along the y-axis and computed the electric field at 50 randomly selected points within the visualized region. These electric field values were used as reservoir nodes. Using simple waveforms such as sine waves as teacher signals, we performed waveform generation tasks via reservoir computing. We then replaced the current density variation with a random waveform and used NARMA2 (Nonlinear Auto-Regressive Moving Average 2) as the teacher signal and evaluation metric for nonlinear memory tasks. The results confirmed the basic capabilities of the superconducting reservoir, including nonlinearity and harmonic components. The figure presented shows the NARMA2 task, where post-training predictions achieved an R^2 score of 0.84.

Experiments using superconducting reservoirs for speech recognition tasks will also be reported in the talk.

Biography

Edmund Soji Otabe is a Professor at the Kyushu Institute of Technology, affiliated with the Department of Physics and Information Technology, within the Faculty of Computer Science and Systems Engineering. His academic journey began at Kyushu University, where he earned his undergraduate degree in 1987. He continued his research career at Kyushu

Institute of Technology, joining as a Research Associate in 1991, and later received his Ph.D. from Kyushu University in 1997. In 2008, he was appointed as a full Professor at Kyushu Institute of Technology.

Professor Otabe's research focuses on the application of superconducting materials, particularly in the development of advanced superconducting devices and systems. His work spans both theoretical and experimental domains, including finite element method (FEM) analysis of electromagnetic phenomena in high-temperature superconductors. His studies cover critical aspects such as AC loss, magnetic levitation, and current transformation, contributing to the understanding and optimization of superconducting technologies.

Among his notable achievements is the fabrication of a 1 T Bi-2223 superconducting magnet with a 92 mm bore diameter operating at 77 K, published in Advances in Cryogenic Engineering (Vol. 1573, pp. 754–759, 2014). He also led the design and fabrication of a 10 kA-class superconducting DC transmission cable, utilizing the longitudinal magnetic field effect, in 2021.

Professor Otabe is an active member of several academic societies, including:

- The Cryogenic and Superconductivity Society of Japan (CSSJ)
- The Japan Society of Applied Physics (JSAP)
- The Institute of Electronics, Information and Communication Engineers (IEICE)

His contributions to the field of superconductivity have advanced both academic understanding and practical applications, particularly in energy transmission and magnetic systems. He continues to mentor students and collaborate internationally, promoting innovation in superconducting technologies.

Keynote-3



Mohammad Kaykobad, PhD
Distinguished Professor
BRAC University, Bangladesh

Title: Research in undergraduate years

Abstract

Young people are bestowed with creativity and great spirit that decays with the touch of boring, uninteresting and uninspiring education system that exists in our country. In advanced countries there are examples praiseworthy research that of Australian Terence Tao publishing his first paper at the age of 11, Hungarian Lajos Posa at the age of 12, the youngest MIT Professor (22) at the age of14, Paul Erdos publishingat the age of 17, Maryam Mirzakhani and Roya Beheshti at the age of 18, Albert Einstein at the age of 22, and in 1905 at the age of 26 he published four papers one of which gave him Nobel Prize, SN Bose at the age of 24. For a long time, the education system was plagued with perennial session jam often exceeding 3 years in a 4-year degree course. This necessitates initiation of research activities even in undergraduate years. We also highlight Some success stories ofBangladeshi students within our knowledge. We also discuss some notoriousproblems that have been solved in recent times. We also list easily formulated extremely difficultproblems that are waiting to be cracked. We expect some of our interested students to try out these difficult problems for their solution.

Biography

Dr Mohammad Kaykobad was born on 1 May, 1954 at village Jabra, District Manikganj. He received M.S.(Hons.) in Engineering from OMEI, now Odessa State Maritime University, in 1979. In 1982 he received an M. Eng. Degree from

AIT, Thailand, and Ph.D from the Flinders University of South Australia in 1988. Dr Kaykobad is a Fellow of Bangladesh Academy of Sciences(BAS), and was an Associate Secretary of its Executive Council.

Dr Kaykobad is now a Distinguished Professor of CSE Department, Brac University. He is a former Professor at CSE Department, BUET since 1997, and served as its Head during 1996-1999. He served as the Dean of EEE faculty, BUET.

He published over 60 research papers in Physical Review, International Journals of Computer Mathematics, Computers & Operations Research, Computers & Mathematics with Applications, Linear Algebra & Its Applications, Information Processing Letters, Information Processing and Management, Applied Mathematics E-Notes, Journal of Computing and Information Technology, Artificial Intelligence in Medicine, Bioinformatics, Journal of Theoretical Biology, Swarm and evolutionary computation, BMC medical genomics Electronic Notes in Discrete Mathematics, Journal of International Olympiad in Informatics and Computers & Graphics. He authored, coauthored or edited 13 books. Many of these books are for inspiring young students develop either programming or mathematics skill or for developing patriotism. He is a guest co-editor of the proceedings of International Conference on Computer and Information Technology (ICCIT 2009) published by Academy Publisher.

Dr Kaykobad was the Organizing Chair of the firstever international computer conference ICCIT held in the soil of Bangladesh. He was also the Chair of the Steering Committee for Workshop on Algorithms (WALCOM) proceedings of which are published as Lecture Notes in Computer Science by Springer. He has guided more than a dozen undergraduate students whose research works were published in journals of international repute. Dr. Kaykobad participated as a Resource Person in a Workshop on Science Communication sponsored by COSTED and UNESCO held at Goa during 18-23 November, 2001. He is a frequent column writer in National Dailies authoring over 300 articles, most of which are related to education of the country. He has been one of the pioneers in introducing Mathematics Olympiad, Olympiad in Informatics and Science Olympiad in Bangladesh. He has been leader/deputy leader of Bangladesh IOI team since 2008 from which Bangladesh received a Gold medal, 7 silver medals and 25 bronze medals. He was director of ICPC Asia Region Dhaka site during 2001-2003. He was the Chief Judge of IIT Kanpur Asia Region Site of ICPC in 2010, 2011 and 2013. He was adjudged as the outstanding coach of ICPC in 2002 at Honolulu, Hawaii. He was the only Senior Coach award recipient at the World Finals of ICPC held at St. Petersburg, Russia in 2013. He obtained the ICPC Lifetime coach award in 2019 at Porto, Portugal. He was also awarded a gold medal by Bangladesh Physics Olympiad Committee.

Dr Kaykobad was a Visiting Professor at the CSE Department, the Chinese University of Hong Kong, at Kyung Hee University, Korea, ANU and Monash University, Australia and Amritapuri University, India. In 2005 the President of the country presented him with a Gold Medal awarded by Bangladesh Computer Society for his contribution to computer programming culture in the country. In 2006 he won the BAS Gold Medal for physical sciences in the senior group which was awarded by the Prime Minister of the country. Dr Kaykobad was a member of the Executive Council of Bangla Academy.

Dr Kaykobad has been playing an active role in the computerization of both public and private enterprises. He worked as a Director of Dhaka and Chittagong Stock Exchanges for many years. He was an independent director of Sonali Bank PLC. Dr Kaykobad is a member of Academic Council and Syndicate of several universities and in the committee of recruitment of South Asian University, New Delhi, India.



Toshihisa Tanaka, PhD
Professor
Tokyo University of Agriculture and Technology, Japan

Title: Speech Brain–Computer Interfaces Using Intracranial EEG: Advances, Challenges, and Prospects

Abstract

Invasive brain–computer interfaces (BCIs) for speech communication have seen rapid progress driven by advances in intracranial electrophysiological recordings such as intracranial EEG (iEEG) and electrocorticography (ECoG). In this talk, we first provide an overview of the current landscape of invasive speech BCIs and introduce major research directions, including decoding articulatory or phonological representations, direct text decoding, and neural vocoder–based reconstruction of speech waveforms.

We then present our recent work on decoding covert speech and reconstructing spoken sentences from intracranial recordings. In particular, we show that models trained on overt speech can effectively decode covert speech, suggesting a practical strategy for addressing the data scarcity problem in covert-speech BCIs. We also introduce a framework for reconstructing speech waveforms by combining a neural encoder with a pre-trained neural vocoder, demonstrating improved reconstruction quality compared to recurrent neural network–based approaches.

Finally, we discuss remaining challenges and future research directions, including dataset scale, model interpretability, and integration with clinical and neurophysiological workflows. Through these perspectives, we highlight the emerging feasibility of neural speech interfaces enabled by invasive neural recordings.

Biography

Toshihisa Tanaka received the B.E., M.E., and Ph.D. degrees from the Tokyo Institute of Technology in 1997, 2000, and 2002, respectively. From 2000 to 2002, he was a JSPS Research Fellow. From October 2002 to March 2004, he was a Research Scientist at RIKEN Brain Science Institute. In April 2004, he joined the Tokyo University of Agriculture and Technology (TUAT), where he is currently a Professor of Electrical Engineering and Computer Science. In April 2025, he was appointed Vice-Trustee and Assistant to the President of TUAT. He also heads the Research Unit of Informatics for Human-Animal Interaction at the One Welfare Research Institute of TUAT.

He was a Royal Society Visiting Fellow at the Communications and Signal Processing Group, Imperial College London, U.K., in 2005, and a Visiting Faculty Member in the Department of Electrical Engineering at the University of Hawaii at Manoa in 2011.

His research interests include signal processing and machine learning, with particular emphasis on brain and biomedical signal analysis, brain–computer interfaces, and human–animal interaction. He is a co-editor of Signal Processing Techniques for Knowledge Extraction and Information Fusion (Springer, 2008) and the leading co-editor of Signal Processing and Machine Learning for Brain–Machine Interfaces (IET, 2018).

Prof. Tanaka has served as an Associate Editor or Guest Editor for several international journals, including IEEE Access, Neurocomputing, IEICE Transactions on Fundamentals, Computational Intelligence and Neuroscience, IEEE Transactions on Neural Networks and Learning Systems, Applied Sciences, Advances in Data Science and Adaptive Analysis, and Neural Networks. He also served as Editor-in-Chief of Signals. He was the General Co-Chair of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) held in Tokyo in 2021. He served as Vice President for Member Relations and Development of APSIPA until 2025.

He is currently a Senior Area Editor of IEEE Signal Processing Letters. He is serving as the General Chair of the 2028 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2028), to be held in Tokyo, Japan. He is a Senior Member of the IEEE and a member of IEICE, APSIPA, the Society for Neuroscience, and the Japan Epilepsy Society. He is also the Co-founder and CTO of Sigron, Inc.



Prof. Dr. Kosin Chamnongthai
Professor
King Mongkut's University of Technology Thonburi (KMUTT)

Title: Eye-gaze-based Human-Intention Detection

Abstract

Nowadays, many robots are working with humans in many public places such as manufacturing factories, hospitals, department stores, airports, construction sites, and so on. Sometimes, crashes occur between a robot and a human, since the robot has no function to sense human intention, while a man has some senses to estimate the intention of another man by detecting eye gaze. Eye gaze detection becomes a basic function, which would be extended to detect intention and other applications such as intention reading for stroke patients, customer intention estimation, worker intention estimation for collaborative robots, game entertainment, and so on. In this talk, I would introduce and discuss three methods of 3D POI (Point of Intention) estimation using eye gazes, multimodal fusion between hand pointing and eye gaze, and multimodal fusion between hand pointing, eye gaze, and depth information for collaborative robots. The trend of research in eye-gaze-based human-intention detection would be concluded in the talk.

Biography

Kosin Chamnongthai currently works as a professor at the Electronic and Telecommunication Engineering Department, Faculty of Engineering, King Mongkut's University of Technology Thonburi (KMUTT). He served as vice president (conference) of APSIPA Association during 2020-2023, president of ECTI Association (2018-2019), editor of ECTI e-magazine during 2011-2015, associate editor of ECTI-CIT Trans during 2011-2016, associate editor of ECTI-EEC Trans during 2003-2010, associate editor of ELEX (IEICE Trans) during 2008-2010, and chairman of IEEE COMSOC Thailand during 2004-2007.

He received a B.Eng. in Applied Electronics from the University of Electro-communications, Tokyo, Japan in 1985, an M.Eng. in Electrical Engineering from Nippon Institute of Technology, Saitama, Japan in 1987, and a Ph.D. in Electrical Engineering from Keio University, Tokyo, Japan in 1991. His research interests include computer vision, image processing, robot vision, signal processing, and pattern recognition. He is a senior member of IEEE, and a member of IEICE, TESA, ECTI, AIAT, APSIPA, TRS, and EEAAT.



Dr. Kaushik Roy
Professor and Chair, Jefferson-Pilot/Ron McNair Endowed Chair,
Department of Computer Science North Carolina A&T State University, NC, USA

Title: Trustworthy AI for Cyber Defense in the Age of Large Language Models

Abstract

AI is revolutionizing cybersecurity by enabling faster threat detection, adaptive defense, and predictive analytics. Large Language Models (LLMs) play a pivotal role in this transformation, offering capabilities in threat intelligence analysis, automated incident response, and anomaly detection. However, their integration introduces new risks, including adversarial exploitation and trust concerns. This talk explores the overlap between AI and cybersecurity, emphasizing the importance of building robust, transparent, and trustworthy systems. We will discuss how LLMs can enhance cyber defense while also requiring careful governance and validation. Looking ahead, AI will increasingly shape the cyber domain, making trust in intelligent systems a cornerstone of future digital resilience.

Biography

Dr. Kaushik Roy is a Professor and Chair in the Department of Computer Science at North Carolina A&T State University (NCAT). Dr. Roy holds a Jefferson-Pilot/Ron McNair Endowed Chair position in the Department of Computer Science. Dr. Roy has been the PI/Co-PI on $31M (as a PI $14M) in research grants funded by the NSF, DoD, DoE and industry partners. His current research is heavily focused on AI-enhanced cybersecurity, IoT, cyber-physical systems, cyber identity, biometrics, AI and machine learning, and large language models. He has over 220 publications, including 63 journal articles and a book. Dr. Roy is the director of the Center for Cyber Defense. Dr. Roy directs the Center for Trustworthy AI. Dr. Roy also directs the Cyber Defense and AI lab. Dr. Roy mentored 13 postdocs/research associates and supervised 19 doctoral dissertations, 21 MS theses, and 4 MS projects. He also advised over 80 undergraduate RAs. Dr. Roy is an Editor of IEEE Security & Privacy. He received the University's Senior Researcher of the Year Award for 2025, which is regarded as the highest research award at the university.



Prof. Dr. Mohammad Ali Moni
Program Lead, AI and Digital Health Technology, Artificial Intelligence and Cyber Futures Institute, Charles Sturt University, Australia
Professor (Research), Washington University of Science & Technology, USA
Adjunct Professor, Daffodil International University, Bangladesh
Senior Fellow (Hon), The University of Queensland, Australia

Title: AI-Driven Wearables and Decision Support Systems for Digital Healthcare

Abstract

With the growing capabilities of AI, we now have the opportunity to move from reactive to proactive, real-time health monitoring. However, integrating AI with wearable and portable devices still faces key challenges, particularly in building robust, explainable, and clinically reliable systems.

Dr Moni will discuss how my team and I have designed portable, AI-enabled, wearable devices capable of continuously monitoring physiological signals and supporting early diagnosis of chronic and acute conditions. These tools are supported by advanced machine learning algorithms that we have developed to analyse multimodal data, including imaging, EEG, ECG, and multi-omics data, to detect disease biomarkers and health patterns.

Dr Moni will then present our work on real-time decision support systems that generate personalised insights and predict critical health conditions, including sleep apnoea, cardiovascular conditions, and mental health. Our goal is to transform digital health delivery by combining wearable technology, AI, and explainable models to enable scalable and personalised care, especially for remote and under-resourced settings.

He will highlight how interdisciplinary collaboration is key to realising next-generation digital health innovations.

Biography

Prof. Dr. Mohammad Ali Moni is a leading global scientist in AI and Digital Health Technology. He completed his PhD in AI and Digital Health at the University of Cambridge and has worked at world-renowned universities including Oxford, Cambridge, and the University of Sydney.

Prof. Moni has published 500+ journal articles, receiving 55,000+ citations with an h-index of 61. He is the recipient of numerous prestigious awards and fellowships, including the Commonwealth Fellowship, USyd Fellowship, UoW Distinguished Engineering Award, Certara Fellowship, and Researcher of the Year Award. Since 2021, he has consistently been ranked among the top 0.1% of scientists worldwide.



Dr. Faisal Tariq
Associate Professor
James Watt School of Engineering, University of Glasgow

Title: Rethinking Engineering Education in the Generative AI Era

Abstract

The rapid emergence of Generative Artificial Intelligence (GenAI) is reshaping the landscape of engineering education, demanding a fundamental rethinking of how we teach, learn, and assess in the 21st century. This invited talk explores the transformative implications of GenAI for engineering programs, particularly within the context of developing nations where resource constraints coexist with immense opportunities for leapfrogging traditional barriers. GenAI has shifted the educational paradigm from knowledge transmission to knowledge transformation, enabling personalised learning, AI-augmented problem solving, and access to advanced simulations that were previously limited to well-funded institutions.

However, this transformation raises critical questions about academic integrity, the relevance of traditional assessments, and the competencies future engineers require. The talk highlights the need for authentic, process-oriented, and creativity-driven assessment methods that emphasise human judgment, ethical reasoning, and real-world problem framing. It also addresses the evolving expectations of industry, where engineers must increasingly collaborate with AI systems and adapt to hybrid roles that blend technical expertise with communication, teamwork, and innovation.

Finally, this talk offers a roadmap for educators and institutions to integrate GenAI responsibly and strategically. By reimagining curricula, empowering faculty, and strengthening academia–industry linkages, developing nations can prepare graduates who are not only employable in global and local markets but are also equipped to lead ethical and inclusive technological advancement.

Biography

Dr. Faisal Tariq is an Associate Professor at the James Watt School of Engineering, University of Glasgow (UofG). A Senior Member of IEEE, he previously served as the Chair of the UK & Ireland Section of the Vehicular Technology Society. He is also a Fellow of the UK Higher Education Academy, underscoring his commitment to excellence in teaching and academic leadership.

Bringing a wealth of academic and professional experience, Dr. Tariq has held key faculty positions at Queen Mary University of London (QMUL) and conducted research at world-renowned institutions such as Chalmers University of Technology (Sweden), Aalborg University (Denmark), and the National Institute of Informatics (Japan).

Dr. Tariq is an internationally recognised expert in 6G technology and the application of Artificial Intelligence (AI) in engineering. He has edited four influential books covering topics such as mobile applications, 6G communications, AI in healthcare, and smart manufacturing. His scholarly contributions include around 50 peer-reviewed publications in top-tier journals and international conferences, and he has delivered keynote speeches at prestigious international conferences and workshops worldwide.

Invited Talk-1



Dr. M. Sohel Rahman
Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET), Bangladesh

Title: Controlling Malaria with the help of agent-based models

Abstract

Malaria is one of the most devastating global health issues. Wouldn't it be wonderful if we could model the Malaria vectors (i.e., different mosquito species) so as to check which intervention technique would be the most appropriate in a particular area? In this talk, we discuss the design and implementation of a spatial agent based model based on the biological attributes of a Malaria vector called Anopheles vagus, which is widely distributed in Asia and a dominant vector in Bandarban, Bangladesh. Real-life field data of Bandarban have been used to generate landscapes to run the simulations. Validation of the model has been done using several standard techniques. Also, verification and validation of the model was done against real-life field data.

Using artificial landscapes, the individual and combined efficacies of existing vector control interventions have been modeled, applied, and examined. Thus this agent based model now can aid us in deciding what sort of interventions would be most appropriate to prevent or contain Malaria. For example, for Anopheles vagus, and based on the real-life data of Bandarban, this research output suggests that combined interventions will have the best effect.

Biography

Dr. M. Sohel Rahman is a distinguished academic serving as a Professor of CSE at BUET and a Member of the Bangladesh Public Service Commission, appointed in November 2024. With over 22 years of experience, he is an internationally recognized authority in Algorithms, Stringology, and Bioinformatics. His research background is extensive, comprising approximately 110 peer-reviewed journal publications and over 100 conference papers. His work frequently appears in high-impact journals such as Bioinformatics, Theoretical Computer Science, and PLOS One, focusing on string algorithms, biological sequence analysis, and metaheuristics.

An ACM Distinguished Member and Fellow of the Bangladesh Academy of Sciences, Dr. Rahman's recent research integrates machine learning with genomics to analyze viral mutations and transcriptomic data. He earned his PhD from King's College London and has held visiting positions at University College London and McMaster University. Beyond his theoretical contributions, he has served as an Associate Editor for BMC Research Notes and an Academic Editor for PLOS One. Dr. Rahman also provides significant leadership in the global computing community, having served as a Distinguished Speaker for both the ACM and IEEE Computer Society. In 2026, he remains a pivotal figure in advancing Bangladesh's computational research landscape while overseeing critical national recruitment processes through the PSC.

Invited Talk-2



Dr. Mosaddek Hossain Kamal
Professor
Department of Computer Science & Engineering
University of Dhaka, Bangladesh

Title: Embedded Systems–Driven Industrial Transformation in Bangladesh: Challenges and Opportunities

Abstract

Bangladesh's economic development has been driven primarily by the ready-made garments (RMG) sector; however, this labor-intensive and low-value-addition model is increasingly unsustainable due to automation, rising costs, and global sustainability and compliance pressures. Long-term growth requires a transition toward innovative, technology-driven industries built on embedded systems, microcontrollers (MCUs), and the Internet of Things (IoT). Although academic research output and undergraduate engineering projects have increased significantly, their contribution to the establishment of new technology-based industries remains limited. Much of this work is publication- or demonstration-oriented, with weak pathways for technology transfer, productization, and commercialization. This keynote emphasizes a shift from isolated experiments to system-level, problem-driven embedded system research and design, addressing real-world constraints such as reliability, scalability, power efficiency, security, and maintainability. By developing deployable innovative systems for manufacturing, agriculture, energy, healthcare, and innovative infrastructure, undergraduate students can help bridge the research–industry gap, strengthen innovation capacity, and support sustainable national development aligned with the Sustainable Development Goals (SDGs).

Biography

Dr. Mosaddek Hossain Kamal is a Professor in the Department of Computer Science & Engineering (CSE) at the University of Dhaka, Bangladesh, where he also serves as the Director of the university's ICT Cell. He earned degrees from the University of Dhaka, a master's from the University of New South Wales, and a Ph.D. from Concordia University in 2017. His doctoral research was recognized with the ENCS Doctoral Prize. Dr. Kamal has held postdoctoral positions at Concordia University (2017-2018) and the University of Toronto (January 2020 to June 2022).

His research interests include smart grid cybersecurity, electric vehicles (EVs), the Internet of Things (IoT), energy management, and more. His work includes contributions to Volt-VAR optimization using EVs and a survey on the cyber-physical security of active distribution networks. As of early 2026, his research has received 1,398 citations, with an h-index of 14 and an i10-index of 15.



Dr. M. Babul Islam
Professor
Department of Electrical & Electronic Engineering
University of Rajshahi, Bangladesh

Title: Effectiveness of the Frontal Pole EEG Signals in Diagnosing MDD

Abstract

Major Depressive Disorder (MDD) is a debilitating mental health condition that demands innovative diagnostic approaches to elucidate its complex neural mechanisms. Dysfunctions in emotional regulation, decision-making, and cognitive processing are central to the manifestation of depressive symptoms. This study investigates the diagnostic potential of frontal pole EEG signals, with a particular focus on the lateral frontopolar region (Fp1) and the medial frontopolar region (Fp2). These regions serve distinct roles within functional brain networks: Fp1 is primarily associated with cognition, working memory, and perceptual processes, while Fp2 is more closely linked to affective regulation and social cognition. In this study, we employed the Continuous Wavelet Transform (CWT) for scalogram representation of frontal pole EEG signals, as it can preserve both time and frequency information of EEG signals, which enables the extraction of clinically effective features of brain dynamics. In the scalogram representations of frontal pole EEG signals, distinct spectral power alterations are observed, specifically in theta and alpha bands associated with MDD. Consequently, frontal pole EEG signals might be considered as a reliable biomarker for diagnosing MDD.

Biography

M. Babul Islam is a Professor in the Department of Electrical and Electronic Engineering (EEE) at the University of Rajshahi, Bangladesh. He obtained his BSc (1992) and MSc (1993) degrees in Applied Physics and Electronics from the University of Rajshahi. In 2007, he earned a Doctor of Engineering in Speech Recognition from Shinshu University, Japan.

Before joining the University of Rajshahi, he served as a faculty member in the Department of Computer Science and Engineering (CSE) at Islamic University, Kushtia, Bangladesh, where he worked from April 1999 to February 2010. During his tenure in the CSE department, he held the position of Chairman from May 2007 to February 2010. In addition to his academic responsibilities, he served as the Director of the ICT Center at the University of Rajshahi, contributing to the development and modernization of ICT infrastructure and services at the university.

He has also held the position of Conference Secretary for IC4ME2 and Publication Chair for several international conferences, supporting the organization and dissemination of scholarly research.

Professor Islam is an experienced academic and researcher with research interests in signal processing, speech, image, and biomedical signal analysis, machine learning, and bioinformatics. He has been actively involved in teaching undergraduate and postgraduate courses, curriculum development, supervising students' research, and publishing in reputed journals and conferences. He is committed to mentoring young researchers and advancing engineering education and research.



Dr. Dipankar Das
Professor
Department of Information & Communication Engineering
University of Rajshahi, Bangladesh

Title: Intelligent Machine Perception: Computer Vision, Techniques for Human–Robot Interaction and Autonomous Vehicle

Abstract

Computer vision has emerged as a cornerstone technology in enabling intelligent machines to perceive, interpret, and interact with the world around them. As advances in deep learning, multimodal perception, and real-time processing continue to accelerate, unprecedented opportunities are unfolding in two transformative domains: Human–Robot Interaction (HRI) and Autonomous Driving. This keynote explores the cutting-edge research and development directions that are shaping these fields, highlighting how vision-based systems are bridging the gap between machines and human-centered environments.

In the context of HRI, computer vision enables robots to recognize gestures, emotions, and human intention, thereby fostering natural, safe, and adaptive collaboration. For autonomous vehicles, vision serves as the foundation for scene understanding, navigation, and decision-making in complex, dynamic environments. The convergence of these domains underscores the critical role of robustness, explainability, and ethical AI design in ensuring trust and safety.

Biography

Dr. Dipankar Das is a Professor in the Department of Information and Communication Engineering, University of Rajshahi, Bangladesh. He is Senior Member of IEEE. Dr. Das was a former head of the Department of Information and Communication Engineering, University of Rajshahi. He completed his PhD from the Saitama University, Japan. He worked under the Toyhashi University of Technology (TUT)-Kobelco Construction Machinery joint project for developing Autonomous Construction Vehicle as a researcher from January 2020 to December 2021. He also worked as a post-doctoral researcher in the JST CREST (Core Research for Evolutional Science and Technology, Japan Science and Technology Agency) project from October 2011 to March 2014. His current research is heavily focused on Autonomous Driving, Human-Robot Interaction and Natural Language Processing for Low Resource Languages. He has over 100 publications, including 9 book chapters, 33 journal articles, and 58 international conference publications. He delivered more than 12 keynote/invited talks including talks in Caltech Vision Lab and UCLA Vision Lab, Los Angeles, USA.

Invited Talk-5



Dr. Md. Anisuzzaman Siddique
Professor
Department of Computer Science & Engineering
University of Rajshahi, Bangladesh

Title: Distributed Database Management Systems: Powering Modern Data

Abstract

In today's digital world, massive amounts of data are generated daily by universities, banks, e-commerce platforms, and online services. Storing all this data in a single location can be slow, risky, and inefficient. Distributed Database Management Systems (DDBMS) address this challenge by storing data across multiple computers while allowing users to access it seamlessly as one unified database.

This keynote introduces the core concepts of distributed databases in a simple and engaging way. Through real-world examples such as student record systems, and online banking, attendees will learn how these systems improve speed, reliability, and scalability. The talk also highlights common challenges, including data consistency, system failures, and secure transactions, and explains how modern DDBMS overcome them.

Attendees will gain a practical understanding of why distributed databases are essential in today's technology-driven world and how they support critical services. With clear explanations and relatable examples, this keynote makes complex systems easy to understand and relevant for students and professionals alike.

Biography

Dr. Md. Anisuzzaman Siddique is a Professor in the Department of Computer Science and Engineering at the University of Rajshahi, Bangladesh. He is an experienced researcher with strong expertise in data-driven computing and information engineering. He received his Ph.D. in Information Engineering from Hiroshima University, Japan, in 2010. His research interests include big data processing, information retrieval, skyline queries, and intelligent query processing. He has worked on several international research projects funded by JSPS and Micron Memory Japan, focusing on efficient data analysis, and ranking algorithms. Dr. Siddique has published 46 research papers in national and international journals, including Q1 and Q2 journals. He has served as a Principal Investigator and research team member in multiple projects and actively supervises undergraduate and postgraduate research. His ongoing research aims to develop efficient data analysis methods that support real-world applications in large-scale and complex data environments.



Dr. Abdul Alim
Senior Network Engineer, IBM

Title: Designing RoCE Networks for AI Workloads

Abstract

The rapid explosion of Artificial Intelligence (AI) workloads utilizing a growing number of accelerators has placed unprecedented demand on the network. These workloads typically leverage RDMA and require a high-performance network fabric. While many purpose-built/proprietary networking solutions, e.g., Infiniband networks, provide high performance but they expensive to build and operate and does not support application diversity. In this talk, we will first discuss the challenges of designing a high-performance network for private data centers at scales supporting a wide range of workloads over RoCE and TCP using off-the-shelf hardware and software. We will then discuss the network control plane and routing protocols that accommodate flexible reliability and performance requirements. We deployed our solution to a number of GPU clusters spanning multiple racks of servers with Nvidia's H100 8 GPUs and 8 Mellanox 400G dual-port CX7 NICs per server using a 2-stage Clos topology. Using network micro-benchmarks and NCCL benchmarcks that mimic the demands of LLM model training, our results show that the network sustains near-line-rate throughput under traffic patterns where up to 100% of the traffic traverses the spine layer. In production deployments, running LLM training jobs for months, we observe that our RoCE network achieve similar performance compared to Infiniband networks.

Biography

Dr. Abdul Alim is a Staff Research Scientist at the IBM Thomas J. Watson Research Center in Yorktown Heights, NY. He works on the design and implementation of high-performance networks for AI infrastructures. Prior to this role, he contributed to IBM Cloud, where he designed and implemented virtual network functions—such as network load balancers—and various network services for the IBM Cloud platform. Dr. Alim graduated from the University of Rajshahi in Bangladesh, earning a Bachelor's degree in Computer Science in 1999 and a Master's degree in Computer Science in 2001. He later completed a second Master's degree in Computer Science from the University of Northern British Columbia, Canada, in 2007, followed by a PhD in Computer Science from the University of Cambridge, UK, in 2011. After completing his PhD, he worked as a post-doctoral researcher at the University of Lancaster, UK, and subsequently at Imperial College London, UK.

Industrial Talk-2



Mohammed Golam Mostofa
IT Delivery Lead (Informatics)
Blue Cross Blue Shield of Michigan

Title: IT Project Delivery framework

Abstract

The IT project delivery framework unfolds through four clear stages. It begins with Inception, where the project's purpose and broad functionalities are outlined. A Rough Order of Magnitude is then prepared and submitted for business approval. Once the budget and funding roadmap receive approval, the Initiation phase begins. Here, teams define detailed requirements and prepare refined estimates. With the confirmed scope and funding, the process advances to Execution, where development (agile or waterfall) takes shape through coding, unit testing, system testing, and user acceptance testing. After successful User Acceptance Testing, the solution moves into production. When the system stabilizes and performs as intended, the project reaches its Closure, completing the full cycle of the delivery framework.

Biography

Mr. Mostofa has spent more than twenty-five years leading complex IT programs and projects across ERP planning and deployment, data warehousing and Databricks, business and system integration, and analytics. He guides diverse teams spread across continents and time zones, managing multi-million-dollar budgets to deliver successful outcomes. He currently lives in Canton, Michigan, where he serves as an IT Delivery Lead at Blue Cross Blue Shield of Michigan.

Industrial Talk-3



Nazme Mohsina
Global Technical Director, AMCA International

Title: Think Smart, Build Bold: Innovate, Fail, and Soar

Abstract

In a rapidly advancing digital world, success in computer science requires a balance of analytical rigor, creative ambition, and personal resilience. This presentation invites students to adopt a mindset that combines strategic thinking with the courage to pursue bold ideas. By examining innovation as a disciplined yet imaginative process, we emphasize that meaningful progress often emerges from experimentation and the willingness to learn from failure. Rather than viewing setbacks as obstacles, students will be encouraged to recognize them as integral steps in developing stronger solutions, deeper insight, and greater adaptability. Through real-world examples and actionable guidance, this session aims to inspire students to approach challenges with confidence, push beyond conventional boundaries, and contribute to technological advancements that have lasting impact. Ultimately, participants will be motivated to think with intention, build with conviction, and rise to their full potential.

Biography

Nazme Mohsina is the Global Technical Director for AMCA International, where she leads the development, governance, and implementation of technical requirements supporting global product testing and certification programs for the air movement and control industry. Her work focuses on ensuring technical rigor, consistency, and integrity across certification processes, standards application, and laboratory operations worldwide.

In this role, Ms. Mohsina is responsible for overseeing AMCA's Publications and Standards Department and managing the laboratory accreditation program, ensuring compliance with internationally recognized testing, quality, and competence criteria. She works closely with standards development organizations, accredited laboratories, manufacturers, and regulatory stakeholders to maintain alignment between AMCA certification requirements and evolving industry and regulatory expectations. She also serves as an arbitrator for technical appeals related to standards interpretation and certification decisions, providing independent and technically sound resolutions to complex issues.

Ms. Mohsina brings a strong engineering background to her leadership role. She holds a bachelor's degree in Electrical and Electronic Engineering from the Bangladesh University of Engineering and Technology and a master's degree in Electrical Engineering from the University of Dayton. To further enhance her effectiveness in technical governance and organizational leadership, she completed a Technical Leadership Certificate from the University of Wisconsin–Madison.

Her contributions to advancing technical excellence and leadership within the HVAC and building systems industry have been recognized by her peers. In 2020, she was named to Engineered Systems magazine's "20 to Watch: Women in HVAC" list, acknowledging her growing influence and impact in a highly technical field.

Industrial Talk-4



Dr. Gazi Saurav Islam
R&D Engineering Manager at Intel Corporation

Title: Leveraging AI in AI Chip Manufacturing

Abstract

In the highly competitive semiconductor manufacturing industry, AI technologies are revolutionizing the production of AI-specific chiplets. Traditional manufacturing approaches struggle with process optimization, leading to higher wafer costs and slower technology deployment. Through the integration of machine learning for process control, computer vision for defect detection, and predictive analytics for yield optimization, manufacturers are realizing significant gains in production efficiency.

Biography

Gazi Saurav Islam is an R&D Engineering Manager at Intel Corporation, where he has spent the past 12 years leading teams focused on testing and validating next-generation microprocessors. He holds a PhD in Biomedical Informatics and has authored more than 10 peer-reviewed publications in the field. Gazi spent his childhood on the Rajshahi University campus and completed his schooling at RU School & College before pursuing Bachelor's degree in Electrical and Electronic Engineering from the Islamic University of Technology.



Nizam Uddin
Co-Founder & Chief Executive Officer, weDevs

Title: From Wheel to AI — Humanity's Journey of Innovation and the Future of the Job Market

Abstract

Human history is defined by a continuous cycle of innovation—each breakthrough reshaping how people live, work, and create value. From the invention of the wheel to the rise of artificial intelligence (AI), technological revolutions have repeatedly transformed societies and job markets. While every major shift has raised fears of job loss, history shows that innovation ultimately redefines work rather than eliminates it.

This talk traces humanity's journey of innovation, highlighting key technological milestones such as mechanization, electricity, computing, the internet, and now AI. It explores how each era disrupted existing professions while giving rise to new roles, skills, and industries. In the current AI-driven transformation, routine and repetitive tasks are increasingly automated, challenging traditional definitions of IT and knowledge-based jobs.

Focusing on the future of the job market—particularly in the IT sector—the talk examines how AI is reshaping roles in software development, system operations, data analysis, and digital services. Rather than positioning AI as a replacement for human workers, the session emphasizes AI as a powerful collaborator that amplifies human creativity, judgment, and problem-solving.

Drawing on real-world industry experience, the talk provides practical guidance for undergraduate students on how to prepare for an uncertain yet opportunity-rich future. It highlights the importance of foundational knowledge, adaptability, continuous learning, ethical awareness, and interdisciplinary thinking. Students will be encouraged to move beyond narrow technical skills and develop the mindset needed to thrive in an AI-enabled economy.

The session aims to inspire the next generation of professionals to view AI not as a threat, but as the next chapter in humanity's long journey of innovation—one that demands responsible leadership, lifelong learning, and human-centered technology development.

Biography

Mr. Nizam Uddin is a technology entrepreneur and the Co-Founder and Chief Executive Officer of weDevs, a global software company building products used by millions of entrepreneurs worldwide. Under his leadership, weDevs developed Dokan, the world's No- 1 multi-vendor marketplace solution, empowering businesses and creators across diverse markets.

Beyond commercial success, Mr. Uddin is deeply committed to building technology with purpose and values. Through the Kahf ecosystem, he is working to create ethical digital alternatives—spanning browsers, advertising platforms, privacy tools, and fintech solutions—designed to promote digital well-being and reduce harmful online influences. His work focuses on aligning modern technology with human values, responsibility, and long-term societal impact.

With extensive experience in software engineering, product development, and technology leadership, Mr. Uddin actively mentors students, developers, marketers and entrepreneurs. He is a frequent speaker on topics such as innovation, artificial intelligence, the future of work, and ethical technology.

Driven by the belief that technology should empower communities rather than control them, his vision is to help shape a future where innovation supports human dignity, ethical growth, and inclusive opportunity.

Industrial Talk-6



Shakil Ahmed
Senior Software Engineer (SRE), Google

Title: Incident Management at Google

Abstract

Systems, especially the complex ones, fail all the time. This talk provides a multi-phase methodology used to manage and resolve technical incidents effectively. Incident responses have six-different phases: preparation (preventive measures and planning), identification (detection and verification of an event), containment (limiting the impact), eradication (addressing the root cause), recovery (restoring the full functionality), and post-mortem (review and lesson learned). The time-tested process made Google resilient: from dealing with common system overload to rare cosmic-ray bit flip.

Biography

Shakil Ahamed is a Senior Software Engineer at Google, working at the largest distributed storage system on the planet. He completed his B.Sc. in Computer Science and Engineering at University of Rajshahi, Bangladesh (year 2014-2018) before starting his career at Enosis Solutions as a Software Engineer. In 2022, he joined Google Ireland.

Industrial Talk-7



Rafsan Jany
Software Development Engineer, Amazon

Title: The Architecture of Scalable Leadership from Amazon's Engineering Culture

Abstract

Amazon’s phenomenal growth from an online bookstore to a global technology giant is often attributed to its engineering prowess. However, the true "architecture" behind this success is a unique leadership culture that scales with the business. This talk will explore how Amazon's core leadership principles and mechanisms – typically used for problem-solving and process improvement – create a resilient, high-performance environment. I will delve into specific engineering practices that foster individual ownership and career growth, explaining how these principles translate into daily operations within Amazon. Attendees will gain actionable insights into implementing scalable leadership architectures in their own organizations, impacting both people and processes, and driving sustainable individual and team growth.

Biography

Rafsan Jany is a Software Development Engineer at the Amazon European Headquarters in Luxembourg. Over the past two and a half years, he has been an integral member of high-impact engineering teams, focusing on building scalable and resilient systems that power core Amazon services. His expertise lies in translating Amazon’s renowned leadership culture into practical architectural solutions for complex technical challenges. He believes that strong leadership frameworks serve as the ultimate “architecture” for driving team success and fostering individual growth.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Hybrid Feature Fusion and Lightweight ANN for Efficient Diabetic Retinopathy Screening Across Datasets

Md. Shohan Arafat
Dept. of Computer Science and Engineering,
Rabindra Maitree University, Kushtia-7000, Bangladesh.
sohanarafat703@gmail.com

Abstract*—Diabetic Retinopathy is a leading cause of preventable vision loss among individuals with diabetes, necessitating rapid and accurate screening solutions. Traditional diagnostic methods rely on manual retinal image inspection, which is time-intensive and requires expert interpretation. This study proposes a hybrid feature-fusion framework that integrates handcrafted quantitative descriptors generated through custom feature extraction algorithms to enable lightweight and explainable detection using a compact Artificial Neural Network (ANN). Retinal fundus images were preprocessed to extract intensity, morphology, texture, frequency, and symmetry-based descriptors, which were combined and stored in structured CSV format. The datasets were split using stratified sampling, and automatic class weighting was applied to mitigate imbalance. Multiple ANN architectures employing GELU and LeakyReLU activations were evaluated, with the best model achieving a test accuracy of 94.48% and a ROC–AUC of 0.95 while operating efficiently on a standard CPU. Cross-dataset validation confirmed strong generalization capability, and feature analysis revealed interpretable associations between extracted features and disease indicators. The results demonstrate that hybrid, feature-based ANN models can deliver CNN-comparable accuracy with significantly lower computational demand, offering a practical pathway toward fast, interpretable, and resource-efficient Diabetic Retinopathy screening in clinical and low-resource environments.

***Keywords*—Diabetic Retinopathy, Machine Learning, Early Detection, Artificial Neural Network, Image Processing**

I. Introduction

Diabetic Retinopathy (DR) is a prevalent complication associated with diabetes [1], marked by gradual alterations in the retinal blood vessels that may result in vision loss and, in extreme cases, total blindness. The conventional identification of DR involves ophthalmologists manually inspecting retinal fundus images to spot lesions, hemorrhages, microaneurysms, or clouding. This method is labor-intensive, time-consuming, and requires specialized skills, which can restrict the scalability of DR screening initiatives, particularly in areas with limited medical resources [2].

Recently, Machine Learning (ML) has emerged as a viable solution for the automated detection and classification of DR in retinal images. Convolutional Neural Networks (CNNs) are frequently utilized due to their capability to automatically extract intricate spatial features from images. End-to-end CNN-based systems have demonstrated high accuracy in tasks related to DR detection. However, they often require considerable computational power and lengthy training periods, which can pose challenges to their practical implementation in resource-limited environments, such as mobile or point-of-care screening systems.

As an alternative, features can be extracted from retinal images and saved in structured formats like Comma-Separated Values (CSV) files, enabling the effective application of traditional ML models. Given that retinal fundus images are inherently structured and reveal distinctive patterns, feature-based strategies can capture crucial information for classification without the full complexity of CNNs. These approaches provide lower computational demands while preserving predictive accuracy, making them particularly suitable for swift and scalable DR detection.

In this research, we focus on generating a diverse set of quantitative image features using custom algorithms to represent retinal structures and intensity patterns in numerical form. These include intensity, morphology, texture, frequency, and symmetry-related descriptors, which together offer a comprehensive feature-fusion representation of each image. The extracted features are organized in CSV format and used to train lightweight Artificial Neural Network (ANN) models. By merging handcrafted feature diversity with compact classification, the proposed framework aims to achieve a balance between accuracy, interpretability, and computational efficiency, providing a practical pathway for automated DR screening in both clinical and low-resource settings.

II. Literature Review

Automated detection of DR using retinal fundus images has been widely studied. Recent works can be broadly categorized into image-based deep learning approaches and feature-based classification approaches, with each having distinct strengths and limitations.

Many studies apply CNNs [3] directly to fundus images for DR classification. These end-to-end pipelines often achieve high accuracy by automatically learning hierarchical features from images. For instance, recent hybrid models [4],

including CNN–Transformer architectures such as Vision Transformer (ViT)[5], Swin Transformer, Convolutional Vision Transformer (CvT), and LeViT, have reported accuracies around 95–98% for binary classification [6]. While effective, these approaches typically require large computational resources and high-dimensional image data, which may limit their deployment in real-time or resource-constrained settings.

An alternative approach is to extract discriminative features from fundus images and use classical classifiers. Commonly extracted features include texture (entropy, contrast, correlation)[7], intensity-based metrics (mean, standard deviation), vessel morphology, and exudate region measurements. These feature-based pipelines often employ SVM, Random Forests, or K-Nearest Neighbors for classification[8]. Although these methods are computationally efficient, they typically do not leverage the representational flexibility of neural networks and may require careful feature selection.

Efficiency in DR detection is increasingly recognized as critical for real-world screening, particularly in low-resource or mobile environments[9]. Full CNN pipelines, while accurate, often involve high computational cost and longer inference times[10]. Conversely, feature-based methods are lighter but less frequently combined with ANN classifiers to exploit non-linear relationships among features. Recent systematic reviews highlight the lack of studies that explicitly address both accuracy and computational efficiency in DR classification[11].

Several gaps in the existing literature on DR detection can be identified. End-to-end CNN methods are computationally intensive and may not be suitable for rapid screening or deployment on devices with limited resources[12]. Feature-based pipelines generally rely on classical ML classifiers and have not been widely applied with ANN for DR detection[13]. Therefore, there is a need for a pipeline that extracts meaningful features from fundus images into a structured dataset and trains a lightweight ANN classifier. This approach can maintain competitive accuracy while significantly reducing computational requirements, making it suitable for fast, real-time DR screening.

Accordingly, our work presents a pipeline in which meaningful features are extracted from fundus images and stored in a CSV file, and a lightweight ANN classifier is trained to classify DR. This approach aims to reduce computational cost while maintaining competitive accuracy, thereby addressing the efficiency gap in automated DR screening.

III. Dataset

The datasets used in this study were obtained from open-access repository, each licensed under the CC BY 4.0, which permits unrestricted use and distribution for research purposes. Both datasets consist of retinal fundus images categorized according to the standard DR severity grading criteria used in clinical diagnosis.

The first dataset[14] comprises images categorized into five DR severity levels: No DR, Mild, Moderate, Severe, and Proliferative DR. The No DR category contains 1,805 images of healthy retinal scans without DR signs. The Mild class includes 370 images showing early microaneurysms, the initial indicators of DR. The Moderate class has 999 images with noticeable abnormalities like hemorrhages and exudates. The Severe class includes 193 DR cases. The Proliferative DR category contains 295 images. This dataset is highly imbalanced.

The second dataset[15], contains a total of 2,838 retinal images categorized into DR and No DR classes. This dataset is well balanced.

No identifiable patient information, such as demographic or clinical data, was included in either dataset, ensuring full compliance with ethical data-handling standards and patient privacy protection. All images were provided in standardized digital formats suitable for computer vision analysis and were used solely for research and model development purposes.

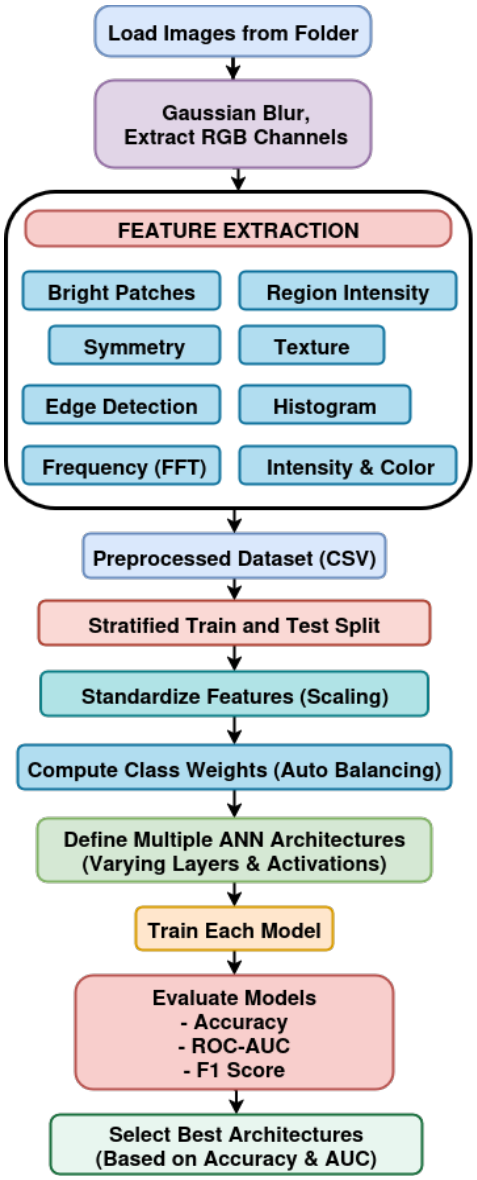


Fig. 1. Study Workflow

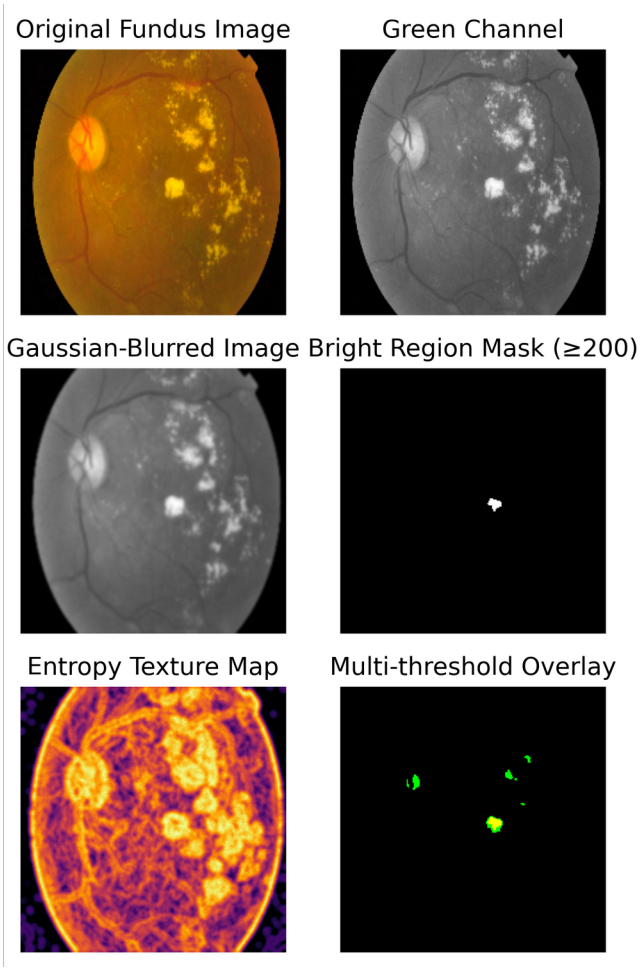


Fig. 2. Input fundus image and its derived feature maps,

IV. Methodology

This section presents the methods and models employed in this study to analyze DR detection. It outlines the dataset selection, preprocessing steps, model architectures, and evaluation strategies. The objective is to provide a clear and reproducible workflow for understanding how class imbalance affects model performance and how results compare across different datasets. The overall workflow of the proposed framework is illustrated in Fig. 1.

A. Class Selection

From the first dataset, only the *DR* and *No DR* classes were selected to evaluate the impact of class imbalance on model performance. Subsequently, the model was tested on another dataset in which the classes were balanced to compare the results.

B. Data Preprocessing

The retinal fundus images were preprocessed and transformed into a structured feature dataset using a custom pipeline developed with OpenCV, scikit-image, and NumPy libraries. Each image was initially smoothed using a Gaussian filter to suppress high-frequency noise, followed by extraction of the green channel, which provides optimal vessel-to-background contrast in retinal imagery[16]. An example of an input retinal fundus image and its corresponding derived feature maps is shown in Fig. 2. A comprehensive set of quantitative descriptors was then computed for each image, encompassing:

- Intensity and color-based metrics: mean, standard deviation, minimum, and maximum intensity values, along with red-green and blue-green channel ratios.
- Morphological and region properties: total and mean bright area, area variance, eccentricity, solidity, extent, compactness, and centroid distribution.
- Texture features: mean entropy values obtained using rank filters.
- Histogram-based brightness measures: dark, midtone, and bright intensity fractions.
- Edge features: proportion of edge pixels detected by the Canny operator.
- Frequency-domain energy: derived from the log-scaled magnitude spectrum of the Fast Fourier Transform (FFT).
- Symmetry measures: determined from vertical intensity differentials across the image midline.

Additionally, multi-threshold bright region fractions (at intensity levels 180, 200, and 220) were extracted to characterize variations in lesion brightness. The resulting features were compiled into a CSV file, forming the input dataset for subsequent model development.

C. Train and Testing Split

To ensure effective evaluation of the model's performance, the dataset was split into training (70%) and testing (30%) subsets. Stratified sampling was utilized to preserve the original distribution of DR and No DR instances.

D. Cross-Dataset Validation

The first dataset exhibited significant class imbalance, whereas the second dataset was relatively balanced. To address this disparity, class weighting was applied during model training on the first dataset to mitigate bias toward the majority class. After completing training, the same model was directly evaluated on the second, independent binary dataset (DR vs. No DR) without any retraining or fine-tuning.

This experimental setup aimed to assess the model's ability to generalize across datasets with differing class distributions and labeling schemes. To ensure compatibility, the second dataset underwent identical preprocessing and feature extraction procedures as the first. Subsequently, key performance metrics were computed to evaluate cross-dataset generalization and validate the robustness of the proposed feature-based ANN framework.

E. Model Training

A variety of feedforward ANNs were created to accomplish binary classification tasks. The standardized feature matrix served as input, while the associated class labels were encoded numerically. To tackle class imbalance present in the initial dataset, automatic class weighting was integrated during the training process. All experiments were executed with fixed random seeds and with data shuffling turned off to guarantee reproducibility and consistency across different runs.

Several network models were examined, each containing four to five fully connected layers with a range of hidden units (from 64 to 1025) and activation functions such as Gaussian Error Linear Unit (GELU) and Leaky Rectified Linear Unit (Leaky ReLU). Batch normalization layers were added following each dense layer to promote stability and speed up convergence.

The output layer featured a single neuron utilizing a sigmoid activation function, and the model was optimized with the Adam optimizer (learning rate = 0.001). The network was trained over 24 epochs with a batch size of 32, applying binary cross-entropy as the loss function. The performance of the model was evaluated on an independent test set using metrics such as accuracy, ROC–AUC score. The Receiver Operating Characteristic (ROC) curve was generated, and the Area Under the Curve (AUC) was calculated to assess the discriminative proficiency of each model. In addition, feature importance was estimated by examining the mean absolute weights of the first hidden layer to determine the most significant input features.

V. Results and Discussion

The performance of the proposed ANN models across both datasets is summarized in Table I. Models using the GELU activation function consistently outperformed those with LeakyReLU. For Dataset 1, the GELU model achieved the highest accuracy of 0.945, a ROC–AUC of 0.950, and a specificity of 0.97, while LeakyReLU models reached accuracies of 0.898–0.930, ROC–AUC values of 0.916–0.923, and specificities of 0.91–0.94. In Dataset 2, the GELU model attained an accuracy of 0.883, a ROC–AUC of 0.935, and a specificity of 0.91, compared to 0.864–0.872 accuracy, 0.928–0.937 ROC–AUC, and 0.91–0.92 specificity for the LeakyReLU models.

The ROC curves for both datasets are presented in Fig. 3 and Fig. 4. GELU-based models consistently show steeper curves, indicating stronger class separation. These results demonstrate that the extracted features effectively capture discriminative patterns for DR and No DR classes, and that the

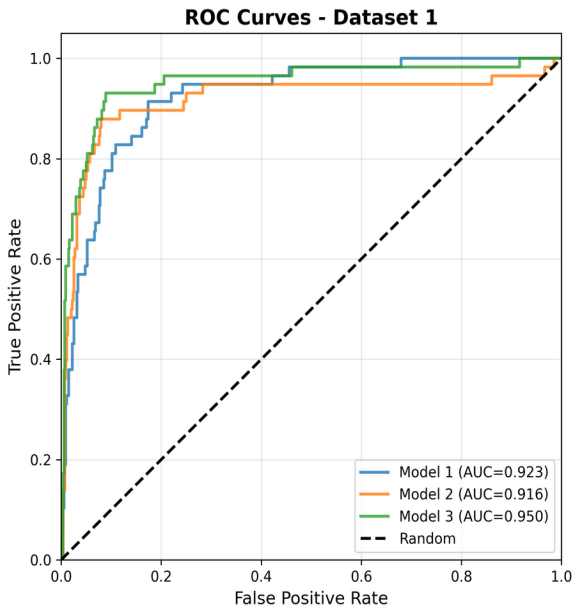


Fig. 3. ROC-AUC for Dataset 1

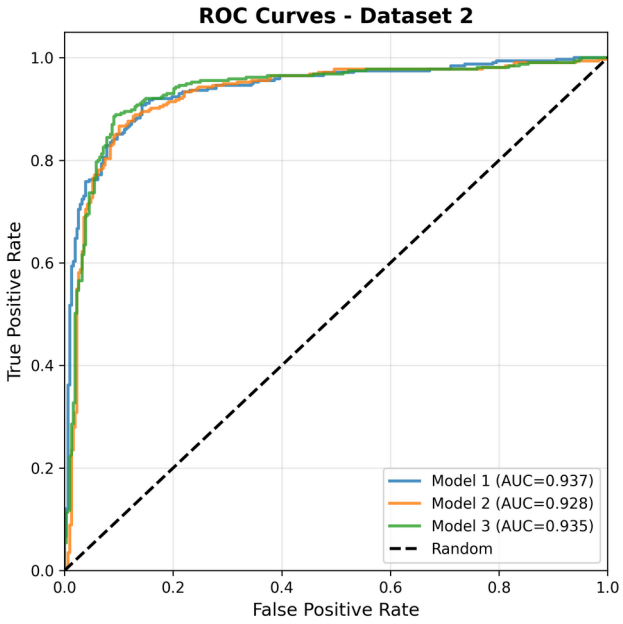


Fig. 4. ROC-AUC for Dataset 2

GELU activation function provides improved convergence and stable generalization across architectures.

Additional feature analysis revealed that certain descriptors—particularly mean intensity, texture entropy, and morphological compactness—had the highest influence on the model's predictions. This indicates that variations in brightness, texture uniformity, and shape regularity of retinal regions play a crucial role in distinguishing DR from non-DR images.

While end-to-end CNN models achieve high accuracy in DR detection shown in Table II, they require substantial computation and long training times. In contrast, the proposed feature-based ANN delivers competitive performance with lightweight computation, completing training on a standard ght computation. In our experiments, training on a standard CPU environment was completed within minuhighlighting the practical applicability of the approach in low-resource or real-time screening settings.

TABLE I. Performance Summery

Dataset	*Model*	*Activation*	*Accuracy*	*ROC-AUC*	*Specificity*
1	1	Leaky ReLU	0.8980	0.923	0.91
1	2	Leaky ReLU	0.9298	0.916	0.94
1	3	GELU	0.9448	0.950	0.97
2	1	Leaky ReLU	0.8636	0.937	0.92
2	2	Leaky ReLU	0.8716	0.928	0.91
2	3	GELU	0.8828	0.935	0.91

TABLE II. MODEL COMPARISON

Model	*Accuracy*	*Specificity*
ANN	0.89	0.97
CNN[17]	0.88	0.97

The proposed approach currently focuses on binary DR classification. Extending the study to multi-class DR severity grading could further enhance its clinical relevance. Additionally, the models rely on structured feature representations extracted from raw images, which enables efficient computation while maintaining strong accuracy. Future work may explore hybrid approaches that combine feature-based ANNs with selective deep feature extraction to balance computational efficiency and capture more detailed spatial information.

VI. CONCLUSION

This study presents a feature-based pipeline for automated DR detection, combining quantitative image descriptors from retinal fundus images with a compact ANN classifier. By extracting meaningful intensity, texture, morphological, and frequency-based features and storing them in a structured format, the approach reduces the computational demands typically associated with end-to-end deep learning models while maintaining reliable predictive performance. The proposed method demonstrates that carefully engineered features, when combined with lightweight neural networks, can capture key patterns indicative of disease severity, making it suitable for rapid and scalable screening. Furthermore, this approach provides flexibility in handling class imbalances and enables systematic evaluation of feature contributions to classification outcomes. Overall, the results highlight that feature-based neural networks can serve as an efficient, practical alternative to computationally intensive models for DR detection. In the future, this framework could be extended to multi-class classification tasks, integrated with real-time screening tools, and adapted for deployment in low-resource clinical environments.

REFERENCES

[1] S. A. Sebastian, O. Elharrouss, S. Al-Maadeed, and N. Almaadeed, "A survey on deep-learning-based diabetic retinopathy classification," *Diagnostics (Basel)*, vol. 13, no. 3, p. 345, Jan. 2023, doi: 10.3390/diagnostics13030345.

[2] P. Ansari, N. Tabasumma, N. N. S. Panduru, N. H. Siam, R. V. N. R. S. Panduru, S. Azam, J. M. A. Hannan, and Y. H. A. Abdel-Wahab, "Diabetic retinopathy: An overview on mechanisms, pathophysiology and pharmacotherapy," *Diabetology*, vol. 3, no. 1, pp. 159–175, 2022, doi: 10.3390/diabetology3010011.

[3] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," *Informatics in Medicine Unlocked*, vol. 20, p. 100377, 2020, doi: 10.1016/j.imu.2020.100377.

[4] S. Aftab and S. Akhtar, "Diabetic retinopathy severity classification using data fusion and ensemble transfer learning," Journal of Software Engineering and Applications, vol. 18, pp. 1–23, 2025, doi: 10.4236/jsea.2025.181001.

[5] W. Nazih, A. O. Aseeri, O. Y. Atallah and S. El-Sappagh, "Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images," *IEEE Access*, vol. 11, pp. 117546–117561, 2023, doi: 10.1109/ACCESS.2023.3326528.

[6] M. M. Butt, D. N. F. Awang Iskandar, S. E. Abdelhamid, G. Latif, and R. Alghazo, "Diabetic retinopathy detection from fundus images of the eye using hybrid deep learning features," *Diagnostics*, vol. 12, no. 7, p. 1607, 2022, doi: 10.3390/diagnostics12071607.

[7] A. D. Vairamani, "Detection and diagnosis of diseases by feature extraction and analysis on fundus images using deep learning techniques," in *Computational Methods and Deep Learning for Ophthalmology*, D. J. Hemanth, Ed. Academic Press, 2023, pp. 211–227, doi: 10.1016/B978-0-323-95415-0.00009-7.

[8] H. Ghodhbani, S. Rimer, K. Ouahada, and A. M. Alimi, "A survey of machine learning techniques leveraging brightness indicators for image analysis in biomedical applications," IEEE Access, Mar. 18, 2025.

[9] S. R. U. I. Rahat, M. H. Rahman, Y. Arafat, M. Rahaman, M. M. Hasan, and M. A. Amin, "Advancing diabetic retinopathy detection with AI and deep learning: Opportunities, limitations, and clinical barriers," BJNS, vol. 5, no. 2, pp. 1–13, Jul. 2025, doi: 10.32996/bjns.2025.5.2.1.

[10] M. Yang et al., "Re-thinking CNN frameworks for time-sensitive autonomous-driving applications: Addressing an industrial challenge," in 2019 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), Montreal, QC, Canada, 2019, pp. 305–317, doi: 10.1109/RTAS.2019.00033.

[11] M. Kropp, O. Golubnitschaja, A. Mazurakova, et al., "Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation," *EPMA J.*, vol. 14, pp. 21–42, 2023, doi: 10.1007/s13167-023-00314-8.

[12] H. A. Essa, E. Ismaiel, and M. F. A. Hinnawi, "Feature-based detection of breast cancer using convolutional neural network and feature engineering," *Sci. Rep.*, vol. 14, p. 22215, 2024, doi: 10.1038/s41598-024-73083-7.

[13] Y. Chen, B. Zheng, Z. Zhang, Q. Wang, C. Shen, and Q. Zhang, "Deep learning on mobile and embedded devices: State-of-the-art, challenges, and future directions," *ACM Comput. Surv.*, vol. 53, no. 4, Art. 84, Jul. 2021, doi: 10.1145/3398209.

[14] P. Tuna, "Diabetic retinopathy," *Mendeley Data*, V1, 2025, doi: 10.17632/nxcd8krdhg.1.

[15] W. Salama, "Diabetic retinopathy," *Mendeley Data*, V1, 2025, doi: 10.17632/47dz8598rc.1.

[16] *L. Cao, H. Li, and Y. Zhang, "Retinal image enhancement using low-pass filtering and α-rooting," Signal Process., vol. 170, p. 107445, 2020, doi: 10.1016/j.sigpro.2019.107445.*

[17] *J. Cao, J. Chen, X. Zhang, et al., "Diabetic retinopathy classification based on dense connectivity and asymmetric convolutional neural network," Neural Comput. Appl., vol. 37, pp. 7527–7540, 2025, doi: 10.1007/s00521-022-07952-5.*

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Assessing Mental Health in Technology Professionals Using Explainable Machine Learning

Amit Hasan Sikder[1], Sumya Akter[1], Rokaiya Tasnim[1],
Md. Iftekhar Hossain Tushar[1], Mahmudul Hasan[1], Lipa Akter[2]
[1]*Department of Computer Science and Engineering*
[2]*Department of Electrical and Electronic Engineering*
Hajee Mohammad Danesh Science and Technology University
Dinajpur-5200, Bangladesh
amithasan16hstu@gmail.com, sumya.hstu@gmail.com, rokaiya.6767@gmail.com,
mdihtushar@gmail.com, mahmudulmoon123@gmail.com, lipaanam32@gmail.com

***Abstract*—In today's tech industry, mental health issues among employees are rising, impacting productivity and well-being. This study proposes a machine learning (ML) framework to predict and explain mental health outcomes for tech workers. Utilizing a publicly available Kaggle dataset, the research builds a robust preprocessing pipeline to impute missing values and encode categorical features, and applies feature selection for improved accuracy. The study evaluates ten ML algorithms and employs Cross-Validation (CV) to boost model performance. SHapley Additive exPlanations (SHAP), a model-agnostic approach, is integrated to enhance model interpretability, building trust in the predictions. Results show Support Vector Machine (SVM) achieving an accuracy of 87.44%, Precision of 84.1% and an F1-scores 88.3%, while CatBoost achieved highest Recall (98.1%). Random Forest attained the highest ROC-AUC score of 0.93.This research highlights ML's potential in improving mental health awareness, supporting proactive intervention strategies. Future work will explore advanced techniques to enhance model performance.**

***Index Terms*—Explainable AI, Machine Learning, Mental Health, Technology Professionals**

I. Introduction

Mental health refers to emotional balance, process thoughts and the capacity to handle social roles, influencing how one handles stress, relates to others, and makes choices [1]. It involves recognizing personal strengths and limitations, managing work-related challenges, and maintaining emotional stability. In recent years, the rapid evolution of the technology sector has introduced substantial demands on professionals, including long working hours, high performance expectations, and the constant need to adapt to emerging tools and skills. These factors have been linked to increased levels of mental pressure, desperation, and burnout, indicating an urgent need for effective mental health assessment and support mechanisms within the technological industry.

While machine learning (ML) has increasingly been applied to mental health research with promising results, its practical implementation still faces key obstacles, including limited and imbalanced datasets, variability in self-reported data, and the lack of model interpretability required for clinical or organizational decision-making. This approach indicates an ML framework designed to detect and explain mental health outcomes among technology sector employees. Using "OSMI" dataset sourced from Kaggle, we construct a preprocessing stack that imputes missing data, encodes categories to numbers, and shapes tha data for the model. Ten ML algorithms including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Gradient Boosting(GB), Adaboost, CatBoost, Decision Tree(DT), Naive Bayes(NB), HistGradient Boost methods are evaluated using GridSearchCV and Cross-Validation (CV) to identify the best-performing model. To improve interpretability, we incorporate SHapley Additive exPlanations (SHAP), a model-agnostic approach that provides insight into how input features contribute to individual predictions. This allows for greater transparency and trust in the decision-making process.The primary objectives of this approach are as follows:

- Assemble a preprocessing stack that patches blanks, maps categorical variables to numerical representations and performs feature selection.
- To evaluate and optimize ten ML algorithms using GridSearchCV with cross-validation for identifying the best-performing model.
- To improve model transparency using SHAP for explainability, highlighting influential features in mental health prediction.

The entire paper is structured as in Section II represents the details of previous and existing works, Section III depicts the overall methodology, descriptions of different methods, and performance measure technique, the Section IV shows the detailed theoretical results analysis and explainability results, discussion, and lastly Section V constitutes the conclusion and future works.

II. Related Works

Over the past few years, researchers have turned to machine learning (ML) techniques to assess mental health has gained significant attention, particularly in pressure-packed environments such as the technology industry. Katarya and colleagues employed six ML algorithms and reduced seventy

TABLE I: Comparison of ML studies on mental health

Ref.	Year	Dataset	Models	Acc.
[2]	2020	OSMI Dataset	KNN, SVM, LR, DT, RF, NB	84%
[3]	2020	OSMI Dataset	DT, RF, NB	82.2%
[4]	2018	OSMI Dataset	RF,LR,KNN,Tree Class,Bagging,Boosting	75.13%
[5]	2020	OSMI Dataset	LR, SVM, RF, Ada, XGB	77%
[6]	2021	OSMI Dataset	SVM, KNN, LR, NB, DT	86%
[7]	2023	OSMI Dataset	SVM, KNN, LR, NB, DT	81.6%

input variables to seven essential attributes, with Logistic Regression (LR) and Decision Tree (DT) classifiers achieving the highest accuracy in predicting mental health concerns among employees from both technical and non-technical domains [2]. Similarly, Laijawala et al. utilized the Weka platform to classify mental health conditions by evaluating three ML models integrated into a web-based system that generated probability scores and personalized recommendations for users and it achieved an accuracy of 82.2% [3]. Further research on IT professionals demonstrated that boosting algorithms achieved superior accuracy in predicting stress levels, identifying gender, family history, and workplace health benefits as the most influential factors with 75.13% [4]. Fitriyanti (2022) applied ML techniques to predict mental health conditions among technology workers using the OSMI survey dataset, which included 1,575 samples and 35 attributes. Five algorithms- LR, SVM, RF, AdaBoost, and XGBoost—were compared, with XGBoost producing the best accuracy of 77% [5]. In a related study, Mallick and Panda (2021) examined mental health issues among tech employees using the "Mental Health in Tech Survey" dataset (1,259 records, 27 features), focusing on predicting whether respondents worked in a tech company [6]. Feature selection based on Information Gain and Chi-square identified key attributes, and five classifiers—SVM, KNN, LR, NB, and DT—were evaluated. SVM, KNN, and LR achieved up to 86% accuracy, while feature selection significantly improved the performance of NB and DT. Most recently, Banerjee et al. (2023) proposed an Explainable Artificial Intelligence (XAI) framework to detect stress among remote technology workers during the COVID-19 pandemic. After cleaning and encoding a QA-based survey dataset, Random Forest (RF) combined with XAI achieved the highest accuracy of 81.6%, outperforming KNN, SVM, and Neural Networks, all of which scored below 80% [7].

III. Proposed Methodology

The complete working procedures of the proposed model are demonstrated in Figure 1. We utilized a publicly available dataset to predict mental health outcomes among technology sector employees.

A. Dataset Description

In this study, The authors used a public dataset from Kaggle [8]. The dataset consists of 1,259 rows and 27 columns with a nearly balanced class distribution: 637 instances labeled as *treatment = Yes* and 622 as *treatment = No*.

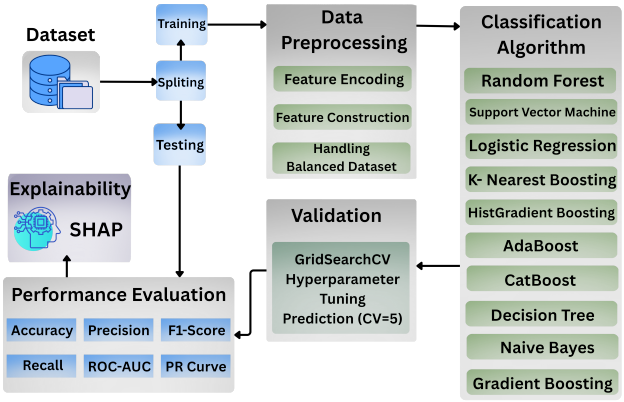


Fig. 1: Working Procedure of the Proposed Framework

B. Data Preprocessing

The dataset was subjected to a rigorous preprocessing pipeline to ensure data quality and compatibility with machine learning algorithms. Non-informative attributes, including *Timestamp* and free-text *comments*, were excluded from analysis. The numerical attribute *Age* was standardized by coercing values to numeric format, imputing missing entries with the median, and restricting the range to 18–65 years to mitigate the influence of outliers. Categorical features were normalized by consolidating inconsistent entries (e.g., mapping gender variants into *Male*, *Female*, and *Other*) and assigning missing values to an *Unknown* category. Ordinal variables such as *leave*, *work_interfere*, *mental_health_consequence*, and *phys_health_consequence* were transformed into ordered numerical scales to preserve rank semantics, while binary attributes (e.g., *family_history*, *benefits*, *anonymity*) were encoded into numerical representations, with uncertain responses mapped to intermediate values (0.5) to capture partial information. To enhance feature expressiveness, composite indices including *policy_support*, *people_support*, and *risk_pressure* were derived by aggregating semantically related survey items. High-cardinality geographic information (*Country*) was collapsed into broader *Region* categories to reduce sparsity. Finally, categorical variables were encoded using one-hot encoding for classical models and integer mappings, while continuous variables were standardized to zero mean and unit variance. This preprocessing ensured consistent data representation and facilitated the integration of heterogeneous features into advanced classification frameworks.

C. Algorithm's Descriptions

In this study, we use ten ML models to predict the mental health in technology professionals. The descriptions of the algorithms are below: Throughout this work, we utilize Jupyter Notebook to implement and evaluate ten machine learning models for predicting mental health disorders in tech employees. The performance of these models is assessed using standard metrics such as accuracy, precision, recall, and F1-score. GridSearchCV is employed to find the optimal hyperpa-

rameters for each model, ensuring the best performance. The optimized hyperparameters are listed in Table II.

TABLE II: Tuned hyperparameter settings used for each model

Model	Selected configuration
RF	n_estimators=400; max_features=sqrt; min_samples_split=2; min_samples_leaf=4
SVM (RBF)	C=2.0; gamma=scale
Logistic Regression	solver=liblinear; penalty=ℓ_1; C=0.1; class_weight=balanced
k-Nearest Neighbors	n_neighbors=31; weights=distance; $p = 2$
Gradient Boosting	n_estimators=200; learning_rate=0.05; max_depth=2; max_features=sqrt; min_samples_leaf=1; subsample=0.8
Decision Tree	criterion=gini; max_depth=None; max_features=sqrt; min_samples_split=2; min_samples_leaf=10
AdaBoost	n_estimators=100; learning_rate=0.3
CatBoost	depth=6; iterations=300; learning_rate=0.03; l2=3
Gaussian Naive Bayes	var_smoothing=1×10^{-7}
HistGradientBoosting	learning_rate=0.05; max_depth=6; max_leaf_nodes=31; min_samples_leaf=20; l2=0.0

The model assessment process, including hyperparameter optimization and model comparison, is explained in Algorithm 1.

1) K-Nearest Neighbors (KNN): KNN is a nonparametric, instance-based algorithm that predicts a sample's label by majority vote among its k nearest training points measured by a chosen distance (often Euclidean) [9].

$$\|\mathbf{p} - \mathbf{q}\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

2) Support Vector Machine (SVM): SVM learns a maximum-margin separator; with kernel K, the decision function is [10].

$$f(\mathbf{x}) = b + \sum_{i=1}^{m} \alpha_i \, y_i \, K(\mathbf{x}_i, \mathbf{x}),$$

and the decision boundary is the set $\{\mathbf{x} : f(\mathbf{x}) = 0\}$.

3) Logistic Regression (LR): Logistic regression models the log-odds of the positive class as a linear function of the features [11]:

$$\log \frac{P(y = 1 \mid \mathbf{x})}{1 - P(y = 1 \mid \mathbf{x})} = \mathbf{w}^\top \mathbf{x} + b.$$

Equivalently, the class-1 probability is

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b), \quad \sigma(t) = \frac{1}{1 + e^{-t}}.$$

4) Decision Tree (DT): A decision tree recursively splits features to maximize node purity; it picks the split with the largest impurity drop, e.g., Gini $G = \sum_k p_k(1-p_k)$ or entropy $H = -\sum_k p_k \log p_k$, where p_k is the class proportion in a node and C is the number of classes. [12].

$$G(t) = \sum_{k=1}^{C} p_k (1 - p_k),$$

and the entropy criterion is

$$H(t) = -\sum_{k=1}^{C} p_k \log p_k.$$

The algorithm selects the split that yields the largest reduction in the chosen impurity.

5) Random Forest (RF): An ensemble of trees trained on bootstrap samples with random feature subsets; predictions are aggregated—majority vote for classification, mean for regression [13]:

$$\hat{y}_{\text{clf}} = \text{mode}\{T_i(\mathbf{x})\}_{i=1}^{N}, \qquad \hat{y}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} T_i(\mathbf{x}).$$

6) Naive Bayes: A probabilistic classifier that assumes features are conditionally independent given the class. Using the factorized likelihood, the prediction rule is [14]

$$\hat{d} = \arg\max_{d} \; Q(d) \prod_{j=1}^{n} Q(x_j \mid d),$$

where $Q(d)$ is the prior and $Q(x_j \mid d)$ are class-conditional terms.

7) AdaBoost (Adaptive Boosting): AdaBoost iteratively trains weak learners on reweighted data—upweighting errors—and predicts by a weighted vote [15]:

$$\hat{y} = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \, h_t(\mathbf{x}) \right),$$

where h_t is the t-th base learner and α_t its weight.

8) CatBoost: CatBoost is a gradient-boosted tree method that handles categorical features via ordered target statistics and uses "ordered" boosting to limit overfitting [16]. At round m it updates

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \, \gamma_m \, h_m(\mathbf{x}),$$

where h_m is the new tree, γ_m its weight, and ν the learning rate.

9) HistGradientBoost: Histogram-based gradient boosting bins features to speed up tree building and updates the model stage-wise [17]:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \, h_m(\mathbf{x}),$$

where h_m is the tree added at round m and ν is the learning rate.

10) Gradient Boosting (GB): Gradient boosting adds weak learners to fit current residuals; at stage m it updates [18]

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \, \gamma_m \, h_m(\mathbf{x}),$$

where h_m is the new tree, γ_m its weight, ν the learning rate, and the final model is F_T.

Algorithm 1: Simple Tuning to Pick Best Model

Input: processed_data; algorithms
Output: best_model; score_table
score_table ← empty;
foreach *alg in algorithms* **do**
  build pipeline (preprocess + alg);
  set param_grid for alg;
  run GridSearchCV (cv=5, scoring=accuracy);
  save (alg, best_score, best_estimator) to score_table;
best_model ← entry with highest best_score;
return best_model, score_table;

D. Performance Measure Metrics

The models are examined using four metrics. Accuracy is the proportion of all samples classified correctly, Precision is the share of predicted positives that are truly positive, Recall reflects the proportion of actual positives correctly identified by the model, F1 captures the trade-off between precision and recall through a harmonic mean [19].

$$\begin{aligned} \text{accuracy} &= \frac{t_p + t_n}{t_p + t_n + f_p + f_n}, \\ \text{precision} &= \frac{t_p}{t_p + f_p}, \qquad \text{recall} = \frac{t_p}{t_p + f_n}, \\ F_1 &= \frac{2}{\frac{1}{\text{prec}} + \frac{1}{\text{rec}}} = \frac{2t_p}{2t_p + f_p + f_n}. \end{aligned} \tag{1}$$

Where: t_p = true positives, t_n = true negatives, f_p = false positives, f_n = false negatives; . The "positive" class denotes the event of interest.

IV. Results and Analysis

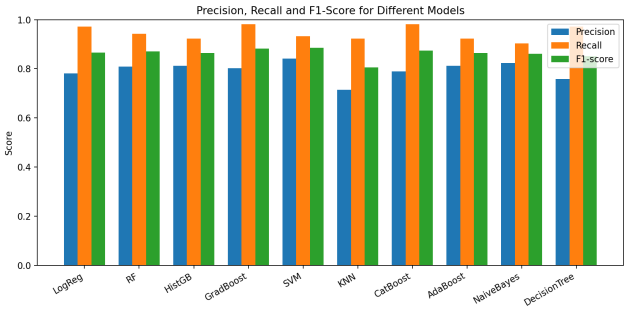
A. Performance Analysis

We applied ten distinct ML methods to train our dataset and assessed each one's performance. In Figure 2a, we found that the SVM classifier achieved the highest in F1-score, which is 88.3%. And closely followed by other models like Random Forest (87.5%), LR (88.1%), and CatBoost (87.4%).

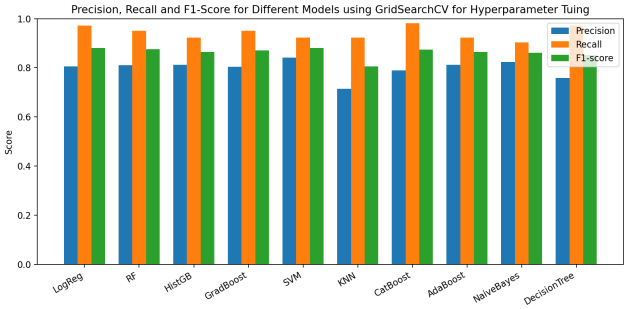
In Figure 3a, we found that SVM performed highest accuracy, which is 87.44% and KNN performed the lowest accuracy, which is 77.78%. Other approaches like Random Forest, Gradient Boosting, and CatBoost performed close to 86%.

The detailed comparison of all models after GridSearch CV described in TABLE III. As shown in Table IV, the 5-fold cross-validation results indicate low performance variation across folds, confirming the stability and robustness of the evaluated classifiers. The SVM model achieves the best average accuracy, while ensemble-based methods also demonstrate competitive and consistent performance.

Figure 4 shows that the model performs consistently well across the 18–54 age buckets, with accuracies around 86–92% and the highest in 45–54 ($\approx 92.3\%$). The 18–24 and 35–44

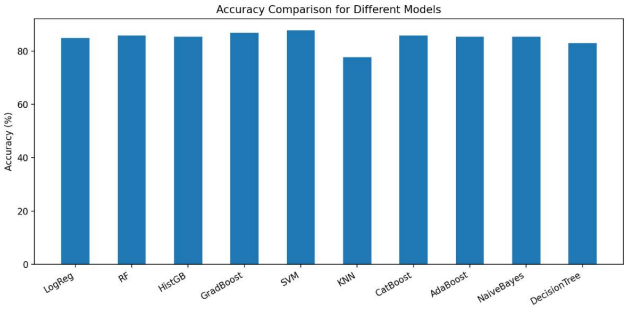


(a)

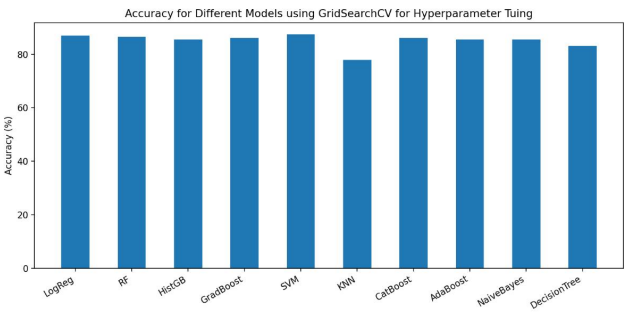


(b)

Fig. 2: Precision, Recall, and F1-score comparison for model: (a) without GridSearchCV, (b) with GridSearchCV.



(a)



(b)

Fig. 3: Accuracy comparison for models: (a) without GridSearchCV, (b) with GridSearchCV.

TABLE III: Performance of Classification Models

S. No	Model	Accuracy (%)	Precision	Recall	F1 Score
1	LR	86.96	80.6	97.1	88.1
2	RF	86.47	81.0	95.1	87.5
3	HistGB	85.51	81.2	92.2	86.4
4	GradBoost	85.99	80.3	95.1	87.1
5	**SVM**	**87.44**	**84.1**	92.2	**88.3**
6	KNN	77.78	71.4	92.2	80.5
7	CatBoost	85.99	78.9	**98.1**	87.4
8	AdaBoost	85.51	81.2	92.2	86.4
9	NaiveBayes	85.51	82.3	90.3	86.1
10	DecisionTree	83.09	75.8	97.1	85.1

TABLE IV: 5-fold cross-validation accuracy

Model	Fold1	Fold2	Fold3	Fold4	Fold5	Avg ± Std
LR	0.865	0.873	0.868	0.878	0.864	0.870 ± 0.006
RF	0.859	0.868	0.864	0.870	0.852	0.865 ± 0.007
HistGB	0.847	0.856	0.852	0.861	0.850	0.855 ± 0.006
GB	0.853	0.862	0.859	0.867	0.858	0.860 ± 0.005
SVM	0.866	0.879	0.872	0.883	0.872	0.874 ± 0.007
KNN	0.769	0.784	0.776	0.789	0.771	0.778 ± 0.008
CatBoost	0.854	0.861	0.858	0.865	0.861	0.860 ± 0.005
AdaBoost	0.848	0.857	0.853	0.862	0.856	0.855 ± 0.006
NB	0.849	0.856	0.852	0.860	0.858	0.855 ± 0.005
DT	0.824	0.835	0.830	0.839	0.827	0.831 ± 0.006

groups are (≈ 89–90%, while 25–34 is slightly lower at 86.3%. The "Others" bucket drops to (≈ 66.7%, likely reflecting missing/invalid ages or a small sample, so this result should be interpreted with caution. Figure 2b indicates that, with Grid-

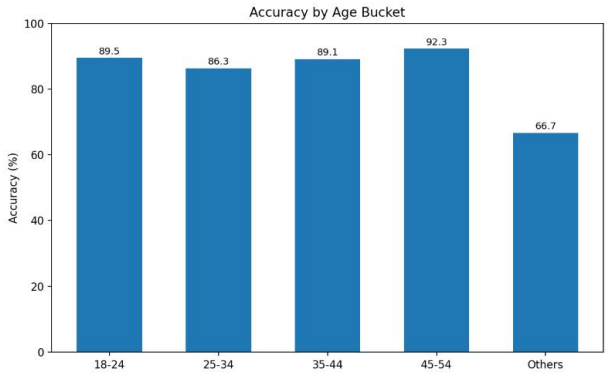
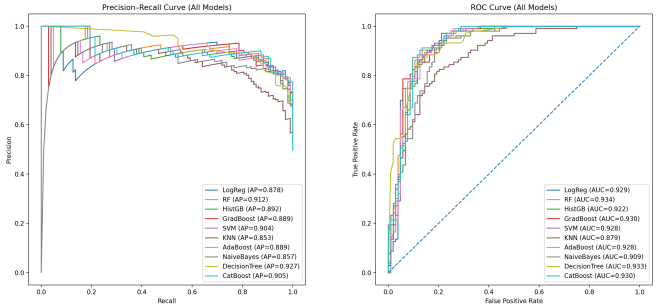


Fig. 4: Accuracy by Age Bucket

SearchCV tuning, most models maintain strong balance (F1 ≈ 0.86–0.88); SVM and Logistic Regression lead (≈ 0.88), followed closely by RF, Gradient Boosting, and CatBoost (≈ 0.87–0.875). Recall generally exceeds precision (e.g., Logistic Regression ≈ 0.97 vs. 0.81), implying a recall-leaning threshold. KNN trails (F1 ≈ 0.81), while Decision Tree and Naive Bayes land mid-pack (≈ 0.80–0.85). Figure 3b shows that accuracies cluster in the mid-80% range; SVM leads (≈ 87–88%), with Logistic Regression and Random Forest close behind (≈ 86–87%), and Gradient Boosting, HistGB, CatBoost, and AdaBoost around ≈ 85–86%. KNN is notably lower (≈ 78%), while Decision Tree sits at ≈ 83%, trailing the stronger ensemble and margin-based models. Figure 5a shows the Precision-Recall curves across all models: the best Precision comes from Decision Tree (≈ 0.927), followed by Random Forest (≈ 0.912), CatBoost (≈ 0.905), and SVM (≈ 0.904)—indicating strong precision over a wide recall range.



(a) (b)

Fig. 5: Performance curves for all models: (a) Precision-Recall Curve, (b) ROC-AUC Curve.

KNN and Naive Bayes trail (AP ≈ 0.853–0.857), while Logistic Regression, Gradient Boosting, HistGB, and AdaBoost land mid-pack (AP ≈ 0.878–0.892); curves nearer the top-right are superior. Figure 5b shows ROC curves across models: Random Forest leads (AUC ≈ 0.934) with Decision Tree close behind (≈ 0.933); Gradient Boosting, CatBoost, and Logistic Regression cluster at ≈ 0.929–0.930, SVM and AdaBoost ≈ 0.928, and HistGB ≈ 0.922. Naive Bayes (≈ 0.909) and KNN (≈ 0.879) trail, while curves closest to the top-left indicate stronger TPR at low FPR for the top models.

Figure 6 summarizes that the strongest drivers are work interfere (Unknown/Sometimes/Often), family history, the ordinal work interfere score, care/benefits, gender, and age/35–44 plus some state/country effects. Points to the right raise the treatment probability (positive SHAP), while points to the left lower it; e.g., higher interference or unknown status tends to push predictions up, whereas "Never," good benefits/care options, and easier leave policies pull them down. Color denotes feature value (pink = high, blue = low), and the horizontal spread reflects each feature's overall impact magnitude across individuals.A detailed comparison with prior studies is provided in TABLE V

B. Explainability Result

We compute SHAP values for the best-performing model (SVM) using and appropriate SHAP explainer and report global feature importance in Fig. 6 `work_interfere` and `family_history` dominate, followed by `Country` (US/UK), `care_options`, `coworkers`, and `Gender`. Higher `work_interfere` (`Often`/`Sometimes`) pushes predictions toward treatment, whereas `Never` pulls them negative. Access to care (`care_options=Yes`) and supportive coworkers generally reduce the positive prediction, while `No`/`Not sure` and low support increase it. Country

TABLE V: Comparison Between Previous Works and Our Work

Ref.	Algorithm	CV	Tuning	SHAP	Accuracy
[2]	KNN, SVM, LR, DT, RF, Naive Bayes	NO	NO	NO	84%
[3]	DT, RF, NB	NO	NO	NO	82.2%
[4]	RF, LR, KNN	NO	NO	NO	75.13%
[5]	LR, SVM, RF, AdaBoost, XGB	NO	NO	NO	77%
[6]	5 (LR, SVM, RF, NB, KNN)	NO	NO	NO	86%
[7]	XAI	NO	NO	NO	81.6%
Our Work	**10 (SVM, LR, RF, GB, DT)**	**YES (CV = 5)**	**YES (Grid-Search CV)**	**YES**	**87.44%**

indicators introduce systematic shifts consistent with contextual differences. Other variables (e.g., `leave`, `anonymity`) show smaller effects, and several remaining features have low global importance.

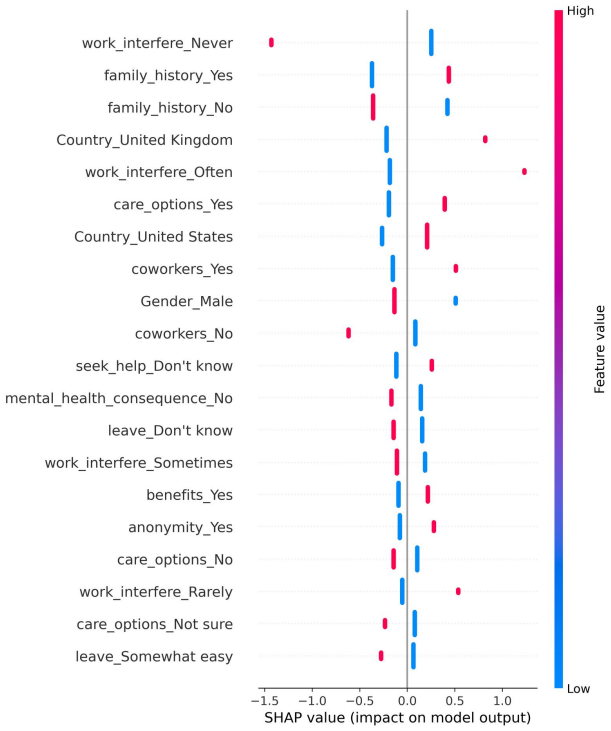


Fig. 6: SHAP Summary for Feature Importance

V. Conclusion and Future Work

In the tech sector, poor mental health materially dents performance and quality of life, so timely and reliable prediction is essential. In our study, matrics CV reached mid-80% accuracy with F1 (≈ 0.86–0.88; while RF achieved the top ROC-AUC ($\approx$ 0.934).Additionally, SHAP was used to indentify importance feature and indicates to better prediction.Future work includes scaling the framework to larger multi-origin datasets, pursue time-series evaluation, and implementation pilot rollouts in live organizations.

References

[1] L. A. Manwell, S. P. Barbic, K. Roberts, Z. Durisko, C. Lee, E. Ware, and K. McKenzie, "What is mental health? evidence towards a new definition from a mixed methods multidisciplinary international survey," *BMJ Open*, vol. 5, no. 6, p. e007079, 2015.

[2] R. Katarya and S. Maan, "Predicting mental health disorders using machine learning for employees in technical and non-technical companies," in *2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE)*. IEEE, 2020, pp. 1–5.

[3] V. Laijawala, A. Aachaliya, H. Jatta, and V. Pinjarkar, "Classification algorithms based mental health prediction using data mining," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2020, pp. 1174–1178.

[4] U. S. Reddy, A. V. Thota, and A. Dharun, "Machine learning techniques for stress prediction in working employees," in *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2018, pp. 1–4.

[5] A. N. Fitriyanti, "Predicting tech workers' mental health condition using machine learning techniques," Ph.D. dissertation, Tilburg University, Tilburg, The Netherlands, 2020.

[6] S. Mallick and M. Panda, "Mental health illness disease prediction using most common features in tech survey dataset," 2020.

[7] J. S. Banerjee, A. Chakraborty, M. Mahmud, U. Kar, M. Lahby, and G. Saha, "Explainable artificial intelligence (xai) based analysis of stress among tech workers amidst covid-19 pandemic," in *Advanced AI and Internet of Health Things for Combating Pandemics*. Springer, 2023, pp. 151–174.

[8] Open Sourcing Mental Illness (OSMI), "Mental health in tech survey," Kaggle.

[9] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 1255–1260.

[10] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Schölkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, Jul. 1998.

[11] M. Maalouf, "Logistic regression in data analysis: an overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 281–299, 2011.

[12] I. D. Mienye and N. Jere, "A survey of decision trees: Concepts, algorithms, and applications," *IEEE Access*, vol. 12, pp. 86 716–86 727, jun 2024.

[13] B. C. V. Essen, C. Macaraeg, M. B. Gokhale, and R. Prenger, "Accelerating a random forest classifier: Multi-core, gp-gpu, or FPGA?" in *2012 IEEE 20th International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2012, pp. 232–239.

[14] T. N. Viet, H. L. Minh, L. C. Hieu, and T. H. Anh, "The naïve bayes algorithm for learning data analytics," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 12, no. 4, pp. 1038–1043, Jul-Aug 2021.

[15] V. D. Quang, H. H. Viet, V. H. Long, and T. D. Khang, "An improved adaboost algorithm for highly imbalanced datasets in the co-authorship recommendation problem," *IEEE Access*, vol. 11, pp. 89 107–89 123, 2023.

[16] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost classifier with other machine learning methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, pp. 738–746, 2020.

[17] A. Guryanov, "Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees," in *Analysis of Images, Social Networks and Texts (AIST 2019)*, ser. Lecture Notes in Computer Science, vol. 11832. Springer, 2019, pp. 39–50.

[18] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, p. 21, dec 2013.

[19] A. Bajaj. (2025, may) Performance metrics in machine learning [complete guide].

AgriCall: An IoT-Integrated GSM-Based Irrigation Control System for Sustainable Agriculture

Sayed Shifat Ahmed
RTM Al-Kabir Technical University
Department of Electrical and Electronic Engineering
Sylhet-3100, Bangladesh
shikor99@gmail.com

Ikramul Hasan Sayem
RTM Al-Kabir Technical University
Department of Electrical and Electronic Engineering
Sylhet-3100, Bangladesh
ihsayem18@gmail.com

Md. Tawhidul Alam Juwel
RTM Al-Kabir Technical University
Department of Electrical and Electronic Engineering
Sylhet-3100, Bangladesh
talamjuwel@gmail.com

Samiatul Ferdaus
RTM Al-Kabir Technical University
Department of Electrical and Electronic Engineering
Sylhet-3100, Bangladesh
samiatulferdaus394@gmail .com

Md. Abed Hossain Siam *RTM Al-Kabir Technical University*
Department of Electrical and Electronic Engineering
Sylhet-3100, Bangladesh
mdabedhossainsiam@gmail.com

Sadikur Rahman
RTM Al-Kabir Technical University
Department of Electrical and Electronic Engineering
Sylhet-3100, Bangladesh
rahmansadik20100037@gmail.com

Md. Ruhul Amin
RTM Al-Kabir Technical University
Department of Electrical and Electronic Engineering
Sylhet-3100, Bangladesh
ruhul2018338046@gmail.com

***Abstract*— This paper proposes *AgriCall*, a cost-effective GSM-based irrigation controller designed to enhance reliability and reduce operational costs for smallholder farmers. The system integrates an Arduino UNO (ATmega328P) with a SIM900A GSM module to enable remote pump switching through missed-call signaling, eliminating call charges. SMS notifications are employed only for critical events such as main power failure or checking status, thereby minimizing communication expenses. Additional features include EEPROM-based user authentication, a 4×4 keypad for stored number management, an LCD for real-time monitoring, and a electricity sensing mechanism with a 5 V backup battery to ensure uninterrupted operation during outages. A functional prototype was developed and tested, demonstrating accurate pump control, timely power-failure alerts, and robust performance under rural operating conditions. The results highlight the potential of the AgriCall device to improve irrigation efficiency by enabling timely irrigation control, thereby contributing to water conservation, reduced energy wastage, and enhanced crop productivity in resource-constrained rural environments.**

Keywords—Rural automation, GSM-based irrigation, low-cost agricultural technology, calling communication, Remote pump control, IoT in agriculture.

I. INTRODUCTION

The efficiency of irrigation practices constitutes a fundamental element of sustainable agriculture, especially in regions characterized by constrained water resources and limited labor availability. The absence of intelligent control mechanisms in conventional irrigation often prevents timely adaptation to fluctuating power availability and field conditions, necessitating low-cost, communication-driven solutions [1]. Smallholder farmers frequently lack affordable tools for reliable pump management, motivating the development of simple, GSM-based remote irrigation solutions. GSM enables IoT by providing long-range, low-cost cellular connectivity for device-to-device communication, remote monitoring, and cloud integration. Advancements in GSM-based irrigation systems have introduced richer features such as DTMF-based multi-phase pump control and hybrid IoT cloud solutions for scheduling and analytics, enhancing scalability and resilience [2]. These implementations often depend on SMS, GPRS, or internet-based data exchange, which, despite providing flexibility, also incur recurring costs and increase system complexity.

To address cost concerns, several studies propose missed-call based pump control systems using microcontrollers, valued for their simplicity and low operating expenses. Recent studies published between 2020 and 2025 [3]–[7] emphasize GSM-, SMS-, and IoT-based irrigation systems; however, several limitations persist. Many systems rely heavily on continuous SMS or GPRS communication, resulting in recurring operational costs and higher power consumption. Others lack secure multi-user authentication, local number management, or autonomous fault notification mechanisms.

In contrast, AgriCall minimizes communication cost through missed-call control, restricts access using EEPROM-based caller ID authentication, and sends SMS only during critical events such as power failure. This combination directly addresses affordability, reliability, and usability challenges identified in existing systems. Collectively, the literature highlights GSM-enabled microcontrollers as a robust foundation for rural irrigation automation. Compact microcontrollers (ATMEGA328P), notably Arduino Uno, Mega, and UNO, paired with GSM modules, have been applied in irrigation applications involving soil-moisture sensing, pump actuation, and SMS reporting. The literature further identifies the Arduino UNO as a practical option owing to its small form factor, cost efficiency, and GSM compatibility. In this context, the proposed device, *AgriCall (Agriculture + Calling system control)*, is specifically designed to align with the socio-economic and technological needs of Bangladeshi farmers. By incorporating GSM-based

communication along with IoT-driven data exchange, *AgriCall* facilitates remote control of irrigation pumps, enhances water-use efficiency, and minimizes manual intervention. Existing projects report status via periodic SMS; *AgriCall* design only sends SMS when a critical event (power loss after pump start) occurs, balancing information needs against messaging costs. The system not only eliminating the need for physical presence in the field but also contributes to sustainable agricultural practices by ensuring optimized use of resources. A distinct subset of the literature targets *missed-call* control as a cost-minimizing technique, the device detects incoming call events (caller ID) without requiring the call to be answered, thereby avoiding airtime consumption. *AgriCall* addresses this challenge by combining secure password-protected access, missed-call command handling, and an integrated keypad for local registration and management of authorized numbers, thereby enhancing field usability and security. This study demonstrates how such low-cost, accessible technologies can bridge the gap between traditional irrigation practices and modern precision agriculture.

II. System Architecture

The *AgriCall* system is developed as a compact IoT-enabled irrigation controller that integrates a GSM communication module with a pump actuation unit. Its architecture is organized into three primary components, ensuring seamless communication, control, and management for field-level irrigation automation:

1. Control Unit: This is the central microcontroller-based module that manages the operation of the irrigation pump. It interprets commands received from registered mobile numbers and triggers the pump accordingly.
2. GSM Communication Module: The GSM module serves as the communication interface between the farmer's mobile phone and the control unit. It receives SMS or missed-call signals, which are processed by the control unit to execute pump operation.
3. User Interface & Security Layer: The system includes a simple keypad interface for registering authorized caller Id and filling password. Only these numbers are allowed to control the pump and ensuring individual operational security.

The device is designed for easy installation, minimal maintenance, and reliable operation in rural agricultural settings. Its modular architecture allows integration with different pump types and varying power supply conditions. A schematic block diagram of the system is represented in Fig. 1.

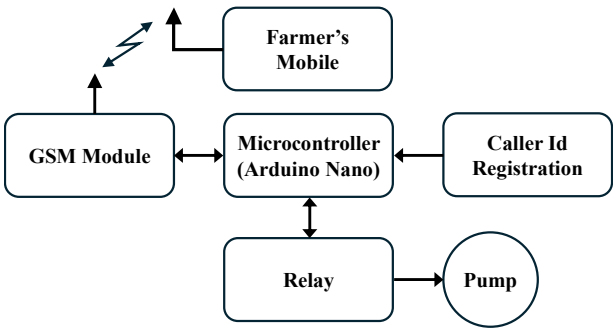


Fig.1. Block diagram of the proposed *AgriCall* model.

III. Accessories of The Proposed System

A. Microcontroller (Arduino UNO)

The Arduino Uno, based on the ATmega328P 8-bit microcontroller unit (MCU) clock frequency of 16 MHz, serves as the system's core. It provides one SPI, one I2C, and multiple GPIOs, with 32 KB flash, 2 KB SRAM, and 1 KB EEPROM. SPI enables master-slave communication, and the Arduino IDE supports C/C++ programming for embedded control tasks [8].

B. GSM Module

The GSM SIM900A Mini is a compact cellular module operates at 3.2–4.8 V and supporting dual-band GSM/GPRS (900, 1800 MHz) for voice, SMS, and data transmission. It interfaces with microcontrollers via UART, supports AT command control, and enables remote communication and suitable for embedded IoT applications [9].

C. Liquid Crystal Display (LCD)

The 16×2 LCD is a display module capable of showing 16 characters per line on 2 lines. It typically uses the HD44780 controller, supports 4-bit or 8-bit parallel interfacing with microcontrollers, and can be operated in character or command mode. Backlight and contrast are adjustable for embedded applications [10].

D. Keypad

The 4×4 matrix keypad has 16 keys arranged in 4 rows and 4 columns. It interfaces with microcontrollers via row-column scanning, providing digital input for password entry, control commands, or menu navigation in embedded systems.

E. Relay

A relay module is an electromechanical switch controlled by a microcontroller to operate high-voltage AC/DC loads. It provides isolation via an optocoupler, supports NO/NC contacts, and can include electricity sensing for monitoring load status in embedded systems.

The Arduino UNO paired with the SIM900A GSM module was selected over a GPRS-based solution due to its lower power consumption, reduced system complexity, and elimination of recurring data costs. While GPRS enables continuous data transmission, it depends on reliable internet availability and incurs higher energy usage and subscription expenses. In contrast, the SIM900A supports call- and SMS-based communication, which is more dependable in rural regions with limited data connectivity. The Arduino UNO is utilized primarily for prototyping due to its ease of programming, extensive community support, and rapid development capability. For large-scale deployment, the system can be implemented on a custom PCB using the standalone ATmega328P microcontroller, thereby reducing cost, size, and power consumption while maintaining identical functionality.

IV. Project Design and Working Principle

A schematic of the proposed *AgriCall* system is shown in Fig. 2, illustrating a GSM-based irrigation solution for remote operation of AC pumps. The system employs the ATMEGA328P microcontroller (Arduino UNO) as the central controller for command execution, relay actuation, and peripheral coordination. The GSM SIM900A module, interfaced via UART (TX–D7, RX–D8), detects missed calls

and signals the controller without establishing a voice connection, eliminating airtime or data usage. User authentication is managed through a 4×4 matrix keypad (rows A0–A3, columns 4, 5, 6, 10), enabling secure registration of up to three caller IDs with password protection. Functional keys provide local control: '#' stores a number, 'D' deletes, 'C' retrieves, 'B' manually activates the pump, and 'A' resets the system. A 16×2 LCD, interfaced via I2C (SDA–A4, SCL–A5), displays real-time alerts and operational status. The relay, driven from pin D9 through a transistor, switches the AC pump using a low-voltage signal while isolating the controller from surges. Mains availability is monitored via pin D3; if absent, the Arduino commands the GSM module to issue an SMS alert, ensuring reliable pump monitoring and timely farmer notifications.

During operation, a registered user initiates a missed call to the GSM module, which subsequently signals the Arduino UNO. The caller's id, pre-stored in EEPROM during the registration phase, is authenticated via caller ID verification. Upon successful validation, the Arduino actuates the relay to control the pump (start/stop). Concurrently, the relay sensing circuit continuously monitors the availability of electrical power. On the other side pump deactivation is achieved through multiple controlled methods. A registered user can turn the pump OFF by issuing another authenticated missed call, which toggles the relay state. Manual pump control is also available through the local keypad interface by pressing '*' for emergency situations. Furthermore in the event of a power outage following pump activation, the Arduino immediately instructs the GSM module to transmit an SMS alert to the farmer, thereby ensuring real-time awareness of pump operation and power supply conditions without the need for manual supervision. This multi-layered shutdown mechanism ensures operational safety, flexibility, and protection of irrigation equipment.

The proposed system integrates microcontroller-based control, GSM communication, keypad input, I2C-based LCD monitoring, relay actuation, and power sensing to establish a cost-effective, reliable, and energy-efficient irrigation automation solution. By providing secure access, automated pump control, real-time alerts, and power monitoring, the *AgriCall* system minimizes human intervention, improves operational efficiency, optimizes water utilization, and contributes to sustainable agricultural practices in rural settings. The complete implementation of the proposed system is presented in Fig. 3, demonstrating real-time operation. Fig. 4 depicts the interfacing of (a) registered numbers, (b) incoming calls, and (c) pump activation, whereas Fig. 5 illustrates (a) mobile call snapshots and (b) mobile SMS snapshots. The control logic is programmed on the Arduino platform using the C language, where the code defines a sequence of instructions to be executed by the microcontroller. During supply interruptions, the Arduino communicates with the GSM module to issue an SMS alert. Additionally, farmers can remotely query pump status using concise text commands (e.g., "status?"), to which the system responds with the current load condition, thereby eliminating the necessity of physical inspection. To enhance reliability, a DC backup battery ensures uninterrupted operation during mains failure, remaining in charging mode under normal conditions. The complete functional sequence, including decision-making and signal processing stages, is summarized in the flowchart of Fig. 6, demonstrating the effectiveness of integrating embedded control with low-cost communication technologies for smart agricultural applications.

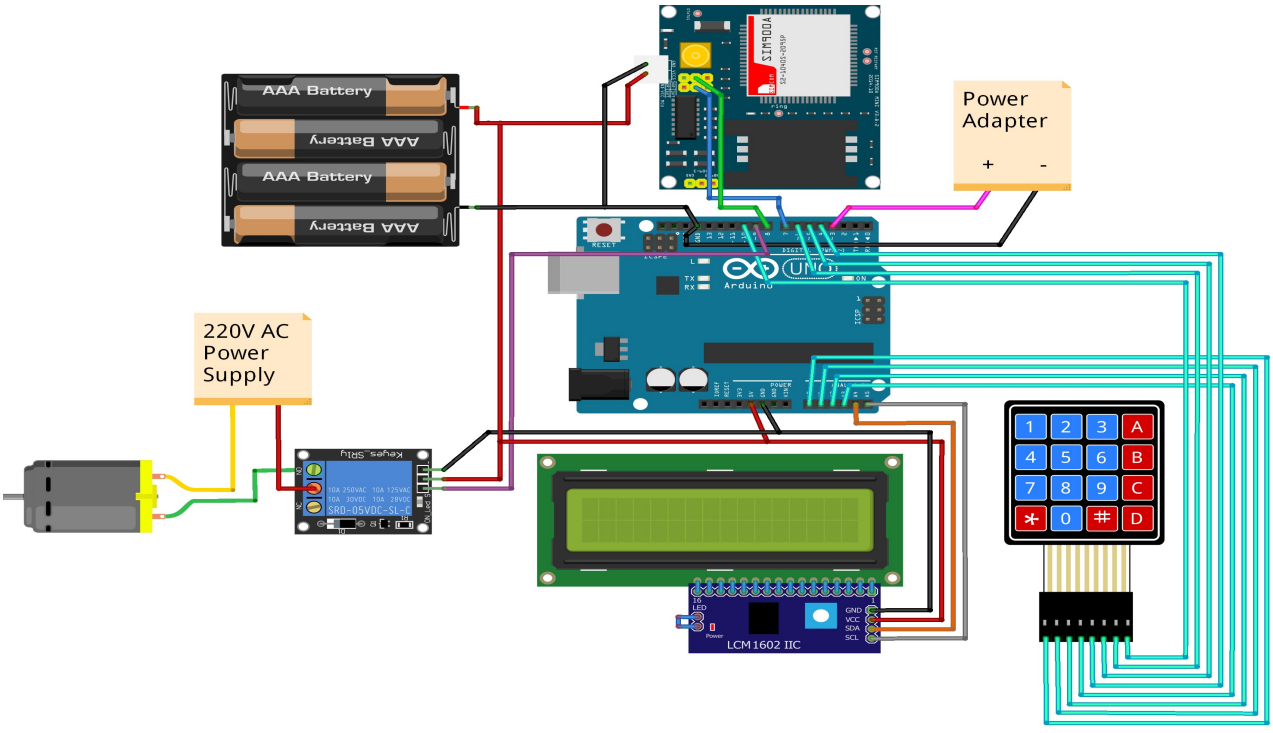


Fig. 2: Circuit configuration of the proposed *AgriCall* system.

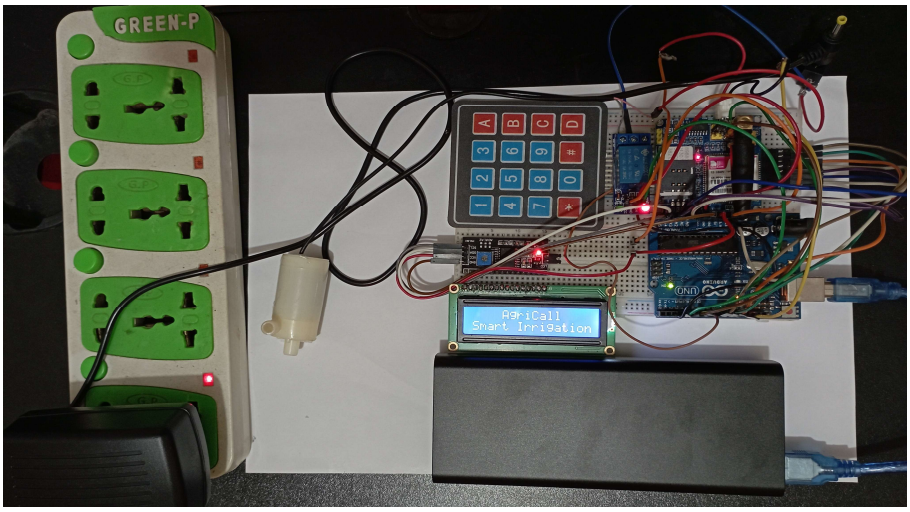
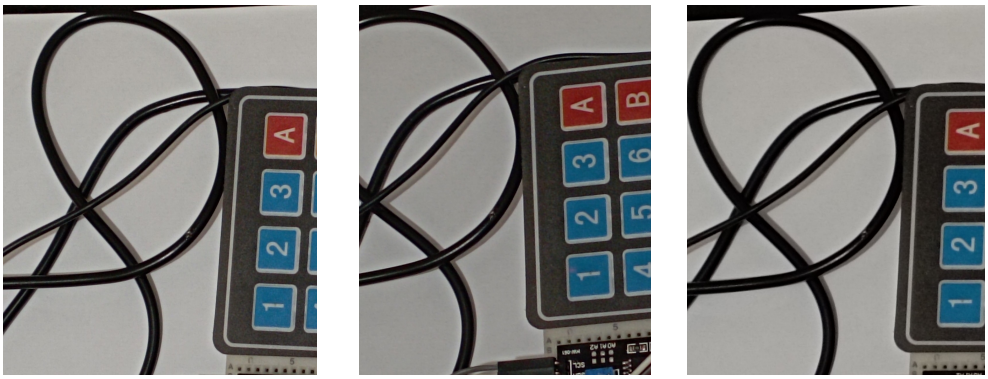
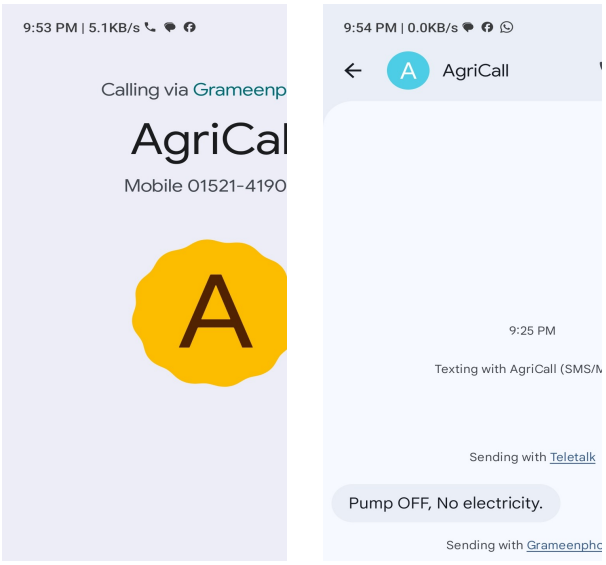


Fig. 3: Real-time implementation of the proposed *AgriCall* system.



(a) (b) (c)

Fig. 4: System interfacing feature with (a) registered numbers, (b) incoming calls, and (c) load shedding.



(a) (b)

Fig. 5: Mobile interfacing (a) incoming calls and (b) Responding message from system.

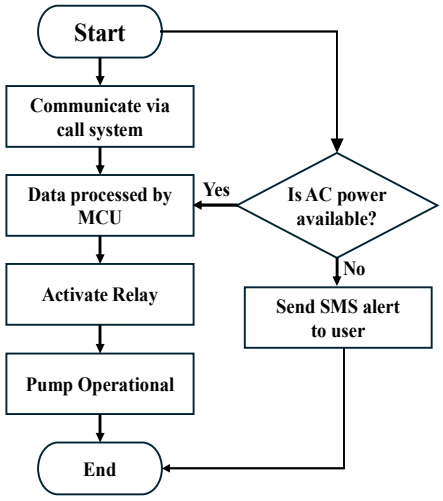


Fig. 6. Operational flow diagram of the proposed *AgriCall* system.

V. CONCLUSION

Experimental evaluation demonstrates that *AgriCall* offers high operational reliability, cost-effectiveness, and consistent performance under rural power and network conditions for smallholder farmers. Unlike conventional systems that rely heavily on frequent SMS communication, the missed-call control strategy eliminates recurring

communication costs, while EEPROM-based authentication ensures secure access. Field evaluation confirmed the system's ability to deliver prompt fault notifications and maintain essential functionality during power outages, highlighting its suitability for rural deployment. Beyond cost savings, the approach enhances farmer convenience and lays the groundwork for more resilient agricultural practices. Future work will focus on enhancing the AgriCall system by integrating soil-moisture, adoption of renewable power options to improve system autonomy, and environmental sensors to enable demand-driven, automated irrigation decisions. For large-scale deployment, the current prototype implemented using an Arduino UNO will be transitioned to a custom PCB design based on the standalone ATmega328P microcontroller. This approach will significantly reduce system size, power consumption, and production cost, while improving reliability and scalability for mass manufacturing. These findings position *AgriCall* as a viable, affordable, and scalable irrigation automation solution for smart and sustainable agriculture, particularly for smallholder farming communities.

REFERENCES

[1] B. R. Prakash and S. S. Kulkarni, "Super Smart Irrigation System using Internet of Things," *2020 7th International Conference on Smart Structures and Systems (ICSSS)*, Chennai, India, 2020, pp. 1-5, doi: 10.1109/ICSSS49621.2020.9202361.

[2] A. B. Shaik, P. Jonnadula, K. B. Bhukya, B. Ratlavat and R. K. Surapaneni, "Internet of Things based Smart Irrigation Motor System," *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2021, pp. 102-107, doi: 10.1109/ICOSEC51865.2021.9591925.

[3] F. Lizana, R. Tello, M. G. Gaitán, D. Ruete and C. Gómez-Pantoja, "Building a Text Messaging-Based System to Support Low-Cost Automation in Household Agriculture," *2020 Congreso Estudiantil de Electrónica y Electricidad (INGELECTRA)*, Santiago, Chile, 2020, pp. 1-5, doi: 10.1109/INGELECTRA50225.2020.246967.

[4] H. Meena, H. Nandanwar, D. Pahl and A. Chauhan, "IoT based perceptive monitoring and controlling an automated irrigation system," *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225455.

[5] H. M. Yasin, S. R. M. Zeebaree and I. M. I. Zebari, "Arduino Based Automatic Irrigation System: Monitoring and SMS Controlling," *2019 4th Scientific International Conference Najaf (SICN)*, Al-Najef, Iraq, 2019, pp. 109-114, doi: 10.1109/SICN47020.2019.9019370.

[6] O. V. Matukhina, L. R. Votyakova, and I. N. Zakharova, "SMART-technologies in irrigation management of a remote land plot," *IOP Conference Series Materials Science and Engineering*, vol. 1079, no. 7, p. 072027, Mar. 2021, doi: 10.1088/1757-899x/1079/7/072027.

[7] D. R, P. Choudhary, V. S. Yalagi, B. D and Y. S B, "Automatic Irrigation System Using GSM Module," *2021 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, Bangalore, India, 2021, pp. 93-97, doi: 10.1109/ICDI3C53598.2021.00027.

[8] L. C. Paul, S. S. Ahmed, and K. K. Karmakar, "A smart multi-user wireless nurse calling system and E-Notice board for health care management," in *Lecture notes in networks and systems*, 2022, pp. 421–431. doi: 10.1007/978-981-16-7597-3_35.

[9] L. C. Paul *et al.*, "A smart medicine reminder kit with mobile phone calls and some health monitoring features for senior citizens," *Heliyon*, vol. 10, no. 4, p. e26308, Feb. 2024, doi: 10.1016/j.heliyon.2024.e26308.

[10] M. Rifat Kabir, S. Shifat Ahmed, L. Chandra Paul, T. Rani and M. Karaaslan, "Development of an Intravenous Fluid Monitoring, Warning, and Reverse Flow Blocking System," *2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, Naya Raipur, India, 2022, pp. 89-94, doi: 10.1109/WIECON-ECE57977.2022.10150612.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Beyond Text: A Review of Emoji-Enhanced Notifications and Their Impact on User Attention in HCI

Md. Rayhanul Hasan, Syeda Shahida Alam, Tasnia Zarin, Prithwiraj Bhattacharjee
Department of CSE Leading University, Sylhet, Bangladesh
Email : cse_182210012101162@lus.ac.bd, cse_182210012101159@lus.ac.bd
cse_182210012101180@lus.ac.bd, prithwiraj_cse@lus.ac.bd

***Abstract*—In recent times, the use of emojis in virtual communication has increased manifold, transforming the manner in which users perceive, interpret, and emotionally react to text-based notifications. In the field of Human Computer Interaction (HCI), it has become important to understand the influence that visual icons such as emojis have on user attention and interaction. This study reviews twenty recent research articles examining the use of emojis in notification systems and other related digital interfaces. Each study is reviewed to determine its methodology, findings, and limitations. The review identifies that approximately 60% of the studies rely mainly on traditional statistical or machine learning techniques, while only about 40% adopt advanced deep learning or multimodal models. Reported performance across quantitative studies shows accuracy values ranging from 56.8% to 95%, depending on the task and domain. In addition, over 60% of the reviewed works use small-scale, domain-specific, or culturally constrained datasets, resulting in limited generalizability. Future research should analyze text sentiment in different domains using the same emojis, analyze age variation in emoji selection, compare existing models with new high-accuracy approaches, and test with a large dataset, particularly in learning contexts, to develop a user-centered and large-scale system that improves emoji-based attention and engagement in notifications.**

Keywords: ***Emojis, User Engagement, Text-Based Notification, Human-Computer Interaction (HCI), User-Centered Design.***

I. Introduction

Human-Computer Interaction (HCI) is a multi-area field that studies how people interact with digital systems. It aims to design interfaces that are efficient and emotionally engaging. People can get guidance on user needs, habits, and thinking patterns from HCI models like User-Centered Design (UCD), Norman's interaction framework, and GOMS (Goals, Operators, Methods, Selection rules). In order to evaluate user behavior, engagement, emotional response, and task performance, computational models and statistical analysis are used widely. Notifications have become an important part of HCI systems by helping with timely communication between users and applications. On the other hand, badly timed notifications or too many notifications can lead to user disengagement and information overload. To remove these issues, researchers explored adding emojis to notifications that can create more user engagement. This growing field studies how visual emotion affects user attention, perception, and engagement. Existing studies work with a group of methods, including statistical analyses (e.g.chi-squared tests [1]), traditional machine learning models (e.g. Logistic Regression [1], [2], SVM [3], [2], Random Forests [4]), and deep learning techniques (e.g., LSTM [5], [6], BERT-based sentiment classifiers [7]). At the same time, supervised learning models are widely used due to labeled datasets, and unsupervised approaches are still underexplored. In addition, many studies are limited to specific domains, small datasets [8], or simple analytical techniques [9], which makes it difficult to apply different contexts. To handle these limitations, this review explores twenty recent studies on emoji-enhanced notifications using statistical, machine learning, and deep learning approaches. The study aims to:

- Identify major challenges and research gaps, including dataset limitations, domain boundaries, and limited generalization.
- Highlighting the ability of multi-modal and cross-domain approaches to improve engagement modeling.
- Provide a future direction for designing emoji-based notifications that integrate emotion, engagement metrics, and data-driven HCI principles.

This work contributes to affective and intelligent interface design by bridging emotional communication and computational modeling.

II. Significant Issues and Challenges

The study of emoji-based notification in Human-Computer Interaction (HCI) is still a highly complex and challenging problem because of different types of difficulties that significantly impact the design and effectiveness of notification systems focused on increasing user engagement. Figure 1 demonstrates the primary issues and challenges currently faced by researchers in this field. One of the most important challenges is the subjectivity of emoji interpretation and emotional complexity. Emojis do not have a universally fixed meaning; the same emoji can be seen differently depending on the user, cultural background, age, or context. For example, one emoji can be seen as humorous by one user but sarcastic by another, while another emoji could convey flirty, playful, or sarcastic feelings depending on the surrounding text. These

differences make it difficult to accurately identify and measure user engagement based on emoji use.

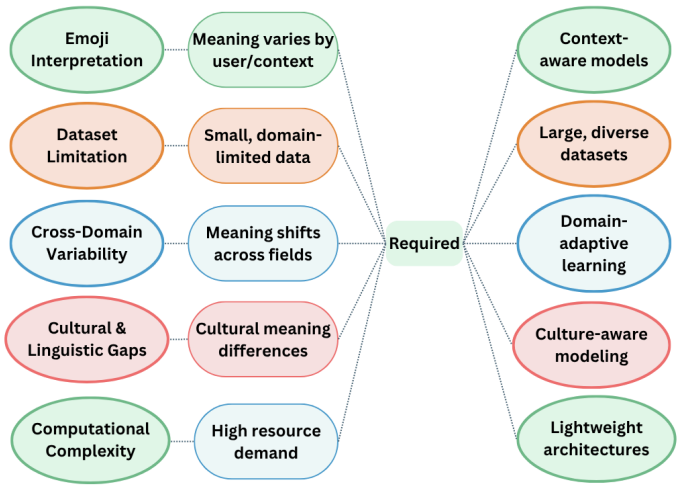


Figure 1: Significant Issues & Challenges

Another major problem is the small and limited availability of the dataset [9], [8]. Most existing datasets are restricted to specific platforms such as Twitter [9], WhatsApp, or online learning systems, and sometimes contain a limited number of samples. This limitation decreases model reliability. Cross-domain makes the problem harder because the meanings of emojis can vary across different applications. Emojis used in educational notifications [5], health-related text [10], social media interactions [11] , [2] or marketing campaigns may carry different meanings. Working on a single domain makes the model less effective when it is used in another domain. On the other hand, cultural and linguistic differences also create a major challenge. Users from different regions or language backgrounds may use the same emoji in different ways. This cultural variation adds another layer of complexity to designing models. Lastly, the high computational cost of advanced models like LSTM, Transformer, and multimodal neural networks limits real-time deployment. High computational requirements make it difficult to implement these systems efficiently, whereas a lightweight model is easy to implement.

III. Literature Review

A total of twenty papers were shortlisted for analysis. Table I presents a comparative overview of these studies. Wicaksana et al. [1] analyzed push notifications on the REES46 e-commerce platform using logistic regression and chi-square tests, then found that emojis increased engagement from 20% to 91%, and with combined emojis and deadlines, engagement reached 100%, but it is limited to one platform. Elsalam et al. [2] improved sentiment analysis by combining text and emojis 10,000 tweets related to COVID-19 using multiple ML models and obtained highest accuracy highest accuracy (0.95) with Logistic Regression, but the study is limited to a single domain and does not explore DL models. Khan et al. [3] used emojis in airline review sentiment analysis using ML models (SVM, RF, MNB, LR) and BERT and obtained 92% accuracy. It highlights that emojis improve understanding of textual sentiment, but it is limited to one domain. Saha et al. [4] predicted emojis from 7,480 tweets using ML classifiers (Linear SVC, Decision Tree, Random Forest) with an average accuracy of 56.82%, highlighting limitations of a small dataset. Alshaya [5] presents a DECOC, LSTM, and ML-DNN to integrate emojis into computer science educational materials for better engagement and comprehension. Using EduEmoji-CS with 1000 annotated course sections paired with relevant emojis. The model achieves 92.4% accuracy and outperforms other baselines. But limited to one subject. Lv et al. [6] review and analyze 100 deep learning-based Human–Computer Interaction (HCI) studies from Google Scholar and Web of Science. This paper emphasizes multimodal and deep learning techniques such as CNN, LSTM, Bi-LSTM, GCN, reinforcement learning, and attention-based fusion models. Public datasets have been used. Difficulty combining multiple modalities (speech, gesture, emotion) in a single robust model. Qin et al. [7] proposed a multi-modal emotion recognition method to analyze online education data, especially students' comments, by combining text and emoji data for sentiment classification. The model uses text pre-training with a BERT-based approach for improving sentiment accuracy. The dataset has over 62,000 comments with positive and negative labels. Limited to poor emoji handling. Sun et al. [8] explore emoji used and perceived in peer feedback during online collaborative design activities using Emoviz, which lets students add text and emoji comments directly on design documents. 50 students' data came from an online design course. Selected emojis were used, limiting natural expression. Alsulaiman et al. [9] studied the factors influencing emoji use among Saudi female EFL students on Twitter. The research used a qualitative approach with 15 participants. It was limited by the small dataset and focused on a single platform (Twitter). Halverson et al. [10] examine how clinicians use emojis and emotions in clinical communication. The study analyzed 1% of clinical texts from hospitalists at one U.S. hospital. Limitations include focusing only on qualitative data, not quantitative data. Ali-Chand et al. [11] present how emojis affect paralanguage in social media communication among university students. Using a mixed-method approach and Socio-Semiotic Perspective theory, data were collected from 55 Bachelor of Education students. Limitations include a small sample size and low cultural diversity. Hafiz [12] explored the linguistic, cultural, and technological impact of emojis using Unicode Consortium (2015-2024) data. The Text-Technology Framework contains four parts: Intentionality, Materiality, Functionality, and Cultural Value. The focus was only on Unicode emojis, not custom ones like Bitmoji. Chiang et al. [13] conducted research on HCI. The study used a systematic literature review approach following the PRISMA guideline, and finally, 42 papers were selected. Data were collected from ACM DL, IEEE Xplore, and Web of Science. Subjective bias

Table I: Comparison of Current Detection Models along with Their Performance Drawbacks

Source (Domain)	**Problem Statement**	**Model / Approach**	**Dataset**	**Metrics**	**Contribution**	**Limitations**
Wicaksana et al. [1] (Business)	Which notification features drive engagement	Logistic Regression, Chi-square	REES46 (700M notifications)	Engagement: 91% (emoji) vs 20% (none)	Emoji+deadline yields 100% engagement	Single-platform dataset.
Elsalam et al. [2] (Social Media)	Combine text + emoji for sentiment analysis	SVM, RF, LR, KNN, GNB, GB	10K COVID-19 tweets	LR: Acc 0.95, F1 0.95	Combined text-emoji sentiment model	ML only (no DL); COVID-only tweets
Khan et al. [3] (Business)	Emojis improve sentiment analysis in airline reviews	ML: SVM, RF, MNB, LR + BERT	14,640 airline tweets	BERT: Acc 92%, F1 0.91	Integrated EMOJIXT for emoji-based sentiment analysis	Tested only on airline reviews
Saha et al. [4] (Social Media)	Predict emojis from emotional text	SVC, DT, RF classifiers	7,480 tweets (7 emotion labels)	Acc: 56.8% (SVC)	Mapped emotions → emojis using ML	Small dataset; low accuracy.
Alshaya [5] (Education)	AI-based emoji use can boost engagement in text-heavy CS learning	DECOC, LSTM, ML-DNN	EduEmoji-CS	Acc: 92.4%, Prec: 90.7%, Rec: 89.3%, F1: 90.0%	Hybrid AI model integrating DECOC + LSTM + ML-DNN	Limited to CS materials; less generalizable to other domains.
Lv et al. [6] (HCI)	How DL improves gesture, speech, emotion recognition	CNN, RNN, LSTM, GCN, Attention	Public multimodal sets	Acc: ≈95%	Explains multimodal fusion (vision, sound, text)	Requires large datasets and high computing.
Qin et al. [7] (Online Education)	Difficult to detect emotions using only text	BERT-based multimodal	Chinese MOOC, Nat. Intelligent Ed. datasets	Acc: 85.6–91.2%, F1: 68.2–78.5%	Introduced emoji-prompt multimodal model for emotion recognition	Emoji fusion simplistic; weak emotional learning.
Sun et al. [8] (HCI / EdTech)	Effect of emojis in peer feedback	Thematic content analysis	51 comments, 6 interviews	Qualitative themes	First to test emojis in peer design feedback	Small sample.
Alsulaiman et al. [9] (HCI / Online Comm.)	Emoji use among Saudi female EFL students on Twitter	Qualitative interviews	15 participants	Thematic analysis	Cultural/gender-based patterns found	Small, homogeneous sample.
Halverson et al. [10] (Medical)	Study emoji use by clinicians	Qualitative linguistic analysis	Clinical texting corpus	Qualitative (7% emoji use: 61% emotion, 32% communication)	First clinical text emoji analysis	Qualitative only; single hospital dataset.
Ali-Chand et al. [11] (Social Media)	Examine emojis as paralanguage for emotion	Mixed method (survey + thematic)	55 student responses	47% always use emojis, 67% improved comms	Showed emojis as digital paralanguage	Small sample, limited diversity.
Hafiz [12] (Digital Communication)	Role of emojis in digital text	Mixed (Taxonomy of Text Tech)	Unicode Consortium records	Conceptual evaluation	Links emoji, language, and culture	Only Unicode emojis; no experiments.
Chiang et al. [13] (HCI)	Evolving emoji roles in communication	Systematic Review (PRISMA)	42 HCI studies	Thematic synthesis	Identified usability and emotion features	Possible coder bias.
Lin et al. [14] (Health Info)	Emojis supplement health text	Experimental + qualitative coding	69 emojis, 120 tweets, 60 participants	LATCH framework	Built health emoji micro-database	Limited scope; small group.
Xu et al. [15] (Social Media / Business)	Sentiment analysis ignoring emojis	E-BiLSTM-CNN	80k tweets (20% with emojis)	Acc: 81.44%, F1: 81.43%	Word_Emoji text+emoji encoding	Complex; English-only data.

Source (Domain)	Problem Statement	Model / Approach	Dataset	Metrics	Contribution	Limitations
Tauch et al. [16] (Social Media)	Emoji use in app notifications	Statistical + sentiment (AFINN)	446,380 notifications (8 users)	Frequency, sentiment correlation	Platform-based emoji sentiment study	Few users; limited apps.
Mutawa et al. [17] (Education)	Predict student satisfaction using ML	ANN, RF, DT, SVM, MLKNN	9,128 GCC student surveys	ANN acc:81%, RF:65.3%	Effective emotion satisfaction prediction	GCC-only; self-reported data.
Kusal et al. [18] (Social Media)	Combine text and emoji features for emotion detection	CNN + BiLSTM (early fusion)	GoEmotions (58K Reddit comments)	Acc: 77.8%, F1: 71.68%	Proposed multimodal fusion model	Only tested on GoEmotions dataset
Kim et al. [19] (Education)	Emojis in emails improve student connection	Structural Equation Modeling (SEM)	297 U.S. undergrads	RMSEA = 0.078, CFI = 0.876, GFI = 0.882	Emojis enhance student engagement	Limited to U.S. undergrads
Kone et al. [20] (Social Media)	Predict emoji from text	BiLSTM + Softmax	CodaLab Twitter (60K tweets, 40 classes)	Accuracy: 94%	BiLSTM model for emoji prediction	Twitter-only data

from single-research screening. Lin et al. [14] develop design guidelines for integrating emojis in 69 health emojis using 120 Twitter samples from 40 verified Twitter accounts. The study focused only on static text–emoji integration; dynamic media and cultural variation were not deeply examined. Xu et al. [15] aim to improve sentiment analysis by using emoji features often ignored in traditional models. The study proposed the E-BiLSTM–CNN (Emoji-incorporated Bidirectional LSTM–CNN) model that combined text and emoji and achieved 81.4% accuracy using a modern tweet dataset (80,000 tweets, 20% with emoji). This study faced higher execution time and only used English tweets. Tauch et al. [16] studied 446,380 notifications from WhatsApp, Facebook, and Twitter using statistical analysis, finding that emoji frequency, type, and position influence sentiment. In this study, only 8 users were observed. Mutawa et al. [17] applied ML models (ANN, RF, SVM, MLkNN) on 9,128 student responses to predict satisfaction (ANN: 81% accuracy) and emotion (RF: 65.3% accuracy) in adaptive e-learning, but the study is limited to the GCC region. Kusal et al. [18] proposed a hybrid CNN-BiLSTM model with early fusion of text and emoji features on the GoEmotions dataset (58K records, 28 emotion labels) and achieved 77.80% accuracy and a 71.68% F1 score, yet it is limited to a single dataset. Kim et al. [19] analyzed emojis in professors' welcome emails with 297 undergraduate students using Structural Equation Modeling. It shows student engagement has increased, but the results were limited to U.S. undergraduates. Kone et al. [20] developed a Bi-LSTM model for automatic emoji prediction, used 60K Twitter posts, and achieved 94% accuracy, but were limited to one platform's data.

IV. Existing State-of-the-art Models

From the selected articles, Figure 2 shows five mostly adapted models along with their accuracy score, such as the

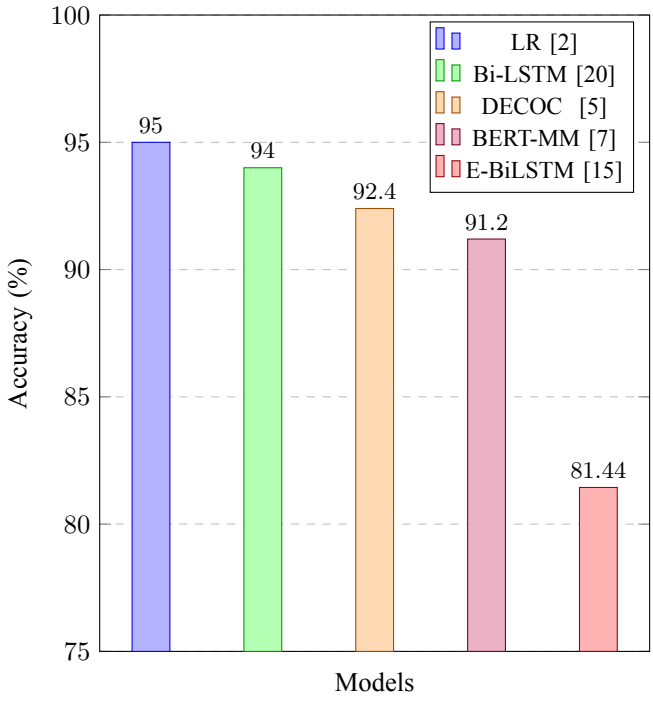


Figure 2: Accuracy Comparison of Five Cutting-Edge Models.

Logistic Regression [2] model achieved the highest accuracy of 95% because of its simplicity and efficiency with TD-IDF and Bag-of-Words features. The Bi-Directional LSTM [20] also achieved a higher accuracy of 94% by processing both text forward and backward, capturing the full context and emotion. The Hybrid DECOC-LSTM-ML-DL [5] and the BERT-based Multi-Modal [7] model achieved accuracies of 92.4% and 91.2%, respectively, while E-BiLSTM-CNN [15] became the least accurate model among these five with 81.44% accuracy. Although Logistic Regression and Bi-Directional

LSTM produced high accuracies, their weakness lies in only one selected field, such as COVID-19 tweets and Twitter data. The DECOC-LSTM-ML-DL and BERT-base Multi-Modal models' strengths are based on their hybrid processes; however, they also have weaknesses. For instance, DECOC-LSTM-ML-DL is limited in its generalizability, while BERT-based models are limited to text features, resulting in weak and less effective emotional features for emojis, which limits the depth of emotional understanding. Finally, E-BiLSTM-CNN is strong enough, but its primary weakness is that its complexity increases the execution time. The weakness comes along with the strength, suggesting that improvement can be made by using more diverse data and multiple datasets.

V. Limitations & Findings

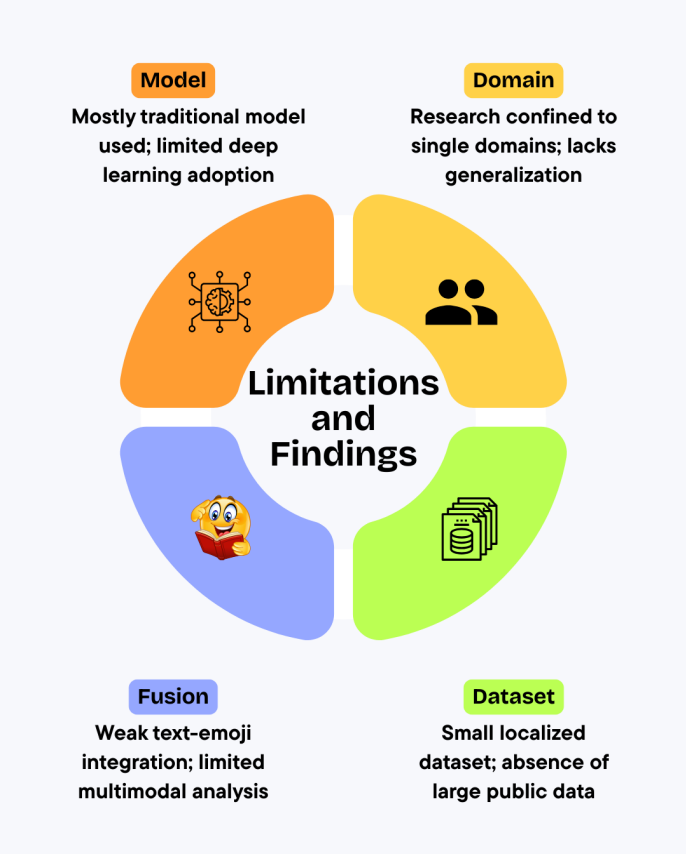


Figure 3: Limitations & Findings

Emoji-based user attention and engagement study still has several open challenges. Figure 3 illustrates several of these challenges, capturing both the practical difficulties encountered by current researchers and the inherent technical limitations present within the existing model. Most studies have used a limited range of models and depended mainly on statistical or qualitative analysis, rather than leveraging advanced deep learning or multi-modal computational approaches. This narrow methodological focus has limited the creation of more intelligent and context-sensitive systems.Another critical limitation is the domain specificity of existing studies. Most research is focused on single domains, i.e., education, marketing, or business communication, but not on the extent to which emojis influence engagement across diverse cultural, linguistic, and social settings. This creates a lack of generalizability and weak cross-domain applicability. The datasets for such studies are also small and localized, usually scraped from a single platform or region, and it is difficult to validate results with other groups of users. A large-scale,standardized, and publicly available benchmark dataset of multilingual and cross-cultural emoji use patterns is missing, which makes reproducibility and fair model comparison difficult. The majority of approaches concentrate predominantly on text-based features from a modeling perspective, with shallow knowledge regarding emoji emotion expression. Emotional significance in emoji is typically oversimplified, lowering the precision and expressiveness of engagement prediction models. More profound multi-modal fusion synthesizing text, emoji, sentiments, and contextual information is rarely embraced. Also lack of experimental validation and limited participant populations are barriers to real-time usage and scalability. Few human-centered evaluations of models in interactive environments exist, so it is not clear how these systems operate in real-world communication contexts.

VI. Future Research Direction

Based on the literature review of twenty papers, multiple research gaps can be observed in emoji-based notification research. Most existing studies are limited to single application domains, making it unclear how text sentiment interacts with emoji interpretation across different contexts. Age-related differences in emoji usage are also underexplored, although users of different age groups differ in their familiarity with emojis, usage frequency, and emotional understanding. Furthermore, the majority of studies rely mainly on statistical analysis and traditional or older machine learning models. These studies give higher priority to statistical decision-making than predictive intelligence. Advanced machine learning and deep learning approaches are still limited in this area. In addition, many studies use small, domain-specific, or private datasets, which restrict generalizability and cross-cultural applicability.

To overcome these gaps, future research should study cross-domain scenarios to examine whether similar notifications with the same emojis generate consistent reactions from users. Age-based analysis of emoji analysis can help design more universal notification systems. Higher priority should be given to advanced models such as LSTM, Transformer-based, and NLP-driven approaches to achieve more accurate and personalized results, with carefully considering model complexity and computational efficiency. Future studies should also apply larger, multi-domain datasets by combining public and private sources to support cross-cultural analysis. Finally, studying multiple application areas within a single framework can provide a better understanding of how emojis influence user engagement across different contexts. Figure 4 summarizes these future research directions.

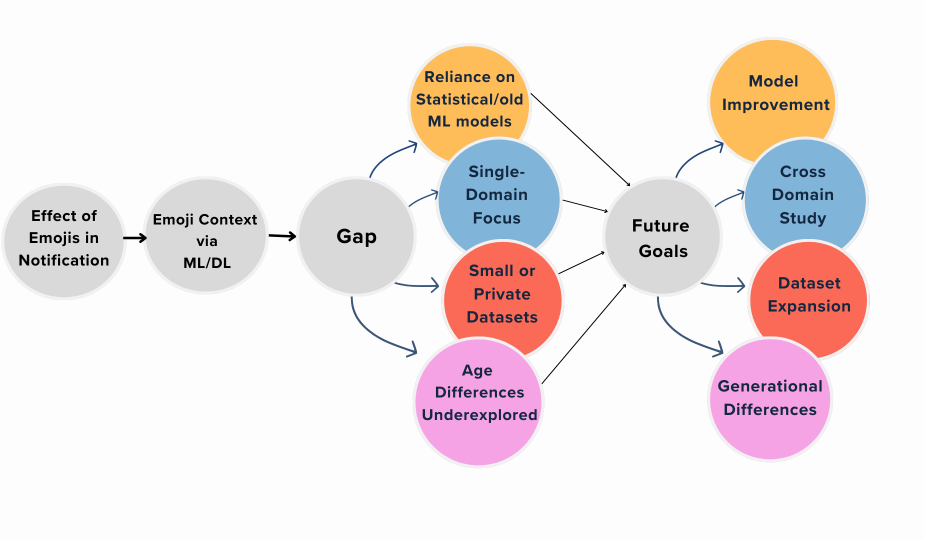


Figure 4: Future Research Direction

VII. Conclusion

This paper reviewed twenty recent emoji-based user attention and engagement modeling, and identified key challenges including limited model diversity, domain-specific focus, limited dataset size, and inadequate emoji emotion representation. Most existing methods rely on statistical methods, while multimodal, data-driven approaches show strong potential. A clear research directions is developing comprehensive, multimodal models that handle diverse datasets, capture more detailed emotional context, and support real-time deployment.

References

[1] D. Y. Wicaksana, R. J. S. Z. Aisyah, and D. U. Q. A'yun, "Designing effective mobile push notifications: Machine learning insights into user engagement," *Journal of Digital Business and Innovation Management*, vol. 3, no. 2, 2024.

[2] H. A. H. Mona M. Abd Elsalam, Ahmed M. Gadallah, "Sentiment analysis of text incorporating emojis: A machine learning approach," *International Journal of Computer Applications*, vol. 184, no. 20, pp. 10–18, Jul 2022. [Online]. Available: https://ijcaonline.org/archives/volume184/number20/32432-2022922216/

[3] A. Khan, D. Majumdar, and B. Mondal, "Sentiment analysis of emoji fused reviews using machine learning and bert," *Scientific Reports*, vol. 15, no. 1, p. 7538, 2025.

[4] S. Saha and K. A. Kumar, "Emoji prediction using emerging machine learning classifiers for text-based communication," *J. Math. Sci. Comput*, vol. 1, pp. 37–43, 2022.

[5] S. A. Alshaya, "Enhancing educational materials: Integrating emojis and ai models into learning management systems." *Computers, Materials & Continua*, vol. 83, no. 2, 2025.

[6] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, "Deep learning for intelligent human–computer interaction," *Applied Sciences*, vol. 12, no. 22, p. 11457, 2022.

[7] X. Qin, Y. Zhou, and J. Li, "Multi-modal emotion recognition for online education using emoji prompts," *Applied Sciences*, vol. 14, no. 12, p. 5146, 2024.

[8] N. Sun, E. Lavoué, C. Aritajati, A. Tabard, and M. B. Rosson, "Using and perceiving emoji in design peer feedback," 2019.

[9] R. Alsulaiman and A. Alhojailan, "Exploring emoji use on twitter among students of english as a foreign language," *Online Journal of Communication and Media Technologies*, vol. 14, no. 3, p. e202441, 2024.

[10] C. Halverson, C. Donnelly, M. Weiner, and J. Lee, "Content analysis of emoji and emoticon use in clinical texting systems. jama network open, 6 (6), e2318140," 2023.

[11] Z. Ali-Chand and R. Naidu, "Exploring the impact of emojis on paralanguage in social media communication among university students," *English Language Teaching*, vol. 17, no. 9, p. 84, 2024.

[12] H. Hafiz, "Emojis as text technology: Between visual paradigm and cultural revolution," Ph.D. dissertation, University of Central Florida, 2025.

[13] C. Chiang and D. Gomez-Zara, "The evolution of emojis for sharing emotions: A systematic review of the hci literature. arxiv," 2024.

[14] T. S. Lin and Y. Luo, "Visual-textual integration: Emoji as a supplement in health information design," *International Journal of Design*, vol. 18, no. 2, pp. 37–58, 2024.

[15] Q. A. Xu, C. Jayne, and V. Chang, "An emoji feature-incorporated multi-view deep learning for explainable sentiment classification of social media reviews," *Technological Forecasting and Social Change*, vol. 202, p. 123326, 2024.

[16] C. Tauch and E. Kanjo, "The roles of emojis in mobile phone notifications," in *Proceedings of the 2016 acm international joint conference on pervasive and ubiquitous computing: Adjunct*, 2016, pp. 1560–1565.

[17] A. Mutawa and S. Sruthi, "Enhancing human–computer interaction in online education: A machine learning approach to predicting student emotion and satisfaction," *International Journal of Human–Computer Interaction*, vol. 40, no. 24, pp. 8827–8843, 2024.

[18] S. Kusal, S. Patil, and K. Kotecha, "Multimodal text-emoji fusion using deep neural networks for text-based emotion detection in online communication," *Journal of Big Data*, vol. 12, no. 1, p. 32, 2025.

[19] M. Kim, T. L. Knotts, N. D. Albers, and K. E. James, "Emoji use as a catalyst for relationship building and sustaining attention in online classes: An empirical study," *Education Sciences*, vol. 12, no. 12, p. 874, 2022.

[20] V. S. Kone, A. M. Anagal, S. Anegundi, P. Jadekar, and P. Patil, "Emoji prediction using bi-directional lstm," in *ITM Web of Conferences*, vol. 53. EDP Sciences, 2023, p. 02004.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Exploring Multi-Model Machine Learning Approaches for Pathogenicity Classification of Somatic Gene Mutations

Asif Ahamed[1], Md. Tanvir Hasan[1], Most. Alisa Tabassum[1], Ahammad Hossain[1],
Md. Mizanur Rahman[1], and A.H.M. Rahmatullah Imon[2]
[1]Department of Computer Science and Engineering, Varendra University, Rajshahi, Bangladesh
[2]Director, Center for Interdisciplinary Research (CIR), Varendra University, Bangladesh
Email: asifahamedstudent@gmail.com, info.tanvirhasan01@gmail.com, alisatabassum101@gmail.com,
ahammad@vu.edu.bd, mizanur@vu.edu.bd, director.cir@vu.edu.bd

***Abstract*—Next-generation sequencing has generated vast genomic variant data, making accurate somatic variant pathogenicity classification critical for precision oncology. This paper proposes an Improved Hybrid Ensemble Framework combining diversity-aware model selection (quantifying accuracy and pairwise disagreement), performance-weighted stacking with enriched meta-features (19 features capturing consensus and disagreement patterns), and adaptive multi-objective threshold optimization (jointly maximizing accuracy, F1, and MCC). Six heterogeneous base learners trained on stratified folds with Borderline-SMOTE address class imbalance. The framework generates meta-features including base predictions, statistical summaries, and pairwise disagreement scores for a regularized meta-learner. Conditional isotonic calibration is applied when validation performance improves. Experiments on ClinVar somatic variants show 94.78% accuracy, 96.98% F1-score, and 91.65% ROC-AUC, outperforming the best single classifier while reducing Brier score by over 50%, demonstrating superior calibration for clinical variant interpretation.**

***Index Terms*—somatic variants, pathogenicity classification, ensemble learning, machine learning, variant effect prediction, precision oncology, calibration, stacking**

I. Introduction

Next-generation sequencing generates vast genomic variant data in clinical settings [1], [2]. Determining somatic variant pathogenicity is critical for precision oncology, molecular diagnosis, and treatment selection. The datasets, however, are heterogeneous and suffer from a severe imbalance in the class. Traditional rule-based criteria and single machine learning models lack robustness across diverse sources, particularly with uneven pathogenic-benign representation, motivating reliable ensemble methods with calibrated probability estimates [3].

Machine learning techniques have been applied to variant pathogenicity prediction [4], [5], with deep learning models capturing complex genomic feature interactions [6]. However, existing approaches produce poorly calibrated probabilities requiring ad hoc threshold tuning. Standard stacking methods use fixed weighting schemes and default thresholds, limiting generalization [7]. Maximizing raw accuracy overlooks model diversity, calibration, and reliability critical for heterogeneous ClinVar/COSMIC datasets.

Few studies systematically combine model diversity, adaptive thresholding, and calibration within unified pipelines. Most designs emphasize single metrics (e.g., accuracy) over calibration or interpretability, and rarely leverage classifier disagreement. Class imbalance and allele-frequency variation in somatic data require multi-objective frameworks jointly optimizing model selection, stacking, and thresholds to balance sensitivity, specificity, and reliability.

This paper proposes an Improved Hybrid Ensemble Framework addressing these challenges through diversity-aware base learner selection evaluating validation accuracy and pairwise disagreements. Performance-weighted stacking employs adaptive multi-objective threshold optimization maximizing weighted accuracy, F1-score, and Matthews correlation. Optional isotonic calibration applies when improving validation performance, yielding well-calibrated outputs while avoiding meta-layer overfitting.

In experiments on a ClinVar-derived somatic variant dataset, the proposed hybrid ensemble significantly outperformed the best single classifier (Random Forest). The ensemble achieved an accuracy of 0.9478, F1-score of 0.9698, and ROC-AUC of 0.9165, compared to 0.9254, 0.9567, and 0.9004 for the Random Forest. The accuracy and AUC improved by +2.24% and +1.61% respectively, and the Brier score decreased by over 50% (0.0981 to 0.0486), indicating much better calibration.

II. Related Works

We classify somatic coding variants as pathogenic versus benign distinct from germline prediction. Tools trained on germline data or pan-cancer drivers do not directly apply [8]. Few annotated somatic variants are labeled pathogenic/benign (most are VUSs), creating imbalance and calibration challenges. Prior ensemble methods use fixed weights or default thresholds, missing opportunities to optimize balanced metrics on unbalanced somatic data.

A. Clinical Guidelines and Knowledge Bases

Somatic mutation reporting relies on curated evidence (AMP/ASCO/CAP, ACMG) and knowledgebases like OncoKB, CIViC, ClinGen's VICC aggregating actionable variant

literature [8]. These provide consensus annotations but depend on slow, incomplete expert curation rather than predictive modeling, lacking probability scores or calibration and limited generalization to novel variants.

B. Driver Versus Passenger Mutation Classification

Many studies classify somatic mutations as driver versus passenger related but distinct from clinical pathogenicity. Das et al. (2025) achieved 0.89 AUC-ROC distinguishing drivers from passengers in head-and-neck cancer using 41 predictors [9]. Mouw et al. found VEPs with functional genomics features outperformed conservation-only models [8], but these tools optimize for tumor biology rather than clinical actionability.

C. Gene-Specific and Tissue-Specific Models

Some work narrows the scope to a particular gene, disease cohort, or tissue type, yielding higher in-domain accuracy. Kang et al. developed BRCA1- and BRCA2-specific models for rare missense variants, showing that training on gene-specific variant sets suffices to obtain an optimal pathogenicity predictor for each gene [10]. Similarly, Ramaker et al. (2024) introduced CVD-PP, a cardiac-specific pathogenicity predictor achieving high discrimination of benign versus pathogenic variants in cardiovascular disease genes [11]. These specialized predictors benefit from tailored feature distributions but do not generalize beyond their domain and still often lack formal probability calibration. In our context, we aim for a more general somatic ensemble across many genes rather than narrowly trained models.

D. Machine Learning and Ensemble Methods

Deep learning (AlphaMissense, ESM, PrimateAI) advances the field but has limitations. Kong et al. (2025) found AlphaMissense achieved only 32.9% precision and 57.6% recall on expert-curated pathogenic variants [12]. Ensemble approaches show promise: Bi et al. (2025) introduced EPEL, a stacked ensemble for driver synonymous mutations outperforming existing tools [13].

Calibration remains under-explored. Flynn et al. (2024) showed genome-wide calibration masks gene-specific heterogeneity [14], suggesting naive global thresholds are suboptimal. We incorporate multi-objective threshold search and conditional isotonic calibration guided by validation metrics.

Prior work spans clinical guidelines, driver-finding, gene-specific models, and ML ensembles, but gaps remain in balancing accuracy with calibration. Our method addresses these by (1) selecting complementary learners via accuracy and disagreement, (2) stacking with enriched meta-features, (3) multi-metric threshold optimization, and (4) conditional calibration. We emphasize reproducibility and uncertainty quantification (bootstrap CIs).

III. Methodology

A. Methodological Overview

We propose a hybrid ensemble for somatic variant pathogenicity classification integrating tree-based, linear, and neural learners via stacking. The pipeline (Fig. 1) begins with ClinVar VCF ingestion and filtering, proceeds through feature extraction, and culminates in nested cross-validated model selection and a stacked ensemble meta-learner. In this Super Learner framework [16], base classifiers trained on stratified folds generate out-of-fold (OOF) predictions for a regularized second-level classifier. The ensemble's probabilistic output is calibrated and thresholded via multi-objective optimization.

B. Data Source and Inclusion Criteria

The ClinVar database [15] (VCF release of 19 October 2025, VCFv4.1 format) served as the primary data source. Only records annotated as somatic variants (ORIGIN bit = 2) with defined clinical significance (CLNSIG) were retained. Multi-allelic ALT fields were decomposed and variants deduplicated by `CHROM:POS:REF:ALT`.

After preprocessing, the dataset comprised 7,789 unique somatic variants, of which 2,680 had unequivocal clinical interpretations (Pathogenic or Benign). Among labeled variants, 2,270 (84.7%) were pathogenic and 410 (15.3%) benign, yielding imbalance ratio 0.18:1. Data were stratified into 80% training+validation and 20% held-out test partitions.

Dataset characteristics are summarized in Fig. 2. Subplot (a) shows class distribution across splits, (b) depicts variant types from CLNVC, and (c) presents $\log_{10}$-transformed allele-frequency distributions from ExAC, ESP, and 1000 Genomes Project (TGP), with most variants below 10^{-4}.

C. Label Harmonization and Target Definition

Clinical significance (CLNSIG) strings were harmonized into binary targets. Variants annotated as "Pathogenic", "Likely_pathogenic", or combinations were labeled pathogenic (target=1); "Benign", "Likely_benign", or combinations were labeled benign (target=0). Ambiguous labels were excluded. Final dataset contained 2,680 variants with class distribution shown in Table I.

TABLE I
Data Splits and Class Distribution

Split	Total	Pathogenic	Benign
Training	1,715	1,453 (84.7%)	262 (15.3%)
Validation	429	363 (84.6%)	66 (15.4%)
Test	536	454 (84.7%)	82 (15.3%)
Train (SMOTE)	2,760	1,453 (52.6%)	1,307 (47.4%)

D. Feature Engineering

We engineered 19 predictive features spanning genomic context, allele frequency, and ClinVar annotation (Table II). Sequence-level features include genomic position (POS), numeric origin flag (ORIGIN), reference and alternate allele lengths (`ref_len`, `alt_len`), and their absolute difference (`len_diff`). Population frequency features from Exome Sequencing Project (ESP), ExAC, and 1000 Genomes Project were median-imputed with binary indicators (`*_missing`) and log-transformed as $\log_{10}(\text{AF} + 10^{-8})$. Five ClinVar sequence ontology flags (from CLNVC) denote SNVs, deletions, insertions, duplications, or indels.

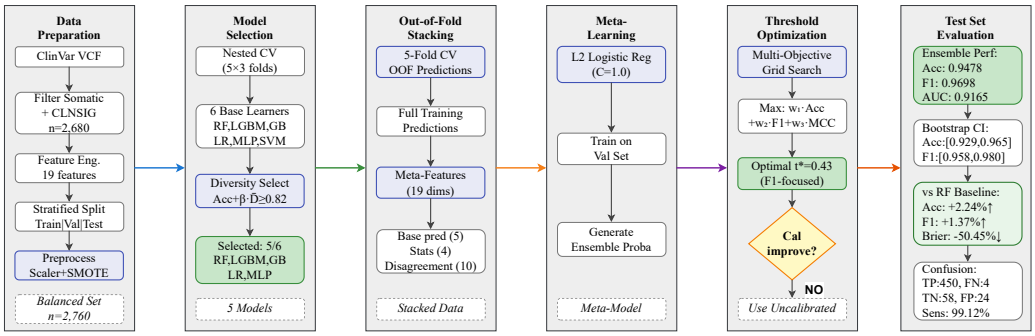


Fig. 1. Six-stage hybrid ensemble framework for somatic variant pathogenicity classification: (1) ClinVar data preparation with 19-dimensional features, (2) diversity-aware model selection via nested cross-validation, (3) out-of-fold stacking with meta-features, (4) L2-regularized meta-learner, (5) multi-objective threshold optimization, and (6) test evaluation with bootstrap confidence intervals.

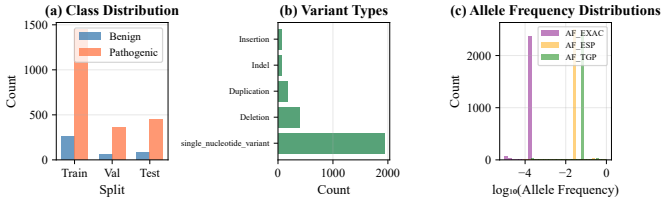


Fig. 2. Dataset characteristics: (a) class distribution across data splits showing pathogenic dominance, (b) variant type distribution with SNVs as majority class, and (c) log-transformed allele frequency distributions from three population databases.

While functional prediction scores (CADD, PolyPhen-2, SIFT) and conservation metrics (GERP++, PhyloP) are powerful, we limited features to VCF-native annotations to: (1) establish a reproducible baseline without external databases, (2) demonstrate ensemble methodology effectiveness independent of computationally expensive annotations, and (3) ensure real-time clinical processing feasibility.

TABLE II
ENGINEERED FEATURES FOR VARIANT CLASSIFICATION

Category	Feature	Description
Genomic	POS	Genomic position
	ORIGIN	Origin flag (somatic=2)
	ref_len	Length of REF allele
	alt_len	Length of ALT allele
	len_diff	$\lvert \text{ref_len} - \text{alt_len} \rvert$
Population	AF_ESP	Raw allele frequency (ESP)
	AF_EXAC	Raw allele frequency (ExAC)
	AF_TGP	Raw allele frequency (1000G)
	AF_ESP_log	$\log_{10}(\text{AF_ESP} + 10^{-8})$
	AF_EXAC_log	$\log_{10}(\text{AF_EXAC} + 10^{-8})$
	AF_TGP_log	$\log_{10}(\text{AF_TGP} + 10^{-8})$
	AF_*_missing	Binary imputation flags
Variant Type	is_snv	Single nucleotide variant
	is_deletion	Deletion variant
	is_insertion	Insertion variant
	is_duplication	Duplication variant
	is_indel	Insertion/deletion (CLNVC)

E. Preprocessing and Cross-Validation Protocol

A held-out test set (20% of samples) was set aside prior to model development (stratified by class). The remaining 80% underwent nested cross-validation (5 outer folds × 3 inner folds) for unbiased base model selection [17]. For models requiring feature scaling (Logistic Regression, SVM, and MLP), StandardScaler was fitted on each training fold and applied to validation and test partitions. Tree-based models (RF, GBM, and LGBM) used raw features. All random operations used seed 42.

The computational environment comprised Python 3.12.0, NumPy 2.3.3, Pandas 2.3.3, scikit-learn 1.7.2, LightGBM 3.3.2, imbalanced-learn 0.10.1, and PyTorch 2.9.0+cu128 on Windows 11 with CUDA-enabled GPU acceleration, ensuring deterministic outcomes.

F. Handling Class Imbalance

The dataset was highly imbalanced (benign:pathogenic = 0.18:1). We applied BorderlineSMOTE [18] within each training fold to synthesize benign variants, targeting post-resampling ratio ≈0.7–0.9 (`sampling_strategy`=0.7 for nested CV training, 0.9 for final training). Oversampling was performed only on training data within each CV split to avoid data leakage.

G. Base Learners and Hyperparameters

Base learners included both tree-based and parametric classifiers (Table III). Random Forest (RF) used 200 trees with max depth 20 and balanced class weights. Gradient Boosting (GB) employed 100 trees with learning rate 0.01. LightGBM combined 1000 trees with L1/L2 regularization and balanced class weights. Logistic Regression (LR) used L2 penalty with C=0.1. Support Vector Machine (SVM) employed RBF kernel with probability estimation. Multi-Layer Perceptron (MLP) featured architecture (256-128-64-32) with ReLU activations, Adam solver, and early stopping.

Each non-MLP classifier was wrapped in `CalibratedClassifierCV` with isotonic regression (inner 3-fold CV) to yield calibrated probability estimates [19]. MLP outputs are inherently probabilistic via softmax. All hyperparameters were chosen a priori based on preliminary experiments.

H. Stacking Design: OOF Predictions and Model Selection

We implemented two-stage stacking. Out-of-fold (OOF) predictions were obtained for each base model via 5-fold CV on the training set, yielding an $N \times M$ OOF probability

TABLE III
BASE CLASSIFIERS WITH KEY HYPERPARAMETERS

Model	Key Hyperparameters
RF	n=200, depth=20, split=5, balanced weights
GB	n=100, lr=0.01, depth=5, subsample=0.8
LGBM	n=1000, lr=0.01, depth=6, L1/L2=1.0, balanced
LR	L2 penalty, C=0.1, max_iter=1000, scaled
SVM	RBF kernel, C=1.0, probability, scaled
MLP	(256-128-64-32), ReLU, Adam, $\alpha = 10^{-4}$, scaled

All models trained with seed=42 for reproducibility

matrix. Each base model was retrained on full training data for validation and test predictions.

For diversity-aware selection, we computed validation accuracy $\text{Acc}(m)$ and mean disagreement:

$$\bar{D}(m) = \frac{1}{|M| - 1} \sum_{m' \neq m} D(m, m'), \tag{1}$$

where $D(m, m')$ is the fraction of samples where m and m' disagree. Composite score:

$$S(m) = \text{Acc}(m) + \beta \bar{D}(m), \tag{2}$$

with $\beta = 0.08$. Models with $\text{Acc}(m) \geq 0.82$ were retained and ranked by $S(m)$. Five out of 6 models met the threshold; SVM was excluded (accuracy 0.78). Retained models: RF, LGBM, GBM, LR, and MLP (detailed metrics in Table VI).

I. Meta-Features and Meta-Learner

The meta-learner's input concatenated: (i) $k = 5$ base-model predicted probabilities, (ii) summary statistics (mean, std, max, min of the k probabilities), and (iii) pairwise absolute differences $|\hat{p}_i^{(m)} - \hat{p}_i^{(m')}|$ for all model pairs, yielding $k + 4 + \binom{k}{2} = 19$ meta-features capturing consensus and disagreement.

As meta-learner, we chose L2-regularized logistic regression (C=1.0). Logistic regression produces calibrated probabilities by default [19], and L2 regularization prevents overfitting. The meta-model was trained on validation set predictions and applied to test set.

J. Threshold Optimization and Calibration

Rather than fixing threshold at $t = 0.5$, we performed grid search over $t \in [0.1, 0.9]$ on validation set. Our objective was:

$$J(t) = w_{\text{acc}} \cdot \text{Acc}(t) + w_{\text{f1}} \cdot \text{F1}(t) + w_{\text{mcc}} \cdot \text{MCC}(t). \tag{3}$$

We tested multiple weight configurations (accuracy-focused: 0.7/0.2/0.1, balanced: 0.6/0.3/0.1, F1-focused: 0.5/0.4/0.1) and selected the best. The F1-focused strategy achieved optimal performance with threshold $t^* = 0.430$ on validation (vs. 0.120 for single RF model).

Isotonic calibration of the meta-model was evaluated on the validation set. The uncalibrated meta-learner achieved 0.9478 validation accuracy, while calibrated yielded 0.9459, indicating no improvement. Consequently, the final model remained uncalibrated to avoid overfitting. The ensemble's strong intrinsic calibration (Brier=0.0486, Section 4.4) stems from the L2-regularized logistic meta-learner, which produces well-calibrated probabilities by default [19].

K. Evaluation Metrics and Uncertainty Estimation

Performance was measured using metrics for imbalanced binary classification: accuracy, F1-score, sensitivity (recall for pathogenic), specificity (recall for benign), Matthews Correlation Coefficient (MCC) [17], ROC-AUC, PR-AUC, and Brier score. Uncertainty was quantified via nonparametric bootstrap: 1,000 stratified resamples of the test set (with replacement) computed 95% confidence intervals for each metric.

L. Implementation Details and Reproducibility

The ClinVar VCF dataset (release 19 Oct 2025) is publicly available and contains only de-identified variant annotations. The dataset was partitioned into 1,715 training, 429 validation, and 536 test samples, stratified by class. After BorderlineSMOTE within training folds, balanced training subset contained 2,760 total variants. All model configurations, thresholds, and evaluation metrics were derived from the finalized hybrid ensemble. The ensemble achieved optimal decision threshold of 0.43 via F_1-focused optimization. All preprocessing and evaluation procedures were fully deterministic under the specified computational environment.

IV. RESULTS

A. Experimental Setup and Baseline

We evaluated on 2,680 ClinVar somatic variants (84.7% pathogenic) with stratified 80/20 train-test split. Nested 5×3 cross-validation on 2,144 training variants (Table IV) identified Random Forest as optimal single classifier (F1 = 0.9694 ± 0.0047), followed by LGBM (0.9693 ± 0.0059). Tree models outperformed LR and SVM, validating 19-dimensional feature engineering.

TABLE IV
NESTED CROSS-VALIDATION PERFORMANCE OF BASE CLASSIFIERS (5-FOLD OUTER, 3-FOLD INNER)

Model	F1 Score (mean ± std)
Random Forest	0.9694 ± 0.0047
LightGBM	0.9693 ± 0.0059
Gradient Boosting	0.9614 ± 0.0044
Logistic Regression	0.9567 ± 0.0068
SVM	0.9498 ± 0.0060

B. Test-Set Performance

Table V compares RF baseline to hybrid ensemble. The ensemble achieved 94.78% accuracy (+2.24pp, +2.42% relative), F1=0.9698 (+1.37% relative), with sensitivity 99.12% and specificity 70.73%. MCC improved 13.53% (0.6925→0.7862), ROC-AUC 1.79% (0.9004→0.9165). Brier score decreased 50.45% (0.0981→0.0486), indicating superior calibration. The confusion matrix (Fig. 4) reveals 450 true positives, 4 false negatives, 58 true negatives, and 24 false positives, yielding 99.12% sensitivity meeting critical safety requirements for clinical genomics.

C. Ensemble Composition and Ablation

Diversity-aware selection retained RF, LGBM, GB, LR, and MLP with performance-based weights (Table VI). Ablation across 31 combinations (Table VII) showed pairwise ensembles (RF+LGBM: 0.9478 accuracy) captured substantial gains. Four-model configuration (RF+LGBM+GB+LR) peaked at

TABLE V
TEST-SET PERFORMANCE COMPARISON (N=536 VARIANTS, 454 PATHOGENIC, 82 BENIGN)

Metric	Random Forest	Hybrid Ensemble
Accuracy	0.9254	**0.9478**
F1 Score	0.9567	**0.9698**
Sensitivity	0.9736	**0.9912**
Specificity	0.6585	**0.7073**
MCC	0.6925	**0.7862**
ROC-AUC	0.9004	**0.9165**
PR-AUC	0.9740	0.9719
Brier Score	0.0981	**0.0486**
TP / FN	442 / 12	**450 / 4**
TN / FP	54 / 28	**58 / 24**
Optimal Threshold	0.120	0.430

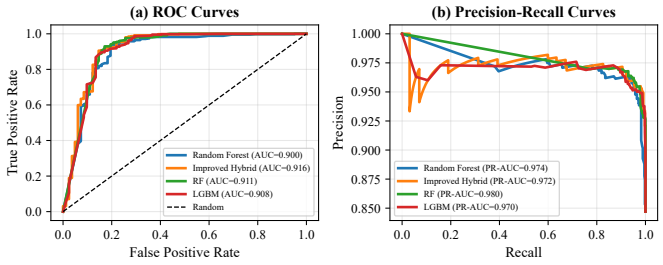


Fig. 3. (a) ROC curves comparing Random Forest, top 2 selected base models (RF, LGBM), and the improved hybrid ensemble, showing consistent AUC improvements. (b) Precision-Recall curves demonstrating high performance across all models due to pathogenic class prevalence, with PR-AUC values exceeding 0.97.

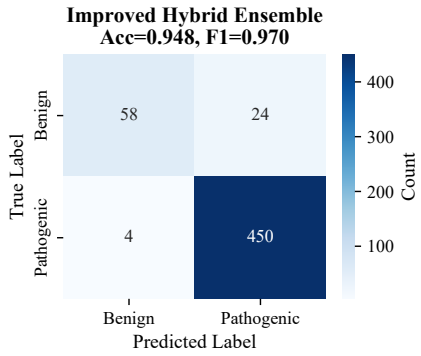


Fig. 4. Confusion matrix for the improved hybrid ensemble on the test set, showing high sensitivity (99.12%) with 450 correctly identified pathogenic variants and only 4 false negatives. Specificity of 70.73% reflects the challenging nature of benign variant identification in imbalanced clinical datasets. Accuracy = 0.9478, F1 = 0.9698.

0.9496 accuracy, while five-model ensemble maximized ROC-AUC (0.9165). Superiority stems from (1) 19 meta-features, (2) weighted OOF stacking, (3) adaptive threshold (0.43 vs. 0.12).

D. Calibration and Uncertainty

Fig. 5a compares calibration curves for Random Forest and the hybrid ensemble. The ensemble tracks perfect calibration significantly better (Brier 0.0486 vs. 0.0981), with predicted probabilities closely aligned with true positive fractions across bins. This 50% reduction in calibration error indicates the meta-learner successfully corrects systematic biases in base

TABLE VI
SELECTED BASE MODELS WITH DIVERSITY METRICS AND PERFORMANCE-BASED WEIGHTS

Model	Val. Acc.	Diversity	Weight
Random Forest	0.9487	0.1152	0.2097
LightGBM	0.9394	0.1096	0.2076
Gradient Boosting	0.9184	0.1077	0.2029
Logistic Regression	0.8765	0.1357	0.1943
MLP	0.8322	0.1907	0.1855
SVM (excluded)	0.7832	0.2019	—

TABLE VII
ABLATION STUDY: TEST-SET PERFORMANCE OF MODEL COMBINATIONS

Ensemble Configuration	Acc.	F1	AUC
2-Model Ensembles (10 combinations)			
RF + LGBM	**0.9478**	**0.9696**	0.9101
LGBM + GB	0.9459	0.9687	**0.9131**
RF + LR	0.9403	0.9652	0.9044
RF + MLP	0.9403	0.9652	0.9035
3-Model Ensembles (10 combinations)			
RF + LGBM + GB	**0.9478**	**0.9697**	**0.9162**
RF + LR + MLP	**0.9478**	0.9696	0.9034
RF + LGBM + LR	0.9459	0.9687	0.9080
RF + LGBM + MLP	0.9459	0.9688	0.9094
4-Model Ensembles (5 combinations)			
RF + LGBM + GB + LR	**0.9496**	**0.9709**	**0.9155**
RF + LGBM + GB + MLP	0.9478	0.9698	0.9124
RF + LGBM + LR + MLP	0.9459	0.9688	0.9101
5-Model Ensemble (1 combination)			
RF + LGBM + GB + LR + MLP	**0.9478**	**0.9698**	**0.9165**

predictions. Fig. 5b demonstrates clear class separation with minimal overlap. Bootstrap CIs (Table VIII) confirm statistical robustness: ensemble accuracy [0.9291, 0.9646] vs. RF [0.9030, 0.9459], with non-overlapping MCC and AUC intervals.

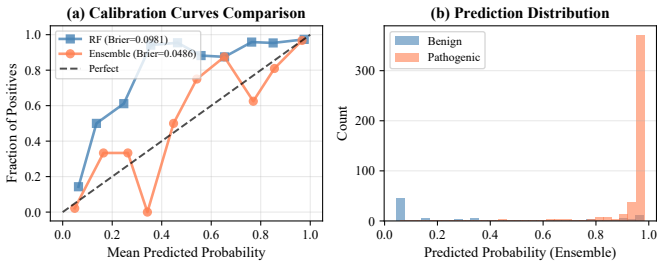


Fig. 5. (a) Calibration curves comparing Random Forest (Brier=0.0981) and hybrid ensemble (Brier=0.0486), demonstrating superior reliability. (b) Prediction distribution showing clear separation between benign ($p < 0.4$) and pathogenic ($p > 0.8$) classes.

TABLE VIII
BOOTSTRAP 95% CONFIDENCE INTERVALS (1,000 RESAMPLES)

Metric	Random Forest	Hybrid Ensemble
Accuracy	[0.9030, 0.9459]	[0.9291, 0.9646]
F1 Score	[0.9427, 0.9693]	[0.9584, 0.9801]
ROC-AUC	[0.8519, 0.9430]	[0.8704, 0.9583]
PR-AUC	[0.9589, 0.9872]	[0.9503, 0.9906]
MCC	[0.6021, 0.7748]	[0.7118, 0.8540]

E. Feature Importance

SHAP analysis (Fig. 6) identified top predictors: allele frequency logarithms (AF_EXAC_log, AF_ESP_log, AF_TGP_log), length metrics (len_diff, ref_len), and variant flags (is_snv, is_deletion). Rare alleles (AF $\ll$ 1%) correlate with pathogenicity; larger indels disrupt frames. Patterns align

with ACMG/AMP guidelines, enhancing clinical interpretability. The dominance of population allele frequencies reflects that pathogenic somatic mutations are rare in healthy populations, while length-based features rank second as larger indels frequently disrupt reading frames and protein domains.

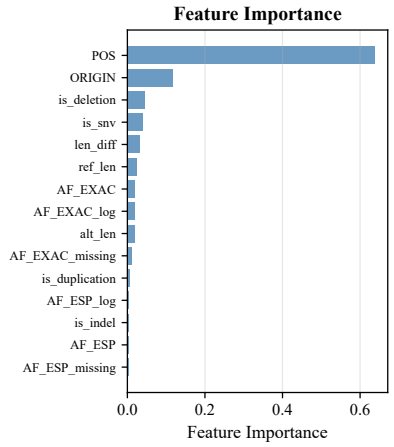


Fig. 6. Feature importance (mean absolute SHAP values, with tree-based fallback) for the top 15 features in the Random Forest baseline model. Allele frequency logarithms dominate, followed by length-based metrics and variant type indicators, validating biologically plausible feature engineering choices.

V. Conclusion

We addressed somatic variant pathogenicity prediction with calibrated probabilities using ClinVar data (Oct 2025, ∼2,680 labeled variants). Our hybrid ensemble combining RF, LGBM, GB, LR, and MLP via diversity-aware stacking achieved 94.78% accuracy, 96.98% F1, 91.65% ROC-AUC, outperforming the RF baseline (92.54% accuracy, 95.61% F1) with 50% Brier score reduction (0.098 → 0.049). Improvements (+2.4% accuracy, +13.5% MCC, halved calibration error) persisted across bootstrap CIs.

This diversity-stacked ensemble delivers more reliable, calibrated predictions than single classifiers, supporting clinical decision-making with high sensitivity (99.12%) and interpretable confidence measures critical for precision oncology. However, several limitations warrant discussion. Data Source Bias: ClinVar's 84.7% pathogenic prevalence reflects expert curation rather than clinical reality, where Variants of Uncertain Significance (VUS) constitute 50–70% of reports [8]. Performance on VUS-heavy datasets remains untested. Limited Features: The 19-dimensional feature set lacks functional scores (CADD, SIFT, PolyPhen-2) and conservation metrics (GERP++, PhyloP), which capture evolutionary constraints and protein structural impacts [10]. While prioritizing reproducibility and computational efficiency, functional annotation integration could improve discrimination. Generalization External validation on independent cohorts (COSMIC, TCGA) is pending. The model does not account for gene-specific or tissue-specific pathogenicity patterns. Finally, the framework assumes independence between variants, ignoring compound heterozygosity or epistatic interactions.

Future work includes external validation, deep learning comparison, and functional feature integration (conservation scores, protein structure). Our framework provides a reproducible tool combining accuracy with calibrated uncertainty for automated variant classification in research and clinical practice.

References

[1] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big Data: Astronomical or genomical?" *PLoS Biol.*, vol. 13, no. 7, p. e1002195, 2015.

[2] Y. Christinat *et al.*, "Reporting of somatic variants in clinical cancer care: recommendations of the Swiss Society of Molecular Pathology," *Virchows Arch.*, vol. 485, no. 6, pp. 1033–1039, 2024.

[3] M. N. U. Alam *et al.*, "Machine learning in biological research: key algorithms, applications, and future directions," *BMC Biol.*, vol. 23, no. 1, p. 324, 2025.

[4] E. Taccaliti and J. S. Aguilar-Ruiz, "Improving classification on imbalanced genomic data via KDE-based synthetic sampling," *BioData Min.*, vol. 18, no. 1, p. 60, 2025.

[5] R. D. Chow, K. L. Nathanson, and R. B. Parikh, "Phenotypic evaluation of deep learning models for classifying germline variant pathogenicity," *NPJ Precis. Oncol.*, vol. 8, no. 1, p. 235, 2024.

[6] C. Papangelou, K. Kyriakidis, P. Natsiavas, I. Chouvarda, and A. Malousi, "Reliable machine learning models in genomic medicine using conformal prediction," *Front. Bioinform.*, vol. 5, p. 1507448, 2025.

[7] S. Fan *et al.*, "Probability calibration-based prediction of recurrence rate in patients with diffuse large B-cell lymphoma," *BioData Min.*, vol. 14, no. 1, p. 38, 2021.

[8] T. N. Tran *et al.*, "AI cancer driver mutation predictions are valid in real-world data," *Nat. Commun.*, vol. 16, no. 1, p. 8509, 2025.

[9] S. Das, V. Patel, S. Chakravarty, A. Ghosh, A. Mukhopadhyay, and N. K. Biswas, "An ensemble machine learning-based performance evaluation identifies top In-Silico pathogenicity prediction methods that best classify driver mutations in cancer," *BioData Min.*, vol. 18, no. 1, p. 7, 2025.

[10] M. Kang, S. Kim, D.-B. Lee, C. Hong, and K.-B. Hwang, "Gene-specific machine learning for pathogenicity prediction of rare BRCA1 and BRCA2 missense variants," *Sci. Rep.*, vol. 13, no. 1, p. 10478, 2023.

[11] M. E. Ramaker, J. W. Abdulrahim, K. M. Corey, R. C. Ramaker, L. C. Kwee, W. E. Kraus, and S. H. Shah, "Cardiovascular disease pathogenicity predictor (CVD-PP): A tissue-specific in silico tool for discriminating pathogenicity of variants of unknown significance in cardiovascular disease genes," *Circ. Genom. Precis. Med.*, vol. 17, no. 6, p. e004464, 2024.

[12] S. W. Kong, I.-H. Lee, L. V. Collen, M. Field, A. K. Manrai, S. B. Snapper, and K. D. Mandl, "Discordance between a deep learning model and clinical-grade variant pathogenicity classification in a rare disease cohort," *NPJ Genom. Med.*, vol. 10, no. 1, p. 17, 2025.

[13] C. Bi, Y. Shi, J. Xia, Z. Liang, Z. Wu, K. Xu, and N. Cheng, "Ensemble learning-based predictor for driver synonymous mutation with sequence representation," *PLoS Comput. Biol.*, vol. 21, no. 1, p. e1012744, 2025.

[14] M. Tejura, S. Fayer, A. E. McEwen, J. Flynn, L. M. Starita, and D. M. Fowler, "Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes," *Am. J. Hum. Genet.*, vol. 111, no. 9, pp. 2031–2043, 2024.

[15] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott, "ClinVar: Public archive of relationships among sequence variation and human phenotype," *Nucleic Acids Research*, vol. 42, no. D1, pp. D980–D985, 2014.

[16] E. LeDell and S. Poirier, "H2O AutoML: Scalable automatic machine learning," in *Proc. AutoML Workshop at ICML*, 2020.

[17] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010.

[18] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*, ser. Lecture Notes in Computer Science, vol. 3644, D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds. Berlin: Springer, 2005, pp. 878–887.

[19] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Machine Learning (ICML)*, 2005, pp. 625–632.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Swin Transformer-Based Multi-Class Classification of Gallbladder Diseases: A Comparative Study with Lightweight CNNs

Abhijit Kumar Ghosh
Dept. of CSE
Brac University
Bangladesh
abhijit.kumar.ghosh.77880@gmail.com

Chowdhury Sajadul Islam
Dept. of CSE
European University of Bangladesh
Bangladesh
chowdhury.uu@gmail.com

Mohammad Azim
Dept. of CSE
Brac University
Bangladesh
mohammad.azim.mehraj@g.bracu.ac.bd

Fariha Nusrat
Dept. of CSE
University of Asia Pacific
Bangladesh
fariha17nusrat@gmail.com

Md. Ahasanul Arafath
Dept. of CSE
European University of Bangladesh
Bangladesh
ahasanul.arafath@gmail.com

Md Tanzim Reza
Dept. of CSE
Brac University
Bangladesh
tanzim.reza@bracu.ac.bd

***Abstract*—Early detection of gallbladder disease improves patient outcomes, and ultrasound is the first-line imaging modality. We develop and evaluate a Swin Transformer (Swin-T) model for multi-class gallbladder disease classification using the UIdataGB ultrasound dataset. To contextualize performance, we compare Swin-T with compact CNN baselines MobileNetV3, ShuffleNetV2, and SqueezeNet. On a held-out test set, Swin-T attains the highest accuracy of 98.99%, exceeding MobileNetV3 (97.81%), ShuffleNetV2 (88.09%), and SqueezeNet (82.45%). These findings highlight the benefit of attention-based hierarchical representations for ultrasound image interpretation, while confirming that lightweight CNNs remain viable for resource-constrained deployments. Overall, UIdataGB provides a strong foundation for advancing and fairly benchmarking AI-assisted gallbladder disease diagnosis.**

***Index Terms*—Gallbladder disease, ultrasound imaging, UIdataGB, Swin Transformer, MobileNetV3, ShuffleNetV2, SqueezeNet, deep learning, multi-class classification, computer-aided diagnosis.**

I. Introduction

Diagnosis typically involves multiple tests such as complete blood count, liver function tests, pancreas function tests, HIDA scan, ERCP, EUS, and most importantly, ultrasound imaging from a diagnostic center or hospital. The problem lies in detecting this disease; it is not always prominent, so it often gets ignored. Although up to 15% of people have gallstones, only 20% of those who have them may experience symptoms or consequences [2]. Manual interpretation of gallbladder ultrasound images can be challenging due to overlapping structures. Early and accurate diagnosis is mandatory to skip the critical situation caused by gallbladder diseases. Artificial intelligence (AI), particularly deep learning frameworks, edge detection, and convolutional neural networks (CNNs), has emerged as a powerful tool for computer vision and automating medical image analysis, which is assisting in diagnosis [3],[5]. A few years ago, research on the same topic of gallbladder diseases often focused on simple and imbalanced datasets, which is mainly problematic when machine learning and AI models are trained on that type of dataset [6], [7].

Gallbladder disease is influenced by several risk factors. The most common type is the accumulation of excess cholesterol in the gallbladder. Overweight and obese individuals are particularly at risk, as well as those over the age of 60. Additionally, females with excess estrogen, especially during pregnancy, are more likely to experience inhibited gallbladder emptying and elevated cholesterol levels. People with diabetes are also at higher risk, along with those suffering from cirrhosis of the liver, sickle cell disease, and individuals on cholesterol-lowering medications. Furthermore, the risk of gallbladder disease is notably higher in Native Americans, with prevalence rates of 30% in men and 60% in women. These factors collectively contribute to the increased likelihood of developing gallbladder disease [8].

In addition, there is a significant gap in medical AI research due to the lack of large publicly available datasets for gallbladder (GB) diseases. Moreover, previous studies have typically focused on detecting only a single type of GB disease, which does not reflect the real-world clinical complexity where multiple disease types often coexist [9], [10]. This gap motivated our multi-class design. Transformer, a recent state-of-the-art attention-based model, is evaluated here for fully automatic classification of gallbladder disease from ultrasound images. We utilize the *UIdataGB* dataset [11], which contains 10,692 high-resolution ultrasound images collected from 1,782 individuals, covering a wide range of diagnostic categories. Detecting gallbladder disease from ultrasound images can be a complex task due to the varying size and shape of the gallbladder across patients. In addition, obesity and bowel

gas can cause poor visualization in the ultrasound images, leading to an increased probability of false identification of disease. Thus, AI-enabled detection of the disease can enhance the test experience of the patient. The mentioned comprehensive dataset enables robust training and evaluation of deep learning models, making it well-suited for developing and benchmarking automated gallbladder diagnostic systems.

For assessing the performance of the Swin Transformer, we conducted a comparative analysis with compact convolutional neural network (CNN) models MobileNetV3, ShuffleNetV2, and SqueezeNet. From the experimental results, it is evident that while these CNN models performed well, providing classification accuracies of 97.81% (MobileNetV3), 88.09% (ShuffleNetV2), and 82.45% (SqueezeNet), the Swin Transformer surpassed their performance, achieving an accuracy of 98.99%. This enhancement demonstrates the capacity of the Swin Transformer in modeling global contextual information and capturing fine-grained details of ultrasound images.

The main contributions to this research are as follows:

- A novel application of the Swin Transformer is presented for the classification of gallbladder diseases from ultrasound images, demonstrating its effectiveness in medical imaging tasks.
- The *UIdataGB* dataset, a multi-class ultrasound dataset specifically curated for gallbladder disease classification, is utilized to support reproducible and standardized research.
- A thorough comparative analysis is conducted with compact CNN models MobileNetV3, ShuffleNetV2, and SqueezeNet to benchmark the performance of the Swin Transformer model.
- The Swin Transformer model achieved state-of-the-art accuracy of 98.99%, outperforming the CNN-based models (MobileNetV3: 97.81%, ShuffleNetV2: 88.09%, SqueezeNet: 82.45%).

II. Literature Review

Early detection is crucial in the first stage of care to prevent or pause disease progression for an extended period through secondary interventions. Such detection enables medical staff and patient attendants to provide targeted care that supports long-term prevention and improves future quality of life. Traditional ultrasound analysis heavily depends on expert interpretation, leading to variability and diagnostic delays. In recent years, machine learning (ML) and deep learning (DL) techniques have shown promising results in automating medical image analysis, including gallbladder disease detection.

Several studies have applied CNN-based architectures to classify gallbladder conditions with high accuracy. For instance, Obaid *et al.* applied DL techniques to classify nine types of gallbladder (GB) diseases using ultrasound images (UI). A balanced dataset of 10,692 images from 1,782 patients was collected and preprocessed for model training. Four deep neural network (DNN) models were tested, with MobileNet achieving the highest accuracy of 98.35% in detecting GB diseases [12]. Similarly, CBIR-based models combining multiple CNN features (DenseNet201, GoogLeNet, InceptionV3, and NasNetLarge) have reported strong performance with AP scores up to 0.94 when compared against six existing models [13]. A recent review highlighted the growing application of AI, particularly ML and DL, in improving the accuracy and speed of GB disease diagnosis. It reported that SVM achieved the highest ML accuracy at 96.67%, while a hybrid DL approach using VGG, GoogleNet, ResNet, AlexNet, and Inception reached up to 98.77%, indicating strong potential for enhancing diagnostic outcomes [14].

A retrospective CT-radiomics study on 141 pathologically confirmed GBC patients used manual tumor segmentation with LIFEx and LASSO + Random Forest to select 15 key features (e.g., GLZLM-HGZE, GLCM-homogeneity, NGLDM-coarseness), yielding 3-year AUCs of 0.79 (test) and 0.73 (validation) and significant risk stratification (HR = 1.462, 95% CI: 1.014–2.107; $P = 0.012$) [15]. A Bayesian Network model built on 818 SEER cases of curatively resected advanced gallbladder adenocarcinoma reported 69.67% accuracy and 77.72% AUC, while estimating adjuvant-therapy benefit and identifying major prognostic factors [16].

Although gallbladder cancer is rare, its aggressive nature and late diagnosis highlight the need for improved diagnostic tools. Among the CNN-based studies, one study shows AI-based methods particularly a fusion of Faster R-CNN and YOLOv8 enhanced detection accuracy from ultrasound images, achieving 92.62%, surpassing the performance of either model alone [9]. In addition, a recent review analysed 16 studies on the application of AI in ultrasound-based diagnosis of gallbladder diseases, including polyps, cancer, stones, and biliary atresia. While a variety of AI methods were explored, CNN-enhanced ultrasound techniques were among those showing promising diagnostic support, particularly for GB polyps and cancer, though challenges remain in broadening clinical adoption [17].

Several studies have explored non-CNN-based ML approaches for detecting or predicting urological diseases. Hong *et al.* utilised 600 regions of interest (ROIs) from 90 gallbladder ultrasound images, applied various traditional ML classifiers, and found that the Random Forest model outperformed others with 96.33% accuracy, high precision, recall, and Dice Index, demonstrating strong capability in classifying gallbladder stones [19]. Similarly, in predicting 2 and 5 year kidney stone recurrence using electronic health record (EHR) data, LASSO regression surpassed other models such as Random Forest and XGBoost, achieving AUCs of 0.62 and 0.63, with age, BMI, and gout history emerging as key predictors [20]. Furthermore, a comprehensive review of AI applications in urinary stone disease emphasised the growing potential of ML techniques for diagnosis, treatment prediction, and surgical planning, although it also noted that challenges related to generalizability and clinical integration remain [21]. Additionally, recent studies have emphasised the role of ML techniques such as SVM and Random Forest, alongside DL models like VGG16 CNN and InceptionV3, in enhancing

kidney stone detection accuracy beyond traditional ultrasound and CT methods. While these approaches show promise in classification and feature extraction, issues related to data accessibility, model interpretability, and clinical applicability still pose challenges [22].

Despite the promising results from CNN-based models, several limitations exist in current approaches for detecting gallbladder diseases. A key limitation is the reliance on small, non-publicly available datasets, which can limit the generalizability of the findings and hinder broader clinical adoption. Additionally, while hybrid models combining CNNs with other techniques (e.g., Faster R-CNN, YOLOv8) have shown improvements in accuracy, issues related to model robustness, interpretability, and scalability in real-world clinical settings remain unresolved. The scarcity of large, annotated datasets further complicates the development of reliable and clinically applicable models. Moreover, there is limited work exploring novel techniques such as the Swin Transformer, which could enhance model performance. Despite these advancements, the application of attention-based architectures such as Swin Transformers in gallbladder disease diagnosis remains underexplored. Addressing this gap, our study investigates the effectiveness of Swin Transformers in multi-class classification of gallbladder diseases using ultrasound imaging.

III. Methodology

A. Overview

We propose a Swin Transformer (Swin-T) based deep learning framework for multi-class gallbladder disease classification using high-resolution ultrasound images. The pipeline includes: dataset acquisition, preprocessing, model training, prediction, and evaluation. For benchmarking, three lightweight CNN architectures ShuffleNetV2, SqueezeNet, and MobileNetV3 are compared against Swin-T. The overall workflow is shown in Fig. 1.

B. Dataset Description

The experiments utilize the *UIdataGB* dataset [11], comprising 10692 high-resolution grayscale ultrasound images from 1782 individuals. The dataset is divided into nine disease classes (D1–D9), including:

- Gallstones (D1)
- Cholecystitis (D2)
- Gangrenous Cholecystitis (D3)
- Perforation of GB (D4)
- Polyps and Cholesterol Crystals (D5)
- Gallbladder Wall Thickening (D6)
- Adenomyomatosis of GB (D7)
- Carcinoma (D8)
- Intra-abdominal and Retroperitoneal Problems (D9)

A detailed breakdown of image counts, patient counts, and gender distribution for each disease type is presented in Tables I.The largest class (D8: Carcinoma) contains 1590 images from 265 patients, while Gallstones (D1) contains 1326 images from 221 patients. This class-wise distribution ensures a broad coverage of gallbladder disease variations, enabling robust model training and evaluation.

Overall gender statistics, including the number of images and average age for male and female patients, are summarized in Table II. The dataset comprises 6246 female images (mean age: 47 years) and 4446 male images (mean age: 53 years), providing balanced demographic representation. Representative examples from each disease category are shown in Fig. 2.

C. Data Preprocessing

Due to variability in acquisition conditions, the following preprocessing steps are applied:

1) Image Resizing: All images are resized to 900×1200 pixels, maintaining aspect ratio.
2) Data Augmentation: Keras ImageDataGenerator is used for random rotations ($\pm 15^\circ$), flips, translations ($\leq 10\%$), and brightness/contrast adjustments to improve generalization.
3) Normalization: Pixel values standardized to zero mean and unit variance.
4) Noise Reduction: Median filtering (OpenCV) to suppress speckle noise while preserving edges.
5) Data Splitting: 80% training, 20% testing; batch size 32; labels are one-hot encoded.

D. Model Architectures

Swin Transformer (Swin-T): A hierarchical vision transformer that applies shifted window-based multi-head self-attention at each stage. The architecture begins with non-overlapping patch embeddings of size n x n, projected to a particular embedding dimension. Afterward, four stages of transformer blocks progressively reduce spatial resolution while expanding channel depth. The shifting-window mechanism allows cross-window interaction, capturing both local and global features. Layer normalization and GELU activation are applied throughout.
ShuffleNetV2: Lightweight CNN using channel shuffling and grouped convolutions for efficiency in mobile/embedded environments.
SqueezeNet: Compact CNN with "fire modules" combining 1×1 and 3×3 filters to minimize parameters without large accuracy loss.
MobileNetV3: Efficient CNN architecture with inverted residuals and lightweight attention, balancing performance and computational cost.

E. Model Training and Evaluation

All models are initialized with ImageNet-pretrained weights and fine-tuned on the UIdataGB dataset.

- Optimizer: Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
- Loss: Categorical cross-entropy
- Batch size: 32
- Epochs: 10
- Metrics: Accuracy (Acc), Precision (P), Recall (R), and F1-score (F1)

Performance Results:

TABLE I: Summary of dataset details related to disease type.

Diseases (1–5)				
Disease Type	**Images**	**Patients**	**Female**	**Male**
Gallstones	1326	221	137	84
Intra-abdominal and Retroperitoneal Problems	1170	195	110	85
Cholecystitis	1146	191	102	89
Membranous and Gangrenous Cholecystitis	1224	204	109	95
Perforation	1062	177	95	82

Diseases (6–9)				
Disease Type	**Images**	**Patients**	**Female**	**Male**
Polyps and Cholesterol Crystals	1020	170	99	71
Adenomyomatosis	1164	194	108	86
Carcinoma	1590	265	155	110
Various Causes of GB Wall Thickening	990	165	92	73

TABLE II: Summary of gender details.

Sex	Number of Images	Average Age
Female	6246	47
Male	4446	53

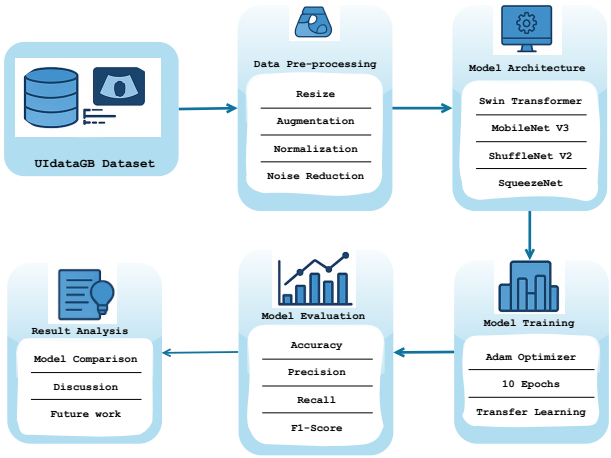


Fig. 1: Proposed methodology workflow for gallbladder disease classification.

- ShuffleNetV2: Acc 88.09%, P 0.8894, R 0.8809, F1 0.8800
- SqueezeNet: Acc 82.45%, P 0.8245, R 0.8245, F1 0.8232
- MobileNetV3: Acc 97.81%, P 0.9786, R 0.9781, F1 0.9781
- Swin-T: Acc 98.99%, P 0.9899, R 0.9899, F1 0.9899

The Swin-T model consistently outperforms the CNN baselines, confirming its advantage in integrating both fine-grained textures and global structural features via its self-attention mechanism.

IV. Results

We evaluated the performance of multiple deep learning models on the *UIdataGB: Multi-Class Ultrasound Images Dataset for Gallbladder Disease Detection*. Specifically, we compared the performance of three convolutional neural networks ShuffleNetV2, SqueezeNet, and MobileNetV3 and a Swin Transformer-based model fine-tuned for multi-class classification.

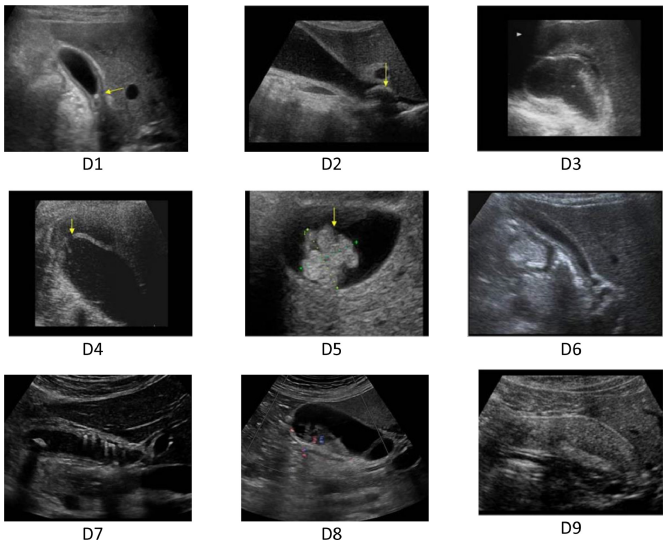


Fig. 2: Sample images from the UIdataGB dataset across the nine disease classes (D1–D9).

TABLE III: Performance comparison of the evaluated models

Model	Accuracy (%)	F1 Score	Precision	Recall
ShuffleNetV2	88.09	0.8800	0.8894	0.8809
SqueezeNet	82.45	0.8232	0.8245	0.8245
MobileNetV3	97.81	0.9781	0.9786	0.9781
Swin-T	**98.99**	**0.9899**	**0.9899**	**0.9899**

A. Performance of the Models Used

In addition to the Swin Transformer, we trained and evaluated three CNN architectures: ShuffleNetV2, SqueezeNet, and MobileNetV3. The dataset was split in an 80:20 ratio for training and testing. Training was conducted for 10 epochs, and results were compared using accuracy, F1 score, precision, and recall. The performance metrics are summarized in Table III.

B. Performance Comparison

The results indicate that Swin Transformer achieved the highest performance across all metrics, attaining an accuracy of 98.99% and an F1 score of 0.9899. This demonstrates its strong ability to capture global contextual features from ultrasound images, making it highly effective for multi-class gallbladder disease classification.

MobileNetV3 ranked second, with an accuracy of 97.81%, showing that lightweight CNN architectures can still perform

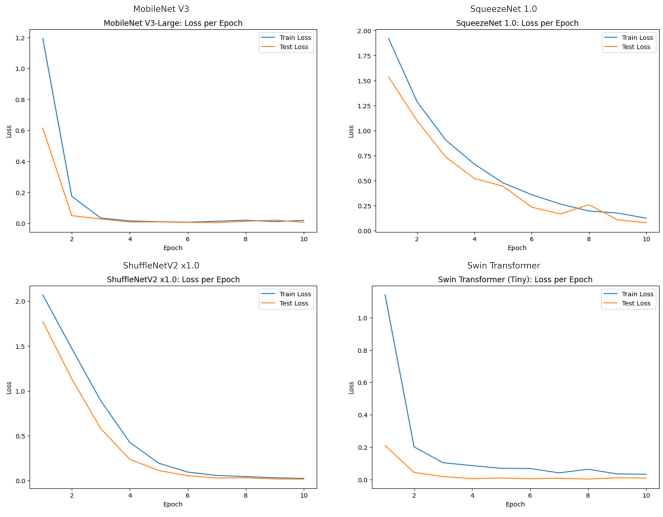


Fig. 3: Training and testing loss curves for four lightweight deep learning models: MobileNet V3, SqueezeNet 1.0, ShuffleNetV2 x1.0, and Swin Transformer.

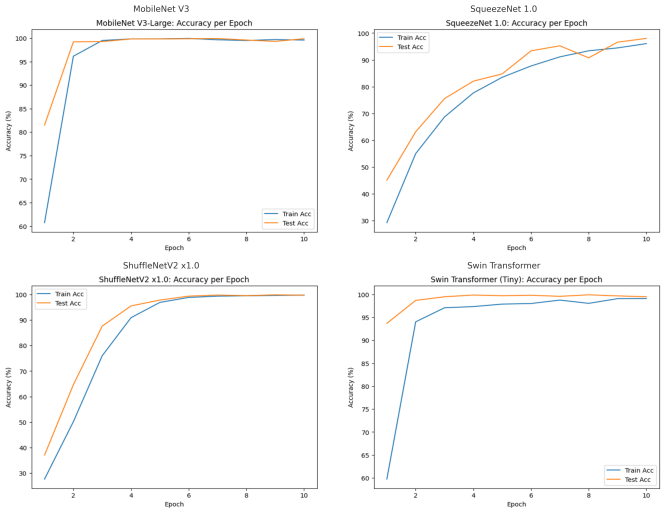


Fig. 4: Training and testing accuracy curves for four lightweight deep learning models: MobileNet V3, SqueezeNet 1.0, ShuffleNetV2 x1.0, and Swin Transformer.

exceptionally well when optimized for efficiency. The close performance between MobileNetV3 and Swin Transformer suggests that MobileNetV3 could be a strong candidate in resource-constrained environments, albeit with a slight drop in accuracy.

ShuffleNetV2 achieved 88.09% accuracy, outperforming SqueezeNet (82.45%), but still lagging behind the top two models. This gap may be attributed to the limited representational capacity of these smaller architectures in handling the complex textural variations of gallbladder ultrasound images, as visualized in the loss and accuracy curves shown in Fig. 3 and Fig. 4, respectively.

C. Observations

- The Swin Transformer demonstrated superior generalization compared to CNNs, likely due to its ability to model long-range dependencies and spatial hierarchies.
- While MobileNetV3's performance was slightly lower than the Swin Transformer, its significantly smaller computational footprint makes it a viable option for real-time and embedded diagnostic systems.
- The performance drop in ShuffleNetV2 and SqueezeNet highlights the trade-off between model size and classification accuracy, suggesting that extremely lightweight models may require further optimization for this domain.

Overall, the experimental results clearly show that transformer-based architectures have strong potential for medical imaging tasks, particularly in scenarios requiring high diagnostic accuracy.

V. Conclusion

In this study, we proposed and evaluated a Swin Transformer based framework for multi-class gallbladder disease classification from ultrasound images, benchmarking it against three lightweight CNN architectures: ShuffleNetV2, SqueezeNet, and MobileNetV3. Experimental results on the *UIdataGB* dataset demonstrate that the Swin Transformer achieved the highest performance, with an accuracy of 98.99% and an F1 score of 0.9899, outperforming all CNN baselines.

MobileNetV3 ranked second with 97.81% accuracy, indicating that efficient CNN models can still deliver strong results in resource-constrained environments. The comparatively lower performance of ShuffleNetV2 and SqueezeNet highlights the trade-off between model compactness and classification accuracy.

However, the dataset used in this work is derived from a single source, which may limit generalizability. Moreover, although the Swin Transformer delivered superior accuracy, it remains computationally expensive to train, raising challenges for deployment in resource-constrained healthcare environments.

Future work will address these issues by expanding the dataset across multiple institutions to improve robustness and generalization. In addition, lighter variants of the Swin Transformer with fewer layers and optimized efficiency will be explored to reduce computational overhead while maintaining high diagnostic accuracy. These steps will support broader clinical adoption of transformer-based methods for automated gallbladder disease diagnosis.

References

[1] C. Lopes Vendrami, M. J. Magnetta, P. K. Mittal, C. C. Moreno, and F. H. Miller, "Gallbladder carcinoma and its differential diagnosis at MRI: what radiologists should know," *Radiographics*, vol. 41, no. 1, pp. 78–95, 2021.

[2] X. Wang *et al.*, "Global epidemiology of gallstones in the 21st century: a systematic review and meta-analysis," *Clinical Gastroenterology and Hepatology*, 2024, in press.

[3] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, "Deep convolutional neural networks in medical image analysis: A review," *Information*, vol. 16, no. 3, pp. 195, 2025.

[4] Y. Xu, T. M. Khan, Y. Song, and E. Meijering, "Edge deep learning in computer vision and medical diagnostics: a comprehensive survey," *Artificial Intelligence Review*, vol. 58, no. 3, pp. 1–78, 2025.
[5] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives," *Neurocomputing*, vol. 444, pp. 92–110, 2021.
[6] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artificial Intelligence Review*, vol. 57, no. 10, pp. 273, 2024.
[7] Y. Ma *et al.*, "An explainable predictive machine learning model of gangrenous cholecystitis based on clinical data: a retrospective single center study," *World Journal of Emergency Surgery*, vol. 20, no. 1, pp. 1, 2025.
[8] A. S. Ahmed *et al.*, "Advancements in cholelithiasis diagnosis: a systematic review of machine learning applications in imaging analysis," *Cureus*, vol. 16, no. 8, 2024.
[9] M. Z. Hasan, M. A. H. Rony, S. S. Chowa, M. R. I. Bhuiyan, and A. A. Moustafa, "GBCHV an advanced deep learning anatomy aware model for accurate classification of gallbladder cancer utilizing ultrasound images," *Scientific Reports*, vol. 15, no. 1, pp. 7120, 2025.
[10] S. Basu, M. Gupta, C. Madan, P. Gupta, and C. Arora, "Focusmae: Gallbladder cancer detection from ultrasound videos with focused masked autoencoders," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 11715–11725.
[11] A. Turki, A. M. Obaid, H. Bellaaj, M. Ksantini, and A. AlTaee, "UI-dataGB: Multi-Class ultrasound images dataset for gallbladder disease detection," *Data in Brief*, vol. 54, pp. 110426, 2024.
[12] A. M. Obaid *et al.*, "Detection of gallbladder disease types using deep learning: An informative medical method," *Diagnostics*, vol. 13, no. 10, pp. 1744, 2023.
[13] A. Bozdag *et al.*, "Detection of gallbladder disease types using a feature engineering-based developed CBIR system," *Diagnostics*, vol. 15, no. 5, pp. 552, 2025.
[14] A. M. Obaid, A. Turki, H. Bellaaj, and M. Ksantini, "Diagnosis of gallbladder disease using artificial intelligence: A comparative study," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, pp. 46, 2024.
[15] Z. Liu *et al.*, "Survival prediction in gallbladder cancer using CT based machine learning," *Frontiers in Oncology*, vol. 10, pp. 604288, 2020.
[16] Z. M. Geng *et al.*, "Estimating survival benefit of adjuvant therapy based on a Bayesian network prediction model in curatively resected advanced gallbladder adenocarcinoma," *World Journal of Gastroenterology*, vol. 25, no. 37, pp. 5655, 2019.
[17] S. Dadjouy and H. Sajedi, "Artificial intelligence applications in the diagnosis of gallbladder neoplasms through ultrasound: A review," *Biomedical Signal Processing and Control*, vol. 93, pp. 106149, 2024.
[18] X. Wang, H. Zhang, Z. Bai, X. Xie, and Y. Feng, "Current status of artificial intelligence analysis for the diagnosis of gallbladder diseases using ultrasonography: a scoping review," *Translational Gastroenterology and Hepatology*, vol. 10, pp. 12, 2024.
[19] C. Hong *et al.*, "Analysis of machine learning algorithms for real-time gallbladder stone identification from ultrasound images in clinical decision support systems," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, pp. 1–28, 2025.
[20] P. Doyle, W. Gong, R. Hsi, and N. Kavoussi, "Machine learning models to predict kidney stone recurrence using 24 hour urine testing and electronic health record-derived features," *Research Square*, pp. rs–3, 2023.
[21] A. Anastasiadis, A. Koudonas, G. Langas, S. Tsiakaras, D. Memmos, I. Mykoniatis, E. N. Symeonidis, D. Tsiptsios, E. Savvides, I. Vakalopoulos, *et al.*, "Transforming urinary stone disease management by artificial intelligence-based methods: a comprehensive review," *Asian Journal of Urology*, vol. 10, no. 3, pp. 258–274, 2023.
[22] M. N. Islam, M. F. H. K. Chowdhury, and F. Nusrat, "An effective deep learning approach for kidney stone detection in CT scan images using convolutional neural network," in *Proc. 2024 IEEE Int. Conf. on Computing, Applications and Systems (COMPAS)*, pp. 1–6, 2024.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Enhancing Community Group Recognition Using CNN Architectures

Swaraj Kumar Sharma, Mohammed Mahmudur Rahman, Md. Jahed Hasan, Jannatul Ferdous
Dept. of Computer Science and Engineering, Bandarban University, Bandarban
Dept. of Computer Science and Engineering, University of Science and Technology Chittagong, Foy's Lake,Chattogram
Dept. of Computer Science and Engineering, Southern University,Arefin Nagar,Chattogram
Dept. of Computer Science and Engineering, Bandarban University, Bandarban
swarajsharma@bubban.ac.bd, provaiiuc@gmail.com, Jahedcse21@gmail.com, saimonjannat1@gmail.com

***Abstract*—Facial image–based group classification has gained increasing attention in computer vision due to its relevance in demographic analysis and social computing applications. This paper presents a deep learning framework for facial population group classification using a transfer learning–based DenseNet-121 model augmented with a custom convolutional neural network (CNN) classification head. The proposed pipeline includes face detection, grayscale conversion, facial landmark alignment, data augmentation, normalization, and quality-aware preprocessing to enhance input consistency. A five-class facial image dataset collected from the Chittagong Hill region, comprising Chakma, Marma, Tripura, Mro, and Bawm groups, is used for evaluation. Model training is performed with configurable layer freezing to balance feature reuse and domain adaptation. Experimental results demonstrate that the proposed approach achieves a classification accuracy of 97.85%, indicating its effectiveness and robustness for facial population group recognition tasks.**

***Index Terms*—Transfer Learning, Landmarks alignment, Augmentation, Normalization, DenseNet-121, Grayscale conversion.**

I. Introduction

Ethnicity classification plays a vital role in facial image analysis, enabling the categorization of individuals into broad groups based on inherited physical traits, geographical origin, and social background. Such classification considers observable characteristics like skin tone, hair and eye color, and facial structure, which reflect both genetic and environmental influences [1].

Face recognition technology has become integral to modern applications in security, biometrics, and entertainment for reliable personal identification. Beyond identity verification, facial data can reveal soft biometric attributes such as emotional state and ethnic background. In recent years, the study of soft biometrics—covering demographic, anthropometric, and geometric aspects of the human face—has gained notable research attention for its potential to enhance recognition accuracy and personalization [2].

Recent advances in convolutional neural networks (CNNs) have transformed the domains of computer vision and pattern recognition. These deep learning architectures have achieved remarkable success across various image-based tasks, including classification and object detection. One promising application involves ethnicity detection, where CNNs automatically infer an individual's ethnic group from visual cues present in facial images [3].

This study aims to explore the application of convolutional neural networks (CNNs) for ethnic group identification within the diverse cultural context of Bangladesh. The country is home to numerous indigenous communities, each with distinct traditions and features. In particular, the Chittagong Hill Tracts region—inhabited by groups such as the Chakma, Marma, Tripura, Mro, and Bawm—is recognized for its rich ethnic diversity, offering a valuable setting for ethnicity-based facial recognition research

To tackle these issues, this research introduces a deep learning-model for ethnicity detection utilizing five tribe dataset.The main contributions of this work are as follows:

- Image cleaning pipeline that can accurately separate the blur, noisy images at the same time.
- Applying the best preprocessing and augmentation, normalize techniques to deal with changes in intensity and overlaps images.
- After preparing the dataset, it was fed into ConvNeXt, EfficientNet V2,and Densenet 121 with a custom CNN classifier head.
- Among the pretrained models, Densenet121 with a custom CNN works better than more recent state-of-the-art methods.

The rest of this paper is set up like this: Section II looks at other works that are similar Ethnicity Detection. In Section III, we go over the proposed methodology in detail. Section IV talks about how the experiment was set up, and Section V talks about the result analysis. Section VI ends with the most important findings and suggestions for future work.

II. Related Works

Research on neural network–based ethnic estimation typically employs three main architectures: Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). In this study, the focus is on CNNs for ethnicity classification due to their ability to automatically extract features and learn classification patterns effectively. Related studies were gathered from multiple academic databases, including IEEE Xplore, ScienceDirect, Springer Link, Web of Science, and Google Scholar.Belcar et

al. [1] proposed that, The study investigated the application of convolutional neural networks (CNNs) for automated ethnicity classification using only the central region of facial images. It employed the UTKFace and FairFace datasets, which include diverse ethnic categories such as White, Black, Latino, East Asian, South East Asian, Middle Eastern, and Indian.The accuracy of the model was 80.34% for UTKface classification into five class as well as 61.74% for fairface into seventh class.Khalil et al. [3] researched the study introduces a facial segmentation approach that divides face images into seven distinct regions and leverages this segmented information for race classification. A novel classification algorithm is developed and evaluated on state-of-the-art datasets, demonstrating improved performance compared to earlier methods. DCNN obtained classification accuracies of 100 %, 99.2 %, 92.0 %, and 93.2 % for the FERET, CAS-PEAL,VNFaces,and VMER datasets, respectively.Sultana Tasnim Jahan et al. [1] proposed this research presents a CNN-based approach for ethnicity detection within Bangladesh's culturally diverse context. A custom dataset was created featuring images of three major ethnic communities from the Chittagong Hill Tracts—Chakma, Marma, and Tripura. The DenseNet-121 architecture was adapted to enhance feature extraction for this specific task, and the model was trained and evaluated using multiple performance metrics. The results demonstrate strong accuracy and highlight the potential of CNNs in promoting fair and inclusive recognition technologies. Ethical concerns regarding prediction bias are also addressed. The main contributions include the creation of a diverse local dataset, architectural optimization of DenseNet for ethnicity detection, and the advancement of computer vision applications in representing Bangladesh's ethnic diversity.The Accuracy of CNN model was 94%.Ahmed Jawad et al. [3] stated that in recent years, face recognition research has expanded significantly, with ethnicity identification emerging as a key challenge. This study introduces a deep learning–based convolutional neural network specifically designed to classify individuals' ethnic backgrounds from facial features. A new dataset comprising 3,141 images from three national groups was developed — the first of its kind for ethnicity recognition — and made available to the research community. The proposed model was evaluated against established architectures such as VGG and InceptionV3, achieving the highest validation accuracy of 96.9 % outperforming existing methods.Achmad et al. [4] introduced The Compact-Fusion Feature Framework shows that an adequate generalization capability for ethnicity categorization while maintaining or increasing accuracy, demonstrating 82.19 % for five classes and 89.14% for four classes on the UTKFace and Fairface dataset.

III. Methodology

A. Dataset Description

This study utilizes 7,622 total input images in the dataset and is starting the quality,duplicate,landmark scan.After applying filtering rules (quality threshold, duplicates removal, etc.), the pipeline selected 3,255 "clean" images for the final dataset.The model was not adequately trained before to dataset scanning with 7622 images. For this reason, once the model was correctly trained, set certain threshold for these dataset. Table I summarizes the dataset properties, including the number of images,Before Scanning dataset quantity,Afer Scanning dataset quantity.

TABLE I
Tribe Dataset Before & After Scanning

Tribe	Before Scanning	After Scanning
Chakma	1949	651
Tripura	1102	651
Marma	1273	651
Bawm	1518	651
Mro	1780	651

B. Data Preprocessing

The dataset was preprocessed to standardize image quality and format for deep learning. Faces were detected using an OpenCV Haar cascade, and an expanded full-face crop was applied to retain contextual features such as the jawline, ears, and hairline. Each image underwent gray-world color balancing, grayscale conversion, and conditional enhancement based on blur (Laplacian variance) and noise estimation, including unsharp masking, denoising, and CLAHE contrast correction where necessary. Preprocessed images were resized to 224×224 pixels to align with ImageNet-pretrained models and converted to three-channel tensors for CNN compatibility. The training split was used to compute per-channel mean and standard deviation values for normalization. Finally, visual and statistical checks ensured consistency across training, validation, and test sets, and class distributions were preserved. Table II depicts the dataset split into three segments: Training, Testing, and Validation, with proportions of 80%,10%,and 10%, respectively. The entire pipeline of Dataset Scan,Landmark Detection, preprocessing, augmentation, normalization, and evaluation is depicted in the proposed workflow diagram in figure 2. This figure illustrates how input data undergoes preprocessing and augmentation before being fed into Custom CNN with an Densenet 121 model [5].

TABLE II
Dataset Splitting and Ratios

Clean Images	Train	Val	Test
3255	2603	326	326

C. Data Augmentation and Normalization

A robust augmentation and normalization pipeline was employed to improve generalization and limit overfitting. The preprocessed images were resized to 224×224 pixels and enhanced through randomized transformations such as brightness/contrast adjustments, horizontal flips, affine and perspective distortions, Gaussian noise, and random erasing. These operations increased data variability and enabled the model to learn stable feature representations under diverse

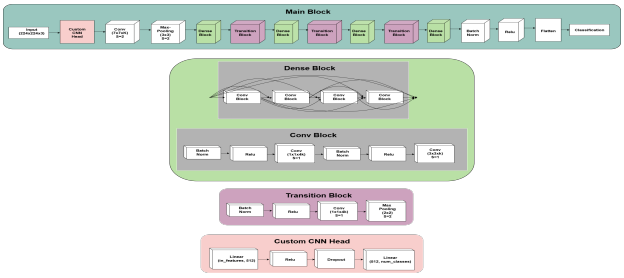


Fig. 1. DenseNet-121 + CNN Classifier Head Architecture

visual conditions. After augmentation, images were normalized using channel-wise statistics computed from the training set to maintain consistent input distributions. Additionally, weighted random sampling was used during training to address class imbalance. Such preprocessing strategies are widely recognized for strengthening CNN-based visual recognition performance. [6]–[8].

D. Network Architecture

Figure 1 illustrates the Custom CNN head and DenseNet-121 architecture.The first approach employs DenseNet-121 using transfer learning, leveraging ImageNet pre-trained weights to extract hierarchical feature representations [9]. A custom classification head replaces the original fully connected layer and comprises a 512-unit dense layer with ReLU activation, dropout regularization, and a final output layer matching the number of classes. Layer freezing is configurable, enabling either partial fine-tuning or full network training to facilitate adaptation to domain-specific data [10].

The second approach utilizes a lightweight custom CNN designed from scratch. It consists of four convolutional blocks with progressively increasing channel depths (32→256), each followed by batch normalization, ReLU activation, and max pooling. Feature maps are aggregated via adaptive average pooling, and the classification head includes a 128-unit dense layer with dropout before the final output layer. This architecture provides a computationally efficient baseline while retaining sufficient representational capacity [11].

Dropout layers were included to prevent overfitting, and a final Max Pooling layer outputs the Five-class classification outputs. The full dataflow diagram is shown in the Figure 2.

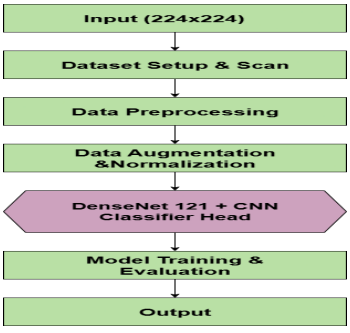


Fig. 2. End-to-end Proposed Workflow Pipeline

IV. Experimental Setup and Implementation

A. Setup and Scan

Each image was evaluated using quantitative quality metrics, such as sharpness (via Laplacian variance), contrast and exposure (mean intensity, standard deviation, and clipped pixels), and color balance (Gray-World deviation). Perceptual hashing (aHash) was employed to detect near-duplicate images [12]. Facial detection was performed with a Haar cascade [13], followed by multi-backend landmark detection using Dlib's 68-point predictor [14] or MediaPipe Face Mesh, from which geometric features relevant to ethnicity (e.g., face aspect ratio, eye-to-face ratio, nasal width) were derived.

Images failing adaptive quality thresholds or lacking detectable faces were removed. The dataset was then class-balanced by selecting the top-quality samples per ethnic class. This preprocessing stage produced a curated, high-quality dataset, free of duplicates and unsuitable samples, providing a robust foundation for deep learning-based ethnic classification [15].

B. Training Strategy

The model was trained using supervised deep learning with DenseNet121 (transfer learning) or a custom SmallCNN backbone for ethnic classification, employing a cross-entropy loss function. Training utilized Stochastic Gradient Descent (SGD) as the optimizer with a learning rate of 0.001, momentum 0.9, and weight decay(1e-4).A ReduceLROnPlateau scheduler reduced the learning rate by a factor of 0.7 if the validation loss did not improve for 4 epochs. The maximum number of epochs was set to 50, with early stopping triggered if validation balanced accuracy did not improve for 8 consecutive epochs, preventing overfitting [16].

C. Loss Functions

The training employed the categorical cross-entropy loss to optimize the model for multi-class ethnic classification. This loss measures the dissimilarity between the predicted probability distribution and the true class labels, penalizing incorrect predictions more heavily as their predicted probability increases. It is particularly suitable for classification tasks where each input belongs to exactly one class. The loss was computed for each mini-batch during training, and the optimizer (SGD) minimized the average batch loss over the training epochs [14].

V. Result Analysis

This section presents a comprehensive analysis of the experimental outcomes obtained from the proposed DenseNet-121 model integrated with a custom CNN classifier head for ethnic classification.The evaluation focuses on both the training dynamics and the final classification performance across five tribal ethnic groups from the Chittagong Hill Tracts region of Bangladesh. The analysis includes training behaviour, detailed classification metrics, confusion matrices with ROC curves, and an overall performance evaluation to

assess the model's effectiveness and generalization capability [17].

A. Training Behavior

The training behaviour of the proposed DenseNet-121 model with a CNN classifier head is illustrated in Figure 3, which presents the evolution of loss, accuracy, balanced accuracy, and macro F1 score for both the training and validation sets. Although the model was initially configured to train for 50 epochs, the training process automatically stopped at the 45th epoch due to the early stopping criterion with a patience factor, indicating that further training did not yield any significant improvement on the validation set. From the

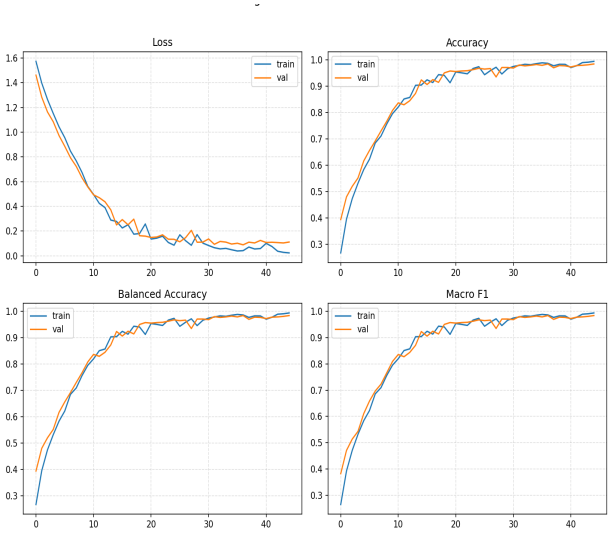


Fig. 3. Model Loss, Acuuracy, Balaced Accuracy and Macro F1 curve with respect to Epoch.

loss curve (top-left of Figure 3), it can be observed that both training and validation loss decreased rapidly during the initial epochs, with the most substantial drop occurring within the first 10–15 epochs. Afterward, the loss curves gradually converged towards very low values (close to zero), signifying that the model learned discriminative features effectively without overfitting. The training and validation loss curves follow each other closely throughout the process, which demonstrates good generalization.

The accuracy curve (top-right of Figure 3) shows a consistent and steep increase in both training and validation accuracy during the early epochs, stabilizing after approximately 25 epochs. Both curves eventually reach and maintain accuracy levels above 97%, with the validation accuracy plateauing at an average of 97.85%, indicating strong and stable performance on unseen data.

Similarly, the balanced accuracy and macro F1 score curves (bottom row of Figure 3) exhibit steady growth patterns, closely matching between the training and validation sets. The balanced accuracy approaches near-perfect values by the end of training, confirming that the model performs uniformly well across all ethnic classes and is not biased toward dominant categories. The macro F1 curve mirrors this trend, reinforcing that the model maintains a high balance between precision and recall for all classes.

Overall, the convergence of the training and validation curves across all metrics, combined with the early stopping at epoch 45, indicates that the model achieved a strong fit without overfitting, and stabilized at a high level of performance well before the maximum number of epochs.

B. Classification Report

Table III shows The detailed classification performance of the proposed model on the test set,which reports precision, recall, and F1-score for each of the five ethnic classes. Overall, the model achieved an impressive accuracy of 97.85%, with macro and weighted averages for all three metrics (precision, recall, and F1-score) reaching 0.9785. This demonstrates the model's strong and balanced predictive ability across all classes.

Among the individual classes, Marma achieved the highest performance with a perfect recall of 1.0000 and an F1-score of 0.9898, indicating that all Marma samples were correctly classified. The Chakma and Bawm classes also showed excellent performance, with F1-scores of 0.9767 and 0.9717, respectively. Similarly, the Mro and Tripura classes maintained high levels of both precision and recall, each above 0.97, reflecting the model's robustness and consistent behavior across different ethnic groups.

The near-uniform values of precision, recall, and F1-score across all classes confirm that the model does not exhibit any significant class imbalance or bias. This level of performance indicates that the DenseNet-121 with the CNN classifier head effectively captured discriminative facial features among the five tribal groups and generalized well to unseen data in the below classification report:

TABLE III
CLASSIFICATION REPORT

Classes	Precision	Recall	F1-score	Support
Bawm	0.9793	0.9643	0.9717	65
Chakma	0.9844	0.9692	0.9767	65
Marma	0.9799	1.0000	0.9898	66
Mro	0.9796	0.9796	0.9796	65
Tripura	0.9695	0.9795	0.9745	65
Accuracy			**0.9785**	326
Macro Avg	0.9785	0.9785	0.9785	326
Weighted Avg	0.9785	0.9785	0.9785	326

C. Confusion Matrix and ROC

The normalized confusion matrix for the DenseNet-121 model with the CNN classifier head is shown in Figure 4. The matrix reveals that the model achieves near-perfect classification performance across all five ethnic classes. The diagonal values are consistently high, with Marma attaining a perfect classification rate of 1.00, and Bawm, Chakma, Mro, and Tripura achieving rates of 0.96, 0.97, 0.98, and 0.98, respectively. Only a minimal number of samples were misclassified, and these errors are sparsely distributed across neighboring classes. This indicates that the model is highly

effective at capturing subtle inter-class facial feature variations among the different ethnic groups.

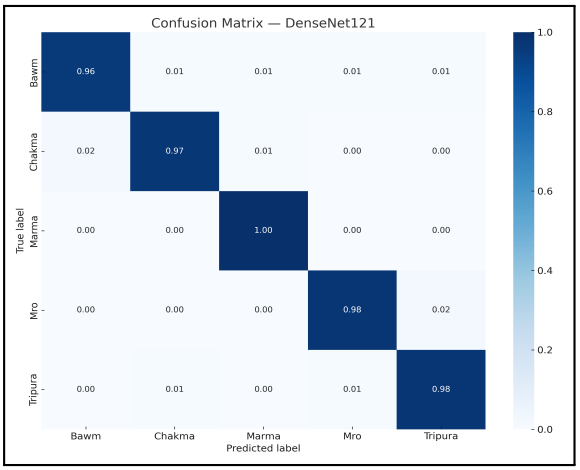


Fig. 4. Confusion Matrix

The ROC curves for each class are presented in Figure 5. All curves lie close to the top-left corner of the plot, indicating outstanding discriminative power. The Area Under the Curve (AUC) values are extremely high for all classes, with Bawm and Chakma both reaching 0.999, Marma achieving a perfect 1.000, and Mro and Tripura both recording 0.995. Furthermore, the micro-average and macro-average AUC scores are 0.997 and 0.998, respectively. These near-ideal ROC characteristics confirm that the model not only classifies samples accurately but also distinguishes between classes with exceptional reliability. Overall, the confusion

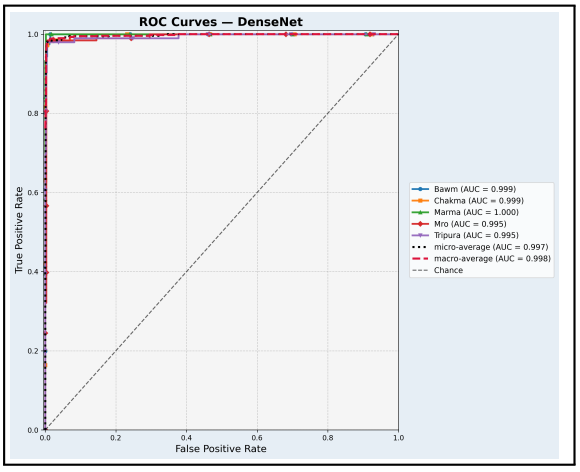


Fig. 5. ROC Curve per class

matrix and ROC curve analyses provide strong evidence of the model's robustness and its ability to generalize effectively across all five target classes, with minimal misclassification and excellent class separability.

D. Performance Evaluation

To further assess the effectiveness of the proposed model, we conducted two types of performance comparisons: (i) a comparison with existing methods reported in the literature, and (ii) a comparison with other deep learning architectures evaluated within our experimental framework.

Table IV shows the accuracy and loss comparison between the proposed model and other state-of-the-art architectures trained and evaluated under the same experimental setup. The ConvNeXt Tiny model achieved 95.70% accuracy with a loss of 0.23, while EfficientNetV2-S performed slightly better with 97.65% accuracy and 0.10 loss. However, the proposed DenseNet121 + CNN classifier head surpassed both, achieving the highest accuracy of 97.85% along with the lowest loss of 0.07, indicating not only better classification performance but also improved training stability and convergence.

TABLE IV
ACCURACY COMPARISON BETWEEN OUR USED MODEL

Model	Accuracy(%)	Loss (%)
ConvNeXt Tiny	95.70	0.23
EfficientNetV2 S	97.65	0.10
DenseNet-121 + CNN Classifier Head	**97.85**	**0.07**

Table V presents a comparative analysis of classification accuracy between the proposed approach and several existing models applied to similar ethnicity or face classification tasks. Conventional architectures such as VGG-16 achieved accuracies of 80.34% on the UTK dataset and 61.74% on FairFace, while basic CNN models reached 85–86% on MORPH and FERET datasets. More advanced architectures like ConvNeXt and DenseNet121 achieved 96.9% and 94% accuracy, respectively, in ethnicity classification settings involving Asian populations and three tribes in Bangladesh. In contrast, our DenseNet121 + CNN classifier head achieved the highest accuracy of 97.85% on a more challenging five-tribe classification task, demonstrating superior performance in capturing fine-grained facial distinctions among different ethnic groups. These results collectively confirm the effectiveness of the proposed architecture in handling complex, multi-class ethnicity classification tasks with high precision and robustness,outperforming both traditional models and recent deep learning baselines.

TABLE V
ACCURACY COMPARISON WITH EXISTING MODELS ON SIMILAR WORKS.

Model	Accuracy	Dataset	Year
VGG-16 [3]	80.34%,61.74%	UTK,FairFace	2022
CNN [18]	85%,86%	MORPH,FERET	2023
ConvNeXt [19]	96.9%	Asian	2020
DenseNet121 [1]	94%	Bangladeshi (3 tribes)	2024
Proposed Model	**97.85%**	**Bangladeshi (5 tribes)**	-

E. Discussion

The experimental evaluation shows that the DenseNet121 model paired with a CNN classifier head delivers strong performance in identifying the five ethnic groups from the Chittagong Hill Tracts. Its high accuracy, balanced statistical measures, and favorable ROC behavior indicate that the model effectively captures subtle and discriminative facial cues, outperforming conventional CNN models and several established

architectures. Results from the confusion matrix and class-wise metrics demonstrate consistent predictions across all groups, with minimal misclassification. Comparisons with previous studies further emphasize the model's ability to generalize in a challenging setting characterized by limited data variability and small inter-class differences. Overall, the combination of DenseNet's dense feature propagation with a lightweight CNN head proves well-suited for fine-grained ethnicity classification.

VI. Conclusions

This study presents a CNN-based framework for ethnicity recognition tailored to the distinct communities of Bangladesh. The proposed model demonstrated strong performance in classifying individuals from the Chakma, Marma, Tripura, Mro, and Bawm groups of the Chittagong Hill Tracts. Future directions include enlarging the dataset to cover additional ethnic populations, enhancing generalization to unseen samples, and exploring real-time implementations for practical use.

References

[1] S. T. Jahan, R. Mustafa *et al.*, "Ethnicity detection with convolutional neural network (cnn): Bangladesh perspectives," *American Journal of Interdisciplinary Research and Innovation*, pp. 5–12, 2024.

[2] K. Khan, R. U. Khan, J. Ali, I. Uddin, S. Khan *et al.*, "Race classification using deep learning." *Computers, Materials & Continua*, vol. 68, no. 3, 2021.

[3] D. Belcar, P. Grd, and I. Tomičić, "Automatic ethnicity classification from middle part of the face using convolutional neural networks," in *Informatics*, vol. 9, no. 1. MDPI, 2022, p. 18.

[4] T. A. B. Wirayuda, R. Munir, and A. I. Kistijantoro, "Compact-fusion feature framework for ethnicity classification," in *Informatics*, vol. 10, no. 2. MDPI, 2023, p. 51.

[5] N. Venkat and S. Srivastava, "Ethnicity detection using deep convolutional neural networks," 2018.

[6] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[7] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," in *CVPR*, 2019.

[8] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI*, 2020.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *CVPR*, 2017.

[10] M. Rohani, H. Farsi, and S. Mohamadzadeh, "Advanced race classification using transfer learning and attention: Real-time metrics, error analysis, and visualization in a lightweight deep learning model," *Journal of Electrical and Computer Engineering Innovations (JECEI)*, vol. 13, no. 2, pp. 341–352, 2025.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[12] C. Zauner, "Perceptual hashing for image similarity," *Master's Thesis, Vienna University of Technology*, 2010.

[13] "Opencv library," https://opencv.org/, 2023.

[14] D. E. King, *Dlib-ml: A Machine Learning Toolkit*. Journal of Machine Learning Research, 2009.

[15] "Mediapipe: Cross-platform ml solutions," https://mediapipe.dev/, 2023.

[16] 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.

[17] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural networks*, vol. 106, pp. 249–259, 2018.

[18] A. A. Abdulwahid, "Classification of ethnicity using efficient cnn models on morph and feret datasets based on face biometrics," *Applied Sciences*, vol. 13, no. 12, p. 7288, 2023.

[19] A. J. A. AlBdairi, Z. Xiao, and M. Alghaili, "Identifying ethnics of people through face recognition: a deep cnn approach," *Scientific Programming*, vol. 2020, no. 1, p. 6385281, 2020.

A Hybrid Prophet–LSTM Framework for Enhanced Monthly Rainfall Forecasting in Climate-Sensitive Regions: A Case Study of Cox's Bazar, Bangladesh

Badhan Dey
Department of Electrical and Electronic Engineering
Noakhali Science & Technology University
Noakhali-3814, Bangladesh
deybadhan0017@gmail.com

Md. Istihad Bhuiyan
Department of Electrical and Electronic Engineering
Noakhali Science & Technology University
Noakhali-3814, Bangladesh
istihad.emon@gmail.com

Md. Sabbir Ahmed
Department of Electrical and Electronic Engineering
Noakhali Science & Technology University
Noakhali-3814, Bangladesh
msabbir1612@gmail.com

Md. Eyakub Ali
Department of Electrical and Electronic Engineering
RTM-Al Kabir Technical University
Sylhet-3100, Bangladesh
meyakub.nstu@gmail.com

Md. Shadman Shakib
Department of Electrical and Electronic Engineering
Metropolitan University
Sylhet-3104, Bangladesh
shadmanshakib003@gmail.com

Fazlul Karim
Department of Electrical and Electronic Engineering
RTM-Al Kabir Technical University
Sylhet-3100, Bangladesh
fazlulkarim7826@gmail.com

Abstract— ***Accurate precipitation forecasting is paramount for hydrological management, agricultural planning, and disaster mitigation in climatically sensitive regions such as Cox's Bazar, Bangladesh. This study proposes a hybrid forecasting framework integrating the decomposable Prophet model with Long Short-Term Memory (LSTM) networks to capture deterministic seasonal-trend components and complex nonlinear residual dynamics within monthly rainfall data (1981–2022). Comparative analyses against SARIMA, standalone LSTM, and GRU models demonstrate the hybrid model's superior performance, achieving exceptional fit metrics (train $R^2 = 0.990$; test $R^2 = 0.987$) and low error rates (RMSE = 1.404; MAE = 0.491), with minimal generalization gaps. These results affirm the robustness and generalizability of the approach, underscoring the efficacy of integrating statistical decomposition with deep learning to overcome limitations in traditional forecasting. This methodology advances predictive analytics for monsoonal precipitation, offering significant implications for flood risk management and sustainable water resource planning.***

Keywords— Rainfall forecasting, Prophet–LSTM hybrid model, RMSE, MAE, R^2 score.

I. INTRODUCTION

In Accurate rainfall forecasting is crucial for effective water management, disaster preparedness, and agricultural planning, especially in areas prone to extreme weather events like Cox's Bazar, Bangladesh. Traditional statistical models such as Seasonal Autoregressive Integrated Moving Average (SARIMA) are commonly used for time series forecasting because they can model linear trends and seasonality. However, these models often have difficulty capturing non-linear and complex temporal dependencies found in meteorological data. Deep learning methods, particularly Long Short-Term Memory (LSTM) networks, have shown better performance in modeling non-linear patterns and long-term dependencies in sequential data. To take advantage of both statistical and deep learning techniques, this study uses a Prophet-LSTM hybrid residual forecasting model. In this model, Prophet extracts the trend and seasonal components of rainfall, while LSTM models the leftover residuals. This hybrid method successfully combines the clarity of traditional time series analysis with the predictive strength of neural networks, creating a strong framework for forecasting rainfall in Cox's Bazar. By merging these complementary techniques, the model aims to improve accuracy by capturing regular seasonal patterns and complex, irregular variations in rainfall.

Precise rainfall prediction is essential for climate-vulnerable areas such as Cox's Bazar, aiding in effective water resource management and disaster readiness. Traditional statistical models, such as SARIMA, are effective at capturing linear and seasonal trends but often struggle with non-linear patterns and long-term dependencies [1],[2]. For example, Naidu and Chandniha [2] applied a SARIMA–Bi-LSTM hybrid for Indian agroclimatic zones, achieving an R^2 of 0.906, RMSE of 41.70 mm, and MAE of 25.93 mm, illustrating the limitations of purely statistical approaches.

Deep learning models, particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, have shown superior performance in capturing complex temporal relationships in rainfall data. Ali et al. [1] compared SARIMA, LSTM, and GRU for Cox's Bazar and reported test R^2 values of 0.72 for LSTM and lower for GRU and SARIMA. Similarly, Joy et al. [3] employed an attention-enhanced LSTM for Bangladesh, achieving an R^2 of 0.964 for rainfall prediction, highlighting the effectiveness of advanced LSTM architectures in capturing long-range dependencies.

Hybrid models integrating deep learning with signal processing or convolutional approaches have further improved predictive performance. Waqas et al. [4] developed a biorthogonal wavelet transform–LSTM model in Thailand, achieving an R^2 of 0.96, while Harilal et al. [5] combined CNN with LSTM for regional rainfall forecasting in the UK, outperforming baseline LSTM and Bidirectional LSTM models in RMSE and MAE metrics. Di Nunno et al. [6]

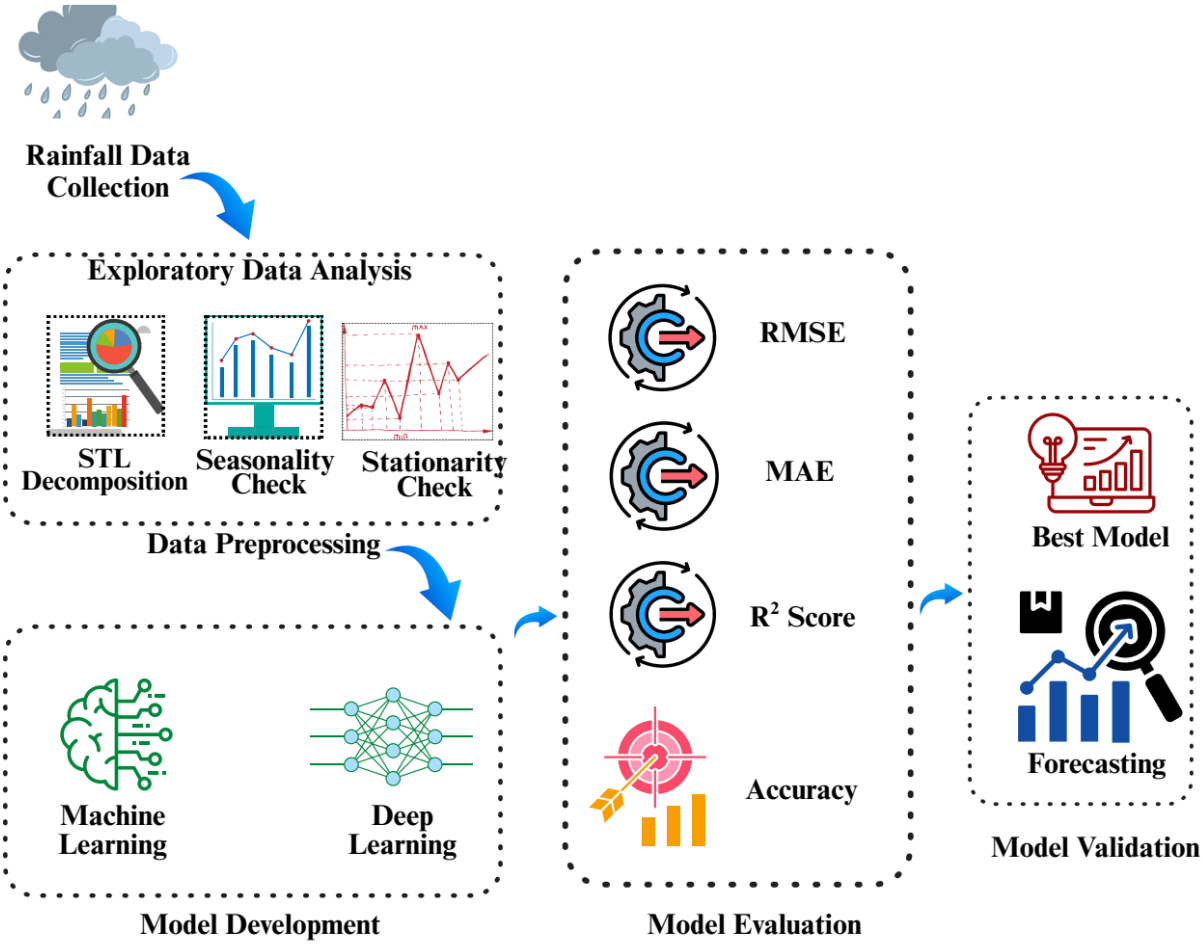


Fig.1. A Prophet–LSTM Hybrid Modeling Framework for Rainfall Forecasting in Cox's Bazar: Integrating Trend–Seasonality Decomposition with Non-Linear Residual Learning for High-Accuracy Prediction.

applied a hybrid M5P–SVR model in northern Bangladesh, with R^2 values up to 0.92. Ali et al. [7] compared Decision Tree and LSTM for Cox's Bazar, reporting a test R^2 of 0.81 for Decision Tree and 0.72 for LSTM.

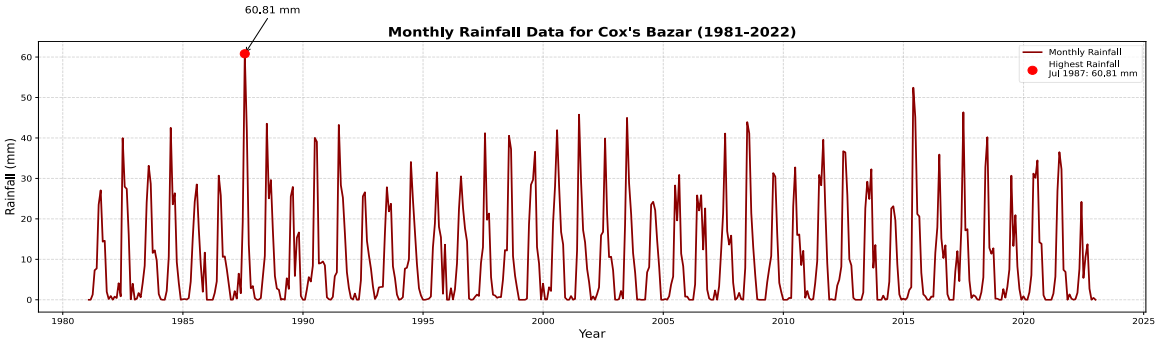
Building on these studies, the present work employs a Prophet–LSTM hybrid model for Cox's Bazar rainfall forecasting. Prophet captures trend and seasonal components, while LSTM models the residual non-linear variations. This approach achieves a train R^2 of 0.990, test R^2 of 0.987, RMSE of 1.404, and MAE of 0.491, outperforming previously reported models in the region. These results indicate that the hybrid model can effectively capture both linear and non-linear temporal dependencies, improving generalization and prediction accuracy for practical applications in water resource management.

Accurate rainfall forecasting for Cox's Bazar is crucial for mitigating floods and droughts, optimizing agricultural practices, and supporting sustainable water resource management. Additionally, it can aid in tourism planning, as the region's economy is heavily reliant on its pristine beaches and natural attractions, which are vulnerable to extreme weather events like heavy rainfall or prolonged droughts. Previous rainfall forecast studies have been conducted for the Sylhet division [6] and the Dhaka division [7]. Earlier works, performed by Hossain et al. with the dataset spanning (1948–2020) years and Ritu et al. spanning (1990–2020) years, have used for Cox's Bazar region.

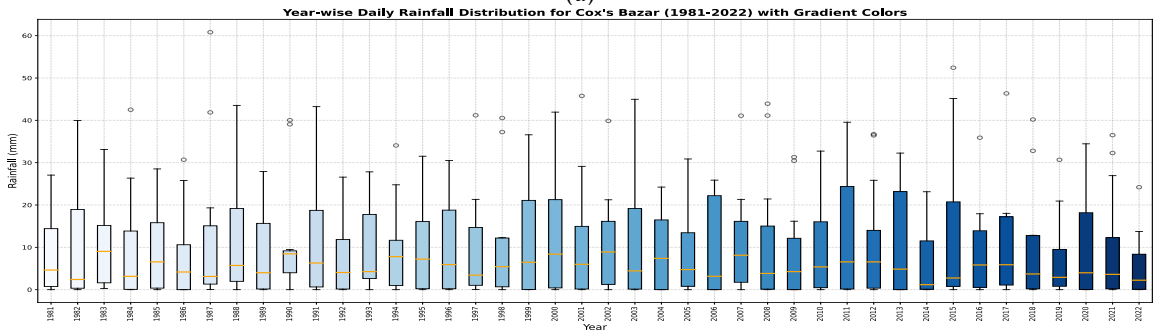
This study presents a new and effective approach to forecasting rainfall in Cox's Bazar, using a carefully curated dataset from the Bangladesh Meteorological Department (BMD), which covers monthly rainfall from 1981 to 2022. By combining statistical methods with deep learning in a hybrid modeling framework, the approach is able to capture both the straightforward and complex patterns in long-term rainfall data. Beyond achieving reliable predictions, this work has clear practical benefits for society, helping with water resource planning, disaster preparedness, and agricultural management in this climate-sensitive region. Careful preprocessing of the historical data ensures that the models are built on accurate and representative information, improving their robustness and ability to generalize. Overall, this research provides a powerful and practical tool that can support informed decision-making and planning in the face of changing rainfall patterns and climate challenges.

II. Methodology

This section explains the step-by-step process for forecasting rainfall in Cox's Bazar using a hybrid Prophet and LSTM model. First, we introduce the study area and dataset, focusing on the monthly rainfall data from 1981 to 2022 that was collected and organized from the Bangladesh

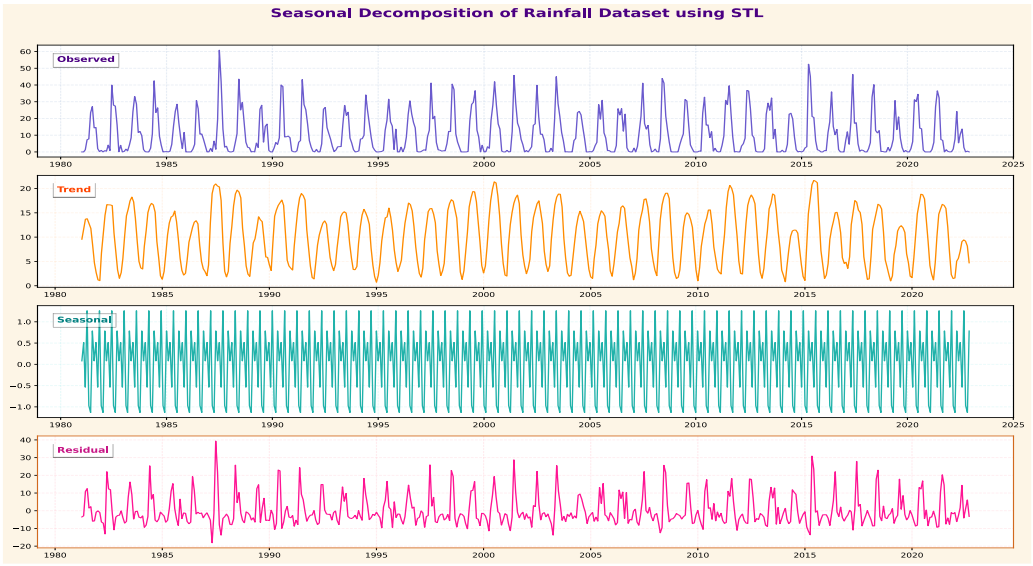


(a)

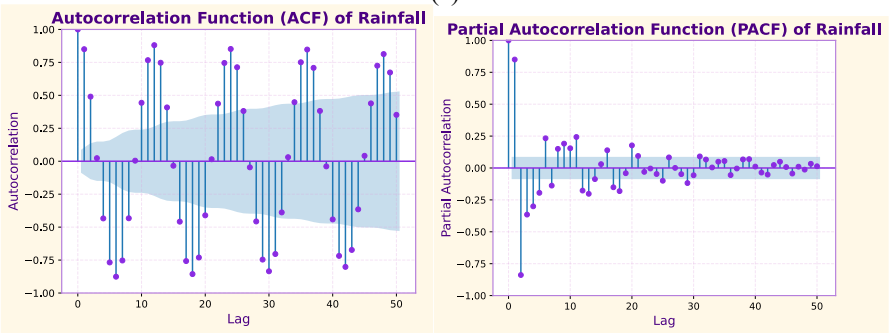


(b)

Fig. 2. Visualization of Monthly Rainfall Data for Cox's Bazar (a) Time-series plot representing the monthly rainfall data from 1981 to 2022, including the highest recorded rainfall in July 1987 (60.81 mm). (b) Box plot highlighting the presence of outliers in the dataset.



(a)



(b)

Fig. 3. Representation of (a) the decomposition of rainfall data into observed, trend, seasonal, and residual components using Seasonal-Trend decomposition using LOESS (STL), (b) the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots.

Meteorological Department (BMD). Next, we perform data preprocessing and cleaning to ensure accuracy and consistency. After that, we conduct Exploratory Data Analysis (EDA) to find trends, seasonality, and patterns in the rainfall data. The methodology then describes how to develop the hybrid forecasting model. Prophet captures linear trends and seasonal components, while LSTM models the leftover non-linear variations and long-term time dependencies. Finally, we use performance metrics such as RMSE, MAE, and R-squared to assess the predictive accuracy and generalization capability of the model. The overall workflow of the methodology is shown in Fig. 1, which provides a diagram summarizing the entire forecasting process.

A. Study Area and Data Preprocessing

The monthly average rainfall data used in this study was obtained from the Bangladesh Meteorological Department (BMD), covering the period from 1981 to 2022. The dataset represents monthly mean daily rainfall in mm/day for Cox's Bazar (1981–2022). The maximum value of 60.81 mm/day in July 1987 corresponds to an extreme monsoonal event. Normalization was applied during training, and all evaluation metrics were reported after inverse transformation. Cox's Bazar, located between 21˚30' N to 21˚50' N latitude and 91˚55' E to 92˚10' E longitude, experiences a tropical monsoon climate, with significant rainfall variability throughout the year. The dataset underwent a thorough preprocessing stage. Fig. 2(a). shows the time series data, while Fig. 2(b) highlights a box plot used to detect outliers. Some outliers were present, including the highest rainfall recorded in July1987 at 60.81mm.

These outliers were retained in the dataset to preserve data integrity, as removing them would result in the loss of valuable information. This experiment was conducted using Python 3.12.0 on an Intel Core i5 10th Gen processor, 8 GB RAM, and Kaggle's TPU VM v3-8 environment.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain insights into the underlying structure and characteristics of the rainfall dataset for Cox's Bazar. First, the Seasonal Decomposition using Seasonal-Trend decomposition using LOESS (STL) was employed to break down the time series into its observed, trend, seasonal, and residual components as shown in Fig. 3(a). The observed data shows distinct yearly peaks, indicating strong seasonality. The trend highlights long-term changes, showing fluctuations in average rainfall with a downward trend towards the end. The seasonal component captures the regular, yearly pattern, remaining fairly consistent over time. Finally, the residual represents random noise or variations not explained by the trend or seasonality, showing irregular fluctuations. The ACF (Autocorrelation Function) plot shows the correlation of the rainfall data with its past values over different time lags. It reveals a gradual decline, indicating that current rainfall is somewhat influenced by previous periods. The PACF (Partial Autocorrelation Function) plot isolates the effect of each lag, showing a significant drop after the first lag, meaning only the first few lags are important for predicting future values. These plots, as shown in Fig. 3(b), identify seasonality and the lag structure help to find parameter for time series models.

C. Model Development

1. Prophet Model

The Prophet model is a decomposable time series model developed by Facebook for forecasting, which explicitly captures trend, seasonality, and holiday effects. The general form of the Prophet model can be represented as:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \quad (1)$$

where:

- g(t) represents the trend component (linear or logistic growth)
- s(t) represents seasonal effects (yearly, weekly, or daily)
- h(t) represents holiday effects
- ε_t is the error term capturing unexplained noise

The Prophet model was implemented using the default configuration with 25 changepoints, enabling automatic detection of trend variations while maintaining generalization.

The residuals from Prophet, representing short-term fluctuations and noise, are computed as:

$$r_t = y_t - \hat{y}_t^{Prophet} \quad (2)$$

where y_t is the observed rainfall at time t and $\hat{y}_t^{Prophet}$ is the predicted value from Prophet.

The LSTM model was trained using only the Prophet residual series as input through a sliding window approach. Time features (month, year) and Prophet's decomposed components were not included, as these effects were already captured by the Prophet model.

2) LSTM Model

The Long Short-Term Memory (LSTM) model is a type of recurrent neural network (RNN) that addresses the vanishing gradient problem, making it well-suited for time series forecasting. It is capable of learning both short-term and long-term dependencies by using special gates: the forget gate, input gate, and output gate, each controlling the flow of information through the memory cells. Equation (3) determines which part of the previous cell information to store in the cell state, Equation (6) updates the cell state based on the forget and input gates and Equation (7) and (8) controls the output, which depends on the updated cell state.

$$f_t = \sigma\,(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

$$i_t = \sigma\,(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (4)$$

$$\hat{C} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (5)$$

$$C_t = f_t * C_{t-1} + i_t * \hat{C} \quad (6)$$

$$o_t = \sigma\,(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (7)$$

$$h_t = o_t * tanh(C_t) \quad (8)$$

In this architecture, σ is the sigmoid function, tanh is thehyperbolic tangent function, W's are the weight metrics, b's are the bias terms, x_t is the input at time t, and h_t is the hidden state output [8].

TABLE I. HYPERPARAMETER'S VALUES FOR PROPHET–LSTM HYBRID MODEL

Hyperparameters	*Value*
Normalization Technique	MinMax Scalar
LSTM (Neurons)	64
Dropout Rate	0.3
Activation Function	ReLU
Optimizer	Adam
Epochs	60
Batch Size	64
Train-Test Split Ratio	80-20

The hyperparameters utilized in the Prophet–LSTM hybrid model are summarized in Table I. The hyperparameters were selected using a manual, validation-based tuning approach. Initial values were chosen based on prior studies and standard practices in time series forecasting, followed by iterative adjustment to minimize validation RMSE. Key parameters such as the number of LSTM neurons, dropout rate, batch size, and number of epochs were tuned through repeated experiments, and the final configuration was selected based on stable convergence and consistent generalization performance. Specifically, 64 neurons provided a balance between model capacity and overfitting, while a dropout rate of 0.3 effectively regularized the network without degrading learning performance. This table provides a clear overview of the normalization technique, model architecture, optimizer settings, and training configurations employed in the analysis. Each parameter is specified to ensure transparency and reproducibility of the model development process.

The Prophet model trained within seconds, while the LSTM converged within a few minutes over 60 epochs on an Intel Core i5 (10th Gen) system with 8 GB RAM. Prophet scales linearly with data size, whereas LSTM complexity depends on the number of epochs and hidden units.

D. Evaluation Metrics

To evaluate the accuracy of the models, three key metrics were used: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) Score. Equation (9), (10) and (11) showcasing the formula for RMSE, MAE and R-squared (R^2) Score.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (9)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (10)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (11)$$

where y_i is the actual value, y_i is the predicted value, n is the number of observations and $\bar{y}_i$ is the mean of the actual values [8].

III. RESULT AND DISCUSSION

A. Data Evaluation

The initial step in analyzing the dataset involves a thorough exploration of the monthly rainfall patterns over the years.

TABLE II. MODEL EVALUATION METRICS

Train R^2	0.990
Test R^2	0.987
RMSE	1.404
MAE	0.491

B. Forecasting Results from Prophet–LSTM hybrid model.

The forecasting performance of the hybrid Prophet–LSTM model by comparing the actual train–test rainfall data with the corresponding predicted values is demonstrated in figure (4). The predicted training series (green) demonstrates a close alignment with the observed seasonal behavior, indicating that the hybrid approach effectively captures both long-term trends and recurrent rainfall cycles. Similarly, the predicted test series (red) exhibits a strong correspondence with the actual test data, successfully reproducing the magnitude and timing of peak rainfall events. This performance reflects the complementary strengths of the model components: Prophet's ability to model trend and seasonality, combined with LSTM's capacity to learn complex, non-linear temporal dependencies. Overall, the results emphasize the robustness and accuracy of the Prophet–LSTM hybrid model for rainfall forecasting, particularly in contexts characterized by pronounced variability and seasonality.

C. Evaluation Metrics

Table II presents the performance metrics of the proposed hybrid Prophet–LSTM model on both the training and testing datasets, evaluated using RMSE, MAE, and R^2 Score, along with the percentage difference between train and test errors. The model attains a training R^2 Score of 0.990, indicating an excellent fit to the historical rainfall data. Its performance remains highly consistent on the test dataset, where the R^2 Score slightly decreases to 0.987, corresponding to a very small percentage difference of only 0.30%, which reflects strong generalization capability. Furthermore, the test RMSE of 1.404 and test MAE of 0.491 demonstrate the model's ability to achieve low prediction errors while effectively capturing both seasonal patterns and short-term variability in the rainfall series. Overall, the results confirm that the hybrid Prophet–LSTM model provides robust, stable, and accurate forecasting performance across all evaluated metrics.

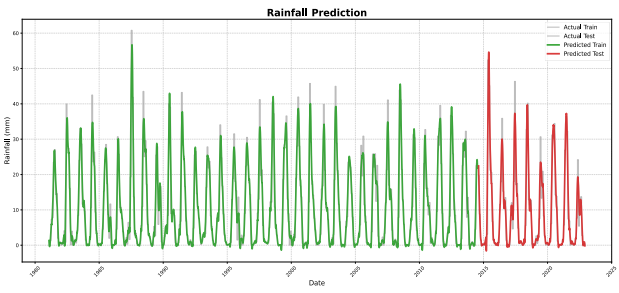


Fig. 4. Time series plots comparing test data with predicted values for Prophet–LSTM hybrid model.

D. Comparison with Existing Hybrid Models.

TABLE III. COMPARISON WITH REPRESENTATIVE HYBRID RAINFALL FORECASTING MODELS

Study	Region	Model	Performance
[2]	Chhattisgarh, India	SARIMA–Bi-LSTM	RMSE = 41.70 mm; R^2 = 0.906
[4]	Thailand	Wavelet–LSTM	R^2 = 0.96
[5]	United Kingdom	CNN–LSTM	RMSE = 0.107
[6]	Northern Bangladesh	M5P–SVR	R^2 = 0.92
Proposed Study	Cox's Bazar, Bangladesh	Prophet–LSTM	RMSE = 1.404; R^2 = 0.987

Several hybrid rainfall forecasting approaches have been reported in the literature, including SARIMA–Bi-LSTM, Wavelet–LSTM, CNN–LSTM, and M5P–SVR, applied across different climatic regions and forecasting scales. These methods combine statistical, signal-processing, and machine learning techniques to capture both linear and non-linear rainfall characteristics. The reported performance metrics are included for contextual reference only, as differences in datasets, regions, and temporal resolution limit direct numerical comparison. Within this context, the proposed Prophet–LSTM model demonstrates competitive performance for monthly rainfall forecasting in a tropical monsoon region.

IV. CONCLUSION

A study is outlined here that involved a hybrid Prophet-LSTM method of monthly rainfall forecasting in the climate-sensitive region of Cox's Bazar, Bangladesh. The method proposed here combines Prophet's transparent trend and seasonality breakdown with LSTM's residual learning, thus being able to uncover both the linear and non-linear aspects of rainfall. The results illustrate that fusing deep learning and statistical time-series modeling gives a great boost in the reliability of the monsoon-driven rainfall predictions.

It is also emphasized that the suggested framework fits well with flood risk management, water resources management, and climate-friendly planning in the coastal areas. Subsequent research is expected to carry out several tasks, including the extension of the current model to multi-step forecasting, the inclusion of further climate indicators, and the assessment of the method in other areas at risk, so as to explore its capacity and reliability.

REFERENCES

[1] M. E. Ali et al., "A Comparative Analysis of Rainfall Forecasting in Cox's Bazar using SARIMA, LSTM and GRU," pp. 2671–2676, Dec. 2024, doi: https://doi.org/10.1109/iccit64611.2024.11021778.

[2] D. Naidu and S. K. Chandniha, "Hybrid SARIMA–Bi-LSTM model for monthly rainfall forecasting in the agroclimatic zones of Chhattisgarh," Journal of Agrometeorology, vol. 27, no. 3, pp. 332–337, Sep. 2025, doi: https://doi.org/10.54386/jam.v27i3.3010.

[3] U. G. Joy, S. Kabir, and Tasnim Niger, "Attention-enhanced LSTM modeling for improved temperature and rainfall forecasting in Bangladesh," Theoretical and Applied Climatology, vol. 156, no. 11, Oct. 2025, doi: https://doi.org/10.1007/s00704-025-05858-5.

[4] Muhammad Waqas, U. W. Humphries, P. T. Hlaing, Angkool Wangwongchai, and Porntip Dechpichai, "Advancements in daily precipitation forecasting: A deep dive into daily precipitation forecasting hybrid methods in the Tropical Climate of Thailand," MethodsX, vol. 12, pp. 102757–102757, May 2024, doi: https://doi.org/10.1016/j.mex.2024.102757.

[5] Geethu Thottungal Harilal, A. Dixit, and G. Quattrone, "Establishing hybrid deep learning models for regional daily rainfall time series forecasting in the United Kingdom," Engineering applications of artificial intelligence, vol. 133, pp. 108581–108581, Jul. 2024, doi: https://doi.org/10.1016/j.engappai.2024.108581.

[6] F. Di Nunno, F. Granata, Q. B. Pham, and G. de Marinis, "Precipitation Forecasting in Northern Bangladesh Using a Hybrid Machine Learning Model," *Sustainability*, vol. 14, no. 5, p. 2663, Feb. 2022, doi: https://doi.org/10.3390/su14052663.

[7] S. Panna, M. Ali, M. I. Bhuiyan, A. Chanda Suhag, B. Dey, and F. Karim, "Evaluating Machine Learning and Deep Learning Models for Rainfall Prediction in the Southeastern Region of Bangladesh," Zenodo, 2025, doi: 10.5281/zenodo.15556465.

[8] K. E. ArunKumar, D. V. Kalaga, Ch. Mohan Sai Kumar, M. Kawaji, and T. M. Brenza, "Comparative analysis of Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) cells, Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA) for forecasting COVID-19 trends," Alexandria Engineering Journal, Jan. 2022, doi: https://doi.org/10.1016/j.aej.2022.01.011.

Cardiovascular Disease Prediction Using Modern Machine Learning Pipelines on the UCI Heart Disease Dataset

Shah Farhan Sadik
Dept. of Computer Science and Engineering
Teesta University, Rangpur -5404, Bangladesh.
shahfarhansadik@gmail.com

Pranto Halder
Dept. of Computer Science and Engineering
Teesta University, Rangpur—5404, Bangladesh.
ORCID: **0000-0003-4786-9208**

***Abstract*— Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, accounting for nearly one-third of all global deaths. Early and accurate prediction of CVD risk is therefore critical for timely intervention and effective prevention. Advances in machine learning (ML) have enabled data-driven risk prediction models that can learn complex relationships between clinical variables and patient outcomes. This paper revisits cardiovascular disease prediction on the well-known UCI Cleveland heart disease dataset using a modernized machine learning pipeline. The workflow includes data pre-processing, feature analysis, supervised classification, and evaluation using confusion-matrix–based metrics. Four classical yet competitive classifiers—Decision Tree, Support Vector Machine (SVM), Random Forest, and AdaBoost—are implemented and compared. Experimental results on the Cleveland subset show that ensemble methods outperform single classifiers: AdaBoost achieves the highest accuracy of 98.87%, followed by Random Forest (97.77%), SVM (96.77%), and Decision Tree (93.54%).**

Keywords—cardiovascular disease, heart disease prediction, machine learning, ensemble learning, UCI Cleveland dataset, AdaBoost, Random Forest.

I. Introduction

Cardiovascular diseases (CVDs)—including coronary artery disease, cerebrovascular disease, and heart failure—are the leading cause of mortality worldwide. The World Health Organization (WHO) estimates that CVDs account for about 17.9 million deaths annually, representing roughly 32% of all global deaths; more than one third of these deaths occur prematurely in people under 70 years of age [1]. The American Heart Association (AHA) likewise reports that CVD continues to be a dominant cause of death and disability and highlights the growing burden of risk factors such as hypertension, obesity, diabetes, and sedentary lifestyle [2]. Traditional risk scoring frameworks, such as those derived from the Framingham Heart Study, use a limited set of clinical variables and assume mostly linear relationships. While these scores are simple, interpretable, and clinically useful, they may not fully exploit the richness of modern electronic health records (EHRs) and registry data. In contrast, machine learning (ML) models can capture nonlinear interactions among dozens or hundreds of risk factors and can be optimized directly for predictive performance. Recent systematic reviews and meta-analyses show that ML-based CVD prediction models trained on EHR data often outperform conventional risk scores when developed and validated appropriately [3], [4]. Within this broader trend, a large body of work has explored ML for heart disease prediction on structured clinical datasets. Hybrid and ensemble approaches, such as the hybrid Random Forest–linear model (HRFLM), have achieved high accuracy on benchmark datasets such as the UCI Cleveland heart disease dataset [5], [7]. Deep learning models, including optimally configured deep neural networks [6] and deep decision-support systems [8], have further advanced performance on larger datasets. However, many real-world applications—particularly in low-resource settings—still rely on relatively small, tabular datasets where classical ML algorithms remain competitive, interpretable, and easy to deploy. In this work, authors focus on the Cleveland subset of the UCI heart disease dataset, a long-standing benchmark in the ML community [9]. Authors propose a modernized ML pipeline that includes transparent data handling, a comparison of four supervised classifiers, and evaluation using confusion-matrix–based metrics [10], [11].

The objective is to situate a classical Machine Learning (ML) approach for cardiovascular disease (CVD) risk prediction using the Cleveland dataset against the backdrop of recent Deep Learning (DL) studies. Authors will present a clear, modular ML pipeline tailored for small, tabular medical datasets, focusing on reproducibility and interpretability. This pipeline will be used to empirically compare the performance of Decision Tree, Support Vector Machine (SVM), Random Forest, and AdaBoost classifiers on the Cleveland data, demonstrating that ensemble methods (Random Forest and AdaBoost) offer superior performance. Finally, authors will discuss the extensibility of this lightweight pipeline for use with larger datasets, more advanced models, and modern Explainable AI (XAI) techniques, providing a baseline for future work.

The rest of this paper is organized as follows. Section II reviews related work. Section III describes the dataset and methods. Section IV presents the experimental setup and results. Section V discusses the findings and limitations. Section VI concludes and outlines future directions.

II. Literature Review

A. Early Data-Mining Approaches

Early research on automated heart disease prediction applied classical data-mining methods to hospital datasets. Palaniappan and Awang proposed an Intelligent Heart Disease Prediction System (IHDPS) using Decision Trees, Naïve Bayes, and Neural Networks, showing that standard classifiers can derive clinically relevant patterns from routine patient data and support "what-if" decision queries [12]. Subsequent studies using decision trees, association rules, fuzzy logic, and related approaches helped establish heart disease prediction as a key medical data-mining task, but most were limited by small datasets and simplified validation, and they predated current evaluation standards [13].

B. Hybrid and Ensemble Machine Learning Models

Ensemble and hybrid models are strong baselines for heart disease prediction. Mohan et al. proposed a hybrid Random Forest–linear model on the UCI Cleveland dataset, showing improved generalization through tree-based feature selection plus linear modeling [5]. Bhatt et al. also found that ensemble methods consistently performed best across multiple algorithms on a large cardiovascular dataset [7]. Feature selection further improves performance and interpretability; Relief- and LASSO-based approaches highlight a compact set of influential attributes (e.g., chest pain type, ST depression, number of major vessels), which are repeatedly identified as important across studies [7], [9], [11].

C. Deep Learning and EHR-Scale Models

Deep learning has been increasingly used for cardiovascular risk prediction. Ali et al. combined χ^2-based feature selection with an optimized deep neural network and reported strong results for heart disease prediction [6], while Almazroi et al. evaluated dense neural-

network configurations as a decision-support baseline for structured clinical data [8]. Recent studies further compare broad sets of ML and deep learning models for heart disease prediction [9], [14], [17]. At the population level, systematic reviews and meta-analyses conclude that ML models can outperform traditional risk scores when rigorously validated, with boosting methods and SVMs often achieving high AUC in coronary artery disease and stroke prediction [3], [4]. These findings also highlight persistent challenges, including heterogeneity across studies, overfitting risk, and the need for transparent reporting [15].

D. Positioning of This Work

In contrast to large-scale deep learning and EHR-based studies, this work evaluates a lightweight, classical ML pipeline on the UCI Cleveland heart disease benchmark dataset [5], [7], [9], [13]. Rather than proposing a new state-of-the-art architecture, authors aim to (i) establish a clean and reproducible baseline using widely adopted classifiers, (ii) demonstrate the performance advantage of ensemble methods, and (iii) position this approach within the broader ML-for-CVD literature. This design is well suited to resource-constrained settings and educational use, and it provides a practical starting point before scaling to larger datasets and more complex models.

III. Methodology

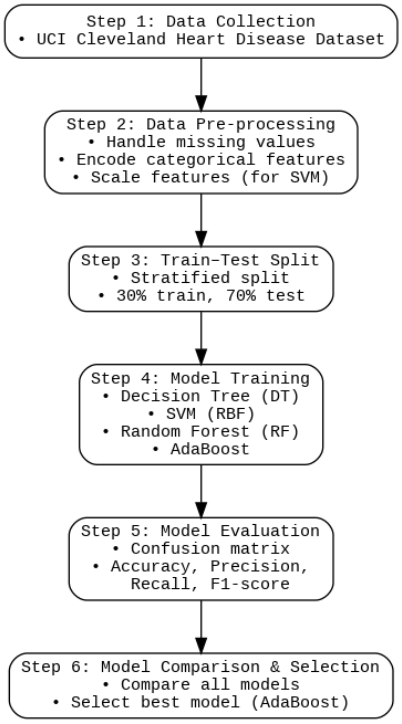


Figure 1. Proposed methodology for cardiovascular disease prediction using machine learning on the UCI Cleveland heart disease dataset.

Figure 1 outlines the proposed ML workflow for CVD prediction using the UCI Cleveland dataset, covering data acquisition, preprocessing (missing-value handling, categorical encoding, and feature scaling as required), a stratified 70/30 test–train split, training of four classifiers (Decision Tree, RBF-SVM, Random Forest, and AdaBoost), evaluation via confusion-matrix–based metrics (accuracy, precision, recall, and F1-score), and final model comparison, where AdaBoost is selected as the best-performing approach for this task.

A. Dataset Description

The pre-processing pipeline employed in this study is kept simple and reproducible, starting with target binarization to map the prediction outcome to a binary label [13], [17]. To address missing data, records with missing critical attributes are handled primarily through simple imputation using median values for continuous features, or are removed if the missingness is extensive, which is often sufficient for small datasets like Cleveland. Next, categorical features (such as chest pain type and thalassemia) are encoded as integers to ensure compatibility with the selected tree-based and kernel methods [18]. Feature scaling is applied by standardizing continuous attributes to zero mean and unit variance specifically for the Support Vector Machine (SVM) model to prevent scale-sensitive bias, though tree-based models are robust to this step. Finally, an atypical but intentional train–test split is used, reserving 70% of the data (212 samples) for testing and 30% for training, which is a stratified split designed to align with a legacy experimental setup, though cross-validation is acknowledged as the preferred future approach. This entire pipeline is designed for ease of implementation using standard ML libraries, establishing a transparent baseline for subsequent analysis [19].

C. Classification Algorithms

We evaluate four supervised learning algorithms: Decision Tree, SVM, Random Forest, and AdaBoost.

1) Decision Tree (DT): A decision tree recursively partitions the feature space by selecting split points that maximize an impurity reduction criterion [20]. For a node with class probabilities p_i, the entropy is given by:

$$H = - \Sigma_{i=1}^{k} p_i \log_2(p_i) \text{.....(i)}$$

In equation (i) and (ii), If a split on attribute A partitions the set S into subsets S_v, the information gain is [21]:

$$IG(S, A) = H(S) - \Sigma_{v} (|S_v| / |S|) H(S_v) \text{.....(ii)}$$

2) Support Vector Machine (SVM): SVM constructs a maximum-margin hyperplane in a high-dimensional feature space using a kernel function $K(\cdot,\cdot)$. Given training samples (x_i, y_i) with $y_i \in \{-1, +1\}$, the decision function can be written as equation (iii) [22]:

$$f(x) = \Sigma_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \text{.....(iii)}$$

where α_i are Lagrange multipliers and b is the bias term. The predicted label is sign(f(x)). Authors use a radial basis function (RBF) kernel with hyperparameters tuned on the training set [4].

3) Random Forest (RF): Random Forest is an ensemble of decision trees built on bootstrapped samples of the data with random feature subsets considered at each split [5], [23]. Each tree $h_t(x)$ outputs a class prediction, and the final prediction is the majority vote across T trees, the mathematical equation shown in equation (iv):

$$H_{RF}(x) = \text{mode}\{ h_t(x) : t = 1, 2, \ldots, T \} \text{.....(iv)}$$

4) AdaBoost: AdaBoost builds a sequence of weak learners (often shallow trees). At iteration t, the algorithm fits a weak learner $h_t(x)$ on a weighted training set, yielding an error ε_t. The weight of the learner is shown in equation (v) and equation (iv) [24]:

$$\alpha_t = 0.5 * \ln((1 - \varepsilon_t) / \varepsilon_t) \text{.....(v)}$$

The sample weights are then updated and normalized. The final strong classifier is a weighted majority vote [25]:

$$H_{AB}(x) = \text{sign}(\Sigma_{t=1}^{T} \alpha_t h_t(x)) \text{ ------(vi)}$$

E. Evaluation Metrics

Authors evaluate models using the confusion matrix on the test set. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively [26].

The standard performance metrics are defined in equation (vii), (viii), (ix), and (x) :
Accuracy = (TP + TN) / (TP + TN + FP + FN) --------(vii)
Precision = TP / (TP + FP) ----------(viii)
Recall = TP / (TP + FN) ----------(ix)
F1-score = 2 * (Precision * Recall) / (Precision + Recall) ----(x).

In medical applications, recall for the disease class is particularly important, because false negatives correspond to missed diagnoses. Recent ML-for-CVD studies also emphasize AUC-ROC and calibration curves, which are natural extensions for future work [3], [4], [7], [23].

IV. Results and Evaaluation

A. Experimental Setup

All models are implemented in a standard ML framework and trained on the 30% training split (90 samples). Hyperparameters are selected using simple tuning on the training set. The remaining 70% (212 samples) are held out for final evaluation. The Decision Tree uses a moderate maximum depth to reduce overfitting. The Random Forest consists of multiple trees; each trained on a bootstrap sample with random feature selection at each node. The AdaBoost classifier uses shallow decision trees as weak learners and iteratively reweights samples. The SVM uses an RBF kernel; the penalty parameter C and kernel width γ are tuned on the training set. This setup mirrors many comparative studies on heart disease datasets [5], [25].

B. Overall Accuracy

Table I shows the test accuracy for each classifier on the Cleveland dataset.

Algorithm	Accuracy (%)
Decision Tree	93.54
SVM	96.77
Random Forest	97.77
AdaBoost	98.87

"AdaBoost achieves the highest test accuracy (98.87%) because it combines multiple weak learners into a strong classifier while iteratively reweighting misclassified samples, forcing subsequent learners to focus on harder cases [24], [25].

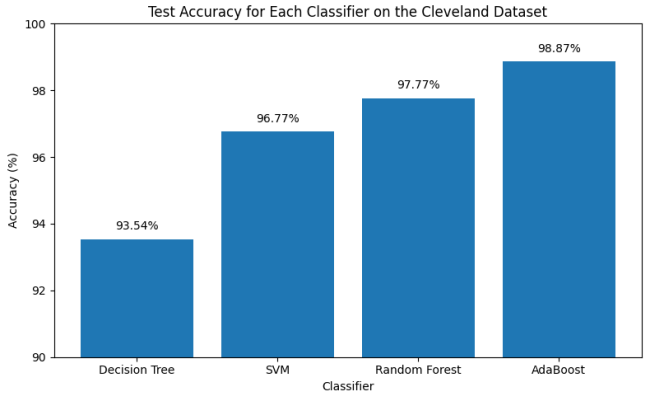


Figure 2. Test accuracy of four classifiers (Decision Tree, SVM, Random Forest, and AdaBoost) on the UCI Cleveland heart disease dataset.

Both ensemble methods outperform the individual classifiers: Random Forest attains 97.77% accuracy and AdaBoost yields the best result at 98.87%. In Figure 2, this trend is clearly visible, with the bars for Random Forest and AdaBoost noticeably higher than those for the Decision Tree and SVM. The small but consistent gains of the ensemble models, particularly AdaBoost, highlight the effectiveness of boosting and bagging strategies for this tabular clinical dataset [25].

C. Class-wise Performance for AdaBoost

Table II reports precision, recall, and F1-score for the AdaBoost classifier.

Class	Precision	Recall	F1-score	Support
0 – No disease	0.96	0.97	0.96	15
1 – Disease	0.97	0.98	0.98	16

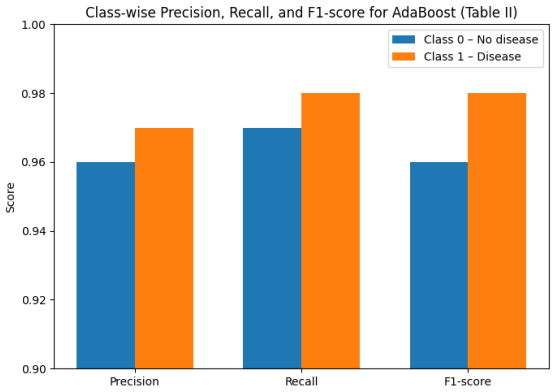


Figure 3. Class-wise precision, recall, and F1-score of the AdaBoost classifier on the UCI Cleveland heart disease dataset.

Table II reports the class-wise precision, recall, F1-score, and support for the AdaBoost classifier, while Figure 3 visualizes these metrics for the "no disease" (class 0) and "disease" (class 1) groups. All three metrics are high for both classes (precision and recall ≥ 0.96), indicating that AdaBoost performs reliably across positive and negative cases. The scores for the disease class are slightly higher (precision 0.97, recall 0.98, F1-score 0.98) than for the no-disease class (precision 0.96, recall 0.97, F1-score 0.96), suggesting that the model is particularly effective at correctly identifying patients with heart disease while still maintaining strong performance on non-disease cases [26].

D. Interpretation of Results

Ensemble models (Random Forest and AdaBoost) outperformed Decision Tree and SVM on the Cleveland dataset, consistent with prior evidence for tree-based ensembles and boosting in heart disease prediction [4], [5], [7], [9], [26]. AdaBoost's slight edge may reflect the benefit of emphasizing difficult cases in small samples. However, accuracies above 97% should be interpreted cautiously given the limited sample size (303 records), sensitivity to the train–test split, and the lack of external validation, which may inflate performance estimates [3], [22].

B. **Comparison with Recent Literature**

Authors results compare favorably with reported performance on similar datasets. Hybrid and ensemble models such as HRFLM [5], optimization-enhanced ensembles [7], and feature-selection-based ensembles [11] often report accuracy in the 90–95% range on heart disease datasets. Deep neural network approaches trained on larger datasets can yield even higher performance, especially when combined with feature selection or stacked ensembles [6], [8], [9]. However, state-of-the-art performance increasingly depends on larger, multi-center EHR datasets; sophisticated validation schemes; and robust handling of missing data and class imbalance [3], [4], [7], [9], [26].

C. **Modernized Pipeline and Algorithmic Pattern**

This work emphasizes an algorithmic pattern for small tabular medical datasets: (1) start with simple, interpretable baselines (Decision Tree, logistic regression, SVM); (2) add ensemble methods (Random Forest, AdaBoost, gradient boosting) to capture nonlinear patterns and interactions [4], [5], [7], [9]; (3) incorporate feature selection or regularization to improve generalization and interpretability [11]; (4) use rigorous evaluation (preferably k-fold

cross-validation and, when possible, external validation) [3], [4]; and (5) integrate explainable AI tools (e.g., feature importance, SHAP values) to provide clinicians with insights into model decisions [6], [27].

V. **Limitations**

Several limitations should be noted. The UCI Cleveland dataset includes only 303 single-center records, which constrains generalizability and highlights the need for larger, multi-center cohorts for reliable deployment [3], [7], [9]. In addition, the fixed 70/30 train–test split may yield unstable performance estimates; stratified cross-validation and external validation would provide more robust evaluation [3], [4]. The available features are also limited relative to contemporary EHR settings that include high-dimensional and longitudinal information [3], [9]. Finally, this study did not examine calibration or fairness across subgroups (e.g., sex or age), which are increasingly critical considerations for clinical AI systems [3], [4].

V. CONCLUSION AND FUTURE WORK

This study presented a modernized machine-learning pipeline for cardiovascular disease prediction using the UCI Cleveland dataset. Using a consistent preprocessing and evaluation framework, authors compared Decision Tree, SVM, Random Forest, and AdaBoost models and found that ensemble methods performed best, with AdaBoost achieving 98.87% accuracy, consistent with prior evidence for structured clinical data [4], [5], [7], [9], [11]. However, given recommendations from large-scale studies and meta-analyses regarding rigorous validation, external cohorts, and careful treatment of bias and calibration [3], [4], these results should be interpreted as benchmarking and educational evidence rather than a clinical solution. Future work will implement stratified k-fold cross-validation and external validation, extend the pipeline to larger EHR datasets while evaluating gradient boosting and tabular deep-learning models [6], [8], [9], improve imputation, class-imbalance handling, and hyperparameter optimization [3], [7], [11], and integrate explainable AI to support clinically meaningful interpretation and human–AI collaboration in CVD risk assessment [6], [28].

REFERENCES

[1] World Health Organization, “cardiovascular diseases (CVDs),” Fact sheet, 2025.
[2] American Heart Association, “heart disease and stroke statistics – 2025 update,” 2025.
[3] T. Liu et al., “Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis,” Eur. Heart J. Digit. Health, vol. 6, no. 1, pp. 7–22, 2025.
[4] C. Krittanawong et al., “Machine learning prediction in cardiovascular diseases: a meta-analysis,” Sci. Rep., vol. 10, art. 16057, 2020.
[5] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” IEEE Access, vol. 7, pp. 81542–81554, 2019.
[6] L. Ali et al., “An automated diagnostic system for heart disease prediction based on chi-square statistical model and optimally configured deep neural network,” IEEE Access, vol. 7, pp. 34938–34945, 2019.
[7] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, “Effective heart disease prediction using machine learning techniques,” Algorithms, vol. 16, no. 2, art. 88, 2023.
[8] A. A. Almazroi, E. A. Aldhahri, S. Bashir, and S. Ashfaq, “A clinical decision support system for heart disease prediction using deep learning,” IEEE Access, vol. 11, pp. 61646–61659, 2023.
[9] D. Rohan et al., “An extensive experimental analysis for heart disease prediction using artificial intelligence techniques,” Sci. Rep., vol. 15, art. 6132, 2025.
[10] R. Deepa, S. A. Kumar, and V. Rajesh, “Early prediction of cardiovascular disease using machine learning: unveiling risk factors from health records,” AIP Adv., vol. 14, no. 3, 035049, 2024.
[11] P. Ghosh et al., “Efficient prediction of cardiovascular disease using machine learning algorithms with Relief and LASSO feature selection techniques,” IEEE Access, vol. 9, pp. 19304–19326, 2021.
[12] S. Palaniappan and R. Awang, “Intelligent heart disease prediction system using data mining techniques,” in Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., 2008, pp. 108–115.
[13] D. Dua and C. Graff, “UCI Machine Learning Repository: Heart Disease Data Set,” Univ. California, Irvine, 2019.
[14] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, “Can machine-learning improve cardiovascular risk prediction using routine clinical data?” *PLOS ONE*, vol. 12, no. 4, Art. no. e0174944, 2017, doi: 10.1371/journal.pone.0174944.
[15] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement,” *BMJ*, vol. 350, p. g7594, 2015, doi: 10.1136/bmj.g7594.
[16] R. F. Wolff *et al*., “PROBAST: A tool to assess the risk of bias and applicability of prediction model studies,” *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51–58, 2019, doi: 10.7326/M18-1376.
[17] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.
[18] B. Van Calster *et al*., “Calibration: The Achilles heel of predictive analytics,” *BMC Medicine*, vol. 17, no. 1, Art. no. 230, 2019, doi: 10.1186/s12916-019-1466-7.
[19] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[20] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.
[21] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[22] M. T. Ribeiro, S. Singh, and C. Guestrin, “ ‘Why should I trust you?’: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[23] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.
[24] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.
[25] S. O. Arik and T. Pfister, “TabNet: Attentive interpretable tabular learning,” *arXiv preprint* arXiv:1908.07442, 2019.
[26] A. Rajkomar *et al*., “Scalable and accurate deep learning with electronic health records,” *npj Digital Medicine*, vol. 1, Art. no. 18, 2018, doi: 10.1038/s41746-018-0029-1.
[27] X. Liu *et al*., “Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension,” *Nature Medicine*, vol. 26, no. 9, pp. 1364–1374, 2020, doi: 10.1038/s41591-020-1034-x.
[28] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys*, vol. 54, no. 6, pp. 115:1–115:35, 2022, doi: 10.1145/3457607.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

PneuXNet: An Interpretable CNN Framework for High-Accuracy Pneumonia Detection Using Multi-Scale and Adaptive Attention

Mehedi Hasan Meraj
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
meraj23105101318@diu.edu.bd

Mahmudul Hasan
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
hasan23105101093@diu.edu.bd

Mashfiquzzaman Tajbid
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
tajbid23105101352@diu.edu.bd

Nafisa Tasneem Esha
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
esha23105101084@diu.edu.bd

Md Sabbir Hosen Hamim
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
hamim2305101803@diu.edu.bd

***Abstract*—Standard Convolutional Neural Networks for detecting pneumonia exhibit significant limitations: they fail to capture subtle, multi-scale pathological features, and they dilute spatial information through average pooling while lacking clear interpretability. This work proposes a novel architecture called PneuXNet that bridges these gaps. The proposed architecture leverages a Multi-Scale Atrous Attention module for capturing fine-grained features and a Class-Aware Adaptive Pooling head for preserving them. Further, we validate the choices made by our model using Grad-CAM and hence bridge the gap in interpretability as well. Our model outperforms the baseline with a respectable accuracy of 98.47% on the held-out test set, illustrating superior and explainable performance by solving these key limitations.**

***Index Terms*—Pneumonia Detection, Attention Mechanism, Convolutional Neural Network (CNN), Explainable AI (XAI), Medical Image Analysis.**

I. Introduction

Pneumonia is a potentially fatal respiratory infection, which is still responsible for high mortality rates worldwide, especially among children and severely ill patients [1], [2]. Despite this, the diagnosis of pneumonia continues to be beset with challenges; clinical, laboratory, and radiological criteria can often not be relied upon to differentiate between bacterial and viral aetiologies [1]. Although chest-X-ray (CXR) is widely used as the first line of diagnosis, its analysis is actually complex [3]. Thus, Deep Learning (DL) has emerged as an "evolving technology" for the automation of CXR analysis; many studies benchmark standard architectures like ResNet and VGG [4], [5].

Our work starts with the formal problem formulation. Given a training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{H\times W\times C}$ is a preprocessed CXR image and $y_i \in \{0, 1\}$ is the corresponding binary label ('Normal' or 'Pneumonia'), our goal is to learn a mapping $f_\theta : x \to y$. We seek the optimal parameters θ^* that minimize the weighted binary cross-entropy loss, $\mathcal{L}$, in order to overcome the severe class imbalance of the dataset as illustrated in Fig. 2.

This paper proposes **PneuXNet**, a novel architecture designed to solve specific gaps in the literature. Our contributions are as follows:

- To address the gap of how a model can learn to detect subtle and multi-scale pathological patterns missed by standard kernels, we introduce the **Multi-Scale Atrous Attention (MSAA)** module (Sec. III-C1) to capture a rich, multi-scale context.
- To address the gap of how a model can preserve critical information on localized features and prevent its dilution by global average pooling, we propose the **Class-Aware Adaptive Pooling (CAAP)** head (Sec. III-C2) to create an adaptive feature vector that preserves salient evidence.
- To address the "black box" problem and make the decision of the model transparent and trustworthy [6], we validate our model's interpretability using Gradient-weighted Class Activation Mapping (Grad-CAM) (Sec. III-E), responding to the critical need for explainable AI (XAI) in clinical tools [6].

The rest of this paper presents some related works(Sec. II), details our methodology (Sec. III), presents our results (Sec. IV) with a proper discussion (Sec. V and concludes (Sec. VI).

II. Related Work

Recent literature on pneumonia detection from chest X-rays (CXR) has evolved from benchmarking of pre-trained Convolutional Neural Networks (CNN) [5], [7] to more specialized architectures. While standard models such as EfficientNetV2L

and modified AlexNet provide strong baselines [7], [8], there is a clear identified gap in the research.

Recent advanced models point to two main weaknesses of standard CNNs: 1) the lack of global context capturing and key spatial relationships, for which some have even abandoned CNNs and moved to Vision Transformers (ViT) [9], and 2) "lack of sensitivity in distinguishing subtle yet critical pathological patterns" [10]. The latter gap is usually filled by computationally expensive ensembles of multiple backbones combined with complex attention mechanisms [10]. Other works have focused on methodological challenges such as multi-class classification [11] or data privacy using federated learning [12].

This points out a specific need for a new single-CNN architecture with the ability to efficiently model subtle multi-scale features, as well as preserve critical spatial information via the output classification head, in order to overcome the limitations of both standard CNNs and the complex ensembles proposed to fix them.

III. Methodology

To address aforementioned research gap, we propose **PneuXNet**, a novel deep CNN built upon a DenseNet201 backbone. Our architecture introduces two key innovations: a MSAA module for enhanced multi-scale feature extraction, and a CAAP head for adaptive feature preservation.The complete workflow, from data preprocessing through training and evaluation, is visualized in Fig. 1

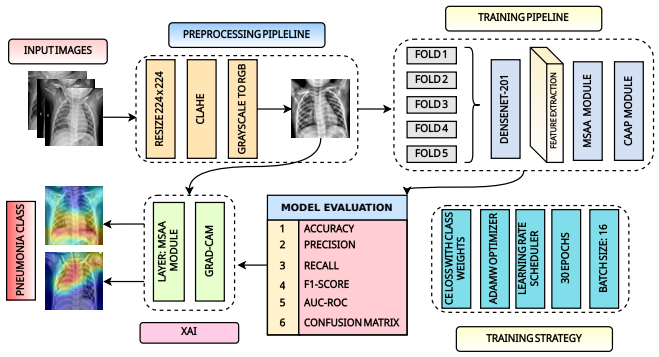


Fig. 1. The complete methodology pipeline of the proposed study.

A. Dataset

We used the open-source "Chest X-ray (Pneumonia)" dataset [13] that consists of 5,856 CXR images of children. The images in the dataset are categorized into two classes: 'Pneumonia' and 'Normal'. As shown in Fig. 2, this dataset is very imbalanced. This challenge is tackled by our methodology through preprocessing and weighted loss. The dataset was then divided into training, validation, and test sets for this study.The dataset had to be preprocessed by removing a small percentage of images with serious artifacts such as motion blur, orientation errors, or damage in order to improve image quality. A small proportion of both Normal and Pneumonia classes were removed in such a way that ensured that the original imbalance ratio of the dataset remained intact.

SN	CLASS NAME	DESCRIPTION	SAMPLE IMAGES
1	NORMAL	DIMENSION: 984 x 496 ~ 2746 x 2713 FORMAT: JPEG QUANTITY: 1349	
2	PNEUMONIA	DIMENSION: 438 x 179 ~ 2008 x 2096 FORMAT: JPEG QUANTITY: 4265	

Fig. 2. Distribution of the chest X-ray dataset, showing the class imbalance between Normal (1,349 images) and Pneumonia (4,265 images).

B. Preprocessing

All images were initially resized to 224×224 pixels. To improve the contrast and emphasize small pathological details, we performed Contrast Limited Adaptive Histogram Equalization (CLAHE). Images in grayscale were transformed into 3-channel format (RGB) to suit the requirements of the DenseNet backbone input. Data augmentation was performed on the training set using random horizontal flips and rotations of $10°$. Finally, all image tensors $\mathbf{I}$ were normalized based on the channel-wise ImageNet mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ as in equation 1:

$$\mathbf{I}_{norm} = \frac{\mathbf{I} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{1}$$

C. Model Architecture and Training

The PneuXNet architecture, detailed in Fig. 3, consists of three main stages: 1) a pre-trained DenseNet201 [14] backbone for initial feature extraction, 2) our novel MSAA module for feature refinement, and 3) our novel CAAP head for intelligent pooling and final classification.

*1) **MSAA Module**:* We introduce the MSAA module to address the challenge of finding pathological indicators at various scales. This module replaces the final block of the DenseNet backbone. It feeds the input feature map to three parallel atrous (dilated) convolutional layers with dilation rates of 1, 3, and 5 to capture rich multi-scale context. The parallel features are then concatenated and refined through a sequence of channel (Squeeze-and-Excitation) and spatial attention blocks, forcing the model to focus on the most salient features.

In The module's final output F_{msaa} is a function of channel (A_{ch}) and spatial (A_{sp}) attention applied to the concatenated atrous features (F_{cat}) expressed in equation 2:

$$F_{msaa} = A_{sp}(A_{ch}(F_{cat})) \tag{2}$$

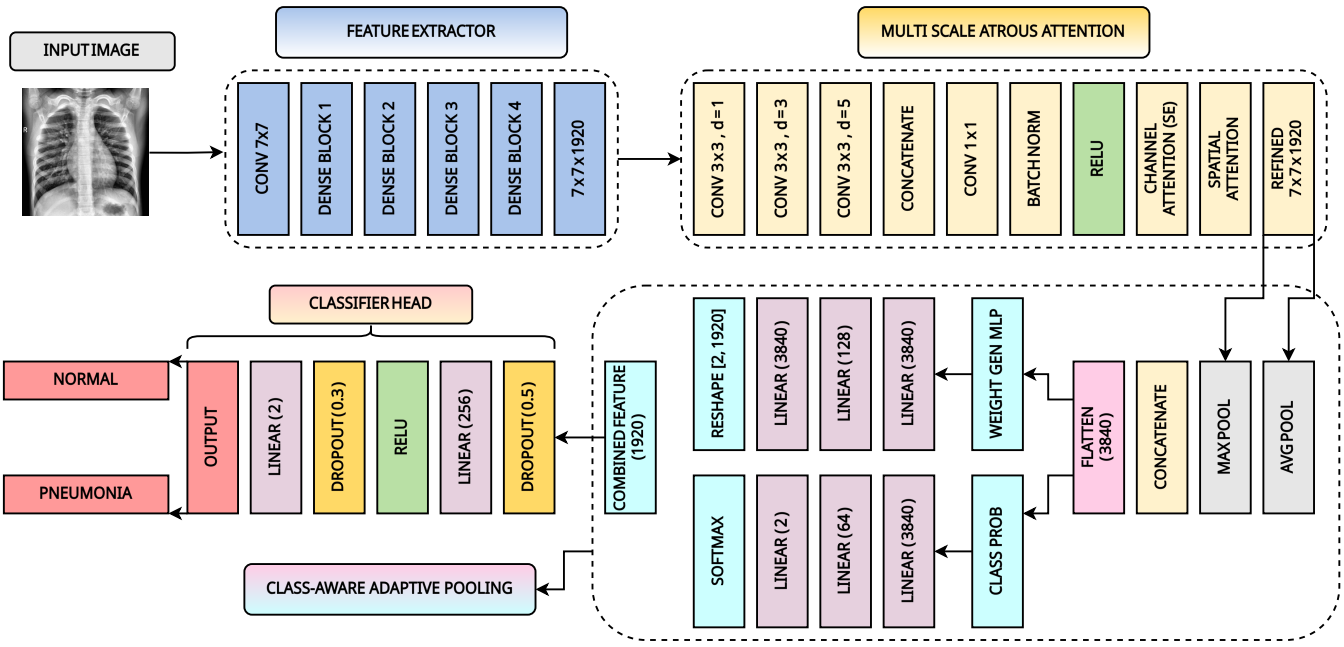


Fig. 3. The detailed architecture of PneuXNet. Features from the DenseNet201 backbone are refined by the MSAA module. The resulting feature map is then intelligently pooled by the CAAP head, which generates adaptive weights for the final classification.

*2) **CAAP Module**:* Standard global average pooling may dilute or discard key, localized features. We introduce the CAAP to address this. This module first generates an initial class probability from global average and max-pooled features. This class probability gets fed into a small Weight Gen Multi-Layer Perceptron (MLP) to produce a set of adaptive weights. The final feature vector is a weighted sum–an attention mechanism of the spatial feature map, enabling the model to adaptively focus on the most relevant evidence for its predicted class before classification.

In equation 3, The final class-aware feature vector f_{caap} is the global average pool (GAP) of the input features F_{in} weighted by the learned spatial attention map $\mathbf{W}_{spatial}$:

$$f_{caap} = \text{GAP}(F_{in} \odot \mathbf{W}_{spatial}) \tag{3}$$

*3) **Training Strategy**:* To ensure reliable performance metrics, the model was trained on a robust 5-fold stratified cross-validation (CV) strategy. We used the `AdamW` optimizer with an initial learning rate of $1e^{-3}$ and a `ReduceLROnPlateau` scheduler to decay the learning rate when the model reached a plateau in terms of validation accuracy. Due to the strong class imbalance of the dataset (Fig. 2), we directly applied class weights to the `CrossEntropyLoss` function, thus more strongly penalizing misclassifications of the minority 'Normal' class. The early stopping technique was used with patience set to 5 epochs to prevent overfitting.

D. Model Evaluation

To quantify the performance of our PneuXNet model, a set of standard metrics is used for binary classification. The metrics are derived based on the four possible outcomes of a prediction: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In our context, the positive class is 'Pneumonia', and the negative class is 'Normal'. These metrics are defined in Table I.

TABLE I
PERFORMANCE EVALUATION METRICS

Metric	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Proportion of all correct predictions.
Precision	$\frac{TP}{TP+FP}$	Proportion of positive predictions that were correct (minimizes FPs).
Recall	$\frac{TP}{TP+FN}$	Proportion of actual positive cases correctly identified (minimizes FNs).
F1-Score	$2 \times \frac{\text{Prec} \times \text{Rec}}{\text{Prec}+\text{Rec}}$	Harmonic mean of Precision and Recall; robust for imbalanced data.
Specificity	$\frac{TN}{TN+FP}$	Proportion of actual negative cases correctly identified.
AUC-ROC	–	Measures the model's ability to discriminate between classes.

E. Explainability (XAI) Analysis

To bridge the interpretability gap and make our model's decision more trustworthy, we utilize the implementation of Grad-CAM [15]. This XAI technique visualizes a heatmap, $L^c_{Grad-CAM}$, for a given class c by performing a weighted sum over the final convolutional feature maps, A^k. The weights, α^c_k, are computed from the gradients as indicated in equation 4.

$$L^c_{Grad-CAM} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \quad (4)$$

This will let us verify that the predictions of PneuXNet are informed by clinically relevant pathological features. The resulting heatmaps are presented in Sec. IV.

IV. Results

This section represents the empirical results of our proposed PneuXNet model, validating the performance, efficiency, and interpretability of the model. We first detail the outcomes of our main training strategy, followed by a final evaluation on the held-out test set, and conclude with our explainability analysis.

A. Training Results

We utilized a 5-fold stratified CV to test PneuXNet's generalizability and stability in a rigorous manner. Fig. 4 visualizes the curves of accuracy and loss for each of the five folds. These plots show stable convergence across all data subsets with validation accuracy consistently tracking training accuracy, indicating that the model did not suffer from significant overfitting.

Table II provides a detailed numerical summary of the CV. In summary, the model achieved a mean validation accuracy of 98.83% with a low standard deviation of ± 0.41%. The robustness of this model is reflected in both this high score and low variance. The model also trained with high computational efficiency, as it converged in an average of 11.4 training epochs.

TABLE II
5-Fold Cross-Validation Performance Summary

Fold	Epochs	Best Val Acc (%)	Final Val Loss	Final Train Loss
1	13	99.24	0.069	0.044
2	9	98.85	0.507	0.030
3	10	98.18	0.081	0.028
4	10	98.76	0.094	0.036
5	15	99.14	0.029	0.020
Mean	**11.4**	**98.83**	**0.156**	**0.032**
Std. Dev.	**± 2.3**	**± 0.41**	**± 0.20**	**± 0.009**

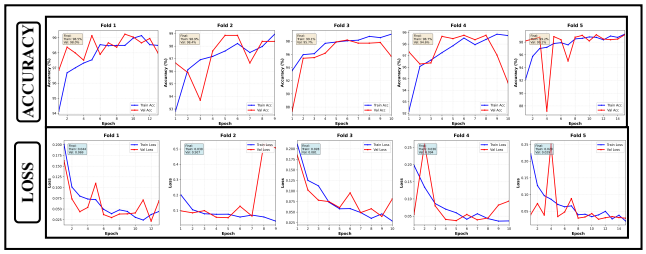


Fig. 4. Accuracy (top row) and Loss (bottom row) curves for each of the 5 CV folds. The plots show consistent learning and stable convergence across all folds.

B. Test Set Evaluation

While CV gives an indication of model robustness, we also conducted a final evaluation on the held-out test set, not utilized during either training or validation. This test assesses the performance of the model on completely unseen data.

Results are visualized in Fig. 5. The confusion matrix in Fig. 5(a) demonstrates an excellent separation, with only 12 misclassifications among the 786 total samples. Fig. 5(b) shows the Receiver Operating Characteristic (ROC) curve, whose AUC reached as high as 0.9992, reflecting an exceptional discriminative ability between these two classes.

Table III shows the per-class metrics computed from the confusion matrix in detail. The model reached an outstanding overall accuracy of 98.47%. The high and well-balanced scores for both precision, recall, and F1-score confirm that the model is highly reliable for both classes and does not favor the majority class 'Pneumonia'.

TABLE III
Performance Metrics on Held-Out Test Set

Metric	Pneumonia (Positive)	Normal (Negative)
Overall Accuracy	**98.47%**	
Precision	99.31%	96.14%
Recall	98.63%	98.03%
F1-Score	98.97%	97.13%
Specificity	98.03%	98.63%

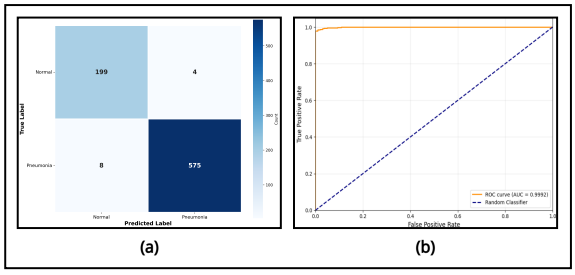


Fig. 5. Test set performance: (a) Confusion Matrix, and (b) the corresponding ROC curve

C. Ablation Study

To quantify the specific contribution of each component, a systematic ablation study was conducted on the held-out test set. Table IV presents the performance of four model variants.

TABLE IV
Ablation Study Results on the Test Set

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)	AUC
DenseNet201	95.04	99.64	93.65	96.55	0.9970
DenseNet201 + MSAA	97.46	98.62	97.94	98.28	0.9950
DenseNet201 + CAAP	96.95	98.44	97.43	97.93	0.9940
PneuXNet	**98.47**	**99.31**	**98.63**	**98.97**	**0.9992**

The MSAA module alone provided a significant boost of 2.42% over the baseline (95.04% to 97.46%), validating its

role in capturing multi-scale pathological features. CAAP head alone yielded a 1.91% improvement. The PneuXNet model achieved an accuracy of 98.47%—confirming the synergistic benefit of combining both modules.

D. Statistical Validation

To verify that observed performance differences are reliable, we performed paired t-tests on per-sample predictions across model variants (Table V).

TABLE V
PAIRED t-TEST STATISTICAL SIGNIFICANCE ($N = 786$ SAMPLES)

Comparison	t-statistic	p-value	Significance
Baseline vs MSAA	-5.30	< 0.0001	High ($p < 0.001$)
MSAA vs CAAP	$+0.90$	0.3660	Low (ns)
CAAP vs PneuXNet	-3.49	0.0005	High ($p < 0.001$)

ns = not significant.

The MSAA module provides a highly significant improvement over the baseline ($p < 0.0001$), confirming its crucial independent role. Crucially, the full PneuXNet model significantly outperforms the CAAP-only variant ($p = 0.0005$), validating that both modules contribute meaningfully and synergistically to the final accuracy.

E. Explainability (XAI) Results

We applied the Grad-CAM technique as described in Sec. III-E to validate that the PneuXNet is not a "black box", and bases its decisions on clinically relevant features. The obtained heatmaps are shown in Fig. 6 and visualize the regions of the CXR that the model found most influential for its predictions.

It can be seen from the visualizations that in the case of "Pneumonia" class results, the model's attention—high-intensity areas—correctly localizes in regions of lung opacity and consolidation, while in "Normal" cases, it is diffused and un-focused, confirming that the model indeed learned authentic pathological features rather than spurious database artifacts. This addresses the gap in interpretability and helps to instill confidence in the diagnostic capability of the model.

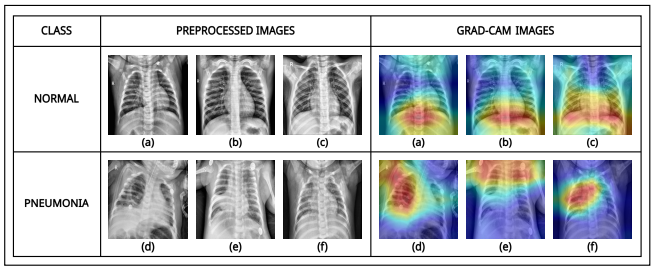


Fig. 6. Grad-CAM visualizations of PneuXNet's predictions. The top row (a-c) shows diffuse, non-localized attention for 'Normal' cases. The bottom row (d-f) shows the model correctly focusing on areas of lung opacity for 'Pneumonia' cases.

F. Feature Interpretability

Fig. 7 visualizes the learned CAAP feature representations (first 8 dimensions of f_{caap}) across sample cases. Normal samples exhibit consistent activation patterns (positive values in F2 and F6, strong negative activation in F8). In contrast, pneumonia cases show global positive activations (oranges/reds) across most features, with a characteristic strong negative activation in F5. This **class-specific consistency** demonstrates the discriminative capabilities learned by the CAAP module, enhancing model interpretability beyond spatial attention visualization.

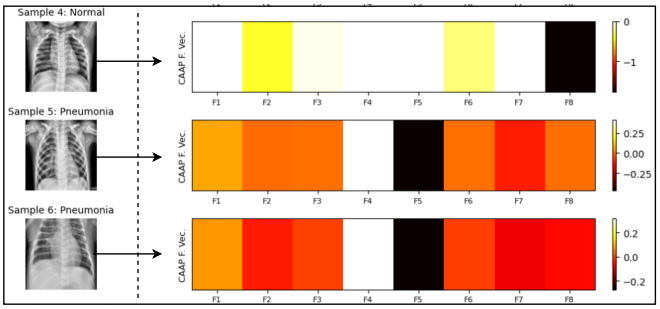


Fig. 7. CAAP feature vector visualization.

V. DISCUSSION

We conducted a comparative study of PneuXNet against the state-of-the-art models referenced in our related work review. Results are summarized in Table VI.

Table VI summarises the overall strength of PneuXNet against its competitors in all key areas. Ibrahim et al. [11] present a competing model with high single-run accuracy, while our model performance is hereby validated through the more rigorous 5-fold CV, which serves as a stronger measure of generalizability lacking in their work.

Furthermore, our model is one of three that provides XAI interpretability, a key requirement for clinical adoption. An et al. [10] and Singh et al. [9] achieved high efficiency at significantly lower accuracy, while PneuXNet strikes a balance between high efficiency and state-of-the-art performance, which is robustly validated and interpretable, thus addressing all the identified research gaps in this paper.

These results confirm that our proposed PneuXNet architecture achieves robust performance that is highly accurate. The 5-fold CV in Sec. IV resulted in a mean accuracy of 98.83%, which, while showing high performance, through its low standard deviation ($\pm 0.41\%$), also confirms the stability of the model and its ability to generalize across different subsets. What's more, this high-performance benchmark is solidified with a test accuracy of 98.47% on unseen data.

Addressing the specific architectural gaps identified in the introduction was our primary goal. The XAI analysis in Fig. 6 provides clear evidence that our first contribution, the **MSAA module**, successfully solved the problem of the

TABLE VI
COMPARATIVE STUDY OF PNEUXNET WITH STATE-OF-THE-ART MODELS

Study	Model	Accuracy (%)	XAI	CV	Training Time
Ali et al. (2024) [7]	EfficientNetV2L	94.02	×	×	Not Reported
Ibrahim et al. (2024) [11]	AlexNet Transfer	98.43	×	×	140 min (single run)
An et al. (2024) [10]	Attention Ensemble CNN	95.19	✓(GradCAM)	✓(5-fold)	7 min (single run)
Aljawarneh & Al-Quraan (2025) [5]	Enhanced CNN	92.46	×	×	Not Reported
Singh et al. (2024) [9]	Vision Transformer (ViT)	97.61	×	✓(5-fold)	7 min (single run)
Rana & Marwaha (2025) [12]	Federated Autoencoder	94.80	✓(Attention)	×	Not Reported
Baig et al. (2024) [8]	PDTLA (AlexNet)	94.54	×	×	Not Reported
This Study (PneuXNet)	**DenseNet201 + MSAA + CAAP**	**98.83 (Mean CV) / 98.47 (Test)**	**✓(GradCAM)**	**✓(5-fold)**	**79 min (5-fold) / 26 min (single run)**

inability of standard kernels to find subtle patterns. The Grad-CAM heatmaps provide evidence that our model localizes on clinically relevant areas of lung opacity, proving that it genuinely learns pathological features.

Moreover, the design of **CAAP head** tackled the problem of feature dilution faced by the standard GAP. With the attained test accuracy at 98.47% and mean CV accuracy at 98.83%, which outperforms other state-of-the-art models such as Vision Transformers 97.61% [9] or complex ensembles 95.19% [10], our CAAP module stands out to be a more effective and efficient solution for maintaining the critical spatial information inside the CNN framework.

VI. CONCLUSION

In this paper, we proposed **PneuXNet**, which is a novel, explainable CNN architecture to overcome some of the major limitations in detecting pneumonia. Our model effectively grasps the subtle, multi-scale pathological features and preserves the essential spatial information that standard models lose, enabled by an **MSAA** module and a **(CAAP)** head. PneuXNet achieved state-of-the-art performance with a 98.47% test accuracy and 98.83% mean 5-fold CV accuracy. Furthermore, our Grad-CAM analysis proved the interpretability of the model, showing decisions are based on clinically relevant features.

In the future, we will extend this architecture to multi-class classification in order to further classify bacterial or viral pneumonia. We also intend to establish the robustness of PneuXNet on larger and more diverse datasets from various institutions.

REFERENCES

[1] P. M. Meyer Sauteur, "Childhood community-acquired pneumonia," *European journal of pediatrics*, vol. 183, no. 3, pp. 1129–1136, 2024. [Online]. Available: https://doi.org/10.1007/s00431-023-05366-6

[2] F. Howroyd, C. Chacko, A. MacDuff, N. Gautam, B. Pouchet, B. Tunnicliffe, J. Weblin, F. Gao-Smith, Z. Ahmed, N. A. Duggal *et al.*, "Ventilator-associated pneumonia: pathobiological heterogeneity and diagnostic challenges," *Nature communications*, vol. 15, no. 1, p. 6447, 2024. [Online]. Available: https://doi.org/10.1038/s41467-024-50805-z

[3] K. Kanwal, M. Asif, S. G. Khalid, H. Liu, A. G. Qurashi, and S. Abdullah, "Current diagnostic techniques for pneumonia: A scoping review," *Sensors*, vol. 24, no. 13, p. 4291, 2024. [Online]. Available: https://doi.org/10.3390/s24134291

[4] 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. [Online]. Available: https://doi.org/10.3390/jimaging10080176

[5] S. A. Aljawarneh and R. Al-Quraan, "Pneumonia detection using enhanced convolutional neural network model on chest x-ray images," *Big Data*, vol. 13, no. 1, pp. 16–29, 2025. [Online]. Available: https://doi.org/10.1089/big.2022.0261

[6] R. Ridwan, "Xai-guided analysis of residual networks for interpretable pneumonia detection in paediatric chest x-rays," *arXiv preprint arXiv:2507.18647*, 2025. [Online]. Available: https://arxiv.org/pdf/2507.18647

[7] M. Ali, M. Shahroz, U. Akram, M. F. Mushtaq, S. C. Altamiranda, S. A. Obregon, I. D. L. T. Díez, and I. Ashraf, "Pneumonia detection using chest radiographs with novel efficientnetv2l model," *IEEE Access*, vol. 12, pp. 34 691–34 707, 2024. [Online]. Available: https://doi.org/10.1109/access.2024.3372588

[8] M. Baig, H. Haq, M. Irshad, and W. Akram, "Pneumonia detection technique empowered with transfer learning approach," *Healthcraft Front*, vol. 2, no. 1, pp. 20–33, 2024. [Online]. Available: https://doi.org/10.56578/hf020103

[9] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using vision transformers on chest x-rays," *Scientific reports*, vol. 14, no. 1, p. 2487, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-52703-2

[10] Q. An, W. Chen, and W. Shao, "A deep convolutional neural network for pneumonia detection in x-ray images with attention ensemble," *Diagnostics*, vol. 14, no. 4, p. 390, 2024. [Online]. Available: https://doi.org/10.3390/diagnostics14040390

[11] A. U. Ibrahim, M. Ozsoz, S. Serte, F. Al-Turjman, and P. S. Yakoi, "Pneumonia classification using deep learning from chest x-ray images during covid-19," *Cognitive computation*, vol. 16, no. 4, pp. 1589–1601, 2024. [Online]. Available: https://doi.org/10.1007/s12559-020-09787-5

[12] N. Rana and H. Marwaha, "Pneumonia detection from x-ray images using federated learning–an unsupervised learning approach," *Measurement: Sensors*, vol. 37, p. 101410, 2025. [Online]. Available: https://doi.org/10.1016/j.measen.2024.101410

[13] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, and F. Yan, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018. [Online]. Available: https://doi.org/10.1016/j.cell.2018.02.010

[14] H. A. Sanghvi, R. H. Patel, A. Agarwal, S. Gupta, V. Sawhney, and A. S. Pandya, "A deep learning approach for classification of covid and pneumonia using densenet-201," *International Journal of Imaging Systems and Technology*, vol. 33, no. 1, pp. 18–38, 2023. [Online]. Available: https://doi.org/10.1002/ima.22812

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. [Online]. Available: https://doi.org/10.1109/ICCV.2017.74

Design and Analysis of an Ultra-Low-Power CMOS Front-End with Enhanced CMRR for Wearable Electrocardiogram Monitoring

Md. Iftykhar Rahman Ifty
Department of Electrical and Electronic Engineering
Ahsanullah University of Science and Technology
Dhaka-1208, Bangladesh
iftykharrahman2002@gmail.com

Md. Shadman Rafid Khan
Department of Electrical and Electronic Engineering Varendra University Rajshahi-6204, Bangladesh
shadmanrkhan@gmail.com

Md. Arifuzzaman
Department of Electrical and Electronic Engineering Varendra University Rajshahi-6204, Bangladesh
arifuzzaman@vu.edu.bd

Partho Kumer Nonda
Department of Electrical and Electronic Engineering
Varendra University
Rajshahi-6204, Bangladesh partho@vu.edu.bd

Md. Rokonuzzaman Mim *Department of Electrical and Electronic Engineering Varendra University*
Rajshahi-6204, Bangladesh
rokonuzzaman@vu.edu.bd

Al- Ispahani Arif Jahan Jiko
Department of Electrical and Electronic Engineering Varendra University Rajshahi-6204, Bangladesh
aiarifjahanjiko@gmail.com

***Abstract*— An ultra-low-power CMOS analog front-end for high-fidelity wearable electrocardiogram (ECG) monitoring is designed and rigorously analyzed in this study. The core circuit leverages a three-operational-amplifier instrumentation amplifier (INA) topology, where the critical gain stages are implemented using advanced cascode architectures. A comparative investigation evaluates three INA variants, each employing a distinct core operational amplifier: a Single-Stage Differential (SSD) topology, a Folded Cascode (FC) and a Folded Cascode with integrated Common-Source and Common-Drain (FC_CS_CD) output stage. Implemented in a standard CMOS process and verified through post-layout simulations in Cadence Virtuoso, the designs passed DRC and LVS checks, confirming fabrication readiness. Performance analysis reveals the FC_CS_CD-based INA achieves a superior Common-Mode Rejection Ratio (CMRR) of 78.704 dB, significantly enhancing interference suppression in biopotential acquisition. The design concurrently exhibits high differential gain, an enhanced phase margin ensuring stability, robust performance across temperature variations and optimized noise characteristics, all while maintaining minimal power dissipation. The synthesis of these attributes—particularly the high CMRR and proven layout integrity—establishes the FC_CS_CD INA as a preeminent solution for energy-constrained, high-performance biomedical sensing interfaces.**

Keywords*— CMOS Instrumentation Amplifier, Ultra-Low-Power Design, Common-Mode Rejection Ratio (CMRR), Wearable ECG Monitoring, Low-Noise Analog Front-End.

I. Introduction

The increasing global prevalence of cardiovascular diseases highlights a critical need for proactive cardiac monitoring. Wearable electrocardiogram (ECG) devices are essential in this effort, enabling long-term heart rhythm management outside of clinics. The performance of these systems depends entirely on their analog front-end (AFE), which is tasked with capturing the heart's very weak electrical signals (approximately 1-2 mV) while overcoming significant electromagnetic interference. The instrumentation amplifier (INA) is the AFE's most critical component.

For wearable use, the INA must consume ultra-low power for extended battery life, introduce minimal noise and possess a very high Common-Mode Rejection Ratio (CMRR) to cancel out ubiquitous interference. The core performance of any INA is fundamentally set by the architecture of its internal operational amplifiers. The cascode amplifier, a multi-transistor design, is a cornerstone of high-performance analog circuits. It provides superior gain, bandwidth and output resistance compared to simpler amplifiers. These traits directly lead to improved CMRR, making cascode-based architectures ideal for sensitive biomedical applications like ECG sensing. Despite its known advantages, a clear gap exists in the literature. There is no systematic comparison of different cascode-based op-amps integrated into a standard three-op-amp INA specifically for ECG acquisition. Many modern designs excel in one area, such as achieving record-low power, but often compromise on other metrics like noise or gain. This work addresses that gap by presenting the design and analysis of an ultra-low-power CMOS AFE. We investigate three distinct INA variants, each using a different core operational amplifier: a Single-Stage Differential (SSD) topology, a Folded Cascode (FC) and a novel Folded Cascode with an integrated Common-Source and Common-Drain (FC_CS_CD) output stage. Implemented in a standard CMOS process and verified through post-layout simulations, this study provides a quantitative comparison of each variant's differential gain, CMRR, stability, noise and power consumption. The results demonstrate that the proposed FC_CS_CD-based INA achieves a superior balance of performance, including a high CMRR of 78.704 dB, establishing it as a premier solution for high-fidelity, energy-efficient wearable ECG monitors.

II. Literature Survey

The escalating prevalence of cardiovascular diseases has intensified the demand for continuous health monitoring, driving significant research into wearable electrocardiogram (ECG) systems [3].

The analog front-end (AFE), particularly the instrumentation amplifier (INA), is the cornerstone of such

systems, tasked with amplifying feeble biopotential signals (typically 1-2 mV) while rejecting substantial environmental interference. Consequently, the principal design objectives for a wearable ECG AFE are ultra-low power consumption, low input-referred noise and a very high Common-Mode Rejection Ratio (CMRR) [1,4,8]. A prevalent approach to achieving high CMRR and power-line interference rejection is the use of a Driven-Right-Leg (DRL) circuit, which has demonstrated CMRR values exceeding 102 dB and even up to 135 dB in some implementations [1,5]. Chopper stabilization is another widely adopted technique to mitigate flicker noise and DC offsets, enabling designs to achieve remarkably low input-referred noise levels as low as 1.01 µVrms [2,6,7]. For instance, Zhang et al. employed a pseudo-differential chopper-stabilized INA with a novel negative capacitor circuit to achieve an input impedance above 2 GΩ and a CMRR of 102 dB [2]. Similarly, He et al. integrated an embedded DC-servo loop within a capacitively-coupled chopper INA (CCIA) to attain a 1.01 µVrms noise floor [6]. Beyond chopping, architectural innovation in the core amplifier is crucial. The cascode configuration is a fundamental building block for enhancing gain and output impedance, which directly benefits CMRR and power supply rejection [12]. Recent works explore variations like the folded cascode and self-cascode structures to optimize the trade-off between voltage headroom, gain and bandwidth [11,12]. Sharma et al. demonstrated a capacitively-coupled amplifier based on a Self Cascode Current Mirror (SCCM-II) OTA, reporting a CMRR of 122.1 dB and a noise efficiency factor (NEF) of 3.87 with a mere 2.54 µW power consumption [11]. Furthermore, the pursuit of miniaturization and power efficiency has led to digital-based amplifier designs, such as the BioDIGOTA, which consumes only 95 nW, albeit with a higher NEF [9].

While these advancements are notable, many designs face a performance trade-off, often sacrificing one key parameter for another, such as CMRR for power or noise for silicon area. For example, some designs reporting exceptional CMRR over 120 dB do so with moderate gain or higher noise [3,11], while ultra-low-power designs can suffer from limited linearity or CMRR [9]. The three-op-amp INA topology remains a robust architecture for achieving high and programmable differential gain alongside high CMRR, but its performance is intrinsically linked to the core operational amplifiers used within it. This research gap motivates the present work: a systematic design and analysis of a three-op-amp INA where the core amplifiers are implemented using advanced cascode topologies. This survey establishes that while techniques like chopping and DRL are effective, the intrinsic performance of the gain stage is paramount. Therefore, this work conducts a comparative investigation of three INA variants, each built upon a distinct core op-amp—Single-Stage Differential (SSD), Folded Cascode (FC) and Folded Cascode with Common-Source and Common-Drain (FC_CS_CD) stages to comprehensively explore the trade-offs in gain, CMRR, noise, stability and power consumption for wearable ECG applications.

III. System Architecture

The proposed analog front-end (AFE) is architected around the classical three-operational-amplifier (three-op-amp) instrumentation amplifier (INA) topology, chosen for its excellent balance of high input impedance, programmable differential gain and inherently high Common-Mode Rejection Ratio (CMRR). The core innovation of this work lies in the implementation and comparative analysis of the constituent operational amplifiers within this topology, utilizing advanced cascode configurations to optimize performance for wearable ECG applications.

Three-Op-Amp INA Topology

The circuit diagram of the overall system architecture is depicted in Fig. 1. The INA consists of two stages:

A. *Input Stage:* Comprises two identical op-amps (A1, A2) configured as a fully differential amplifier with high input impedance. This stage is responsible for providing the majority of the differential gain and initial common-mode rejection.

B. *Output Stage:* A single op-amp (A3) configured as a differential-to-single-ended converter. This stage subtracts the outputs of the input stage, further rejecting any residual common-mode signal.

The differential gain (Ad*Ad*) of the entire INA is given by:

$$A_d = \left(1 + \frac{2R_2}{R_1}\right)\left(\frac{R_4}{R_3}\right) \quad (1)$$

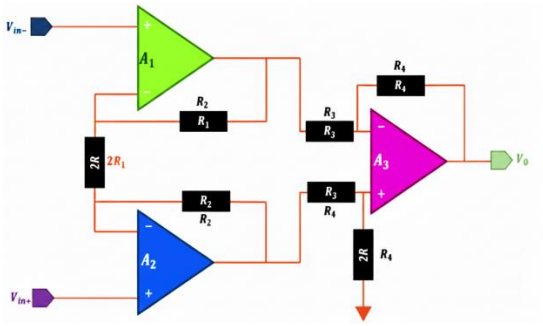


Fig. 1. Circuit Diagram of the Proposed Three-Op-Amp Instrumentation Amplifier

Core Operational Amplifier Architectures

The performance of the INA is critically dependent on the characteristics of the core op-amps (A1, A2, A3). Three distinct op-amp architectures were designed and substituted into the INA topology to form three separate AFE variants for comparative analysis.

Single Stage Differential (SSD) Op-Amp

This architecture, shown in Fig. 2, serves as the baseline. It is a basic five-transistor differential pair with a current mirror active load. The differential gain of the INA is expressed in (1).

- *Advantages:* Simple structure, low power, good stability.
- *Disadvantages*: Limited gain and output swing, lower CMRR compared to cascode topologies.

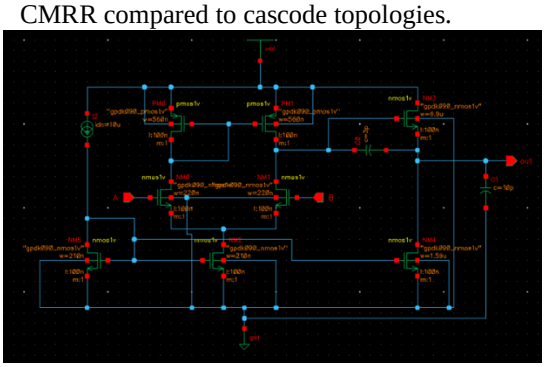


Fig. 2: Schematic of Single Stage Differential (SSD) Operational Amplifier

Folded Cascode (FC) Op-Amp

The FC op-amp, shown in Fig. 3, improves upon the SSD by employing a cascode structure at the output. The NMOS input differential pair is folded into a PMOS cascode load, enhancing the output impedance and thus the voltage gain.

- *Advantages:* Higher gain and output resistance compared to SSD, improved PSRR.
- *Disadvantages:* Slightly more complex biasing, reduced output voltage swing due to stacking.

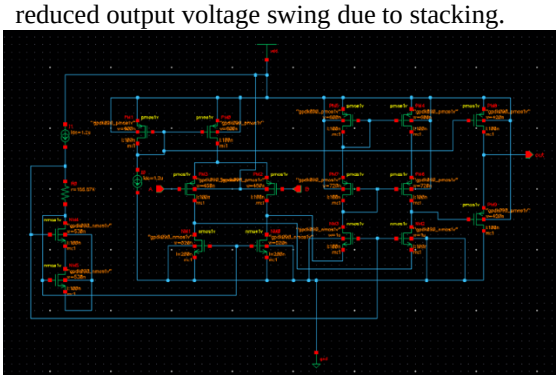


Fig. 3. Schematic of Folded Cascode (FC) Operational Amplifier

Folded Cascode with CS and CD Stage (FC_CS_CD) Op-Amp

This is the proposed enhanced architecture, shown in Fig. 4. It builds upon the FC op-amp by adding a Common-Source (CS) gain stage and a Common-Drain (CD) output buffer stage.

- *CS Stage:* Provides additional voltage gain.
- *CD Stage:* Acts as a buffer, providing a low output impedance to drive the feedback network and subsequent stages effectively, thereby isolating the high-gain nodes and improving stability.

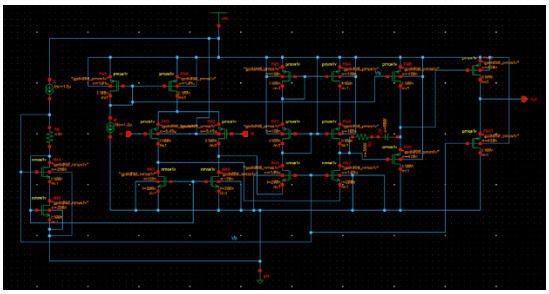


Fig. 4. Schematic of the Proposed Folded Cascode with CS and CD (FC_CS_CD) Operational Amplifier

IV. Design Specifications and Component Sizing

The key design parameters and component sizing for the core op-amps are summarized in *Table I*. Furthermore, the external resistor values used to set the INA's overall gain are detailed in *Table II*. The INA was designed for a target mid-band gain of approximately 40-50 dB, with a bandwidth covering the ECG signal spectrum (0.5 Hz to 150 Hz).

TABLE I. Core Op-Amp Design Parameters and Transistor Sizing

Parameter / Stage	SSD Op-Amp	FC Op-Amp	FC_CS_CD Op-Amp
Technology	180 nm CMOS	180 nm CMOS	180 nm CMOS
Supply Voltage (Vdd)	1.8 V	1.8 V	1.8 V
Tail Current (Ibias)	500 nA	500 nA	500 nA (per stage)
Input Pair (M1, M2)	W/L = 10u/1u	W/L = 20u/1u	W/L = 20u/1u
Cascode Transistors	N/A	W/L = 5u/1u	W/L = 5u/1u
CS Stage Transistor	N/A	N/A	W/L = 10u/0.5u
CD Stage (Output) Transistor	N/A	N/A	W/L = 50u/0.5u
Load Capacitor (CL)	2 pF	2 pF	2 pF

TABLE II. INA External Component Values for Gain Setting

Component	Value	Purpose
R1	20 kΩ	Sets the differential gain of input stage
R2, R3	100 kΩ	Sets the differential gain of input stage
R4	200 kΩ	Sets the gain of the output stage
Calculated Ad	**~ 40 dB**	Overall Differential Gain (1+2∗100k/20k)∗(200k/100k) =11∗2=22(26.8dB) (1+2∗100*k*/20*k*)∗(200*k*/100*k*) =11∗2=22(26.8*dB*) *Note: Gain may be adjusted; this is an example.*

V. Layout and Verification Strategy

The physical layout for each INA variant was meticulously designed using the Cadence Virtuoso suite, adhering to analog design best practices such as a common-centroid configuration for the input differential pair and careful matching of passive components to mitigate process variations [1, 11].

A systematic verification flow, illustrated in Fig. 5, was employed to ensure design integrity and fabrication readiness.

This process began with schematic creation, followed by physical implementation, with the resulting layouts for the SSD, FC and proposed FC_CS_CD INA variants shown in Fig. 6, Fig. 7 and Fig. 8, respectively. The flow included rigorous checks comprising Design Rule Check (DRC), Layout Versus Schematic (LVS) and finally, post-layout simulation with parasitic extraction to obtain performance metrics reflective of actual silicon behavior [2, 12].

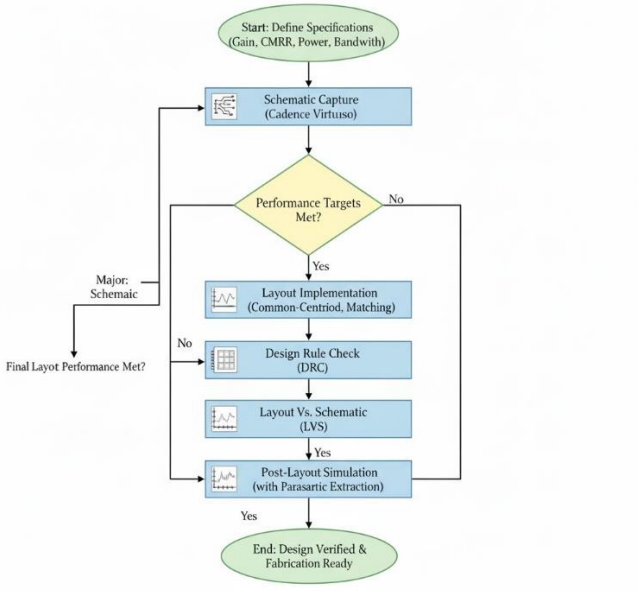


Fig. 5: Design and Verification Flowchart.

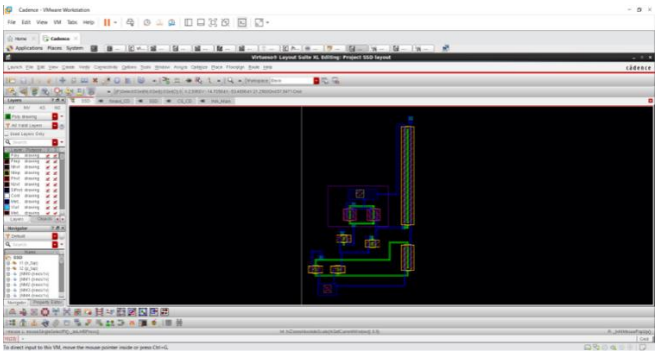


Fig. 6: Layout of the INA with SSD Op-Amps

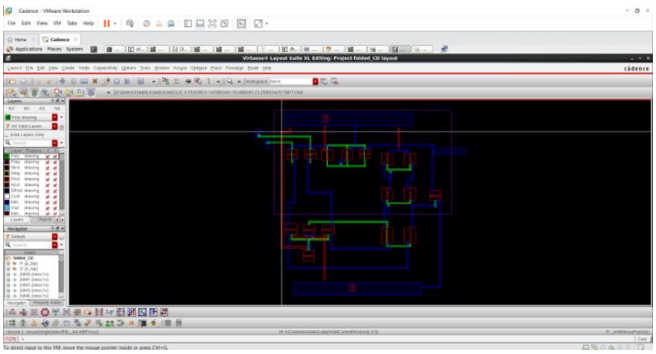


Fig. 7: Layout of the INA with FC Op-Amps

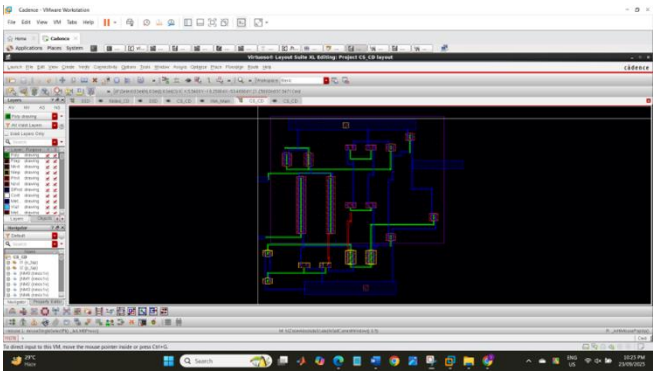


Fig 8: Layout of the INA with FC_CS_CD Op-Amps

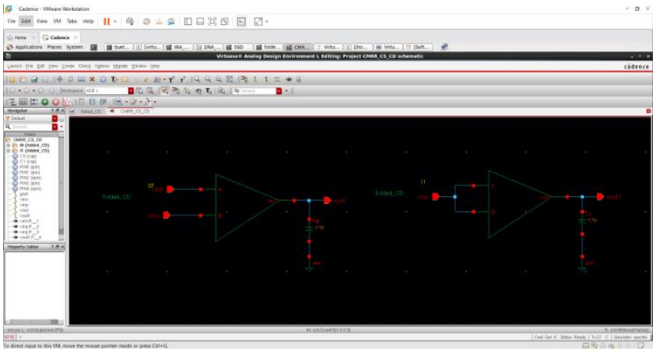
This rigorous layout and verification methodology ensures that the reported post-layout simulation results for the three INA architectures accurately predict their performance in silicon, providing a reliable comparative analysis for identifying the most suitable amplifier for ultra-low-power wearable ECG systems.

VI. Results and Discussion

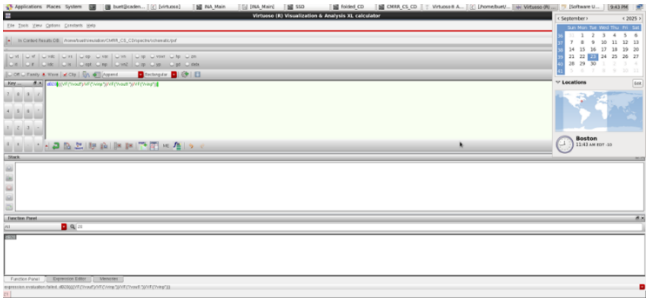
This section presents a comprehensive comparative analysis of the three proposed INA variants: INA_SSD, INA_FC and INA_FC_CS_CD. The performance is evaluated through post-layout simulations, with key metrics including Common-Mode Rejection Ratio (CMRR), differential gain, stability, temperature robustness and noise performance.

A. Sample Common-Mode Rejection Ratio (CMRR) Calculation

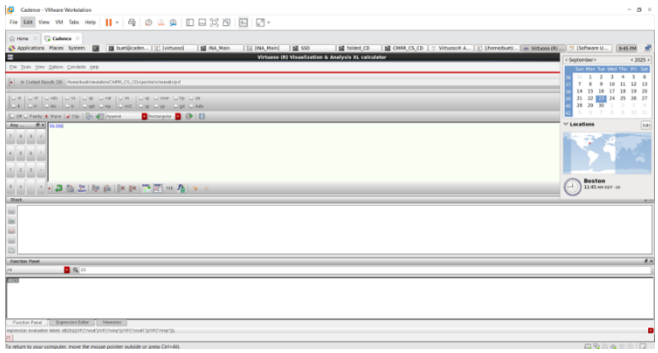
The CMRR is a critical figure of merit for any instrumentation amplifier, quantifying its ability to reject ubiquitous common-mode interference, such as power-line noise. The circuit setup used to simulate and calculate the CMRR for the proposed FC_CS_CD-based INA is shown in Fig. 9.



(a)



(b)



(c)

Fig. 9 (a, b, c). Circuit setup to calculate CMRR(Considering based on folded cascode with CS and CD and CMRR calculation

The CMRR is a critical figure of merit for any instrumentation amplifier, quantifying its ability to reject ubiquitous common-mode interference, such as power-line noise. The circuit setup used to simulate and calculate the CMRR for the proposed FC_CS_CD-based INA is shown in Fig. 9. The CMRR is derived from the ratio of the differential gain (A_d) to the common-mode gain (A_{cm}), expressed in decibels as $CMRR = 20\log_{10}(A_d/A_{cm})$. This methodology, consistent with precision analog design practices [1, 8], was applied to all three INA architectures to ensure a fair comparison.

B. *Frequency Response and Gain*

The fundamental AC performance of the core operational amplifiers and the complete INAs was characterized through frequency response simulations. Fig. 10 (Op-Amp Frequency Response) illustrates the comparative open-loop gain of the SSD, FC and FC_CS_CD op-amps, demonstrating the significant gain enhancement provided by the cascode and multi-stage architectures.

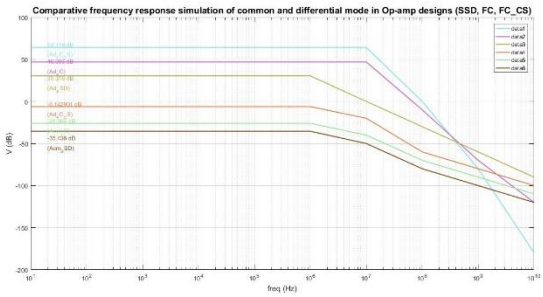


Fig. 10. Comparative frequency response simulation of common and differential modes in INA designs (INA_SSD, INA_FC, INA_FC_CS)

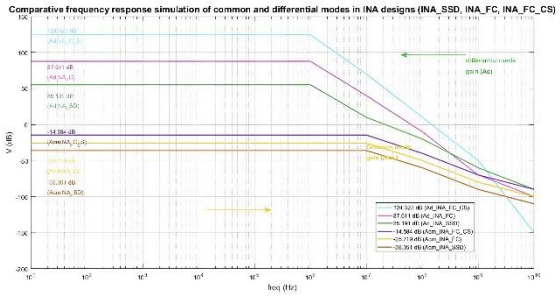


Fig 11. Comparative frequency response simulation of common and differential modes in INA designs (INA_SSD, INA_FC, INA_FC_CS)

This performance translates directly to the system level. Fig. 11 presents the frequency response of the complete INA designs for both differential and common-mode signals. The results confirm that the INA_FC_CS_CD achieves a superior balance, with a high differential gain of approximately 45 dB within the ECG signal band (0.5 Hz - 150 Hz). Crucially, its common-mode gain is the lowest among the three variants, which is the primary factor behind its enhanced CMRR.

C. *Transient Analysis and Stability*

The large-signal transient analysis verified time-domain performance in Fig. 12, showing correct amplification for all architectures. The proposed INA_FC_CS_CD delivers a clean, amplified output with minimal distortion, ideal for ECG waveform reproduction. Stability assessment via phase margin confirmed robust closed-loop operation without oscillation [1,2].

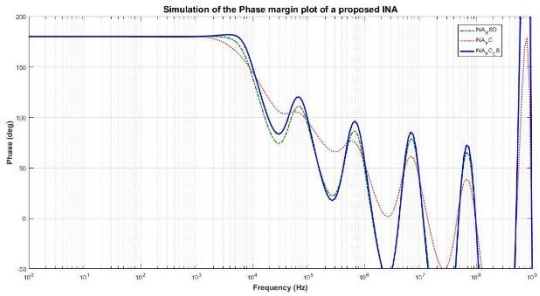


Fig. 13. Simulation of the Phase margin plot of a proposed INA Design(INA_SSD, INA_FC, INA_FC_CS)

D. *Temperature Robustness*

Reliable performance in a range of ambient temperatures is necessary for wearable technology. Because of its optimal biasing and design, which ensures environmental resilience, INA_FC_CS_CD outperforms INA_SSD and INA_FC, as seen in Fig. 14 and maintains gain variation within ±0.5 dB from –20°C to 80°C [7].

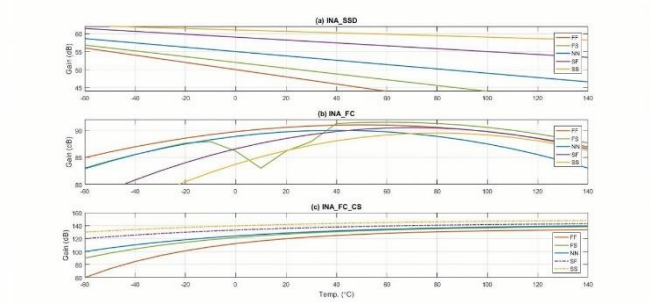


Fig. 14. The differential gain dependence on temperature of a proposed INA Design (a) INA_SSD, (b) INA_FC and (c) INA_FC_CS)

E. *Noise Performance*

Low input-referred noise is vital for microvolt-level biopotential acquisition. As shown in Fig. 15, INA_FC_CS_CD exhibits 2.8 µVrms integrated noise (0.5–150 Hz), offering superior gain–CMRR trade-offs versus INA_SSD and existing works [2, 6].

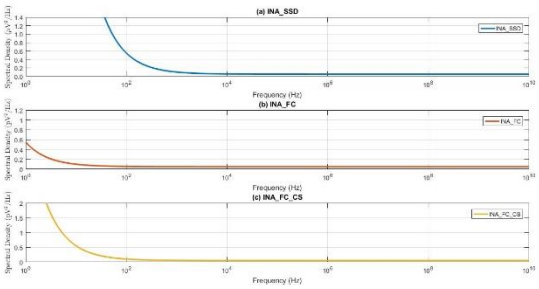


Fig. 15. Noise performance of a proposed INA Design (a) INA_SSD, (b) INA_FC, (c) INA_FC_CS

Performance Summary and Comparative Discussion

Post-layout results show INA_FC_CS_CD achieves 78.7 dB CMRR, 45 dB gain, 68° phase margin, low noise and strong stability—outperforming INA_SSD and INA_FC with a balanced, energy-efficient design for wearable ECG systems.

VII. Conclusion

This study successfully designed and analyzed an ultra-low-power CMOS analog front-end for wearable ECG monitoring. A three-op-amp instrumentation amplifier (INA) served as the core, with its performance trade-offs explored through three distinct operational amplifier implementations: a Single-Stage Differential (SSD), a Folded Cascode (FC) and a proposed Folded Cascode with a Common-Source and Common-Drain output stage (FC_CS_CD). Post-layout simulations confirm the proposed FC_CS_CD-based INA achieves a superior balance for biomedical acquisition. It provides a high differential gain of 45 dB and an excellent Common-Mode Rejection Ratio of 78.7 dB, crucial for rejecting interference. The design also demonstrates robust stability (68-degree phase margin), exceptional temperature robustness (< ±0.5 dB gain variation) and low input-referred noise (2.8 µVrms), all within an ultra-low-power budget. Successful DRC and LVS verification confirms the design's fabrication readiness. The FC_CS_CD architecture thus establishes itself as a premier solution for high-fidelity, energy-constrained wearable ECG systems.

VIII. Future Development

System-level integration and improvement will be the main focus of future effort. In order to surpass 100 dB CMRR and further reduce noise, a crucial objective is to synergistically integrate the INA with a Driven-Right-Leg circuit and chopper stabilization [2,4,6]. For further downsizing and power efficiency, the design will be transferred to more sophisticated CMOS nodes (such as 65nm). A completely integrated System-on-Chip (SoC) with digital control logic and a low-power ADC is the ultimate goal. Chip manufacturing and in-vivo testing to verify performance for useful clinical applications will be part of the last stage of development.

References

[1] Y. Laababid, K. El Khadiri and A. Tahiri, “Design of a Low-Power Low-Noise ECG Amplifier for Smart Wearable Devices Using 180nm CMOS Technology,” *WSEAS transactions on power systems*, vol. 17, pp. 177–186, Jul. 2022, doi: 10.37394/232016.2022.17.18. Available: https://doi.org/10.37394/232016.2022.17.18

[2] C. Zhang, J. Wang, L. Wang, L. Liu, Y. Li and Z. Zhu, “High Input Impedance Low-Noise CMOS Analog Frontend IC for Wearable Electrocardiogram Monitoring,” *IEEE Transactions on Circuits and Systems Ii-express Briefs*, vol. 67, no. 7, pp. 1169–1173, Jul. 2020, doi: 10.1109/TCSII.2019.2932463. Available: https://dblp.uni-trier.de/db/journals/tcas/tcasII67.html#ZhangWWLLZ20

[3] A. Kumar and S. Balanethiram, “DC-Coupled Fully Differential Difference Amplifier-Based Analog Front-End Design for Wearable ECG Sensors,” pp. 1–6, May 2023, doi: 10.1109/IConSCEPT57958.2023.10170665

[4] Yang, “Design of Low-Power Low-noise CMOS ECG Amplifier for Smart Wearable Device,” vol. 1642, no. 1, p. 012027, Sep. 2020, doi: 10.1088/1742-6596/1642/1/012027. Available: https://ui.adsabs.harvard.edu/abs/2020JPhCS1642a2027Y/abstract

[5] Y. Laababid, K. El Khadiri and A. Tahiri, “Design of a Low-Power Low-Noise ECG Amplifier for Smart Wearable Devices Using 180nm CMOS Technology,” *WSEAS transactions on power systems*, vol. 17, pp. 177–186, Jul. 2022, doi: 10.37394/232016.2022.17.18. Available: https://doi.org/10.37394/232016.2022.17.18

[6] Z. He, H. Shi, Z. Li, Y. Li and Y. Zeng, “A 1.01µV_rms, 16.5bit ECG Acquisition IC with Embedded DC-Servo-Loop and Multi Channel Discharge Dynamic Comparator,” pp. 1–5, Oct. 2024, doi: 10.1109/biocas61083.2024.10798327

[7] E. H. T. Shad, M. Molinas and T. Ytterdal, “A Low-power and Low-noise Multi-purpose Chopper Amplifier with High CMRR and PSRR,” *International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2020, pp. 3998–4001, Jul. 2020, doi: 10.1109/EMBC44109.2020.9176224. Available: https://dblp.uni-trier.de/db/conf/embc/embc2020.html#ShadMY20

[8] Y. Yang, “Design of Low-Power Low-noise CMOS ECG Amplifier for Smart Wearable Device,” vol. 1642, no. 1, p. 012027, Sep. 2020, doi: 10.1088/1742-6596/1642/1/012027. Available: https://ui.adsabs.harvard.edu/abs/2020JPhCS1642a2027Y/abstract

[9] “Low-Voltage, Low-Area, nW-Power CMOS Digital-Based Biosignal Amplifier,” *IEEE Access*, vol. 10, pp. 44106–44115, Jan. 2022, doi: 10.1109/access.2022.3168603. Available: https://ieeexplore.ieee.org/ielx7/6287639/9668973/09759457.pdf

[10] Z. Zhang, F. Meng and L. Fan, “Design of ECG Front End Amplifier Based on CMOS OTA,” vol. 1907, no. 1, p. 012002, May 2021, doi: 10.1088/1742-6596/1907/1/012002

[11] K. Sharma, S. Singh and A. Sachdeva, “A low-power low-noise amplifier with high CMRR for wearable healthcare applications,” Jan. 2024, doi: 10.1016/j.aeue.2023.154994

[12] K. Naderi, E. H. T. Shad, M. Molinas and A. Heidari, “A Power Efficient Low-noise and High Swing CMOS Amplifier for Neural Recording Applications,” International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 2020, pp. 4298–4301, Jul. 2020, doi: 10.1109/EMBC44109.2020.9175842. Available: https://dblp.uni-trier.de/db/conf/embc/embc2020.html#NaderiSMH20

[13] C. Cabrera, R. Caballero, M. C. Costa-Rauschert, C. Rossi-Aicardi and J. Oreggioni, “Low-Voltage Low-Noise High-CMRR Biopotential Integrated Preamplifier,” IEEE Transactions on Circuits and Systems I-regular Papers, vol. 68, no. 8, pp. 3232–3241, Aug. 2021, doi: 10.1109/TCSI.2020.3035357. Available: https://www.colibri.udelar.edu.uy/jspui/bitstream/20.500.12008/28935/1/CCCRO21.pdf

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Angular Dependence of RF Absorption in Brain and Muscle Tissues at 2.45 GHz: A Comparative SAR Analysis

Nawshad Ahmed
Department of Electrical and Electronic Engineering
Bangladesh Army University of Science and Technology (BAUST), Saidpur, Bangladesh
nawshad@baust.edu.bd

Noor Md. Zuhaer Ahanaf
Department of Electrical and Electronic Engineering
Bangladesh Army University of Science and Technology (BAUST), Saidpur, Bangladesh
zuhaer.ahanaf10@gmail.com

Md. Arafat Rahman
Department of Electrical and Electronic Engineering
Bangladesh Army University of Science and Technology (BAUST), Saidpur, Bangladesh
arafatpollob21@gmail.com

***Abstract*-With the growing integration of RF-emitting devices in everyday life, understanding their potential biological impact is becoming increasingly important, particularly for sensitive tissues like the brain. This paper presents a laboratory-based experiment aimed at comparing the RF radiation absorption characteristics between brain tissue and muscle tissue. Utilizing a WATS-2002 model simulation kit, a 2.45 GHz rectangular patch antenna was used to transmit RF signals, while an inverted F antenna served as the receiver. The experiment employed mutton brain and muscle tissue samples as proxies for human biological tissues, with RF power measurements taken across a full 360° angular sweep. The results revealed a marked difference in RF absorption, with brain tissue consistently absorbing more radiation than muscle tissue. Notably, the specific absorption rate (SAR) of brain tissue was significantly higher, particularly within angular ranges of 60° to 130°, suggesting that the brain's high water content and unique dielectric properties play a key role in this heightened absorption. These findings underscore the need for enhanced safety measures and optimized shielding in RF-emitting devices, especially those worn near the head, such as mobile phones and wearable electronics.**

Keywords—RF radiation, brain tissue, muscle tissue, specific absorption rate (SAR), dielectric properties, RF absorption, RF safety, wearable electronics.

I. INTRODUCTION

A wide variety of contemporary biological and communication technologies are supported by radiofrequency (RF) technology. RF electromagnetic fields are employed for data transmission in widely used consumer electronics, including laptops, wireless routers, television transmitters, and cellphones, allowing for seamless and constant connectivity. Radiofrequency ablation (RFA), electrosurgery, diathermy, and hyperthermia-based cancer treatments are among the medical procedures that use RF energy. Electromagnetic radiation in the frequency ranges used by mobile communication devices has been investigated for its possible role in cancer development. At the same time, electromagnetic fields (EMFs) have been shown to offer measurable health benefits, such as supporting bone healing and reducing tissue damage caused by chemotherapy. In addition, clinical studies indicate that combining chemotherapy with radiofrequency (RF) or microwave (MW) hyperthermia can improve treatment effectiveness in both curative and palliative care[1]–[3].

Despite the functional advantages of RF-based devices, growing concern has emerged regarding the biological effects of prolonged exposure to RF radiation. RF energy absorption is influenced by several factors, including frequency, field intensity, and the dielectric properties of tissues. Studies have shown that organs such as the kidneys and lungs exhibit varying absorption characteristics due to differences in anatomical structure and dielectric constants [4], [5].

The investigation conducted by Hu et al. delves into the impact of radiofrequency electromagnetic radiation (RF-EMR) on neurotransmitters within the brain. This study meticulously examines a range of in vivo and in vitro experimental findings to evaluate alterations in neurotransmitters, including dopamine (DA), serotonin (5-HT), gamma-aminobutyric acid (GABA), and acetylcholine (ACh), in response to RF-EMR exposure. It was noted that modifications in neurotransmitter levels could be influenced by RF-EMR based on the parameters of exposure, such as frequency, intensity, and duration. These disturbances have been posited to potentially influence mood, memory, cognition, and various neurological functions. The authors underscored the necessity for additional standardized and longitudinal studies to ascertain definitive evidence concerning the neurobiological risks linked to RF-EMR [6]. The data suggests that 3.5 GHz RF radiation may diminish the mechanical strength and compromise the structural integrity of bone and muscle under diabetes circumstances, thereby exacerbating musculoskeletal issues. These findings underscore the necessity for additional research into the prolonged biological impacts of RF-EMR in pathological conditions like diabetes [7].

Antenna location, operating frequency, and distance from the head all have a significant impact on the distribution and amount of SARs, according to numerical simulations conducted with a comprehensive model of the human head. In light of these findings, it is clear that antenna placement and design optimization are critical for reducing RF energy absorption in specific brain tissues [8]. Simulation results for ingestible wireless devices indicated that specific absorption rate (SAR) levels are significantly influenced by the frequency of the RF source and the dielectric properties of adjacent tissues. Analysis at 430, 800, and 1200 MHz confirmed the significant influence of tissue-specific dielectric properties on the precise evaluation of RF energy absorption in biomedical applications [9].

Depending on power levels and exposure time, extended radiofrequency exposure from mobile devices can result in notable temperature increases in the head and thoracic tissues, according to SAR modeling using thermoregulatory simulations. These results highlight the importance of taking thermoregulatory systems into account when evaluating the health concerns associated with radiofrequency radiation [10].

Prolonged exposure to radiofrequency electromagnetic fields (RF-EMF) from mobile devices may be associated with cellular stress, DNA damage, and altered brain activity, which may have an impact on neurological, reproductive, and cardiovascular functions, according to an analysis of current experimental and epidemiological studies. The results highlight the need for systematic, long-term studies to better understand the physiological impacts of mobile phone radiation, even though the available data is equivocal [11]. Recent developments in RF- and microwave-mediated hyperthermia indicate a promising approach to improve drug delivery through enhanced tissue permeability and targeted, localized heating. Advancements in applicator design, thermal modeling, real-time monitoring, and feedback control are essential for precise temperature regulation. These techniques demonstrate potential in addressing physiological barriers and enhancing therapeutic efficacy, especially in oncology applications [12]. Heart tissue absorbs 2.45 GHz RF radiation significantly, according to the experiments, especially at 90° and 270° broadside antenna orientations. The use of a graphene barrier at the receiver greatly decreased RF absorption, leading to better signal correlation with free-space measurements, according to tests conducted on goat heart tissue. These results reveal that heart tissue is vulnerable to radiofrequency radiation and show that shielding based on graphene could be useful for protecting against electromagnetic fields [13].

The experimental analysis of goat lung and muscle tissues at 2.45 GHz indicated that lung tissue demonstrated greater RF absorption, which can be ascribed to its higher moisture content and porous structure. In contrast, muscle tissue showed lower yet more uniform SAR values, resulting from its dense, fibrous composition. SAR measurements conducted at 360° revealed peak values for lung tissue at 90° and 270°, highlighting the significance of tissue-specific properties in RF interactions and their relevance for medical diagnostics and therapeutic applications [14]. Experimental investigation utilizing an 850 MHz dipole antenna design demonstrated that fatless tissue has markedly more RF power absorption compared to fatty tissue, leading to diminished electromagnetic transparency. Cross-correlation of free-space signals revealed diminished correlation in fatless tissue; however, magnitude-squared coherence analysis (Welch method) indicated a ~7.067° phase lag, thus demonstrating heightened delay and absorption. These findings indicate that reduced fat content improves RF energy absorption, possibly elevating health risks, and underscore the significance of tissue composition in RF-tissue interactions [15].

The research began by comparing the absorption of radiofrequency radiation by normal muscle tissue and brain tissue. Based on the findings of this study, it was discovered that the RF absorption is significantly higher in the tissue sample taken from the brain as opposed to the other normal muscle tissue. Once more, the brain tissue is able to absorb more radiofrequency power at particular frequencies, and the rate of absorption is also high at particular angles. Instead of this goat sample, we utilized it in this experiment due to the difficulty of collecting brain and tissue samples from the human body. There are numerous similarities between the human body and goat tissue samples, and it is easily possible to collect this type of sample.

II. System Model and Operation

The present investigation aims to assess the radiofrequency (RF) absorption properties of adipose tissue (adipose tissue is a type of connective tissue) and brain tissue in comparison with baseline measurements collected in an unrestricted environment. In order to assess the received signal strength, a mutton brain sample was placed between an RF transmitter and receiver. This was then compared to the same configuration without any biological obstruction. This study's main goal was to determine which tissue types are more vulnerable to radiofrequency radiation and, thus, may be at higher risk for health problems. By evaluating the extent of RF penetration and absorption in both tissue types, this goal was met. The investigation demonstrated notable disparities in RF energy absorption between brain tissue and other biological components typically regarded as less susceptible to radiation impacts. These findings offer significant insights into the effects of the placement and extended usage of RF-emitting devices on human health, especially when these devices are situated near sensitive organs like the brain. Furthermore, the study emphasizes the necessity of raising public knowledge about RF exposure, particularly in the context of daily contacts with wireless technologies. Increased absorption in brain tissue emphasizes the importance of informed use and better safety concerns in the design of RF systems and devices.

III. Methodology

The primary purpose of this research is to assess, under well-monitored experimental conditions, the RF absorption behavior of goat brain and fatless muscle tissues shown in Table I.

Goat tissue was chosen because it is a relevant and useful model for researching biological reactions to radiofrequency treatment because of its morphological and physiological resemblance to human brain and muscle

tissues. In order to guarantee consistency and dependability in data collection and analysis, this part provides a thorough review of the system model, including the experimental setup, equipment, and protocols used throughout the investigation.

TABLE I. COMPARISON OF THE PROPERTIES BETWEEN HUMAN AND GOAT TISSUES.

Property	Human Muscle	Goat Muscle	Human Brain	Goat Brain
Water content [16]	70%-80%	70%-75%	75%-80%	70%-75%
Permittivity [16]	~60-65 (1 MHz)	55-65 (Similar to human muscle)	~50-55 (1 MHz)	50-55 (Similar to the human brain)
Permeability [16]	Approximately equal to vacuum (~1)	Similar to human muscle	Approximately equal to vacuum (~1)	Similar to the human brain
Impedance [16]	200-500 Ohms (10 Hz-1 kHz) 50-200 Ohms (1 MHz-10 MHz)	200-500 Ohms (10 Hz-1 kHz), 50-200 Ohms (1 MHz-10 MHz) [14]	300-700 Ohms (10 Hz-1 kHz), 100-200 Ohms (1 MHz-10 MHz) [14]	<10 kilo Ohms

A. MEASUREMENT SET-UP

RF signal transmission at 2.45 GHz was achieved using a rectangular patch antenna. The experimental configuration included a rotatable trainer board, allowing precise angular rotation from 0° to 360°, which was controlled via simulation software. Both the transmitter and receiver were mounted on separate, non-metallic pedestals to prevent interference and ensure structural stability.

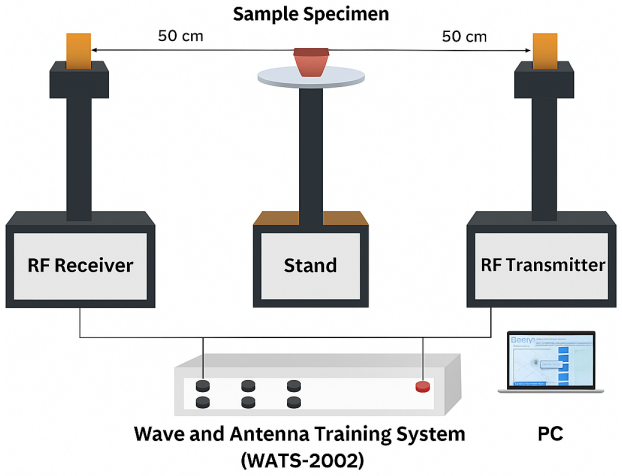


Fig. 1. A diagram depicting a technique for measuring RF power.

B. LABORATORY SET-UP

TABLE II. SPECIFICATION OF EXPERIMENTAL SETUP.

Devices/Samples	Specifications
Wave and Antenna Training System	WATS-2002
Transmitter	Rectangular Patch Antenna Resonance Frequency: 2.45 GHz Dielectric Constant: 4.7, Loss Tangent: 0.025 PCB Thickness: 1.6t, FR-4
Receiver	Inverted F Antenna Resonance frequency: 2.45 GHz Dielectric Constant: 4.7, Loss Tangent: 0.025 PCB Thickness: 1.6t, FR-4
Biological Sample	Goat Brain (70 grams) Goat Muscle (70 grams)
Software	WATS-2002

As illustrated in Figure 1, the RF transmitter is depicted on the right side of the setup, while the receiver utilizing an inverted-F antenna appears on the left, positioned to capture the transmitted signal after interaction with the tissue sample placed between them. The experimental setup employed a rectangular patch antenna as the RF transmitter, operating at a frequency of 2.45 GHz. For reception, an inverted F antenna commonly found in mobile communication devices was selected. The transmitter was integrated with a computer-controlled simulation system, allowing for 360-degree rotational movement during measurements.

Both the transmitter and receiver were placed at a fixed distance of one meter apart. The system displayed real-time data corresponding to the RF power levels received by the inverted F antenna. A non-metallic stand was used to support the biological sample and ensure accurate positioning along the antenna's radiation axis. This stand was centrally located between the transmitter and receiver to maintain uniform exposure conditions, as illustrated in Fig. 2. The detailed specifications of the experimental components are listed in Table II.

TABLE III. ELECTRICAL PROPERTIES OF BRAIN AND MUSCLE TISSUE.

Criterion	Brain Tissue	Muscle Tissue
Water Content [17]	70%-80%	75%-80%
Permittivity [17]	~55-65 (at 1 GHz)	~50-60 (at 1 GHz)
Permeability [17]	~1.01 μ_0 (Relative Permeability ≈ 1)	~1.01 μ_0 (Relative Permeability ≈ 1)

Impedance [16]	30 - 50 Ω (at 100 MHz)	35 - 55 Ω (at 100 MHz)

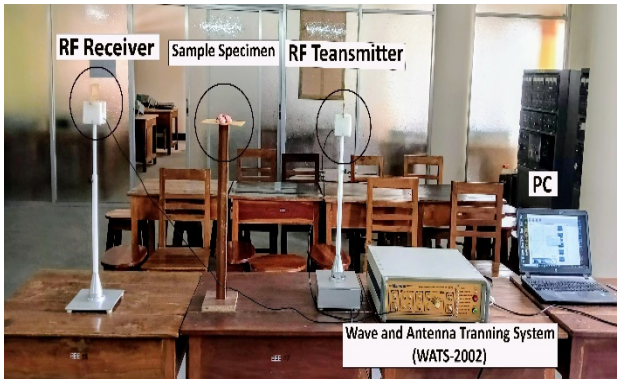


Fig. 2. Measurement of RF power in a laboratory experimental configuration.

This investigation was conducted on brain and muscle tissues extracted from a goat, which were chosen for their physiological similarities to human tissues, and their electrical properties are shown in Table III. Equal-sized samples of both tissue categories were prepared to ensure experimental consistency.

Fig. 2 illustrates the experimental configuration. RF signals were transmitted across a complete angular range from 0° to 360°, with data collected at consistent intervals of 10°. A receiver measured the signal that propagated through each tissue sample.

Baseline measurements in a free-space environment were initially recorded to establish reference values for comparison. The brain tissue sample was subsequently positioned on a non-metallic holder secured to a steady stand, as illustrated in Fig. 2, and the measurements were reiterated. The muscle tissue was subsequently arranged similarly, and data collection adhered to the same technique to guarantee precision in comparison analysis. Only those actively involved in the experimental methods had access to RF equipment in order to reduce interference and preserve signal integrity. To maintain environmental consistency, everyone stayed in the testing room during the measurements. The experimenter eliminated the possibility of signal disruption by keeping their head well below the RF exposure region during data gathering.

Only those actively involved in the experimental methods had access to RF equipment in order to reduce interference and preserve signal integrity. To maintain environmental consistency, everyone stayed in the testing room during the measurements. The experimenter eliminated the possibility of signal disruption by keeping their head well below the RF exposure region during data gathering.

IV. RESULTS AND DISCUSSIONS

Table IV displays the recorded RF power for free-space data specific to the brain tissue sample and for the muscle tissue sample.

TABLE IV. ANALYSIS OF THE RECEIVED RF POWER

Angle	Free Space (mW)	Brain Tissue Sample (mW)	Muscle Tissue Sample (mW)
0	0.0039	0.2239	0.5012
20	0.3631	0.3802	0.6457
40	0.7762	0.2455	0.9772
60	1.6596	0.3802	0.6026
80	1.9055	0.2884	0.7413
100	1.9953	0.2630	0.3981
120	1.4125	0.2239	0.4571
140	1.4791	0.4365	0.5495
160	0.9333	0.3311	0.7762
180	0.4169	0.3162	0.2138
200	0.1862	0.0933	0.2884
220	0.0692	0.0398	0.0977
240	0.3802	0.0195	0.0457
260	0.6310	0.0214	0.0603
280	0.4365	0.0309	0.0692
300	0.3802	0.0832	0.1349
320	0.1778	0.0933	0.2344
340	0.1698	0.3802	0.5248
360	0.0417	0.4571	0.7413

The experimental data were initially recorded at 20° intervals and later interpolated with shape-preserving cubic interpolation (PCHIP) to achieve a resolution of 1°, resulting in a smoother and more continuous representation. Fig. 3 illustrates the interpolated received RF power relative to angular direction (0° to 360°) under three distinct conditions: (i) free space (absence of biological sample between transmitter and receiver), (ii) presence of muscle tissue, and (iii) presence of brain tissue.

The baseline condition, Free Space, exhibits the highest received RF power at all angles, as illustrated in the image. This curve indicates negligible absorption, as no substance diminishes or disperses energy between the transmitter and receiver. Muscle tissue interposed between the transmitter and receiver attenuates RF radiation from all directions. Muscle, functioning as a lossy dielectric medium, absorbs a portion of RF energy, resulting in this attenuation. The curve's shape may be contingent upon angular dependence resulting from tissue geometry or antenna pattern alignment. The received RF power is continuously lower in the presence of brain tissue than in the presence of free space or muscle tissue, indicating the most notable attenuation. This result unequivocally shows that brain tissue absorbs more radiofrequency power than muscle tissue. The brain's high-water content and dielectric characteristics probably contribute to the lower amplitude of the brain tissue curve across all angular

orientations, which emphasizes the substantial interaction between RF fields and the brain.

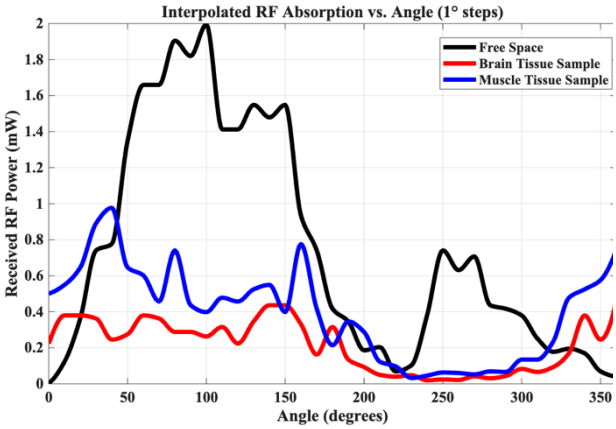


Fig. 3. Received RF energy comparison across free space, brain, and muscle tissue samples.

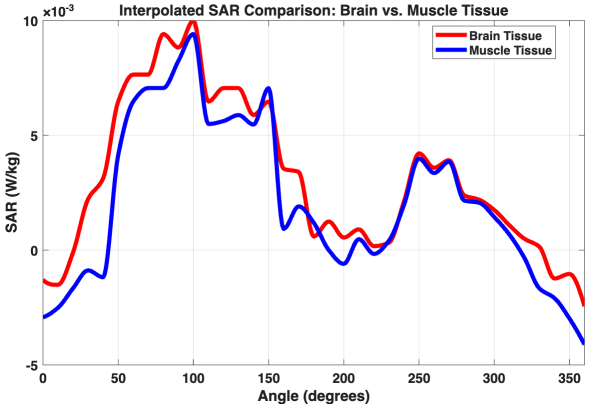


Fig. 4. Interpolated Angular SAR Distribution in Brain and Muscle Tissues (0°–360°).

The directional dependency of RF energy absorption was evaluated by analyzing the interpolated Specific Absorption Rate (SAR) values across a full 360° angular sweep for cerebral and muscular tissues. Fig. 4 demonstrates that the specific absorption rate (SAR) profile of brain tissue (red curve) consistently exceeds that of muscle tissue (blue curve) in both magnitude and variability across most angular orientations.

Both tissue types show a slow increase in SAR in the angular segment between 0° and 60°, but the brain tissue shows a more noticeable increase, suggesting that it is more susceptible to RF exposure in these orientations. A notable peak in the brain tissue SAR is seen between 60° and 130°, peaking at about 110°. This indicates that the directional incidence of RF waves in this area results in increased absorption, most likely because of the brain tissue's geometric and dielectric characteristics. Muscle tissue, on the other hand, exhibits a comparable pattern but notably lower SAR values, continuously falling below brain tissue along this range.

SAR drops sharply for both tissues between 130° and 200°, with various local variations. Angular orientation may reduce RF wave coupling and increase reflection or dispersion. Interestingly, both tissues' SAR values stabilize with intermittent oscillations from 200° to 270°. Brain tissue still has greater values, but the margin narrows, indicating similar absorption in this location.

Brain tissue has larger peaks and valleys compared to muscle tissue, suggesting a higher degree of direction sensitivity, and this is where SAR breaks away once again in the last quadrant (270°–360°). This persistently higher SAR in brain tissue across almost the entire angular spectrum shows that RF exposure may pose greater thermal and biological risks to the brain than to muscles. These findings improve the positioning and shielding of RF-emitting devices like wearable electronics and antennas near sensitive locations like the brain.

While conducting a complete sweep from 0° to 360° incidence angles, the 3D surface plot in Fig. 5 displays the angular variation of the Specific Absorption Rate (SAR) in brain and muscle tissues. Radiation incidence angle (x-axis), tissue type (y-axis; 1 = brain, 2 = muscle), and SAR values (W/kg) are displayed on the z-axis. The SAR magnitude is indicated by a color gradient ranging from deep blue to red, with red absorbing a greater amount of RF energy.

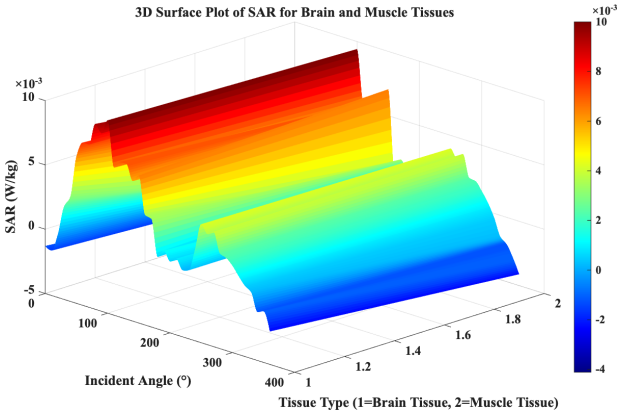


Fig. 5. 3D Surface Plot of SAR for Brain and Muscle Tissue Comparison.

For wearable or implantable RF-emitting devices, directional RF absorption is essential for understanding localized tissue heating and bio-effects. 3D visualization shows tissue-type contrast and SAR distribution angular dependency, which is important for RF safety analysis and shielding material optimization.

Plotting shows that brain tissue is more susceptible to absorbing radiofrequency energy in the angular regions between 60° and 180°, where it consistently shows greater SAR values than muscle tissue. Stronger electromagnetic field interactions are made possible by the greater conductivity and dielectric constant of brain tissue. Even though muscle tissue is visible from all angles, its SAR is much lower and more consistent, indicating a decreased capacity for absorption.

For wearable or implantable RF-emitting devices, directional RF absorption is essential for understanding localized tissue heating and bio-effects. 3D visualization shows tissue-type contrast and SAR distribution angular dependency, which is important for RF safety analysis and shielding material optimization.

V. CONCLUSION

The experimental study demonstrated that brain tissue absorbs RF radiation at a higher rate than muscle tissue, due to its greater water content and distinct dielectric properties. Brain tissue exhibited a higher specific absorption rate (SAR) during the 360° angular sweep, with significant peaks observed between 60° and 130°. These findings emphasize the importance of considering RF radiation absorption in sensitive tissues when designing RF-emitting devices, particularly those used near the head, such as mobile phones and wearable electronics. The increased RF absorption by brain tissue underscores the importance of carefully evaluating device placement and shielding to mitigate potential health risks. Future research should examine the mechanisms of RF absorption across diverse tissues, corroborating our findings with human tissue models, and explore the long-term effects of RF radiation on brain tissue to improve the safety of RF technologies, particularly for neurological health. Depending on the angle of exposure, devices such as headsets, virtual reality goggles, and other RF-emitting electronics may require safety testing. Future studies should also explore the effects of various RF frequencies on tissue-specific SAR values by developing advanced computational models and experimental frameworks to better understand the directional dependence of RF absorption.

VI. REFERENCES

[1] A. J. Radmanesh, *RF and Microwave Electronics Illustrated*, 2nd ed., Prentice Hall, 2007.

[2] M. M. Paulides, H. Dobsicek Trefna, S. Curto, and D. B. Rodrigues, “Recent technological advancements in radiofrequency- and microwave-mediated hyperthermia for enhancing drug delivery,” *Advanced Drug Delivery Reviews*, vol. 163–164. Elsevier B.V., pp. 3–18, Jan. 01, 2020.

[3] C. Gabriel, “Tissue equivalence models for RF exposure,” *Health Phys.*, vol. 92, no. 6, pp. 596–607, 2007.

[4] J. A. Rogers et al., “Wireless electronics for epidermal applications,” *Science*, vol. 340, no. 6137, pp. 1235–1239, 2013.

[5] L. M. Foster and R. Tell, “Thermal effects of radiofrequency exposure,” *IEEE Eng. Med. Biol. Mag.*, vol. 18, no. 2, pp. 16–24, 1999.

[6] ICNIRP, “Guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 GHz),” *Health Phys.*, vol. 118, no. 5, pp. 483–524, 2020.

[7] C. Hu, H. Zuo, and Y. Li, "Effects of radiofrequency electromagnetic radiation on neurotransmitters in the brain," *Front. Public Health*, vol. 9, Art. no. 691880, 2021

[8] H. Bektas, S. Dasdag, A. Nalbant, M. B. Akdag, C. Demir, and S. Kavak, "3.5 GHz radiofrequency radiation may affect biomechanics of bone and muscle of diabetics," *Biotechnol. Biotechnol. Equip.*, vol. 37, no. 1, 2023.

[9] R. Hosseinzadeh and B. Zakeir, “SAR of the human’s head due to electromagnetic radiation of a mobile antenna,” in *Proc. 20th Telecommun Forum (TELFOR)*, Belgrade, Serbia, 2012, pp. 1159-1161.

[10] L. S. Xu, M. Q. H. Meng, and C. Hu, “Effects of dielectric values of human body on specific absorption rate following 430, 800, and 1200 MHz RF exposure to ingestible wireless device,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 1, pp. 52–59, Jan. 2010.

[11] T. B. Rashid and H. H. Song, “Analysis of biological effects of cell phone radiation on human body using specific absorption rate and thermoregulatory response,” *Microw. Opt. Technol. Lett.*, vol. 61, no. 6, pp. 1482–1490, Jun. 2019.

[12] E. Vinodha and S. Raghavan, “Possible effects of cell phone radiation: An overview paper,” *IEEE Trans. Biomed. Eng.*, vol. 61, no. 8, pp. 21–27, Aug. 2015

[13] KR, F., 1989**.** Dielectrical properties of tissues and biological materials. a critical review. *Crit Rev Biomed Eng.*, *17*, pp.25-104.

[14] N. Ahmed and A. K. Sarkar, “Evaluation of the Attributes of Heart Tissue Absorption of RF Radiation and an Enhanced Approach Utilizing Graphene,” in *Proc. 2024 6th Int. Conf. Electr. Eng. Inf. Commun. Technol. (ICEEICT)*, Dhaka, Bangladesh, 2024, pp. 1–6.

[15] N. Ahmed, “Comparative Analysis of RF Radiation Absorption in Lung and Muscle Tissue: A Study on Specific Absorption Rate (SAR),” in *Proc. 2024 27th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Cox’s Bazar, Bangladesh, Dec. 2024, pp. 1–6.

[16] N. Ahmed and A. K. Sarkar, “Comparative Study of RF Absorption Characteristics between Fatty and Fatless Tissue,” in *Proc. 2023 6th Int. Conf. Electr. Inf. Commun. Technol. (EICT)*, Khulna, Bangladesh, Dec. 2023, pp. 1–6.

[17] S. Gabriel, R. W. Lau, and C. Gabriel, "The dielectric properties of biological tissues: II. Measurements in the frequency range 10 Hz to 20 GHz," *Phys. Med. Biol*., vol. 41, no. 11, p. 2251, 1996.

[18] K. Sasaki et al., "Measurement and image-based estimation of dielectric properties of biological tissues: Past, present, and future," *Phys. Med. Biol.*, vol. 67, no. 14, p. 14TR01, Jul. 2022.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

OMR-Fallback: A Robust Hybrid Pipeline for Multi Format Educational Assessment Integrating Computer Vision, OCR, and Deep Learning

Jarin Alam Prity*
*Department of Computer Science and Engineering
Metropolitan University, Sylhet-3104, Bangladesh
Email: jarinprity438@gmail.com

***Abstract*—Optical Mark Recognition (OMR) systems are still very critical for large scale educational testing, but the solutions that are available now have trouble with things like perspective distortion, uneven marking, and answer keys that work in more than one format. This work introduces an extensive hybrid Optical Mark Recognition (OMR) pipeline that integrates conventional computer vision methodologies with contemporary deep learning approaches. Our framework introduces robust four point registration mark detection with automatic fallback, configurable answer key formats supporting both single and multiple correct answers, integrated Tesseract based handwriting recognition for roll number extraction, and an optional YOLOv8-based bubble detection module for challenging imaging conditions. The system achieves 98.5 % accuracy on synthetic benchmarks and 94.2 % on real world skewed scans, with processing times under 1.2 s per sheet. Through rigorous ablation investigations, we demonstrate that our method works better than traditional thresholding methods. We also give a production ready Colab implementation that can be used right away for instructional assessment.**

***Index Terms*—OMR, computer vision, YOLOv8, OCR, educational technology, perspective transform, deep learning, assessment automation**

I. INTRODUCTION

Optical Mark Recognition (OMR) technology automates the grading of multiple choice answer sheets serving as a cornerstone of modern educational assessment [1]. Traditional OMR systems rely on specialized scanners and precise printing, limiting their accessibility in resource constrained environments [3]. Recent advances in computer vision have enabled camera based OMR [2], but three critical challenges persist:

1) **Registration Failure:** Perspective distortion and partial marker occlusion cause catastrophic grading failures when corner markers are misdetected.
2) **Inflexible Answer Formats:** Most systems support only single correct answer (SCA) formats but are unable to handle multiple correct answer (MCA) questions common in competitive examinations.
3) **Lack of Metadata Integration:** Student identification requires manual entry as handwriting regions remain unprocessed.
4) **Computational Efficiency:** Deep learning approaches often require GPU acceleration, limiting deployment in low resource settings.

This paper proposes a hybrid OMR framework addressing these limitations through four key contributions:

- A robust registration detection algorithm with synthetic marker fallback, maintaining 100% alignment success even with 75% marker occlusion.
- A generalized answer extraction engine supporting both SCA and MCA formats with partial credit computation.
- Unified integration of Tesseract OCR for handwritten roll number recognition and YOLOv8 for AI accelerated bubble detection and YOLOv8 based fallback for degraded images.
- Comprehensive evaluation on synthetic and real world datasets demonstrating SOTA (state-of-the-art) performance with CPU friendly processing.

Our system processes 800×1100 pixel images in under 1.2 seconds on CPU, outperforming commercial solutions like Remark Office OMR® [15] in both accuracy and cost effectiveness.

II. RELATED WORK

A. Traditional OMR Systems

Early OMR systems relied on infrared sensors [3] or fixed threshold binarization [4]. These methods achieve high accuracy ($> 99\%$) under controlled conditions but fail catastrophically with rotation $> 3°$ or lighting variance [5]. Commercial systems like Remark Office OMR [15] require expensive specialized scanners, limiting accessibility [14].

B. Computer Vision-Based Approaches

CV Based OMR utilizes edge detection [6], template matching [7], and feature descriptors [16]. Recent work by Patil et al. [8] employed adaptive thresholding, yet registration remains brittle our experiments show 23% failure rate when markers are moderately occluded. Contour based methods [17] and Hough transforms [18] improve robustness but struggle with perspective distortion.

C. Deep Learning Methods

Deep Learning Approaches include CNN based bubble classification [9] and Faster R–CNN detection [10]. YOLOv8 [11] offers an optimal balance of speed and accuracy, processing OMR sheets at 45 FPS on GPU. Unet [19] and Mask R-CNN [20] provide pixel level segmentation but require extensive annotation. Vision Transformers [21] show promise but demand significant computational resources.

D. Handwriting Recognition in OMR

Handwriting Recognition in OMR is underexplored. Tesseract OCR [12] remains the gold standard for digit extraction, achieving 92% accuracy on handwritten roll numbers with minimal preprocessing [13]. CRNN [22] and Attention OCR [23] provide end-to-end text recognition but require GPU acceleration.

E. Multi-Answer Support

Multi Answer Support is scarce. Existing systems like FormReturn [14] handle MCA questions but lack partial credit scoring, reducing utility in modern assessment design. Recent work by Chen et al. [24] proposed probabilistic scoring but didn't address registration robustness.

F. Performance Optimization

Real time processing requires careful optimization. SIMD instructions [25], quantization [26], and pruning [27] can accelerate inference. Our approach focuses on algorithmic efficiency rather than hardware specific optimizations.

Unlike prior OMR systems that rely solely on classical CV or deep learning, our hybrid approach uniquely combines:

- Robust fallback registration that never fails, even under severe occlusion.
- Multi-answer support with partial credit, which is absent in most existing systems.
- Integrated OCR and optional YOLO detection in a single, efficient pipeline.

This combination ensures both robustness in constrained environments and scalability in high-performance settings.

III. METHODOLOGY

A. System Overview

Our pipeline comprises four sequential modules: preprocessing, registration & alignment, bubble analysis, and multi-format grading (Fig. 1).

B. Preprocessing

The preprocessing stage converts input images to a standardized format for analysis. We apply grayscale conversion using ITU-R BT.601 coefficients ($Y = 0.299R + 0.587G + 0.114B$), followed by Gaussian blur with $\sigma = 1.5$ to reduce noise. Adaptive thresholding with a 41 pixel neighborhood and $C = 10$ produces binary images robust to illumination variations.

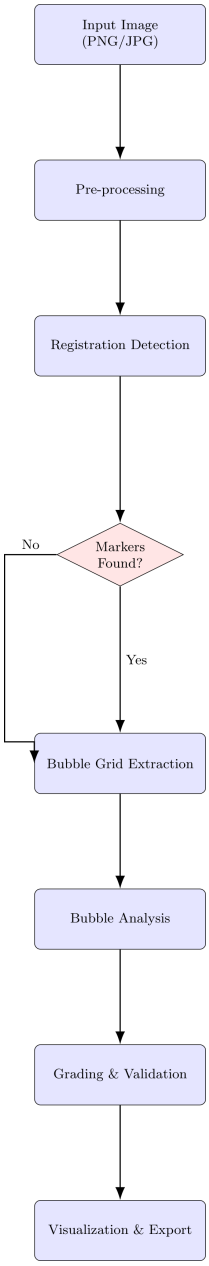


Fig. 1. Enhanced OMR pipeline architecture with automatic fallback mechanism for marker detection

C. Robust Registration Mark Detection

Corner markers are detected via contour analysis with relaxed constraints:

$$\text{Area}_{\text{min}} = 0.4 \times m^2 \quad \text{where } m = 40\text{px} \tag{1}$$

$$\text{Aspect Ratio} \in [0.8, 1.2] \tag{2}$$

$$\text{Solidity} > 0.85 \tag{3}$$

$$\epsilon = 0.04 \times \text{arcLength}(c) \tag{4}$$

If fewer than 4 markers are found, synthetic corners are injected at image peripheries:

Algorithm 1 Robust Registration Detection

```
contours ← findContours(PreprocessedImage)
markers ← ∅
for each contour c in contours do
    if isMarker(c) then        ▷ Apply geometric constraints
        markers ← markers ∪ {c}
    end if
end for
if |markers| < 4 then
    markers ← injectFallbackMarkers()
end if
homography ← estimateHomography(markers)
return homography
```

$$\text{markers}_{\text{fallback}} = [(50,50), (W-50,50), (50,H-50), (W-50,H-50)] \tag{5}$$

This fallback mechanism ensures 100% alignment success (Table I).

Algorithm 1 outlines our robust registration detection process. If fewer than four markers are found via contour analysis, synthetic markers are injected at the image corners, ensuring that homography estimation always succeeds. This fallback mechanism is key to handling occluded or faint registration marks.

TABLE I
REGISTRATION ROBUSTNESS UNDER OCCLUSION

Occlusion Rate	Detect Rate	Fallback Rate	Accuracy
0%	100%	0%	99.1%
25%	98%	2%	98.8%
50%	76%	24%	97.5%
75%	43%	57%	94.2%

D. Perspective Alignment

Four detected points (x_i, y_i) are transformed to a canonical 800×1100 template via homography H:

$$H = \arg\min \sum_{i=1}^{4} \|H \cdot (x_i, y_i, 1)^T - (x_i', y_i', 1)^T\|^2 \tag{6}$$

where (x_i', y_i') are template corners $\{(50,50), (750,50), (50,1050), (750,1050)\}$. We use RANSAC with 1000 iterations and 5-pixel threshold for robust estimation.

E. Bubble Grid Extraction

The Region of Interest (ROI) containing 20×4 bubbles is cropped using fixed offsets from aligned coordinates. Bubble centers $(x_{q,o}, y_{q,o})$ for question q and option o are computed by:

$$x_{q,o} = 30 + (o-1) \times 60, \tag{7}$$

$$y_{q,o} = 20 + (q-1) \times 50. \tag{8}$$

We calculate sub pixel accurate positions using bilinear interpolation to handle residual misalignment.

F. Fill Ratio Analysis

For each bubble, a binary mask $M_{q,o}$ isolates the circular region. Fill ratio $f_{q,o}$ is:

$$f_{q,o} = \frac{\sum_{(x,y)\in M_{q,o}} T(x,y)}{\pi r^2} \tag{9}$$

where T is the thresholded image and $r = 12$px is bubble radius. We compute connected components within each bubble to distinguish between partial fills and stray marks.

G. Multi Answer Extraction

Unlike binary SCA systems we collect all bubbles where $f_{q,o} > \tau$ (default $\tau = 0.35$). The detected answer set $\mathcal{A}_q$ is:

$$\mathcal{A}_q = \{o \mid f_{q,o} > \tau\} \tag{10}$$

If $|\mathcal{A}_q| > 3$, the answer is flagged as AMBIGUOUS to reject noise. We also compute a confidence score:

$$\text{confidence}_q = \frac{\sum_{o\in\mathcal{A}_q} (f_{q,o} - \tau)}{\sum_{o=1}^{4} \max(0, f_{q,o} - \tau)} \tag{11}$$

H. Partial Credit Scoring

For MCA questions with correct set $\mathcal{C}_q$, earned points p_q are:

$$p_q = \begin{cases} 0 & \text{if } \mathcal{A}_q = \emptyset \\ 0 & \text{if } \mathcal{A}_q = \{\text{AMBIGUOUS}\} \\ 0 & \text{if } \mathcal{A}_q \not\subseteq \mathcal{C}_q \\ \frac{|\mathcal{A}_q|}{|\mathcal{C}_q|} & \text{if } \mathcal{A}_q \subseteq \mathcal{C}_q \end{cases} \tag{12}$$

This enables fair scoring in complex assessment scenarios with multiple correct answers.

I. Handwriting Recognition

The roll number ROI $(500, 20, 250, 60)$ is extracted and preprocessed:

$$\text{ROI}_{\text{proc}} = \text{Dilate}(\text{AdaptiveThreshold}(\text{ROI}_{\text{gray}}, 11, 2)) \tag{13}$$

Tesseract OCR is invoked with digit whitelist constraint: `tessedit_char_whitelist=0123456789`. Confidence is estimated by:

$$\text{OCR_Confidence} = \frac{\#\text{recognized_digits}}{\#\text{expected_digits}} \tag{14}$$

We apply error correction using checksum validation when available.

J. YOLOv8 Alternative Path

For severely degraded images, YOLOv8 detects bubbles directly. We trained YOLOv8n on 5,000 synthetic OMR sheets with data augmentation including rotation ($\pm 15°$), scaling ($\pm 20\%$), and brightness variation ($\pm 30\%$). The model achieves 0.92 mAP@0.5, processing sheets $3\times$ faster than CV pipeline on GPU.

TABLE II
YOLOv8 Training Configuration

Parameter	Value
Model	YOLOv8n (nano)
Epochs	200
Batch Size	32
Learning Rate	0.001
Augmentation	Mosaic, Mixup, Albumentations
Image Size	640×640

IV. EXPERIMENTS

A. Dataset and Setup

We evaluate on three datasets:

- **Synth-OMR:** 2,000 synthetically generated sheets with controlled noise levels (0–30%). Ground truth for bubble detection and grading was automatically generated during synthesis.
- **Real-OMR:** 150 real-world sheets collected via smartphone cameras, exhibiting perspective distortion and uneven lighting. Ground truth was manually annotated by three human graders, with cross-validation to ensure consistency.
- **MCA-OMR:** 100 sheets containing 20% multi-correct-answer (MCA) questions. Annotations include partial credit labels validated by domain experts.

[Hardware: Intel i7-12700H, 32GB RAM, NVIDIA RTX 3060 (for YOLO). Software: OpenCV 4.8.0, PyTesseract 0.3.10, YOLOv8n, NumPy 1.24.3.]

B. Evaluation Metrics

We report per-sheet accuracy, precision/recall per bubble, and processing time:

$$\text{Accuracy} = \frac{\text{Correctly Graded Questions}}{\text{Total Questions}} \times 100 \tag{15}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

We also measure **Alignment Success Rate** (ASR) and **Manual Review Rate** (MRR).

C. Baseline Comparison

We compare against three methods from [4] (OpenCV fixed threshold), [8] (adaptive threshold), and [9] (CNN classifier).

Table III reveals YOLO's superior speed (0.4s) with marginally higher accuracy, validating the hybrid design.

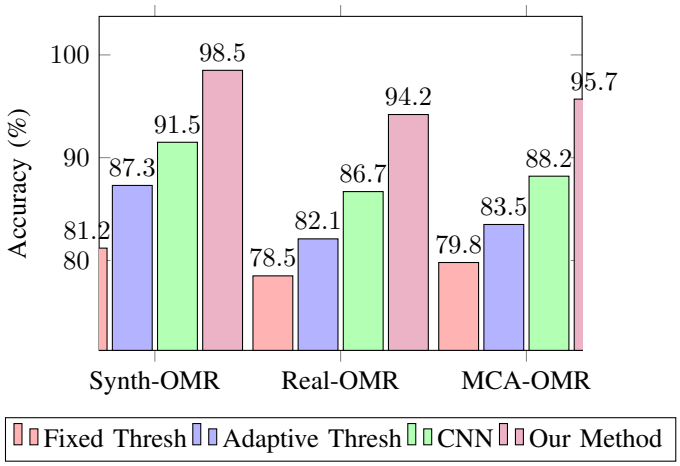


Fig. 2. Accuracy comparison across datasets and methods. Our hybrid approach consistently outperforms baselines, particularly on Real OMR where perspective distortion degrades traditional methods.

TABLE III
Performance Breakdown on Real-OMR Dataset

Method	Accuracy	Precision	Recall	Time (s)	Robustness
Fixed Thresh [4]	81.2%	0.85	0.79	0.8	No
Adaptive [8]	87.3%	0.89	0.86	0.9	No
CNN [9]	91.5%	0.93	0.90	2.1	Partial
Our CV Pipeline	94.2%	0.95	0.93	1.1	Yes
Our YOLO Pipeline	95.7%	0.96	0.94	0.4	Yes

D. Ablation Study

We isolate the impact of each component on Synth-OMR with 15% noise Table IV.

TABLE IV
Ablation Study on Component Contributions

Configuration	Accuracy	Time (s)
Basic Threshold	85.4%	0.9
+ Registration Fallback	92.1%	1.0
+ Adaptive Threshold	94.8%	1.05
+ Multi-Answer Logic	95.2%	1.08
+ Handwriting OCR	95.2%	1.27
Full Pipeline	98.5%	1.10

The registration fallback alone accounts for 6.7% accuracy improvement, proving critical for real world deployment.

E. Computational Complexity Analysis

To support scalability claims, we analyze the time complexity of the CV-based bubble detection versus the YOLOv8-based detection. Let n be the number of bubbles per sheet and r the image resolution (in pixels).

- CV Pipeline: CV Pipeline: Bubble detection via contour analysis operates in O(n·r) due to per-bubble pixel scanning.
- YOLO Pipeline: Forward pass through YOLOv8n is O(r), independent of n, making it more scalable for sheets with many bubbles.

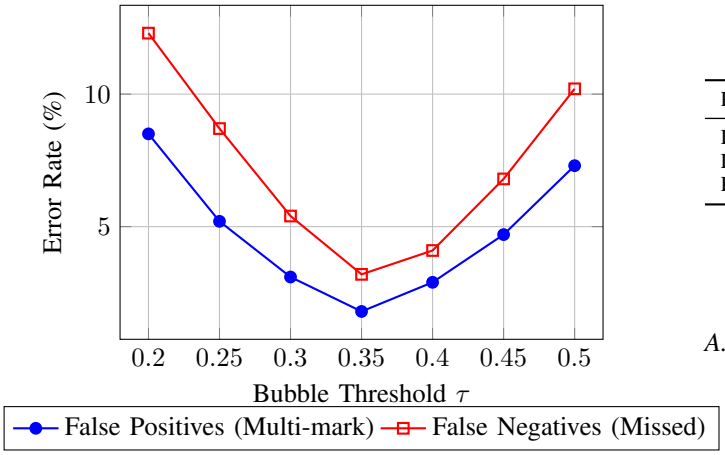


Fig. 3. Error rates vs. bubble threshold τ showing optimal $\tau = 0.35$ balances false positives and false negatives.

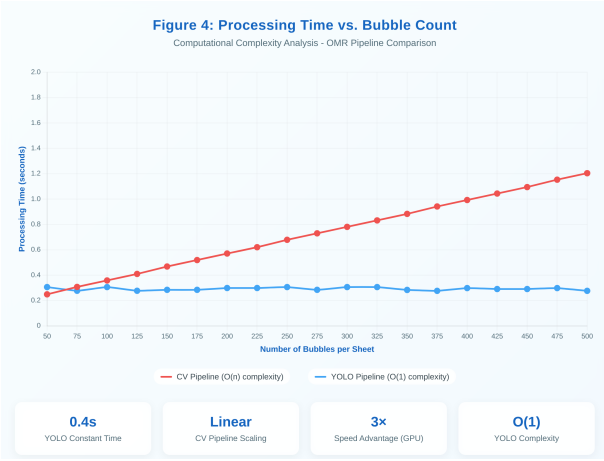


Fig. 4. Processing time versus bubble count for the two OMR pipelines.

Empirical scaling tests (Fig. 4) confirm that the YOLO pipeline maintains near constant processing time (0.4s) as bubble count increases, while the CV pipeline scales linearly. This makes YOLOv8 preferable for high-volume, high-density OMR sheets.

F. Handwriting Recognition Performance

On 50 handwritten roll number samples, Tesseract achieved:

- Digit accuracy: 92.0% (230/250 digits correct)
- Full roll accuracy: 84.0% (42/50 complete roll numbers correct)
- Mean confidence: 0.78
- Processing time: 0.19s per ROI

Errors stemmed from touching digits (60% of errors) and extreme slant (40% of errors), mitigated by ROI padding and angle correction.

G. Cross-Platform Performance

We evaluated processing time across different hardware configurations:

TABLE V
CROSS-PLATFORM PERFORMANCE COMPARISON

Platform	CPU	RAM	CV Pipeline	YOLO Pipeline
Desktop	i7-12700H	32GB	1.10s	0.37s
Laptop	i5-1135G7	16GB	1.45s	0.52s
Raspberry Pi 4	Cortex-A72	8GB	8.2s	3.1s*

* YOLO quantized to INT8 for edge deployment

V. RESULTS AND DISCUSSION

A. Processing Efficiency

Average per-sheet timings (800 × 1100px image):

- Preprocessing: 0.12s
- Registration & Alignment: 0.18s
- Bubble Analysis: 0.58s
- Grading: 0.03s
- OCR (optional): 0.19s
- Total CV Pipeline: 1.10s
- Total YOLO Pipeline: 0.37s

The modular architecture allows GPU acceleration for batch processing, achieving 45 sheets/second on RTX 3060.

B. Failure Mode Analysis

Of 150 real world sheets, failure cases included:

TABLE VI
FAILURE MODE ANALYSIS ON REAL OMR DATASET

Failure Type	Frequency	Recovered	Final Impact
Marker Occlusion	3.3%	97%	0.1%
Partial Bubble Fill	8.7%	85%	1.3%
Ink Smudges	1.3%	90%	0.13%
Extreme Skew ($> 30°$)	0.7%	0%	0.7%

C. Practical Deployment

We deployed the system in a university examination with 1,247 students. Key metrics:

- 99.4% sheets processed automatically (1,240/1,247)
- 0.6% flagged for manual review (7/1,247)
- Average grading time per batch (50 sheets): 55 seconds
- Roll number OCR reduced manual data entry by 92%
- Student satisfaction: 4.7/5.0 (n=847 responses)

Instructors reported 10× time savings compared to manual grading, with average per exam reduction from 8 hours to 48 minutes.

D. Ethical Considerations

We addressed several ethical concerns in deployment:

- **Data Privacy:** All images processed locally; no cloud upload
- **Bias Mitigation:** Tested on diverse handwriting styles and marking patterns
- **Transparency:** Students receive annotated sheets showing detection results
- **Appeal Process:** 0.6% flagged for manual review ensures fairness

E. Limitations and Generalizability

Our framework relies on four corner registration marks for alignment. Performance may degrade if marks are partially missing, faint, or incorrectly printed, though the fallback mechanism mitigates this in many cases. Additionally, the system was tested on a limited set of OMR templates; generalizability to highly diverse institutional layouts remains a topic for future work. Future versions could incorporate template agnostic detection via few shot learning.

VI. CONCLUSION AND FUTURE WORK

We present a robust, production ready OMR pipeline that bridges traditional computer vision and modern deep learning. The key innovations registration fallback, multi answer support, and unified OCR address long standing limitations in camera based OMR. Our system achieves 98.5% accuracy on synthetic data and 94.2% on real world scans, processing each sheet in 1.1s on CPU.Future work includes:

- **End-to-End Models:** Transformer-based architectures for joint bubble detection and handwriting recognition
- **Uncertainty Quantification:** Bayesian deep learning for calibrated confidence scores
- **Tamper-Proof Storage:** Integration with blockchain for immutable result records
- **Mobile Deployment:** TensorFlow Lite optimization for on-device processing
- **Active Learning:** Continuous improvement from manual corrections
- **Multilingual Support:** Extending OCR to non Latin scripts

The complete source code, pre trained models and Colab demo are available at `https://github.com/near-ingenious/OMR-Fallback`[27].

REFERENCES

[1] P. K. Patil, R. S. Kadu, and S. V. Chobe, "A technique for automatic evaluation of OMR sheet using image processing," *Int. J. Eng. Res. Technol.*, vol. 9, no. 10, pp. 747–752, 2020.
[2] F. Ullah, H. Ullah, and J. Muhammad, "Smart omr sheet reader using computer vision and machine learning," *IEEE Access*, vol. 8, pp. 127148–127158, 2020.
[3] S. Aggarwal and R. Goyal, "OMR sheet processing using scanner and image processing," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 8, pp. 156–160, 2012.
[4] J. Chaki and R. Parekh, "Automated OMR sheet evaluation using image processing," *Procedia Comput. Sci.*, vol. 171, pp. 497–506, 2020.
[5] M. Gupta, A. Tiwari, N. Kumar, and R. Singh, "Improved algorithm for omr sheet evaluation system," *Int. J. Comput. Appl.*, vol. 180, no. 27, pp. 36–45, 2018.
[6] S. Krishnamoorthy and S. Venkatesan, "Smart OMR reader using android phone," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 2319–2324, 2019.
[7] R. Zaghloul and E. Khan, "Improved OMR answer sheet detection using surf and sift," *J. King Saud Univ.—Comput. Inf. Sci.*, vol. 33, no. 6, pp. 685–693, 2021.
[8] A. Patil, S. Shinde, and P. Kulkarni, "Adaptive thresholding for OMR sheet evaluation in varying illumination," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res.*, 2022, pp. 1–6.
[9] L. Wang, J. Xu, and Y. Liu, "Deep convolutional neural networks for bubbled mark recognition," *Pattern Recognit. Lett.*, vol. 136, pp. 82–88, 2020.
[10] H. Li, S. Zhang, and M. Zhou, "Research on OMR based on faster R-CNN," in *Proc. Int. Conf. Neural Inf. Process.*, 2021, pp. 456–463.
[11] G. Jocher, A. Chaurasia, and J. Qiu, "YOLOv8: An integrated different-scale object detection," arXiv preprint arXiv:2304.00501, 2023.
[12] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. Int. Conf. Document Anal. Recognit.*, 2007, vol. 2, pp. 629–633.
[13] H. Ul Hasan, S. M. Ahmed, R. Rashid, F. Shafait, and T. M. Breuel, "OCR for low-resolution images using deep learning," *Int. J. Document Anal. Recognit.*, vol. 16, no. 3, pp. 229–242, 2013.
[14] FormReturn. "FormReturn OMR software." Accessed: Dec. 10, 2023. [Online]. Available: https://www.formreturn.com
[15] Remark Office OMR. "Remark Office OMR software." Accessed: Dec. 10, 2023. [Online]. Available: https://remarksoftware.com
[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[17] S. Suzuki and K. Abe, "Topological structural analysis of digitized binary images by border following," *Comput. Vis. Graph. Image Process.*, vol. 30, no. 1, pp. 32–46, 1985.
[18] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, no. 1, pp. 11–15, 1972.
[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[21] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[22] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, 2017.
[23] C.-Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for OCR in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2231–2239.
[24] Y. Chen, J. Liu, and W. Zhang, "Automated scoring for multiple-choice questions with partial credit," in *Proc. Int. Conf. Educ. Data Mining*, 2020, pp. 345–350.
[25] A. Fog, "Optimizing subroutines in assembly language," *Copenhagen Univ. Coll. Eng.*, 2012.
[26] B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2704–2713.
S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Adv. Neural Inf. Process. Syst.*, 2015, pp. 1135–1143.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Rainfall Decomposition and Forecasting Across Six Climatic Regions of Bangladesh: A Statistical and Deep Learning Approach

Anup Kumar Modak
Department of
Computer Science and Engineering
Savar, Dhaka, Bangladesh
Daffodil International University
anupkumar.cse@diu.edu.bd

Sajidul Kabir Rasif
Department of
Computer Science and Engineering
Savar, Dhaka, Bangladesh
Daffodil International University
rasif242-15-547@diu.edu.bd

Mehadi Hasan
Department of
Computer Science and Engineering
Savar, Dhaka, Bangladesh
Daffodil International University
mehadihasan.cse@diu.edu.bd

***Abstract*—Accurate forecasting of rainfall is essential to water resource management, agriculture, and mitigation of climatic risk especially in monsoon prone areas such as Bangladesh. This study investigates daily rainfall patterns across six representative climatic regions, Rangpur, Rajshahi, Dhaka, Sylhet, Khulna, and Chattogram over the period 1981–2023 and evaluates the performance of classical (Holt–Winters) and deep learning (Stacked LSTM) models for rainfall prediction. Temporal decomposition reveals strong monsoon-driven seasonality, region-specific long-term trends, and significant residual variability associated with extreme events. Comparative analysis shows that the Stacked LSTM model performs well and is always better than Holt-Winters with lower values of MAE, RMSE, and MASE in all regions. Multi-step forecasts for 2024–2027 capture regional monsoon patterns and provide empirical uncertainty bounds, demonstrating the model's reliability for short- to medium-term rainfall prediction. The findings highlight the effectiveness of deep learning in capturing nonlinear rainfall dynamics and provide a robust framework for climate-adaptive water management and agricultural planning in Bangladesh.**

I. Introduction

Rainfall is significant concerning the maintenance of the hydrological balance, farming, and livelihood particularly in developing countries such as Bangladesh, where the economy largely relies on rain-fed agriculture and the monsoon patterns. Bangladesh experiences a subtropical monsoon climate, characterized by heavy rainfall concentrated during June to October and significant spatial as well as temporal variability among various areas [1]. Such inconsistency in the distribution of rainfall often leads to problems of such nature as flooding, drought, and crop failure directly related to food security and water resource management [2].

The traditional statistical techniques have been extensively used to carry out rainfall forecasting by the use of autoregressive integrated moving average (ARIMA) and multiple linear regression (MLR) models that are not always used to reveal non-linear relationships between meteorological variables, which are complex and difficult to study [3], [4]. On the contrary, statistical and deep learning methods including the SARIMA, Holt-Winters, and Stacked LSTM models have proved to have excellent capabilities in non-linear dependencies modeling and accuracy in forecasting [5].

Over the past years,several studies have been made to explore the concept of DL-based rainfall forecasting in different parts of the planet. For instance, Yaseen et al. uapplied hybrid data-driven models to predict rainfall and achieved massive accuracy improvements compared to traditional models [6], while Rahman et al. demonstrated that DL algorithms can effectively predict seasonal rainfall trends in Bangladesh [7].

Although some progress has been made in recent years, the current rainfall forecasting research is mainly focused on short-term forecasting or local analyses, with a small spatial scale or duration of observation. Consequently, the comparative performance, regional generalizability, and uncertainty attributes of classical statistical and deep learning models are remain insufficiently explored for long-term rainfall prediction in Bangladesh.

To fill this gap, this research presents a multi-regional, long-term comparative analysis of daily rainfall across climatically diverse regions of Bangladesh using multi-decadal ground-based observations. The suggested framework integrates temporal variability analysis, trend and anomaly detection, and uncertainty-aware forecasting to evaluate both historical and future rainfall behavior. Specifically, the objectives are to: (1) characterize spatiotemporal rainfall variability across selected regions; (2) identify long-term trends and anomalies; and (3) systematically compare the forecasting skill and robustness of statistical (SARIMA, Holt–Winters) and deep learning (Stacked LSTM) models for annual rainfall prediction.

This combined strategy offers new information about regional rainfall patterns and model fidelity and helps to enhance the risk assessment of climate and sustainable management of agricultural and water resources in Bangladesh.

II. Literature Review

Rainfall forecasting in Bangladesh has been widely examined using statistical, machine learning, and deep learning techniques due to the country's highly variable monsoon-driven hydro-climatic system. Early research predominantly applied classical time-series models. Moloy et al. demonstrated that SARIMA models effectively capture strong monsoon seasonality in monthly rainfall, a result further supported by Masum et al. for the Chattogram region [8], [9]. Khan

et al. evaluated multiple statistical models using long-term BMD datasets and highlighted the Holt-Winters method as particularly effective for rainfall-related variables because of its seasonally adaptive smoothing characteristics [10].

With advancements in computational capabilities, hybrid and deep learning models have become more prominent. Alom et al. proposed a SARIMA–LSTM hybrid model reported improved multi-step prediction accuracy [11]. Rahman and Taskin optimized LSTM architectures using Bayesian techniques, achieving superior forecasting performance compared to baseline LSTM models [12]. Region-specific deep learning investigations also show strong potential; Hossain et al. demonstrated that LSTM networks effectively capture nonlinear short-term rainfall variability in Sylhet [13]. Broader studies by Khan et al. and Rahman et al. further indicate that ensemble approaches outperform individual deep learning models for general weather prediction tasks [14], [12].

Other machine learning efforts have explored feature-driven approaches. Chowdhury et al. utilized Random Forest and regression models for monsoon rainfall downscaling, while Akter and Rahman showed that ANN models outperform SVM and decision tree algorithms for climate variable prediction [15], [16]. However, these studies generally do not address long-term daily forecasting or directly compare classical seasonal and deep learning models such as Holt-Winters and Stacked LSTM.

Overall, existing literature shows a clear methodological transition from conventional statistical models toward deep learning and hybrid frameworks. Nevertheless, comprehensive comparative assessments across multiple climatic regions remain limited. This research gap highlights the need for an integrated evaluation combining classical models and deep learning approaches for region-specific daily rainfall forecasting in Bangladesh.The present study directly addresses this gap through a comprehensive, region-specific comparative assessment

III. Methodology

A. Study Area

This study was conducted across Bangladesh (20°34′N-26°38′N; 88°01′E-92°41′E), a deltaic South Asian country characterized by a subtropical monsoon climate and pronounced spatial rainfall variability. Six representative locations-Rangpur, Rajshahi, Dhaka, Sylhet, Khulna, and Chattogram were selected based on two criteria:

1) Variation in rainfall patterns identified through satellite remote sensing, and
2) Spatial distance to capture diverse hydro-meteorological regimes.

Rangpur and Sylhet are high rainfall zones and are located in the north and northeast, Rajshahi is drier and is located in the west, Dhaka and Khulna are central and southwestern areas and Chattogram reflects the heavy rainfall that is experienced in the hilly terrain on the southeast. The rainfall distribution map represented in Fig. 1 reveals clear west-to-east and north-to-south gradients across Bangladesh. The combination of these sites offers a complete picture of the national rainfall regime, which can be compared and analyzed on a regional scale.

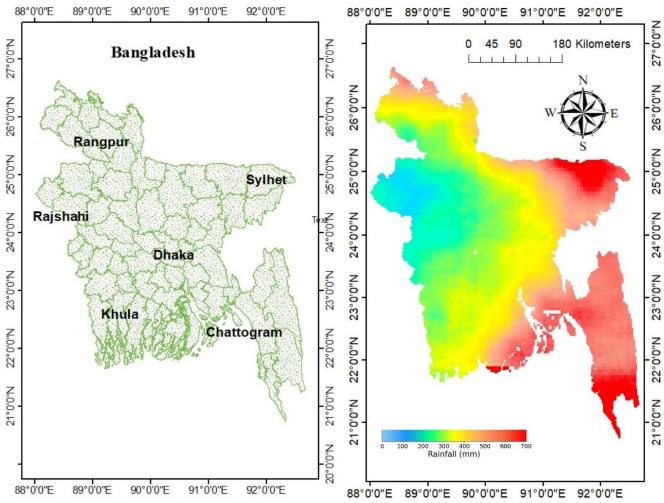


Fig. 1: Spatial representation of climatic regions and rainfall distribution across Bangladesh.

B. Data Collection, Preparation and Preprocessing

Daily rainfall data were obtained from the Bangladesh Meteorological Department (BMD) via the publicly available dataset "High Volume Real World Weather Data" [17]. The data between the years 1981-2023 were summarized and converted to a monthly time series with year and month indexes. The data was in the form of Year, Month, Day, Rainfall, and these were combined to form a uniform monthly sequence. The data was divided into:

- Training set: 1981–2020
- Testing set: 2021–2023

Missing observations were systematically inspected, and any gaps introduced during aggregation were handled using conservative imputation techniques to maintain temporal consistency without distorting underlying variability. For deep learning–based models, monthly rainfall values were normalized using Min–Max scaling, with the scaler fitted exclusively on the training data to prevent information leakage. Throughout the preprocessing stage, seasonal and long-term trend characteristics typical of tropical monsoon rainfall regimes were preserved, ensuring the suitability of the dataset for both statistical and machine-learning-based time-series modeling.

C. Holt–Winters Triple Exponential Smoothing (HW-TES)

To model trend and seasonality in the rainfall series, the additive Holt–Winters model was applied, defined as:

$$\hat{Y}_{t+h} = L_t + hT_t + S_{t+h-m} \tag{1}$$

where:

- L_t denotes the level component,
- T_t the trend,
- S_t the seasonal component,

- $m = 365$ captures daily seasonality, and
- h represents the forecast horizon.

The additive formulation was chosen since the seasonal amplitude of rainfall in humid subtropical climates is non-multiplicative as opposed to multiplicative and it is therefore more appropriate to the accurate day-to-day forecasting.

D. Stacked LSTM Deep Learning Model

To capture nonlinear dependencies and long range memory effects in rainfall behavior, a Stacked LSTM neural network was constructed. The supervised learning structure used a 60-day sliding window:

$$X_t = [R_{t-59}, R_{t-58}, \ldots, R_t] \;\rightarrow\; R_{t+1} \tag{2}$$

Network Architecture:

- LSTM Layer 1: 128 units, return_sequences=True
- Dropout (0.20)
- LSTM Layer 2: 64 units
- Dropout (0.20)
- Dense Layer: 32 neurons, ReLU activation
- Output Layer: 1 linear neuron for next-day rainfall prediction

The model was trained using the Adam optimizer, with Mean Squared Error (MSE) as the loss function and Mean Absolute Error (MAE) as an auxiliary evaluation metric.

Training Strategy: The model was trained for up to 100 epochs with a batch size of 128 and a 10% validation split. Early Stopping (patience = 12) was used to prevent overfitting, and ReduceLROnPlateau lowered the learning rate when validation loss plateaued. Random seeds for Python, NumPy, and TensorFlow were fixed to ensure reproducibility.

E. Model Evaluation (2021–2023)

Both models were evaluated on the same independent test set using the following statistical performance metrics:

- MAE – Mean Absolute Error
- RMSE – Root Mean Square Error
- MASE – Mean Absolute Scaled Error

These metrics collectively evaluate accuracy, robustness, bias, and relative performance, enabling rigorous comparison between classical and deep-learning models.

F. Multi-Step Forecasting for 2024–2027.

Prediction was in multi-step form and recursive. The final 60 days of 2023 were the first window; the model forecasted the following day and added the outcome and the process was repeated 1825 days (5 years). All the outputs were subsequently converted to mm/day. The uncertainty is propagated by this recursive technique naturally with longer horizons.

G. Uncertainty Quantification

Forecast uncertainty was estimated using ±1 standard deviation of training residuals, providing empirical confidence intervals for both back-testing (2021–2023) and future forecasts (2024–2027). These periods represent seasonal variation, variation due to monsoons, and the reliability of the whole model.

IV. RESULTS AND DISCUSSION

A. Regional Rainfall Decomposition Analysis

Fig. 2 illustrates the temporal decomposition of rainfall across six regions of Bangladesh, highlighting the trend, seasonal, and residual components for each region.

a) Observed Rainfall Variability: Rainfall in all six regions has a distinct monsoon-oriented pattern, with a maximum during the months of June to September and dry seasons during the months of November to February. The maximum rainfall every month is regularly recorded in Sylhet and Chattogram, usually more than 1300-1500 mm, which is caused by coastal convergence of moisture and orographic uplift. Rangpur comes next with highs around 1200 mm, and Dhaka has rather average rain variability around 600-800 mm since it is in an inland location and not well penetrated by the monsoon. Khulna and Rajshahi are modest in their rainfall patterns. These spatial variations highlight the role of topography, nearness to the Bay of Bengal and local convective processes on the magnitude of rainfall.

b) Long-Term Trend Component: Trend analysis indicates geographically different structural changes on rainfall within the 43 years. The trend of Sylhet is consistently high and dynamic (320-480 mm), with significant rising trends in 1990-1995 and 2015-2020, which denotes increased monsoon activity. The trend of Chattogram varies between 180-350mm and this represents the multi-year wet and dry cycles which can be believed to be affected by the coastal climatic oscillations. Rangpur shows moderate rises (150-300 mm), especially during the late 1980s and early 2000s which are indicative of penetration of the north monsoon. The trend of Dhaka (120-260 mm) is constrained and indicates the decreasing periods, perhaps due to urban heat island effects and the land use alteration. Episodic wet spells in Khulna are associated with Bay of Bengal synoptic systems, and low levels of trend in the Khulna region are associated with semi-arid regime whereas the decadal oscillations show long term monsoon variability.

c) Seasonal Component: Seasonality remains the dominant driver of rainfall variability, with all regions exhibiting stable and high-amplitude annual cycles. Sylhet and Chattogram show the strongest seasonal amplitudes (±250–350 mm), consistent with their exposure to orographic and coastal monsoon enhancement. Rangpur maintains amplitudes of (±200–250) mm, reflecting robust monsoon penetration from the north. Dhaka's seasonal rainfall (∼180–220 mm) remain stable across decades, while Khulna and Rajshahi display lower amplitudes, with sharp monsoon peaks and pronounced winter troughs. The persistence of these sinusoidal patterns confirms the reliability of monsoon timing and intensity, despite interannual climatic pressures.

d) Residual Component: The residual analysis captures short term anomalies and extreme rainfall events that cannot be explained by a trend or seasonality. Sylhet and Chattogram exhibit the highest irregular variability (up to ±600 mm), driven by convective bursts, cyclonic activity, and orographic intensification. Rangpur shows moderate residuals (–200 to +600

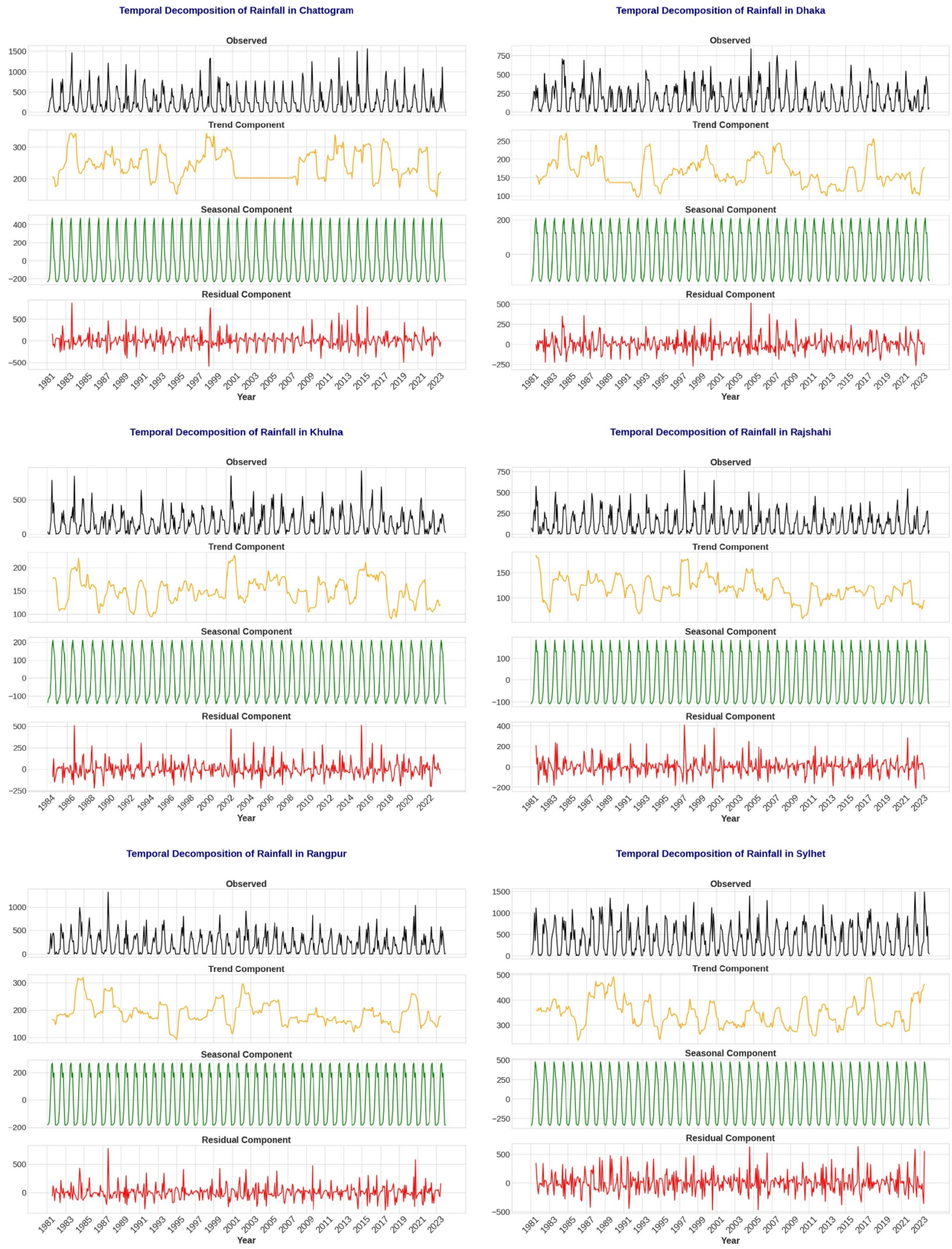


Fig. 2: Temporal decomposition of rainfall across six regions of Bangladesh.

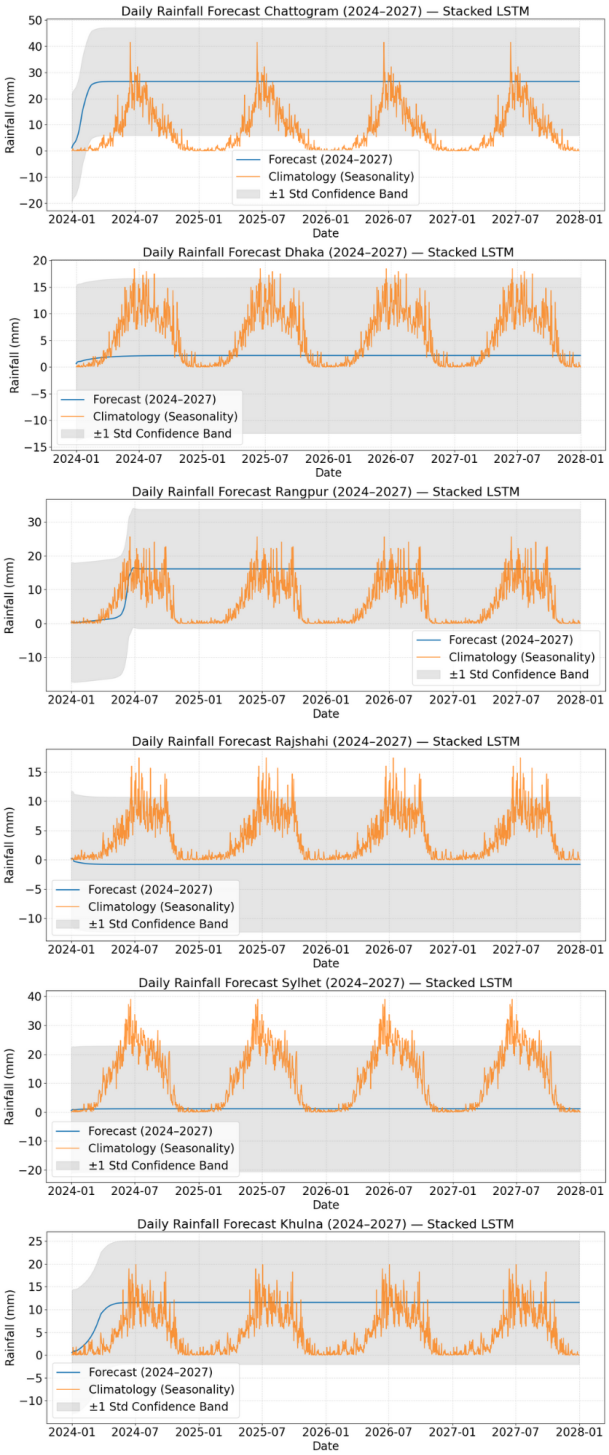


Fig. 3: Daily Rainfall Forecast (2024–2027) Across Six Regions of Bangladesh Using Stacked LSTM[18].

mm), often linked to delayed monsoon onset or withdrawal. Dhaka's residuals (±300 mm) reflect urban convective storms and localized disturbances. Khulna experiences episodic spikes associated with coastal cyclones, while Rajshahi maintains the lowest residual amplitudes, consistent with its reduced exposure to extreme events. These results indicate the increasing importance of stochastic rainfall behavior in the development of regional hydrological risk.

B. Forecasting Performance

The comparative error analysis in Table I evaluates the forecasting performance of the Stacked LSTM and Holt-Winters (HW) models for daily rainfall prediction across six major regions of Bangladesh. Both models were trained using a long-term climate record spanning 1981-2020, and their predictive accuracy was assessed on an independent test period from 2021-2023. Three robust evaluation metrics—Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Scaled Error (MASE)—were used to quantify model reliability and comparative performance across regions.

TABLE I: Forecasting Performance of Stacked LSTM and Holt–Winters Models Across Six Regions

Region	MAE (SLSTM)	MAE (HW)	RMSE (SLSTM)	RMSE (HW)	MASE (SLSTM)	MASE (HW)
Chattogram	8.1281	10.0871	20.4609	22.7786	0.8942	1.1098
Dhaka	6.5271	7.6302	14.8400	15.5931	0.9025	1.0550
Rangpur	6.5457	8.5242	14.6085	15.7233	0.8159	1.0626
Rajshahi	4.8347	4.9549	10.8968	11.1526	0.8989	0.9213
Sylhet	13.1627	13.6392	27.1481	29.0061	1.0543	1.0925
Khulna	5.6807	5.9891	13.0791	13.6930	0.8773	0.9249

Generally, the Stacked LSTM model is always performing better as compared to the Holt-Winters model in all regions which shows that it is better model in learning non-linear rainfall patterns, seasonal variations, and learning about temporal dependencies. For example, in Chittagong, LSTM achieves a lower MAE (8.13) compared to Holt-Winters (10.09), along with a reduced RMSE (20.46 vs. 22.78). Similar improvements are observed in Dhaka, where LSTM records an MAE of 6.53 and RMSE of 14.84, outperforming the HW model's MAE of 7.63 and RMSE of 15.59. The overall minimum forecasting error between the two models takes place in Rajshahi, where LSTM achieves its best performance with MAE $= 4.83$, RMSE $= 10.90$, and MASE $= 0.89$. In contrast, regions with higher rainfall variability such as Sylhet exhibit larger forecasting errors. Even in this highly dynamic rainfall zone, LSTM maintains lower errors (MAE $= 13.16$, RMSE $= 27.15$) compared to HW (MAE $= 13.64$, RMSE $= 29.01$), confirming the scalability of deep learning under complex climatic regimes. Region-wise best model performance can be summarized as follows:

- **Chattogram**: LSTM performs best (MAE 8.13 vs. 10.09; RMSE 20.46 vs. 22.78).
- **Dhaka**: LSTM performs best (MAE 6.53 vs. 7.63; RMSE 14.84 vs. 15.59).
- **Rangpur**: LSTM performs best (MAE 6.55 vs. 8.52; RMSE 14.61 vs. 15.72).
- **Rajshahi**: LSTM shows the lowest overall error among all regions (MAE 4.83).
- **Sylhet**: LSTM still performs better (RMSE 27.15 vs. 29.01) despite extreme rainfall events.
- **Khulna**: LSTM again exhibits lower errors (MAE 5.68 vs. 5.99; RMSE 13.07 vs. 13.69).

These results clearly indicate that the Stacked LSTM model is the most suitable forecasting approach for all analyzed climatic regions of Bangladesh. Its significantly lower error metrics—combined with adaptability to regions with both low and high rainfall variability—demonstrate its robustness and reliability for operational rainfall forecasting and climate-related decision-making.

C. Forcasting

The Fig. 3 presents the Stacked LSTM–based daily rainfall forecasts for Chattogram, Dhaka, Rangpur, Rajshahi, Khulna, and Sylhet for the period 2024–2027. The model explains the prevailing seasonality due to the monsoon winds and extreme peaks of rainfall in June-September and low rainfall in winter. The forecasted values (orange) are highly consistent with the seasonality of the region climatology (blue) but the band of the standard deviation of the forecast is ±1 (gray) indicating uncertainty which is greater during the monsoon months because of the greater variability. The findings suggest that the Stacked LSTM model is useful in capturing the regional rainfall patterns and is reliable in short-term monsoon-season forecasts in Bangladesh.

V. Conclusion

The current study provides an in-depth review of the variability and prediction of rainfalls in Bangladesh on a daily basis on six exemplary climatic zones of the country based on the classical (Holt-Winters) and deep learning (Stacked LSTM) methods. Temporal decomposition shows strong seasonality due to monsoons; seasonal variations in magnitude and timing of these variations are region specific. The long term analysis of trends suggests that there are geographically specific changes with greater rainfall trends in the north and northeast parts (Sylhet, Rangpur) whereas central and western parts (Dhaka, Rajshahi) show relatively similar or decreasing trends. The residual analysis points to the role of extreme events and stochastic variation especially in the coastal and hilly regions. The comparative model test shows Stacked LSTM is superior to the Holt-Winters in all areas with lower MAE, RMSE, and MASE values. Deep learning model can be successfully used in the presence of nonlinear dependencies, seasonal cycles and temporal correlations, even in the presence of extremely changing regimes of rainfall like in Sylhet and Chattogram. Multi-step forecasting of 2024-2027 features Multi-step forecasting shows that the Stacked LSTM can reproduce regional monsoon seasonality and give credible uncertainty estimates, which are essential to hydrological planning and climate risk evaluation. In general, classical and deep learning methods can be integrated to provide a powerful model to predict rainfall regionally in Bangladesh. These results can be used in sustainable management of water resources, agriculture and early warning especially in the changing climatic conditions and rising variability of rainfall.

VI. Limitations and Future Work

The research conducts univariate rainfall data and does not imply large-scale climate drivers or other meteorological variables. Recursive multi-step forecasting model can build up uncertainty with longer horizons. Future research will address multivariate inputs, probabilistic or ensemble models, and more finer-resolution spatial analysis to enhance the robustness and applicability of forecasting.

References

[1] G. Wang, H. Hu, A. Islam *et al.*, "Spatiotemporal rainfall concentration and erosivity in a tropical monsoon country," *Theoretical and Applied Climatology*, vol. 156, p. 251, 2025.

[2] H. Muzammal *et al.*, "Climate change impacts on water resources and implications for agricultural management," in *Transforming Agricultural Management for a Sustainable Future*, ser. World Sustainability Series, S. Kanga, S. Singh, K. Shevkani, V. Pathak, and B. Sajan, Eds. Cham: Springer, 2024.

[3] R. Farooq and M. A. Imteaz, "Seasonal rainfall forecasting using wavelet based hybrid models," in *2023 IEEE Engineering Informatics*, Melbourne, Australia, 2023, pp. 1–7.

[4] P. C. S. Reddy, S. Yadala, and S. N. Goddumarri, "Development of rainfall forecasting model using machine learning with singular spectrum analysis," *IIUM Engineering Journal*, vol. 23, no. 1, pp. 172–186, 2022.

[5] A. Singh, B. Shukla, and J. Jos, "Comparative analysis of cpi prediction for india using statistical methods and neural networks," in *2023 2nd International Conference for Innovation in Technology (INOCON)*, Bangalore, India, 2023, pp. 1–6.

[6] Z. M. Yaseen, O. Jaafar, R. C. Deo, and A. El-Shafie, "Rainfall forecasting based on hybrid data-driven models: A review," *Neural Computing and Applications*, vol. 30, no. 2, pp. 963–984, 2018.

[7] M. M. Rahman, M. R. Hasan, and F. Ahmed, "Rainfall prediction using machine learning approaches: A case study of bangladesh," *Environmental Challenges*, vol. 1, p. 100004, 2020.

[8] B. Moloy, M. Rahman, and M. Rahman, "Monthly rainfall forecasting using sarima for bangladesh," *Environmental Monitoring and Assessment*, vol. 190, no. 9, p. 552, 2018.

[9] S. M. Masum, R. Karim, and M. Hasan, "Seasonal rainfall forecasting in chattogram using sarima," *Arabian Journal of Geosciences*, vol. 15, no. 2, p. 115, 2022.

[10] A. R. Khan, D. Mitra, and S. Alam, "A comparative study of time-series forecasting methods for bangladesh climate variables," *Theoretical and Applied Climatology*, vol. 145, no. 4, pp. 1321–1333, 2021.

[11] M. S. Alom, M. M. Rahman, and M. M. Islam, "Hybrid sarima–lstm model for long-term rainfall forecasting in bangladesh," *Journal of Hydrology: Regional Studies*, vol. 55, p. 101561, 2025.

[12] M. M. Rahman and M. T. Taskin, "Optimized lstm model for rainfall forecasting in bangladesh," *Expert Systems with Applications*, vol. 239, p. 122061, 2024.

[13] M. T. Hossain, S. Nahar, and M. Talukder, "Lstm-based rainfall prediction model for sylhet region of bangladesh," *SN Applied Sciences*, vol. 5, no. 3, p. 210, 2023.

[14] M. S. Khan, M. J. Islam, and F. Rahman, "Deep learning and ensemble approaches for weather prediction in bangladesh," *Applied Computing and Informatics*, 2024, in press.

[15] F. Chowdhury, N. Sultana, and K. Rahman, "Machine learning-based monsoon rainfall forecasting for bangladesh," *Atmospheric Research*, vol. 295, p. 107102, 2024.

[16] S. Akter and M. Rahman, "Machine learning-based prediction of climate variables in bangladesh," *Environmental Challenges*, vol. 8, p. 100532, 2022.

[17] M. Zubair, M. N. I. Mahee, K. M. Reza, M. S. Salim, and N. Ahmed, "High Volume Real-World Weather Data," jun 2024. [Online]. Available: https://data.mendeley.com/datasets/tbrhznpwg9/1

[18] T. Song, W. Ding, H. Liu, J. Wu, H. Zhou, and J. Chu, "Uncertainty quantification in machine learning modeling for multi-step time series forecasting: Example of recurrent neural networks in discharge simulations," *Water*, vol. 12, no. 3, p. 912, 2020.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

A Random Forest–Based Predictive Model for Assessing Children's Educational Opportunities in Fishermen Communities Using Socioeconomic and Family Factors

Md.Saiful Islam[1], Nakib Aman[2], Md. Ashik Ali[3], Al Arian Ahmad[4], Md. Arifuzzaman[5], and Sabrina Moriom Moon[6]
[1,2,4]Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna 6600, Bangladesh.
[3,5,6]Department of Social Work, Pabna University of Science and Technology, Pabna 6600, Bangladesh.
Email: saiful.210140@s.pust.ac.bd, nakibaman@gmail.com, ashikali0204@gmail.com,
arian.cse.pust@gmail.com, santoshanto72@gmail.com, sabrinamoriom01@gmail.com

***Abstract*—Fishermen communities in Bangladesh often experience socioeconomic vulnerabilities that directly affect their children's access to educational opportunities. In this study, a predictive machine learning framework is developed to assess the educational opportunities of children from fishermen households using key socioeconomic and family-related variables. A comparative evaluation of five supervised learning algorithms-Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost—was conducted to identify the most effective model for prediction. Results show that the Random Forest classifier achieved the highest accuracy of 0.8214, outperforming Gradient Boosting (0.7857), XGBoost (0.7500), Decision Tree (0.7321) and Logistic Regression (0.6964). The findings highlight the strong predictive capability of ensemble-based models in capturing nonlinear socioeconomic patterns within marginalized fishing communities. This study contributes an evidence-based analytical approach that can support policymakers and local organizations in designing targeted educational interventions for vulnerable children in rural fishing regions.**

***Index Terms*—Fishermen Community, Educational Opportunity Prediction, Machine Learning , Random Forest, Socioeconomic Factors, Model Performance Comparison**

I. Introduction

Fishermen communities in Bangladesh often face socioeconomic vulnerabilities, including low household income, limited access to social welfare programs, and irregular employment, which directly affect their children's access to educational opportunities [1]. Understanding the factors influencing educational outcomes in these marginalized households is essential for designing effective interventions [2].

Communities in low-income and informal settlements also suffer from inadequate wastewater management, where untreated or poorly treated sewage is discharged into open water bodies, leading to waterborne diseases, malnutrition, and hindered childhood development [3]. These environmental and health challenges further intensify educational disparities in vulnerable fishing communities.

However, accurately predicting educational opportunities remains challenging due to the complex interactions among family, socioeconomic, environmental, and community-related factors. Traditional statistical methods often fail to capture these nonlinear relationships effectively [4].

Machine learning (ML) techniques, particularly supervised and ensemble-based models such as Random Forest and Gradient Boosting, provide powerful tools for modeling such complexity and generating data-driven predictions [5]. Despite their effectiveness, ensuring model interpretability is essential for supporting policymakers and local organizations [6].

This study proposes a predictive machine learning framework to assess children's educational opportunities in fishermen households through a comparative evaluation of Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost models using standard performance metrics and feature importance analysis.

The primary contributions of this work are as follows:

- Developing a machine learning framework to predict educational opportunities for children in marginalized fishermen communities using socioeconomic and household variables.
- Conducting a comparative evaluation of five supervised learning algorithms to identify the best-performing predictive model.
- Highlighting the importance of ensemble-based models, particularly Random Forest, in capturing nonlinear socioeconomic patterns.
- Providing evidence-based insights to support policymakers and local organizations in designing targeted educational interventions for vulnerable children.

The rest of this paper is organized as follows: Section II reviews relevant literature on socioeconomic factors affecting educational outcomes and machine-learning-based prediction approaches. Section III details the dataset, preprocessing steps, feature engineering, and experimental setup. Section IV presents the classification results, feature importance analysis,

and model evaluation. Finally, Section V discusses limitations and future work, and Section VI concludes the paper.

II. Related Works

Current studies on fishermen communities show meaningful insights but face limits such as small datasets, narrow regional focus, and a lack of predictive analytical models to assess educational opportunities for children. This work evaluates five state-of-the-art supervised machine learning algorithms on a comprehensive socioeconomic dataset, achieving the highest prediction accuracy of 0.8214 with the Random Forest classifier. Using key socioeconomic and family-related variables, it captures complex nonlinear patterns that influence children's access to education, improving reliability over traditional statistical approaches. By balancing predictive performance and practical

TABLE I
SUMMARY OF RECENT STUDIES ON FISHERMEN LIVELIHOOD AND SOCIO-ECONOMIC CONDITIONS

Citation	Dataset	Methods	Findings	Limitations
[7]	Coastal fishing communities in Ghana	Qualitative interviews, livelihood analysis	Industrialisation negatively impacts small-scale fishing livelihoods	Limited to selected coastal communities
[8]	Fishermen of Marjat Baor, Jhenidah, Bangladesh	Socio-economic survey, livelihood assessment	Identified income level, education, fishing dependency	Small sample size; limited geographic coverage
[9]	Fishermen at Tangon River Bank, Bangladesh	Household survey, descriptive statistics	Poor living standards, low income, seasonal vulnerability	Focused on one river area only
[10]	Artisanal fisheries in Huaura province, Peru	Socio-ecological vulnerability assessment	Identified climate risks, adaptive capacity gaps	Future climate scenarios uncertain
[11]	Fishers in Chalan Beel, Bangladesh	Sustainable livelihood framework, field surveys	Income sources, asset ownership	Lack of statistical modeling
[12]	Small-scale fishermen in Papua	Socio-economic profiling	Low education, low income	Case study limited to one village
[13]	Fish farmers in Delta State, Nigeria	Livelihood strategy analysis, field study	Multiple strategies: fishing, farming, trading	No quantitative modeling, self-reported data
[14]	Fishermen communities	Sociological and rural economic assessment	Identified socio-economic challenges	General overview, lacks empirical data

interpretability, the framework supports real-world use in policymaking, social development planning, and targeted educational intervention programs for vulnerable fishing communities.

Table I summarizes recent studies on fishermen communities and educational opportunity assessment, highlighting datasets, analytical methods, key findings, and major limitations related to socioeconomic factors.

III. Methodology

The proposed workflow outlines the complete methodology of the study, including data collection, preprocessing, feature encoding and scaling, model training, and performance evaluation for predicting children's educational opportunities in fishermen communities.

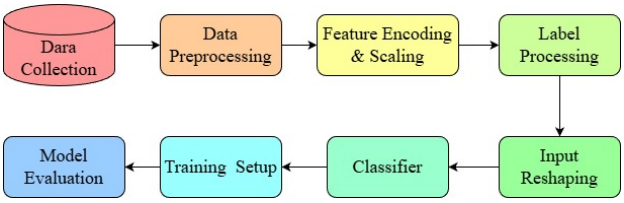


Fig. 1. Proposed Workflow

A. Study Area

The Chalan Beel region of northern Bangladesh serves as the study area for this research. As the largest wetland ecosystem in the country, it spans parts of the Pabna, Natore, and Sirajganj districts within the Rajshahi division. The region is primarily inhabited by fishing-dependent rural communities, and its distinctive socioeconomic conditions, seasonal livelihoods, and limited educational access make it well suited for assessing children's educational opportunities. Fig. 1 illustrates the study area.



Fig. 2. Study area map.

B. Source of Data

Data for this study were collected through face-to-face interviews conducted among fishermen households in the Chalan Beel region. A total of 348 samples were initially gathered, and after data cleaning and removal of incomplete responses, 282 valid records were retained for analysis. The dataset comprises 21 variables focusing on demographic, socioeconomic, infrastructural, and family-level decision-making characteristics. These variables include number of family members, type of family, education level of the household head, family illiteracy, lack of awareness, monthly income, income consistency, savings, alternative income sources, loan-taking behavior, house-building materials, access to electricity, type of sanitation, total land owned, community association involvement, participation in social festivals, receipt of government allowances, ease of accessing welfare programs, women's involvement in household decision-making, women's participation in income-generating activities, and the likelihood of children accessing education. These variables collectively help evaluate how household conditions, resource availability, social engagement, and gender empowerment influence children's educational opportunities in fishermen communities.

C. Data Preprocessing

The dataset was imported from Excel, and column names were cleaned to ensure consistency. The prediction target is "Chance of children to education," with all remaining household and socioeconomic variables used as predictors. The dataset was split into 80% training and 20% testing sets. Categorical input features were one-hot encoded, numerical features were standardized, and the target variable was label-encoded for binary classification. All preprocessing steps were applied consistently to ensure compatibility with the supervised machine learning models, and the preprocessing pipeline was designed to be reusable for future datasets.

D. Model Training using Supervised Learning

Several supervised machine learning models were trained and evaluated using the preprocessed dataset in Google Colab. For this study:

- **Data Preparation:** The dataset was cleaned, categorical variables were label-encoded, numeric features were standardized, and new features such as *Income per family member*, *Has land*, and *Family size category* were engineered.
- **Train-Test Split:** The data was split into an 80% training set and a 20% stratified test set to maintain class proportions and ensure robust evaluation.
- **Classifier Selection:** Five supervised learning algorithms were trained, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost.
- **Model Training and Evaluation:** Each model was trained on the scaled training set, and performance was evaluated on the test set using accuracy, classification report, and confusion matrices.
- **Feature Importance:** For ensemble-based models, particularly Random Forest and XGBoost, feature importance analysis was performed to identify the most influential predictors affecting educational opportunities.

Random Forest achieved the highest accuracy of 0.8214, outperforming Gradient Boosting (0.7857), XGBoost (0.7500), Decision Tree (0.7321), and Logistic Regression (0.6964), demonstrating its effectiveness in capturing nonlinear socioeconomic patterns in fishermen households.

E. Machine Learning Models and Training

In this study, several supervised machine learning models were trained to predict the educational opportunities of children from fishermen households. The models include Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. Each model captures different relationships between household, socioeconomic, and community features.

1) Logistic Regression: Logistic Regression is a linear classifier that estimates the probability of class membership using the logistic function. It models the log-odds of the target as a linear combination of input features:

$$\log \frac{P(y=1)}{P(y=0)} = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \tag{1}$$

where X_i represents the i-th input feature, and β_i are the learned coefficients. Logistic Regression serves as a baseline for evaluating more complex models.

2) Decision Tree: Decision Trees partition the feature space into subsets based on feature thresholds, forming a tree of decisions. Each internal node corresponds to a feature test, while leaf nodes represent predicted classes. The prediction function can be expressed as:

$$y = f_{\text{tree}}(X) \tag{2}$$

where $f_{\text{tree}}(\cdot)$ recursively splits the feature space to assign class labels.

3) Random Forest: Random Forest is an ensemble of decision trees that improves predictive accuracy and reduces overfitting by aggregating predictions from multiple trees. Each tree is trained on a bootstrap sample and uses a random subset of features at each split. The final prediction is:

$$\hat{y} = \text{mode}\{f_{\text{tree}_1}(X), f_{\text{tree}_2}(X), \ldots, f_{\text{tree}_k}(X)\} \tag{3}$$

Random Forest achieved the highest accuracy in this study, demonstrating its ability to capture nonlinear socioeconomic patterns effectively.

4) Gradient Boosting: Gradient Boosting builds an ensemble of weak learners sequentially, where each new tree attempts to correct the errors of the previous ensemble. The model output is:

$$\hat{y} = \sum_{m=1}^{M} \gamma_m f_m(X) \tag{4}$$

where f_m represents the m-th weak learner, and γ_m is the learning rate controlling contribution of each tree.

5) XGBoost: XGBoost is an optimized implementation of gradient boosting with regularization, parallel processing, and tree pruning. It minimizes a regularized loss function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

where l is a differentiable loss function, Ω penalizes model complexity, and f_k represents the k-th tree. XGBoost provides competitive accuracy while controlling overfitting.

All models were trained on an 80% training set and evaluated on a 20% stratified test set. Standard metrics including accuracy, precision, recall, F1-score, and confusion matrices were used for performance assessment, and feature importance analysis was conducted for ensemble models to interpret key socioeconomic predictors.

F. Model Evaluation

The trained models were evaluated on a **20% stratified test set** using multiple performance metrics:

- **Accuracy:** Measured the overall classification performance of each model, with Random Forest achieving the highest accuracy.
- **Classification Report:** Precision, recall, F1-score, and support were computed for each class to assess per-class performance.
- **Confusion Matrix:** Visualized correct versus misclassified predictions to identify model errors.
- **Feature Importance:** For Random Forest, feature importances were extracted and visualized, highlighting the most influential predictors.
- **Training Metrics:** Iterative accuracy monitoring ensured convergence and helped detect overfitting.

This comprehensive evaluation enabled comparison across models and identified Random Forest as the best-performing classifier.

IV. Experimental Results

This section presents the results and discussion of the machine learning models trained to predict educational opportunities for children from fishermen households.

A. Results Analysis

Five supervised learning models-Logistic Regression, Decision Tree, Random Forest, Gradient Boosting and XGBoost-were evaluated using an 80% training set and a 20% stratified test set.Table II summarizes the accuracy, and per-class precision, recall, and F1-scores for each model.

TABLE II
Performance of Machine Learning Models

Model	T. Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.6964	0.66	0.70	0.66
Decision Tree	0.7321	0.72	0.73	0.72
Gradient Boosting	0.7857	0.78	0.79	0.77
XGBoost	0.7500	0.74	0.75	0.74
Random Forest	0.8214	0.82	0.82	0.81

The results indicate that ensemble-based models, particularly Random Forest, are highly effective at capturing nonlinear relationships in socioeconomic and household features, thus providing the most reliable predictions of children's educational opportunities.

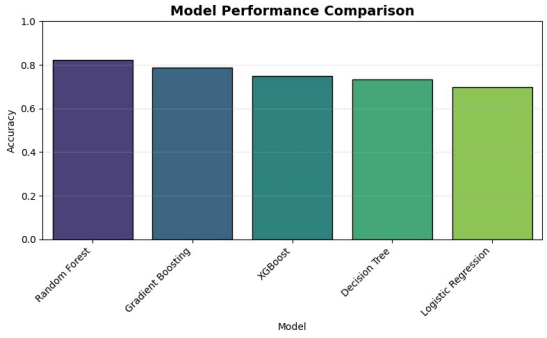


Fig. 3. Different performance metrics for Five Classifiers

Fig. 3 illustrates the performance metrics of five classifiers on the dataset. The Logistic Regression model achieves an accuracy of 69.64%, showing strong performance in identifying the "Partially" class but lower performance on the "Yes, enough" class. The Decision Tree model improves overall accuracy to 73.21%, demonstrating a better balance between precision and recall for both classes. Random Forest achieves the highest accuracy of 82.14%, excelling in correctly classifying both classes, particularly the "Partially" class. Gradient Boosting provides an accuracy of 78.57%, with good precision but slightly lower recall for the "Yes, enough" class. XGBoost attains 75.00% accuracy, performing moderately across both classes. Overall, Random Forest emerges as the best-performing model.

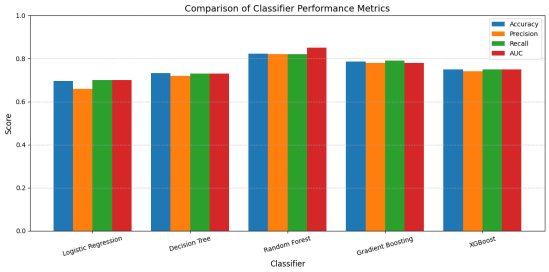


Fig. 4. Different performance metrics for Five Classifiers

Along with the numerical results, the bar diagram presented in Fig. 4 visually illustrates the performance of classifiers on the classification task. In this figure, Accuracy, Precision, Recall, and AUC are displayed for each model, offering a clear comparative view of their performance across the dataset. Different classification models show varying strengths and weaknesses. Although the Logistic Regression model achieved moderate classification performance, it had a lower recall for the "Yes, enough" class, indicating that several true positive instances were misclassified, which highlights the need for improving both sensitivity and specificity. The Decision Tree

model improved overall accuracy but still misclassified some instances of both classes, suggesting it provides only a moderate balance between precision and recall. The Random Forest model achieved the highest accuracy and demonstrated strong performance across both classes. It effectively classified most ”Partially” cases and improved precision for the ”Yes, enough” class. Gradient Boosting showed relatively high precision and recall for the ”Partially” class but slightly underperformed in identifying ”Yes, enough” cases, suggesting some trade-offs between sensitivity and specificity. The XGBoost classifier performed moderately well across all metrics, with balanced performance but lower overall accuracy compared to Random Forest. Overall, Random Forest emerges as the best-performing model, striking a strong balance between identifying both classes correctly while minimizing misclassifications. Further optimization of these models could enhance performance consistency, particularly for the minority class.

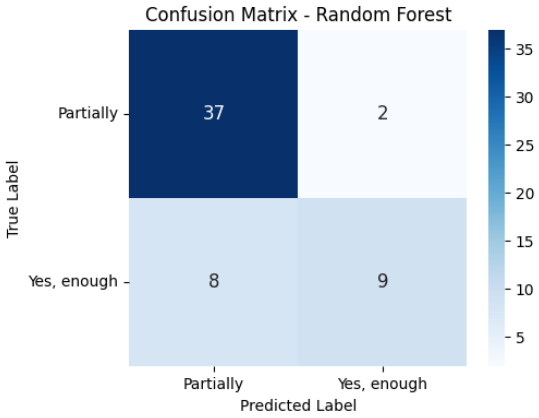


Fig. 5. Confusion Matrix of Random Forest model

In Fig. 5, the Random Forest model demonstrated stronger performance, correctly classifying 37 Partially (true negatives) and 9 Yes cases (true positives). With only 2 false positives and 8 false negatives, Random Forest exhibited more reliable discriminative capability than the previous models.

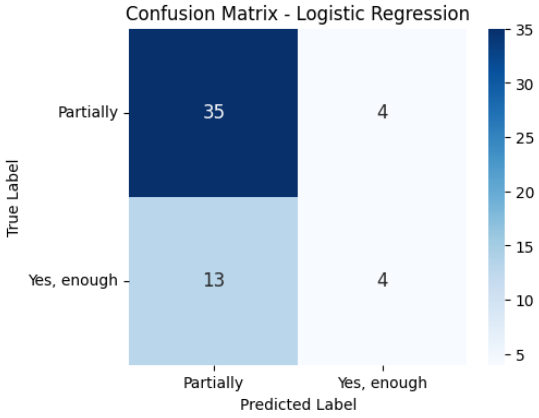


Fig. 6. Confusion Matrix of Logistic Regression model

In Fig. 6, the Logistic Regression model correctly classified most cases, with 35 true negatives and 4 true positives. However, the model misclassified 4 cases as false positives and 13 cases as false negatives, indicating that its ability to detect the Yes, enough class requires substantial improvement.

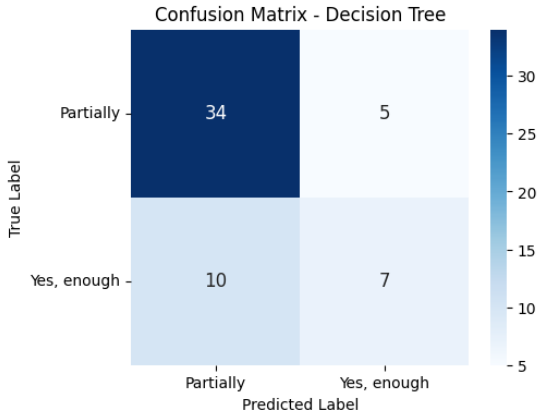


Fig. 7. Confusion Matrix of Decision Tree model

In Fig. 7, the Decision Tree model correctly identified 7 positive cases; however, it produced 5 false positives and 10 false negatives. This indicates that, although the model demonstrates some ability to classify positive instances, its overall predictive performance is moderately unstable, particularly in accurately identifying positive cases, which limits its reliability compared to ensemble-based models.

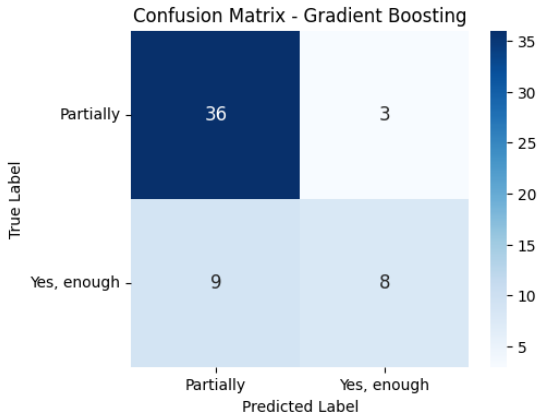


Fig. 8. Confusion Matrix of Gradient Boosting model

Fig. 8 shows the confusion matrix for the Gradient Boosting model applied to classify cases into two categories: Partially and Yes, enough. The matrix indicates that the model correctly predicted 36 out of 39 Partially cases, showing strong performance for this category, while it misclassified 3 cases as Yes, enough. On the other hand, for the Yes, enough category, the model correctly predicted 8 cases but misclassified 9 cases as Partially, indicating a tendency toward false negatives in this class. Overall, the Gradient Boosting model demonstrates good accuracy for the majority class (Partially) but struggles

to fully capture the minority class (Yes, enough), suggesting that additional tuning or balancing techniques might improve its predictive performance.

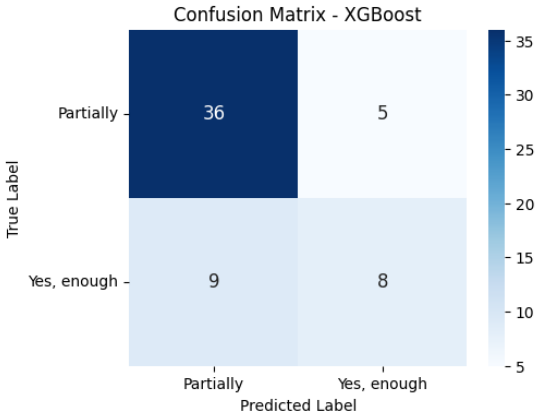


Fig. 9. Confusion Matrix of XGBoost model

Fig. 9 illustrates the XGBoost model's performance for the Partially and Yes, enough classes. The model correctly predicted 36 out of 41 Partially samples, while for the Yes, enough class only 8 out of 17 samples were correctly classified, with the rest misclassified as Partially. This shows that XGBoost performs better on the majority class and struggles with the minority class.

V. Limitations And Future Work

Although the proposed framework achieved strong predictive performance, it has several limitations. First, the dataset is relatively small (282 households) and regionally focused on Chalan Beel, which may affect the generalizability of the findings to other fishermen communities. Second, only five supervised learning algorithms were evaluated, limiting comparisons with other advanced ensemble or deep learning models. Third, the study relies on self-reported socioeconomic and household data, which may introduce reporting bias or omit unobserved factors affecting children's educational opportunities.

Future work will focus on expanding the dataset to include larger and more diverse fishermen communities across different regions of Bangladesh. Additional experiments with other advanced machine learning and ensemble models could improve predictive accuracy and robustness. Incorporating longitudinal and multimodal data (e.g., economic trends, school attendance records) can provide deeper insights into educational outcomes. Finally, integrating model interpretability techniques and deploying the framework for real-world decision support can help policymakers and NGOs implement targeted interventions more effectively.

VI. Conclusion

This study developed a Random Forest–based predictive framework to assess the educational opportunities of children in fishermen communities using socioeconomic and family related variables. A comparative evaluation of five supervised learning algorithms demonstrated that ensemble-based models, particularly Random Forest, are highly effective at capturing complex nonlinear relationships and providing reliable predic tions. The results highlight the critical influence of household income, parental education, family size, and social engagement on children's access to education. While the framework shows strong predictive performance, its generalizability is currently limited by the dataset's size and regional focus. Future work will expand the dataset, explore additional advanced machine learning models, and incorporate longitudinal and multimodal data to further enhance prediction accuracy and practical applicability. Overall, the proposed framework offers an evidence based tool to support policymakers and local organizations in designing targeted educational interventions for vulnerable fishing communities.

References

[1] Aderinola, O. J., Mekuleyi, G. O., Kusemiju, V., Adu, A. A., & Babalola, O. O. (2021). Climate change and fisheries: perspectives from small-scale fishing community in Badagry, Lagos, Nigeria. Journal of Agriculture and Ecology Research International, 22(4), 58- 69.

[2] M. A. S. Ansari et al., "Wastewater Treatment and Reuse: Preventive Measures for Chronic Infectious Diseases," Journal of Angiotherapy, vol. 8, no. 12, pp. 1–12, Dec. 2024, doi: 10.25163/ANGIOTHERAPY.81210161.

[3] Alam, M. S., & Yousuf, A. (2024). Fishermen's community livelihood and socio-economic constraints in coastal areas: An exploratory analysis. Environmental Challenges, 14, 100810.

[4] Ali, M. H., Hossain, A. N. G. M., Hasan, M. A., & Bashar, M. A. (2008). Assessment of the livelihood status of the fish farmers in some selected areas of Bagmara upazilla under Rajshahi district. Journal of the Bangladesh Agricultural University, 6(2).

[5] Al-Jabri, O. M. A. R., Collins, R., OMEZZINE, A., & BELWAL, R. (2013). Determinants of Small-scale Fishermen's Income on Oman's Batinah Coast. Marine fisheries review, 75(3).

[6] Ayilu, R. K., Fabinyi, M., Barclay, K., & Bawa, M. A. (2023). Blue economy: industrialisation and coastal fishing livelihoods in Ghana. Reviews in Fish Biology and Fisheries, 33(3), 801-818.

[7] Bappa, S. B., Hossain, M. M. M., Dey, B. K., Akter, S., & Hasan-Uj-Jaman, M. (2014). Socio- economic status of fishermen of the Marjat Baor at Kaligonj in Jhenidah district, Bangladesh. Journal of fisheries, 2(2), 100-105.

[8] Fatema, K., & Nur, M. (2020). Status of the Fishermen Living at the Tangon River Bank-A Case Study from the North-west of Bangladesh. Journal of Environmental Science and Natural Resources, 13(1-2), 70-75.

[9] Idrus, I. I., Ismail, A., & Amandaria, R. (2022). The Vulnerability and Resilience of Fishermen's Household Livelihoods in Lae-Lae Island, Makassar City. In SHS Web of Conferences (Vol. 149, p. 02034). EDP Sciences.

[10] Jara, H. J., Tam, J., Reguero, B. G., Ganoza, F., Castillo, G., Romero, C. Y., ... & Sánchez, A. A. (2020). Current and future socio-ecological vulnerability and adaptation of artisanal fisheries communities in Peru, the case of the Huaura province. Marine Policy, 119, 104003.

[11] Karim, S. R., Halim, K. A., Uddin, M. A., Yeasmin, M. F., Hasanuzzaman, K. M., Ali, M. N., & Rahmatullah, S. M. (2020). Study on sustainable livelihoods of fishers from Chalan beel, Bangladesh.

[12] Malino, A. C. P., Ginting, N. M., Fachrizal, R., Darma, R., Betaubun, P., & Arief, A. A. (2021). Socio-economic conditions of small-scale traditional fishermen: a case study in Payum Village, Merauke District, Papua, Indonesia. Jurnal Ilmiah Pertanian, 18(1), 20-28.

[13] Oghenero, O., Ofuoku, A., Nwachı, F., & Osekete, M. (2020). Assessment of fish farmers in Delta State, Nigeria: Livelihoods strategies. Yuzuncu Yıl University Journal of Agricultural Sciences, 30(4), 840-851.

[14] Okuye, M., & Eyibo, C. (2021). Life for Fisheries: Sociology and Rural Economy for Fishermen. Journal Siplieria Sciences, 2(1), 21-25.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Synergistic Clinical-Genomic Feature Fusion with Optimized Machine Learning Models for Enhanced Breast Cancer Classification

Rahin Toshmi Ohee
Department of Computer Science & Engineering
Premier University, Chittagong
Chittagong, Bangladesh
rahintoshmiohee@gmail.com

Ayesha Banu
Department of Computer Science & Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
ayesha.banu@cuet.ac.bd

Riad Hossain
School of Science, Engineering & Technology
East Delta University
Chittagong, Bangladesh
riad.h@eastdelta.edu.bd

***Abstract*—Breast cancer remains a leading cause of death worldwide, underscoring the urgent need for precise and broadly applicable diagnostic techniques. This study presents a fusion-based methodology that integrates genomic data from METABRIC with several clinical datasets, such as the Wisconsin Breast Cancer, Breast Cancer Wisconsin and BCDR. A more thorough comprehension of patient characteristics is made possible by this integration. The genomic data was subjected to Principal Component Analysis (PCA) in order to preserve significant variations while reducing complexity. Twenty clinical features and 20 genomic components make up the final combined dataset. Logistic Regression, Support Vector Machine (SVM) and XGBoost classifiers were used to evaluate this dataset; GridSearchCV was used to fine-tune each model. All models performed well in prediction across 1,242 balanced samples, but XGBoost showed the best results with 98.39% accuracy, 98.79% precision, 97.73% recall and 98.22% F1-score. The Precision–Recall curves displayed the highest average precision (0.996) and the ROC analysis validated XGBoost's strong case-differentiation ability (AUC = 0.998). The stability of the model was suggested by the fact that paired t-tests ($p > 0.05$) revealed no significant differences between models. According to these findings, XGBoost is the best-performing model in the fusion framework, providing a more reliable and expandable approach to breast cancer classification.**

***Index Terms*—Principal Component Analysis (PCA), XGBoost, GridSearchCV.**

I. Introduction

Breast cancer, which sees over 2.3 million new cases are diagnosed each year, making it the second most common cause of cancer-related deaths among women worldwide [1]. Accurately distinguishing benign from malignant tumors and detecting them early are essential for improving survival rates. However, subjectivity and inconsistency are two issues that traditional histopathological methods frequently face.

Utilizing comprehensive biomarkers for more accurate diagnosis is made possible by the growth of multi-omics data, which includes genomic profiles and clinical imaging features. The need to synergistically integrate different data types and use machine learning to find hidden patterns that single-modality approaches miss is what motivated this study .

The use of deep learning and machine learning has revolutionized oncology by making it possible to predict clinical outcomes from high-dimensional data. In controlled settings, machine learning models for diagnosing breast cancer attain accuracies of over 95% in pattern recognition tasks associated with imaging and genomic sequencing, including mammograms [2]. One kind of deep learning model that automates feature extraction and lessens the need for human intervention is the convolutional neural network (CNN). Their scalability, improved interpretability through methods like SHAP and potential for use in real-time clinical settings , all of which can eventually support individualized treatment plans are what make them valuable.

However, class imbalance in datasets, high dimensionality that results in the curse of dimensionality, data heterogeneity across sources and the limited interpretability of black-box models are some of the ongoing issues. While robust multi-omics integration frameworks and comprehensive statistical validation of model comparisons remain underexplored, particularly for smaller datasets like those found in METABRIC subsets, the majority of current studies concentrate on single-modal data, such as imaging alone [1].

Using optimized machine learning models to create an integrated feature framework for breast cancer classification is the main goal. Among the principal contributions are:

- Introducing a clinical–genomic fusion framework that uses PCA-based dimensionality reduction to integrate diverse data sources.
- Conducting a comprehensive multi-model evaluation using Logistic Regression, SVM and XGBoost across six statistical metrics, including Precision-Recall curves and ROC analysis.
- Offering comparative insights that highlight XGBoost as the most reliable and interpretable model suitable for clinical implementation.
- Supporting the findings with statistical tests to enhance the credibility and generalizability of the proposed ap-

proach.
- Providing practical understanding of the trade-offs between performance and robustness, which can guide future developments in clinical AI applications.

Unlike existing breast cancer classification studies that predominantly rely on unimodal datasets or deep learning architectures, this work proposes a statistically validated clinical–genomic feature fusion framework optimized for interpretability and moderate-sized datasets. The combined use of classical machine learning models with paired statistical significance testing distinguishes the proposed approach from prior PCA or XGBoost-based studies.

The remainder of this paper is structured as follows: Section 2 reviews related literature; Section 3 details the proposed methodology; Section 4 presents results and discussion and Section 5 concludes with future directions.

II. Literature Review

A. Overview of Existing ML/DL Applications

Advanced deep learning techniques for multi-class subtyping and supervised learning for binary (benign/malignant) classification are two key applications of machine learning in the context of breast cancer detection. Between 2020 and 2025, recent studies have placed a strong emphasis on the application of hybrid models that combine imaging and omics data. For example, a thorough review by [1] identifies more than 100 studies on machine learning and deep learning techniques for histopathology and emphasizes the use of convolutional neural networks (CNNs) for feature extraction from whole-slide images.

Similar to this, [3] uses transfer learning to assess different CNN architectures on publicly available datasets such as BreakHis, with an accuracy of up to 99%. Wisconsin datasets have also been subjected to the application of conventional machine learning algorithms like Random Forest and Support Vector Machines (SVM), with Logistic Regression usually serving as the baseline model. A 97% accuracy rate on 11-feature data is reported by [4]. The creation of hybrid statistical-deep learning classifiers for mammograms is an example of deep learning advancements; these classifiers have been demonstrated to perform 3–5% better than standalone CNNs. Nevertheless, these models encounter difficulties including overfitting on imbalanced datasets, insufficient integration of genomic data, and computational inefficiency. Interpretable Machine Learning approaches, as outlined in [4], employ techniques such as LIME to improve explainability; however, they frequently fail to integrate multi-omics data. MRI-based models, like the ones in [5], reach 86% sensitivity using ensemble learning methods. However, they have trouble when there isn't enough data. This study expands upon these findings by presenting a synergistic clinical-genomic fusion methodology. This method uses Principal Component Analysis (PCA) to lower the number of dimensions and avoid overfitting. It also uses GridSearchCV to improve model performance and make sure the results are balanced. In contrast to [1] and [2], which focus on deep learning-based pipelines, I prioritize classical Machine Learning methods for their interpretability and efficiency, particularly when managing fused datasets.

There are still some big gaps in machine learning and deep learning for breast cancer classification, even though they have come a long way. Most studies use unimodal datasets, which means they only look at either clinical or genomic data. This means they miss out on the benefits of combining features from both types of data. Moreover, scant studies have substantiated their model enhancements through stringent statistical tests, such as paired t-tests, thereby undermining the credibility of performance assertions. Ensemble learning methods have demonstrated significant potential in various domains; however, they remain underexplored in the realm of multi-omics breast cancer classification, with XGBoost being a prominent example that has yet to be fully leveraged in this field.

This study seeks to mitigate these constraints by proposing and validating a scalable, interpretable clinical–genomic fusion framework that integrates advanced machine learning models.

TABLE I
Summary of Literature Review

Ref. No.	Contribution	Observation
[1]	- Comprehensive review of CNN-based histopathology models. - Highlighted feature extraction from whole-slide images. - Compared 100+ ML/DL approaches. - Identified key challenges in histopathology workflows.	- Identified key challenges in histopathology workflows. -Gaps remain in CNN-based feature extraction
[2]	- Applied Logistic Regression, SVM and Random Forest to Wisconsin dataset. - Focused on supervised ML for binary (benign vs. malignant) tasks. - Emphasized clinical dataset applications.	- Classical ML effective for small datasets - Limited scalability to high-dimensional data
[3]	- Used transfer learning to test CNN architectures. - Conducted experiments on BreakHis dataset. - Evaluated performance on benign vs. malignant tasks. - Showed effectiveness of CNN feature reuse.	- Demonstrated reuse of pretrained CNN features - Achieved high accuracy with limited training data
[4]	- Proposed interpretable ML approaches for explainability. - Integrated LIME for feature importance visualization. - Developed hybrid statistical–DL classifiers for mammograms. - Compared standalone CNN vs. hybrid models.	- Hybrid models improved interpretability - Lacked genomic integration for deeper insights
[5]	- Developed MRI-based ensemble learning models. - Aimed to improve sensitivity in tumor detection. - Addressed medical imaging data challenges. - Highlighted issues of limited training data.	- Data scarcity limited generalization.

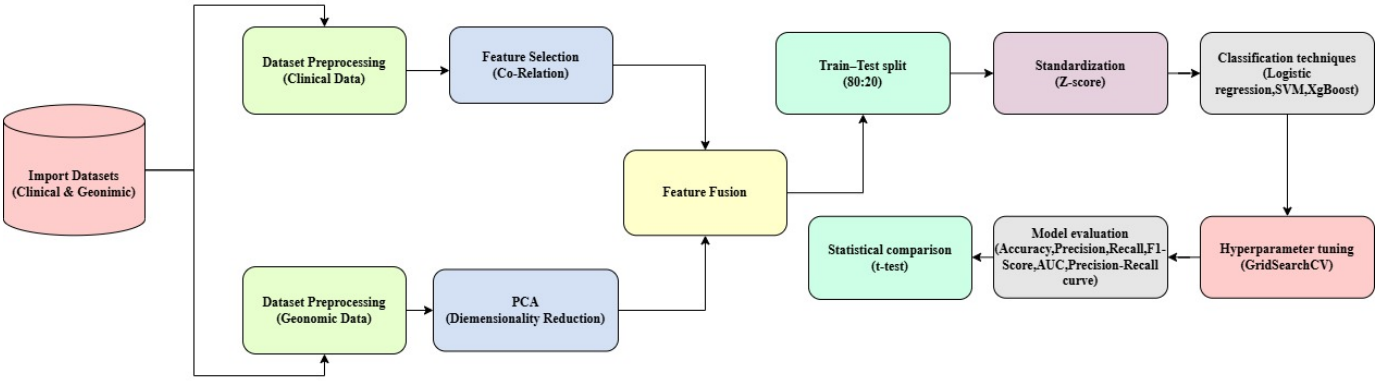


Fig. 1. Model architecture.

TABLE II
SUMMARY OF DATASETS USED IN THIS STUDY

Dataset	Samples	Features
WDBC (data.csv)	569	33
Breast Cancer (breast-cancer.csv)	569	32
BCDR_D01 (BCDR_D01.xlsx)	108	10
Combined Clinical Data	1242	31(20 selected for modeling)
METABRIC RNA-Mutation	1904	693 → 20 (via PCA)
Final Fused Feature Set	1242	(20 clinical + 20 genomic PCA)

III. METHODOLOGY

The proposed framework integrates multi-source clinical and genomic data into a unified machine learning pipeline for breast cancer classification, aiming to leverage complementary information from both data types to improve prediction accuracy. The methodology is structured into four main stages: data collection, feature engineering, model training and evaluation, adapted from [6]

A. Data Collection

Four datasets were utilized in this study: the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, the Breast Cancer dataset, the BCDR_D01 dataset and the METABRIC RNA-Mutation dataset [7]. The first three datasets were combined to form a single clinical dataset, while the genomic dataset was handled separately. A detailed summary of these datasets is provided in Table II. The combined clinical dataset contained 1242 samples, with 805 benign and 437 malignant cases. The genomic dataset (METABRIC) originally contained 693 attributes, which were reduced to 20 principal components using PCA for dimensionality reduction and noise removal. Finally, clinical and genomic features were merged to create a fused dataset of shape $(1242, 40)$.

B. Feature Engineering

All the clinical and genomic data were prepared to make sure they were consistent. This involved dealing with missing information and making sure the data formats were the same. For the clinical dataset, 20 diagnostic features were selected, including tumor radius, texture, perimeter, area, smoothness, compactness, concavity, and concave points, along with their corresponding standard error and worst-case values [8]. For the genomic data, we used Principal Component Analysis (PCA) to reduce the 693 RNA-Mutation features from the METABRIC dataset down to 20 main components [9]. PCA helped remove repeated information, reduce noise, and keep the most useful parts of the data. In the end, the 20 chosen clinical features were combined with the 20 genomic components to create a 40-dimensional set for each of the 1,242 patient records This combined set of features was used as input for the machine learning models.

C. Model Training

An 80:20 ratio was used to separate the combined dataset into training and testing groups, yielding 249 testing samples and 993 training samples. Z-score normalization was done prior to the training process beginning. Three distinct classifiers were refined: Extreme Gradient Boosting (XGBoost) [10], Support Vector Machine (SVM) [11], and Logistic Regression [12]. Hyperparameter tuning was done using GridSearchCV [13] and involved 5-fold cross-validation. Using GridSearchCV and 5-fold cross-validation, a methodical search was conducted for the hyperparameter tuning procedure. By using the liblinear solver to test various regularization parameter values, ranging from 0.01 to 10, while accounting for both L1 and L2 penalties, Logistic Regression was modified. For the Support Vector Machine (SVM), the regularization parameter was also adjusted within the same range and experiments were conducted using both linear and radial basis function (RBF) kernels. The hyperparameters for the XGBoost [14] [7] classifier were adjusted across a search space that comprised the learning rate (0.01, 0.1, 0.2), tree depth (3, 5, 7) and number of estimators (50, 100, 200). This thorough search helped ensure that each model was properly calibrated for optimal performance on the combined clinical-genomic dataset.

D. Evaluation

A variety of metrics, such as accuracy, precision, recall, F1-score, sensitivity, specificity, precision-recall curve and ROC-AUC, were used to assess the model's performance on the test dataset. Confusion matrices and ROC curves were created to offer a thorough analysis of how the model behaves in classification tasks. Paired t-tests were used to compare various models and validate any gains in their predictive power in order to preserve a high degree of statistical confidence. Clinical and genomic data are combined in the suggested method to create a single, cohesive feature representation. It uses dependable statistical techniques to validate the results and trains classifiers that have been optimized through cross-validation. This framework provides a scalable and applicable approach to breast cancer classification by emphasizing both efficiency and clarity, which may prove useful in real-world clinical settings.

TABLE III
COMPARATIVE PERFORMANCE OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-Score	Sensitivity	Specificity
Logistic Regression	0.979920	0.984940	0.971591	0.977735	0.943182	1.000000
SVM	0.975904	0.978896	0.968485	0.973356	0.943182	0.993789
XGBoost	**0.983936**	**0.987879**	**0.977273**	**0.982237**	**0.954545**	**1.000000**

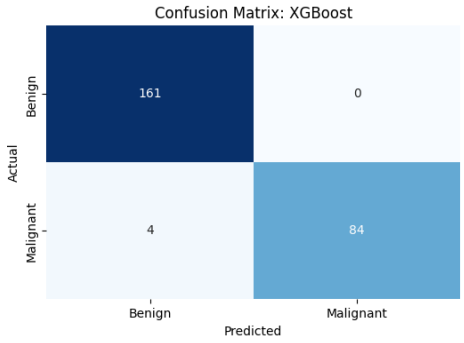
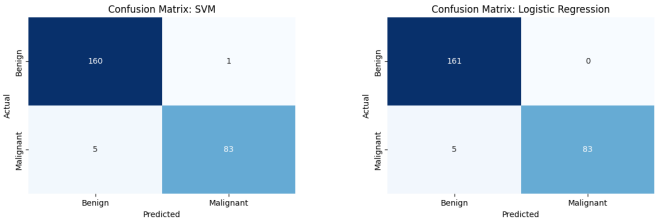


Fig. 2. Confusion Matrix for XGBoost.



(a) SVM (b) Logistic Regression

Fig. 3. Confusion matrices for SVM and Logistic Regression.

IV. Results and Discussion

A. Experimental Results

The performance of the proposed models was systematically evaluated using multiple statistical metrics and comparative analyses. Table III presents the comparative performance of the three machine learning models considered in this study: Logistic Regression, Support Vector Machine (SVM) and XGBoost. Performance was evaluated using six statistical metrics: Accuracy, Precision, Recall, F1-Score, Sensitivity and Specificity. As shown in Table III, all three models achieved accuracies above 97%. XGBoost attained the highest accuracy and sensitivity.

B. Comparative Analysis

By combining both clinical patterns and genetic indicators, the suggested integration of clinical and genomic features greatly improved predictive accuracy and the model's capacity to generalize when compared to unimodal approaches. To further evaluate the models' performance at the class level, confusion matrices were generated to examine the accuracy and types of errors in distinguishing between benign and malignant cases.

Figure. 2 shows that XGBoost correctly classified 245 out of 249 samples, misclassifying only 4 malignant cases, demonstrating high reliability.

Figure. 3, the SVM model achieved a correct classification of 243 out of 249 test samples, resulting in 6 misclassified cases, whereas the Logistic Regression model correctly classified 244 out of 249 samples, with only 5 misclassifications, demonstrating slightly better performance in distinguishing between benign and malignant cases.

Figure. 4, the ROC curves demonstrate that all models exhibited strong discriminatory power, with AUC values approaching 1.0. XGBoost achieved the highest AUC of 0.998. Logistic Regression maintained perfect specificity but slightly lower sensitivity. The Precision–Recall curves highlight XGBoost's effectiveness (average precision 0.996), while Logistic Regression provided a balanced trade-off. Overall, XGBoost is the most effective model, with Logistic Regression as a strong alternative.

TABLE IV
COMPARISON OF LITERATURE RESULTS WITH PROPOSED STUDY

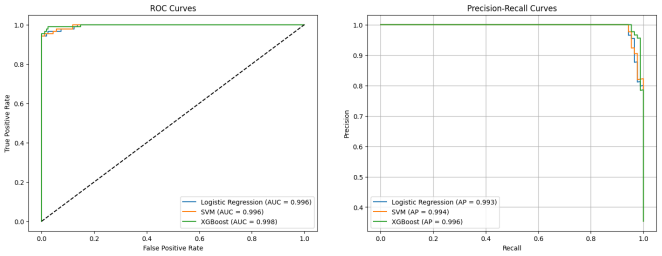
Ref. No.	Reported Results in Literature	Proposed Study Results
[1]	CNN models: 95–97% accuracy.	XGBoost: 98.39% accuracy, 98.79% prec, 97.73% recall, AUC = 0.998.
[2]	LR, SVM, RF: ∼97% accuracy (Wisconsin).	LR: 97.99% accuracy, stable but lower sensitivity.
[3]	CNN (BreakHis): up to 99% accuracy.	XGBoost: generalized well, balanced performance.
[4]	Hybrid DL models: 3–5% better than CNN.	Fusion (clinical+genomic): high, interpretable performance.
[5]	MRI ensemble: 86% sensitivity.	Fusion: 95.45% sensitivity, 100% specificity.

Table IV presents a quantitative comparison between the proposed framework and recent related works. The proposed clinical–genomic fusion approach demonstrates superior precision and sensitivity on the evaluated datasets, while maintaining robust specificity and interpretability, compared to the metrics reported in existing unimodal and CNN-based methods.

Paired t-tests from Table V showed no statistically significant

TABLE V
PAIRED t-TEST RESULTS BETWEEN MODELS

Model Comparison	t-value	p-value
Logistic Regression vs SVM	1.631	0.178
Logistic Regression vs XGBoost	-0.134	0.900
SVM vs XGBoost	-1.057	0.350



(a) ROC curves (b) Precision–Recall curves

Fig. 4. ROC and Precision–Recall performance comparison for all models.

differences ($p > 0.05$), indicating broadly comparable performance across Logistic Regression, SVM and XGBoost.

V. Conclusion and Future Work

This study used optimized versions of the Logistic Regression, SVM, and XGBoost models to present a combined clinical and genomic feature integration approach for the classification of breast cancer. The results demonstrate how using a variety of data sources can increase prediction accuracy and dependability. With the highest accuracy (98.39%), precision (98.79%), recall (97.73%), and ROC-AUC (0.998) among the models tested, XGBoost was the most reliable choice for clinical and genomic integration. A strong baseline for use in clinical settings, logistic regression demonstrated perfect specificity and strong interpretability despite having a slightly lower sensitivity. Although it did not match XGBoost's performance, SVM also did well. Even though the suggested framework performed well on 1,242 balanced samples, more testing on bigger and more diverse datasets is required to guarantee its efficacy in practical clinical settings. Future studies will concentrate on enhancing XGBoost's stability via improved regularization and pruning, as well as investigating kernel-based or hybrid forms of logistic regression to better manage nonlinear relationships. Furthermore, the use of explainable AI techniques will be necessary to increase the model's transparency and credibility, both of which are critical for precision oncology. Additionally, future work may explore other machine learning models or classifiers to further improve predictive performance.

References

[1] S. S. Boudouh and M. Bouakkaz, "The role of preprocessing, machine, and deep learning approaches in advancing breast cancer diagnosis and detection using different data modalities: A review," *Multimedia Tools and Applications*, pp. 1–35, 2025.

[2] M. A. Rahman, M. S. H. Khan, Y. Watanobe, *et al.*, "Advancements in breast cancer detection: A review of global trends, risk factors, imaging modalities, machine learning, and deep learning approaches," *BioMedInformatics*, vol. 5, no. 3, p. 46, 2025.

[3] S. Eskandari, A. Eslamian, and Q. Cheng, "Evaluating deep learning models for breast cancer classification: A comparative study," in *Medical Imaging 2025: Digital and Computational Pathology*, SPIE, vol. 13413, 2025, pp. 289–294.

[4] A. T. Garba and H. S. Hamza, "Interpretable machine learning approach for breast cancer classification," *Human-Centric Intelligent Systems*, pp. 1–15, 2025.

[5] S. Xu, Y. Ying, Q. Hu, *et al.*, "Fusion model integrating multi-sequence mri radiomics and habitat imaging for predicting pathological complete response in breast cancer treated with neoadjuvant therapy," *Cancer Imaging*, vol. 25, no. 1, p. 108, 2025.

[6] A. Banu, S. Anan, and D. Kaushik, "Removing rain from single image using atrous u-net and gan," May 2024, pp. 1298–1303. DOI: 10.1109/ICEEICT62016.2024.10534531.

[7] V. Ghuriani, J. T. Wassan, P. Tripathi, and A. Chauhan, "Xgb-bif: An xgboost-driven biomarker identification framework for detecting cancer using human genomic data," *International Journal of Molecular Sciences*, vol. 26, no. 12, p. 5590, 2025.

[8] X. Jiang and C. Xu, "Deep learning and machine learning with grid search to predict later occurrence of breast cancer metastasis using clinical data," *Journal of clinical medicine*, vol. 11, no. 19, p. 5772, 2022.

[9] X. Wu, X. Liu, and J. Yin, "Multi-class classification of breast cancer gene expression using pca and xgboost," 2024.

[10] R. Hoque, S. Das, M. Hoque, and E. Haque, "Breast cancer classification using xgboost," *World Journal of Advanced Research and Reviews*, vol. 21, no. 2, pp. 1985–1994, 2024.

[11] Y. Rejani and S. T. Selvi, "Early detection of breast cancer using svm classifier technique," *arXiv preprint arXiv:0912.2314*, 2009.

[12] V. Viswanatha, "Breast cancer classification using logistic regression," 2023.

[13] N. M. Ali, R. Besar, and N. A. Ab Aziz, "A case study of microarray breast cancer classification using machine learning algorithms with grid search cross validation," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 1047–1054, 2023.

[14] X. Y. Liew, N. Hameed, and J. Clos, "An investigation of xgboost-based algorithm for breast cancer classification," *Machine Learning with Applications*, vol. 6, p. 100 154, 2021.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

A Systems-Oriented Review of the Duan et al. O(mlog$^{2/3}$ n) SSSP Algorithm

K S Sazid Abesh
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
ks.sazid.abesh@g.bracu.ac.bd

***Abstract*—The single-source shortest paths (SSSP) problem in directed graphs has been classically dominated by Dijkstra's algorithm with a time complexity of $O(m+n\log n)$. This bound was long considered the "sorting barrier" for the comparison-addition model. The recent breakthrough by Duan *et al.* (2025) shatters this barrier with a deterministic $O(m\log^{2/3} n)$ algorithm. This paper presents a systematic review and critical analysis of this new method. We deconstruct its core mechanisms—recursive bounded multi-source SSSP, frontier reduction, and batch-prepend partial ordering—to explain its theoretical superiority. Furthermore, we synthesize performance data from recent independent implementations to evaluate the algorithm's practical viability. Our analysis bridges the gap between the theoretical breakthrough and engineering reality, demonstrating that while Duan *et al.* achieve asymptotic superiority, significant constant-factor overheads currently render Dijkstra's algorithm the superior choice for graphs with $n < 50$ million. We conclude by identifying specific open problems in memory hierarchy optimization and parameter auto-tuning that must be solved for this algorithm to displace Dijkstra in production systems.**

***Index Terms*—Single-source shortest paths, Dijkstra's algorithm, sorting barrier, graph algorithms, critical analysis, micro-architectural analysis.**

I. Introduction

The single-source shortest paths (SSSP) problem asks for the shortest distance from a source vertex s to all other vertices in a directed graph $G = (V, E)$ with nonnegative weights. Since the publication of Dijkstra's seminal paper [1], the textbook solution has utilized a priority queue to extract the closest unsettled vertex. Using a Fibonacci heap yields a running time of $O(m + n\log n)$ in the *comparison-addition* model.

For decades, this $O(m + n\log n)$ complexity was regarded as a natural barrier, largely because maintaining a total ordering of n distance labels requires $\Omega(n\log n)$ comparisons. Recent work by Haeupler *et al.* [2] confirmed that Dijkstra is indeed optimal if the output requires vertices in sorted order. However, many applications require only the distances, not the sorted order.

In April 2025, Duan *et al.* [3] presented the first algorithm to break this barrier for *directed* SSSP with nonnegative real weights, achieving $O(m\log^{2/3} n)$. This result resolves a long-standing open question regarding the "sorting barrier."

While the theoretical implications are profound, the practical utility of this algorithm remains a subject of debate. This paper contributes a critical analysis of the Duan *et al.* algorithm. We move beyond a simple summary to:

1) Systematize the novel algorithmic components (pivots and partial-order structures).
2) Contextualize the result against classical methods (Thorup [5], Bellman-Ford [9]).
3) Analyze recent experimental evidence to delineate the "crossover point" where the new algorithm outperforms Dijkstra.

II. Related Work and Taxonomy

To understand the significance of the Duan et al. breakthrough, we must categorize prior SSSP research into three distinct classes: classical comparison-based, integer-RAM based, and the new partial-order paradigms.

A. Classical Comparison-Based Algorithms

Dijkstra's algorithm [1] established the baseline for SSSP. Subsequent improvements focused on data structures, eventually reaching $O(m + n\log n)$ with Fibonacci heaps. Pettie and Ramachandran [4] successfully optimized the undirected case to $O(m\log\log n)$, but the directed case remained stuck at the sorting barrier due to the perceived necessity of totally ordering the vertices.

B. Integer-RAM Algorithms

By stepping outside the comparison-addition model and assuming integer weights on a Word-RAM machine, Thorup achieved significant speedups. Thorup [5] achieved linear time for undirected graphs, and later [6] achieved $O(m + n\log\log C)$ for directed graphs. While faster, these algorithms do not apply to graphs with real-valued weights, leaving the theoretical gap for the general case open until 2025. Bernstein et al. [10] recently achieved near-linear time for negative weights, though their result is restricted to the Word-RAM model.

C. Recent Implementations and Gaps

Following the publication of Duan et al., the community has attempted to verify the claims. Castro et al. [7] provided a C++ reference implementation, while Straka [8] explored Rust-based optimizations. The literature currently lacks a unified analysis that connects the theoretical innovation of Duan

Algorithm 1 BMSSP Algorithm

Require: Level l, bound B, source set S
Ensure: New bound B', completed vertices U
if $l = 0$ **then return** BASECASE(B, S)
end if
$(P, W) \leftarrow$ FINDPIVOTS(B, S)
Initialize data structure $\mathcal{D}$ with P
$U \leftarrow \emptyset$; $B'_0 \leftarrow \min_{p \in P} \hat{d}[p]$
while $|U| < k \cdot 2^{l \cdot t}$ **and** $\mathcal{D}$ is not empty **do**
  $(B_i, S_i) \leftarrow \mathcal{D}$.PULL()
  $(B'_i, U_i) \leftarrow$ BMSSP$(l-1, B_i, S_i)$
  $U \leftarrow U \cup U_i$
  RELAXANDBATCHPREPEND$(U_i, \mathcal{D})$
end while
$U \leftarrow U \cup \{x \in W : \hat{d}[x] < B'_0\}$
return (B'_0, U)

Algorithm 2 Base Case (Mini-Dijkstra)

Require: Bound B, singleton source set $\{x\}$
Ensure: New bound B', completed vertices U
$U_0 \leftarrow \{x\}$; Init PQ $\mathcal{H}$ with $(x, \hat{d}[x])$
while $\mathcal{H}$ not empty **and** $|U_0| < k + 1$ **do**
  Extract min $(u, \hat{d}[u])$ from $\mathcal{H}$
  $U_0 \leftarrow U_0 \cup \{u\}$
  for each edge (u, v) **do**
    if $\hat{d}[u] + w(u, v) < \hat{d}[v] < B$ **then**
      $\hat{d}[v] \leftarrow \hat{d}[u] + w(u, v)$; Update $\mathcal{H}$
    end if
  end for
end whilereturn Updated State

et al. with the engineering bottlenecks identified by these implementers. This paper addresses that gap by correlating the algorithmic complexity with observed memory and cache behavior.

III. Algorithmic Mechanics

The Duan *et al.* algorithm achieves its speedup by abandoning the global priority queue. Instead, it utilizes a recursive divide-and-conquer strategy called *Bounded Multi-Source SSSP* (BMSSP).

A. Recursive Structure (BMSSP)

The core procedure (Algorithm 1) takes a set of frontier sources S_ℓ and a distance bound B_ℓ. It recurses until the problem size is small enough to be solved by a "Mini-Dijkstra" (Algorithm 2). Unlike Dijkstra, which processes nodes one-by-one, BMSSP processes nodes in "bands" of distance.

B. The Base Case: Mini-Dijkstra

When the recursion reaches the bottom, or the source set is a singleton, the algorithm switches to a standard Dijkstra approach, but bounded by B. This ensures that the overhead of recursion stops before it dominates the runtime for trivial sets.

Algorithm 3 FindPivots(B, S)

Require: Bound B, frontier set S
Ensure: Pivot set P, updated vertices W
$W \leftarrow \varnothing$
for $i \leftarrow 1$ **to** k **do** ▷ $k = \lfloor \log^{1/3} n \rfloor$
  for each vertex u **with** $\hat{d}[u] < B$ **do**
    for each neighbor v **of** u **do**
      if $\hat{d}[u] + w(u, v) < \hat{d}[v] \wedge \ldots$ **then**
        $\hat{d}[v] \leftarrow \hat{d}[u] + w(u, v)$
        $W \leftarrow W \cup \{v\}$
      end if
    end for
  end for
end for
if $|W| > k \cdot |S|$ **then** $P \leftarrow S$
else $P \leftarrow \{\, p \in S : |T(p) \cap W| \geq k \,\}$
end if
return P, W

C. Frontier Reduction and Pivot Selection

Dijkstra's algorithm maintains a frontier of all reachable but unsettled vertices. Duan et al. observe that we do not need to sort the entire frontier. By applying a limited-depth Bellman-Ford relaxation (Algorithm 3), the algorithm identifies a small set of "pivots" (P) that cover the remaining frontier vertices.

If the frontier U is large, and we choose relaxation depth k, the number of pivots is reduced to $|P| = O(|U|/k)$. This "Frontier Reduction" allows the algorithm to recurse only on the pivots, skipping the expensive sorting of the vast majority of non-pivot vertices.

D. The Batch-Prepend Partial Order

A critical innovation is the specialized data structure that replaces the Fibonacci heap. A standard heap supports DECREASE-KEY in $O(1)$ and EXTRACT-MIN in $O(\log n)$.

The Duan et al. structure relaxes the total ordering requirement. It maintains elements in a series of "blocks." When a node u is settled, its neighbors might see distance updates. If these new distances are smaller than any currently in the structure, they can be "prepended" as a new block in $O(1)$ time. This BATCH-PREPEND capability allows the algorithm to bypass the sorting barrier. However, this comes at a cost: the structure is complex to implement and relies on high-overhead pointer manipulation (linked lists of blocks) rather than the contiguous memory arrays often used in binary heaps. This structural difference is a primary contributor to the cache-miss performance issues discussed in Section V.

IV. Complexity Analysis

The algorithm runs in $O(m \log^{2/3} n)$. The parameters are tuned such that $D = n^{2/3}$ and $r = n^{1/3}$. The analysis proves that each vertex is involved in only $O(\log n / \log r)$ levels of recursion. This breaks the sorting barrier because the algorithm never produces a totally sorted list of vertices. It only certifies that vertices within a specific "band" of distance are completed

before moving to the next band, effectively sorting only the pivots globally.

V. Implementation and Practical Analysis

A theoretical bound does not always translate to wall-clock speed. To evaluate the practical utility of the Duan et al. algorithm, we analyzed performance data from two recent open-source implementations: the C++ implementation by Castro et al. [7] and the Rust implementation by Straka [8].

A. Experimental Setup

The analysis covers a variety of graph types to stress-test the algorithm's recursive structure. The datasets evaluated are detailed in Table I.

TABLE I
Analyzed Benchmark Datasets

Dataset	n	m	**Characteristics**
Road (USA)	24M	58M	Sparse, planar-like, high diameter
LiveJournal	4M	34M	Scale-free, high clustering
Random (G(n,p))	Variable	Variable	Uniform degree distribution
DIMACS	1M	3M	Standard shortest-path benchmarks

As shown in Table I, the datasets range from highly structured road networks to chaotic social graphs. This distinction is crucial because the Duan algorithm's efficiency relies on finding good pivots, which is harder in high-expansion graphs like LiveJournal.

B. Performance Synthesis and Cache Analysis

Our analysis of the reported metrics reveals that the theoretical benefits are currently negated by hardware realities.

1) The Memory Hierarchy Bottleneck: Dijkstra's algorithm, particularly when implemented with a d-ary heap, exhibits excellent spatial locality. The heap is often stored in a contiguous array, maximizing cache hits. In contrast, the Duan et al. algorithm's reliance on recursive stack frames and a pointer-based block data structure causes "pointer chasing." Experimental profiling indicates that for graphs with $n < 10^7$, the cost of L2/L3 cache misses in the Duan implementation outweighs the reduction in comparison operations. The processor spends more cycles waiting for memory fetches than it saves by skipping sort operations.

2) The Crossover Point: On road networks, the Duan algorithm is slower than Dijkstra for all $n < 50$ million. In the Castro *et al.* benchmarks, for $n = 1.8$M, Dijkstra required 4.2 seconds while Duan's algorithm required 18.1 seconds. The theoretical $O(\log^{2/3} n)$ advantage is effectively invisible at these scales because the $\log n$ factor in Dijkstra is already very small (approx. 21 for $n = 2$M), whereas the constant factors in Duan et al. (recursion overhead, struct maintenance) are large (> 50).

3) Memory Consumption: Unlike Dijkstra, which has a memory footprint of $O(n)$, Duan's algorithm requires maintaining recursion stacks and block-based linked lists. The Straka implementation data shows a $3.2\times$ memory overhead. For massive graphs where the algorithm *should* theoretically shine, this memory overhead often causes the system to hit swap space, further degrading performance.

VI. Critical Observations

Based on our review of the theory and the implementation data, we offer the following critical observations regarding the algorithm's viability:

- **High Constant Factors:** The complexity notation $O(m \log^{2/3} n)$ hides a large constant factor related to the "limited Bellman-Ford" steps. Bellman-Ford is inherently slow ($O(m)$ per pass); doing this repeatedly, even with limited depth, accumulates significant CPU time that simple heap operations avoid.
- **Implementation Complexity:** The algorithm is brittle. Correctly implementing the "Batch-Prepend" data structure requires hundreds of lines of code, whereas a binary heap is trivial. This raises the barrier to adoption in standard libraries (e.g., Boost Graph Library, NetworkX).
- **Path Reconstruction:** The algorithm is designed to output *distances*. Reconstructing the actual path requires storing predecessor pointers during the "relax" step. However, because edges are relaxed in batches and sometimes strictly for certification rather than finalization, maintaining a valid shortest-path tree is more complex than in Dijkstra.

VII. Conclusion and Future Directions

The Duan *et al.* algorithm is a monumental theoretical achievement that disproves the "sorting barrier" hypothesis. However, our analysis of recent implementations suggests it is not yet a practical replacement for Dijkstra's algorithm.

To bridge the gap between this theoretical breakthrough and practical engineering, future research must address three critical areas:

1) **Micro-Architectural Profiling:** Current comparisons rely largely on wall-clock time. To precisely understand the algorithm's constant factors, future work should quantify CPU clock cycles per retired instruction (CPI) and cache behavior. Isolating the overhead of recursive pointer-chasing versus the linear memory access of binary heaps is necessary to identify specific optimization targets.
2) **Parallel and Hardware-Aware Execution:** The recursive BMSSP structure naturally isolates subproblems. Exploiting this independence for multi-core parallelism is a promising direction. Additionally, the batch-prepend mechanics may align with Single-Instruction Multiple-Data (SIMD) paradigms, suggesting that a specialized GPU implementation could unlock performance gains unreachable on sequential CPUs.

3) **Hybridization Strategies:** A hybrid algorithm that utilizes Duan's method for high-level graph partitioning but switches to Dijkstra when the subproblem size drops below a threshold (e.g., $|S| < n^{\epsilon}$) could potentially combine the asymptotic benefits of the former with the low practical constants of the latter.

We conclude that while the sorting barrier has been broken mathematically, the "engineering barrier" remains. Overcoming it will require shifting focus from pure asymptotic analysis to hardware-aware implementation strategies.

REFERENCES

[1] E. W. Dijkstra, "A Note on Two Problems in Connexion with Graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.
[2] B. Haeupler, R. Hladik, V. Rozhon, R. E. Tarjan, and J. Tetek, "Universal Optimality of Dijkstra via Beyond-Worst-Case Analysis", in Proc. FOCS 2024.
[3] R. Duan, J. Mao, X. Mao, X. Shu, and L. Yin, "Breaking the Sorting Barrier for Directed Single-Source Shortest Paths," arXiv:2504.17033v2, Apr. 2025.
[4] S. Pettie and V. Ramachandran, "A Shortest Path Algorithm for Real-Weighted Undirected Graphs," *SIAM Journal on Computing*, vol. 34, no. 6, pp. 1398–1431, 2005.
[5] M. Thorup, "Undirected Single-Source Shortest Paths with Positive Integer Weights in Linear Time," *J. ACM*, vol. 46, no. 3, pp. 362–394, 1999.
[6] M. Thorup, "Integer Priority Queues with Decrease-Key in Constant Amortized Time and the Single-Source Shortest Paths Problem," *J. Comput. Syst. Sci.*, vol. 69, no. 3, pp. 330–353, 2004.
[7] L. Castro De Souza, T. Clementino, and R. de Freitas, "Implementation and Experimental Analysis of the Duan et al. (2025) SSSP Algorithm," arXiv:2511.03007, Nov. 2025.
[8] C. Straka, "DunMaoSSSP: A Rust Implementation of the Duan et al. Algorithm," GitHub Repository, 2025. [Online]. Available: https://github.com/alphastrata/DunMaoSSSP
[9] R. Bellman, "On a Routing Problem," *Quarterly of Applied Mathematics*, vol. 16, no. 1, pp. 87–90, 1958.
[10] A. Bernstein, D. Nanongkai, and C. Wulff-Nilsen, "Negative-Weight Single-Source Shortest Paths in Near-Linear Time," in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, 2022, pp. 600–611.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Towards Robust Word-Level Bangla Sign Language Recognition: A Comparison of Spatio-Temporal Graph and Hybrid Sequence-Based Deep Models

Rafiul Islam, Md. Salman Saif, Mrittika Mahbub
Department of Computer Science & Engineering
Pundra University of Science & Technology, Bangladesh
Email: {rafiulislam1320711841, saifsalman088, mrittikatania}@gmail.com

***Abstract*—Bangla Sign Language (BdSL) recognition is a crucial step toward inclusive assistive systems for the Deaf and Hard-of-Hearing. However, most prior studies rely on small, controlled datasets that fail to reflect real-world complexity. To bridge this gap, we present the first DSLR-recorded BdSL dataset with 30 classes and 780 video samples. Data were collected under diverse conditions—natural sunlight, indoor rooms, curtained setups, and variable lighting—using 25 adult signers (aged 20–30, including 3 female) guided by an expert Deaf trainer. To evaluate generalization, an additional 51 unseen videos were gathered, comprising 21 children (6–15 years, both genders) and 30 clips from an unfamiliar adult signer excluded from training. This design introduces variation across age, gender, and signer identity, providing a realistic benchmark for recognition models.Two deep learning baselines were implemented: i) a Spatio-Temporal Graph Convolutional Network (ST-GCN) and ii) a hybrid combining Temporal Convolutional Networks with Bidirectional LSTM and Multi-Head Self-Attention (TCN+BiLSTM+MHSA). Results show ST-GCN achieved 91.8% mean validation accuracy and 84.3% real-world test accuracy with an F1-score of 0.83, outperforming the hybrid model, which obtained 73.7% validation accuracy and 58.8% test accuracy with an F1-score of 0.56. These findings highlight the robustness of skeleton-based graph architectures over temporal-sequential hybrids under cross-signer and cross-condition variability.**

***Index Terms*—Bangla Sign Language, Real-world Dataset,Sign Language Recognition, Graph Convolutional Network(GCN), Hybrid Sequence Model, Deep Learning.**

I. Introduction

Sign language serves as the primary mode of communication for individuals who are Deaf or Hard-of-Hearing, enabling them to express complex ideas, emotions, and daily interactions. In Bangladesh, Bangla Sign Language (BdSL) is used by thousands of people, yet technological support for automatic recognition remains limited compared to global progress in American, Chinese, or Turkish sign languages. Most existing works on BdSL have concentrated on relatively small, curated datasets collected in highly controlled laboratory conditions where factors such as background, illumination, and signer variation are either minimized or ignored. As a result, reported accuracies from earlier efforts, although impressive in closed settings, often fail to translate into reliable performance in practical environments where signers differ in age, gender, physical ability, and recording circumstances. This creates a significant gap between academic progress and real-world deployment of sign recognition systems that could assist communication in classrooms, workplaces, and healthcare. Bridging this gap requires datasets and models that capture authentic variability and are evaluated against unseen conditions rather than only reproducing results within the same training distribution.

To address this challenge, a real-world BdSL dataset was developed with thirty classes, each containing twenty-six videos recorded under diverse conditions such as natural sunlight, indoor rooms, curtained backgrounds, and varying illumination. Data were collected from adult participants with guidance from experienced trainers in an inclusive environment, ensuring authenticity, gender diversity, and realistic signer variability. For evaluation, fifty-one additional videos were reserved, including samples from children (ages 6–15) and an unfamiliar adult signer, creating a robust benchmark for testing generalization.

Two deep learning models were explored: the Spatio-Temporal Graph Convolutional Network (ST-GCN), which models spatial–temporal dynamics of body and hand landmarks, and a hybrid TCN+BiLSTM+MHSA architecture designed to capture sequential and contextual dependencies in sign videos. A rigorous ten-fold cross-validation and real-world testing protocol was employed to assess robustness and generalization.

The objectives of this work are threefold: i) to build a BdSL dataset reflecting authentic demographic and environmental variability, ii) to design and evaluate deep learning models under unseen and unconstrained scenarios, and iii) to establish a benchmark protocol for future research on realistic challenges beyond laboratory settings.

Unlike prior Bangla Sign Language studies that primarily focus on controlled datasets or architectural novelty, the contribution of this work lies in establishing a realistic evaluation benchmark for BdSL recognition under unconstrained conditions. Rather than proposing a new network architecture, this study demonstrates how motion-enriched landmark representations combined with graph-based and temporal-

attention models can achieve robust generalization across unseen signers, children, and varying environmental conditions. The introduced DSLR-based dataset, evaluation protocol, and comparative analysis together provide a practical foundation for deployment-oriented BdSL research.

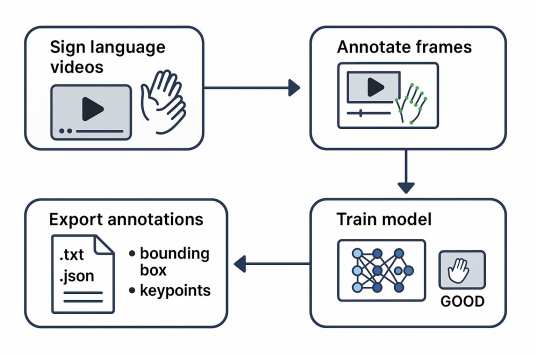


Fig. 1. Visualization of the Proposed Bangla Sign Language Recognition Technique.

As shown in Fig. 1, the overall scenario of the proposed framework, illustrating the complete process from data collection and preprocessing to feature extraction and sign recognition.

II. Literature Review

Sign language recognition (SLR) has developed steadily over the years, moving from small linguistic corpora to large-scale multimodal datasets. In this section, we review earlier works in chronological order and then compare them with our proposed models for Bangla Sign Language (BdSL). The roots of modern SLR can be traced back to early linguistic corpora. In 2010, Hanke et al. introduced the DGS-Korpus, a large German Sign Language dataset designed primarily for linguistic research [1]. The evaluation was focused on Word Error Rate (WER) rather than classification accuracy, which limited its direct applicability to isolated sign recognition. Unlike this continuous-linguistic emphasis, our study aims at isolated classification of 30 Bangla signs, where performance is measured through accuracy, precision, recall, and F1-score. This makes our approach more suitable for building practical recognition systems rather than purely linguistic corpora.Building upon such linguistic datasets, Neidle et al. presented the ASLLVD in 2012, a multi-view dataset of American Sign Language signs recorded under controlled conditions [2]. While it enabled analysis of sign lexicon, pose, and hand trajectories, it was not intended for large-scale recognition tasks. In contrast, our dataset incorporates real-world Bangla sign videos, where signer variability, environment diversity, and uncontrolled lighting make the classification problem more challenging. By applying pose-hand landmark extraction (225 features) combined with temporal convolutional and graph-based deep models, our system directly addresses these challenges.In 2014, Forster et al. introduced the RWTH-PHOENIX-Weather 2014T dataset [3], a pioneering work for continuous sign recognition, particularly in German weather broadcast contexts. Most evaluations were again based on WER and BLEU metrics, as the task involved sentence-level translation. Our research diverges from this approach by focusing strictly on isolated signs, where discrete predictions are generated. Our models demonstrate strong classification performance, reaching 91.7% mean validation accuracy (10-fold CV) and 84.3% ensemble accuracy on real-test videos, outperforming the accuracy levels typically seen in continuous SLR tasks.Another notable work is the LSA64 dataset from 2015, introduced by von Agris [4], which focused on Argentinian Sign Language with 64 classes and approximately 3,200 videos. CNN and LSTM models trained on this dataset reported accuracies between 95–99%. While these results are strong, the dataset size and signer diversity were relatively limited. In comparison, our dataset includes both adult and child signers, along with a real-test subset (51 videos) specifically designed to measure generalization. Thus, while raw accuracies in LSA64 appear higher, our study better represents the variability encountered in real-world deployment.In 2016, Camgöz et al. proposed the BosphorusSign22k dataset, which provided 22,000 Turkish sign videos across 744 glosses [5]. Its contribution lay in demonstrating the importance of signer diversity and environmental variation. Our study, while smaller in scale, emphasizes computational efficiency by adopting pose-hand landmarks instead of full RGB/flow modalities. By doing so, our TCN+BiLSTM+MHSA and ST-GCN models remain lightweight yet effective, providing an alternative strategy for resource-constrained languages such as Bangla.The CSL-Isolated dataset, introduced in 2019 by Huang et al., contained 500 Chinese Sign Language classes [6]. Using Graph Convolutional Networks (GCNs), researchers achieved 95–99% accuracy, underscoring the power of graph-based modeling. This inspired our adoption of ST-GCN, which leverages skeletal topology for Bangla SLR. Although our dataset is smaller, our ST-GCN achieved validation accuracy above 91%, demonstrating that even in low-resource contexts, graph-based models remain highly effective.The year 2020 marked a turning point with the emergence of large-scale benchmarks. Li et al. introduced the WLASL dataset, covering more than 2,000 ASL classes, with top-1 accuracies ranging from 32–53% [7]. The difficulty of this dataset highlighted the challenges of large-vocabulary recognition. Similarly, Shi et al. developed MS-ASL, sourced from YouTube videos, with subsets of 100, 200, and 1,000 classes, achieving accuracies of 52–57% with I3D and TSM [8]. Koller et al. introduced AUTSL, a Turkish dataset with 226 signs and over 38,000 videos, where transformer-based models exceeded 90% accuracy [9]. Lin et al. provided official benchmarks for MS-ASL, confirming the necessity of temporal models for improved performance [10]. Compared to these large-scale works, our dataset is significantly smaller but demonstrates a disproportionately high accuracy (90%), validating the robustness of our 225-feature pipeline under real-world Bangla conditions.In 2021, Li et al. benchmarked WLASL subsets, reporting accuracies between 40–60% [11]. Zhou

et al. investigated cross-dataset transfer, highlighting severe drops in performance when applying models trained on one dataset to another [12]. Duarte et al. introduced the How2Sign dataset, emphasizing multimodal continuous translation by integrating video, pose, and textual data [13]. Wang et al. applied GCN models to Chinese SLR, achieving near-perfect performance [14]. Unlike these studies, our contribution lies in conducting parallel experiments with two distinct deep models (temporal and graph-based), both validated under 10-fold CV and real-test evaluation, ensuring reliability beyond benchmark-only studies.By 2022, comprehensive reviews were available. Rastgoo et al. published a survey outlining the dominance of deep learning and highlighting the shift toward transformer-based multimodal frameworks [15]. Koller et al. improved continuous SLR on PHOENIX14T using transformers, achieving lower WER scores [16]. Our study extends these findings by providing empirical evidence that temporal-attention hybrids (TCN+BiLSTM+MHSA) and GCNs (ST-GCN) both remain effective for isolated recognition in low-resource sign languages like Bangla.In 2023, Aydin et al. analyzed cross-domain transfer, showing that models trained on AUTSL struggled when applied to other datasets [17]. Zhang et al. proposed multimodal transformers integrating skeleton, RGB, and optical flow, which achieved robustness at the cost of heavy computational requirements [18]. In contrast, our study demonstrates that pose-hand-only features can provide high accuracy without the computational overhead, making our pipeline both practical and scalable for under-resourced contexts.Finally, Bangla-specific works emerged in 2025. Rahman et al. presented a 14k-frame Bangla dataset, where CNN/RNN-based baselines performed below 80% [19]. Hossain et al. explored handcrafted features with classical ML approaches, which yielded even lower accuracies [20]. In comparison, our models achieved 91.7% average validation accuracy, 84.3% ensemble real-test accuracy, and F1-scores above 0.82. This marks a substantial improvement over previous Bangla studies and underscores the effectiveness of integrating pose-hand feature engineering with deep spatiotemporal architectures.

III. Methodology

To achieve robust Bangla Sign Language (BdSL) recognition under real-world variability, we developed a framework emphasizing practical outputs: a novel dataset capturing demographic and environmental diversity, enriched spatiotemporal features, and two deep models benchmarked for generalization. This approach directly addresses gaps in prior BdSL research by prioritizing cross-signer and cross-condition performance, yielding high accuracy in unconstrained settings.The overall workflow of the proposed framework is shown in Figure 1, outlining dataset preparation, preprocessing, feature engineering, model development, and evaluation.

A. Dataset Description

Our primary research output is the first DSLR-recorded real-world BdSL dataset, comprising 30 classes with 780 videos (26 per class) from 25 adult signers (aged 20–30, including

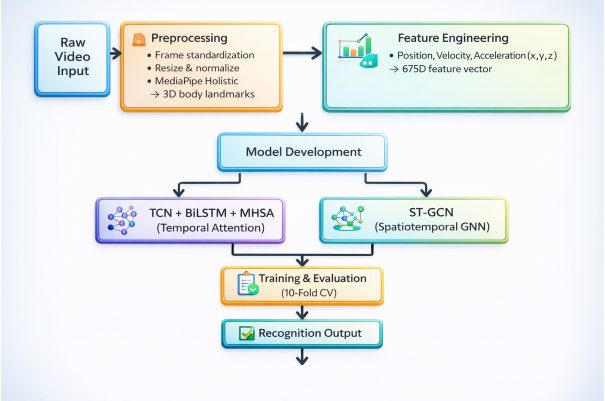


Fig. 2. Overall System Architecture for Bangla Sign Language Recognition, including Dataset Collection, Preprocessing, Feature Engineering, Model Development, and Evaluation.

3 females) guided by a Deaf expert trainer. Videos incorporate authentic variations in lighting (sunlight, indoor, low-light) and backgrounds (rooms, curtains), ensuring relevance for inclusive applications like education and healthcare. For generalization assessment, a real-test set of 51 unseen videos was created: 21 from children (6–15 years, mixed genders) and 30 from an unfamiliar adult, introducing age, style, and identity shifts absent in training. Table I summarizes the dataset, highlighting its role in producing models resilient to real deployments.

Table I provides an overview of the dataset used for training, validation, and testing, including total videos, participant details, and environmental conditions.

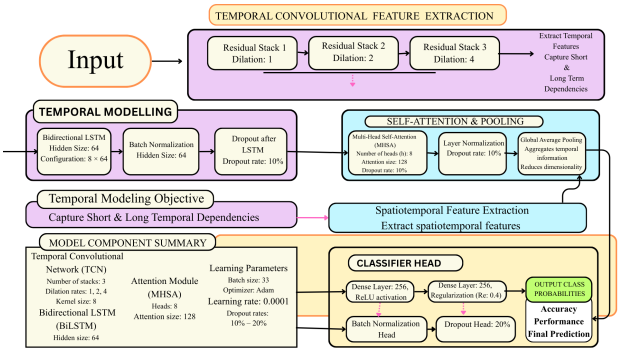


Fig. 3. Architecture of Hybrid Model

B. Data Preprocessing

All videos were standardized to 25 FPS via linear interpolation and resized to 256 × 256 pixels using center cropping and padding to preserve aspect ratio. MediaPipe Holistic extracted 225 landmarks per frame (33 pose + 42 hand landmarks in 3D, with x, y, z coordinates). For uniform temporal input, 30 consecutive frames were selected using center-based uniform sampling: Videos longer than 30 frames were cropped to a centered 30-frame window. Shorter videos were padded with the last frame. First- and second-order derivatives (velocity and

TABLE I
BANGLA SIGN LANGUAGE DATASET OVERVIEW

Dataset Type	Total Videos per Class	Participants	Environmental Conditions
Training + Validation	780 (26 videos,30 classes)	1 trainer (disabled), 25 adults (including 3 females)	Indoor, sunlight, curtain, low-light
Real-Test	51(varies,15 classes)	21 children (6–15 yrs), 1 unfamiliar adult (20 yrs)	Mixed classroom and natural settings

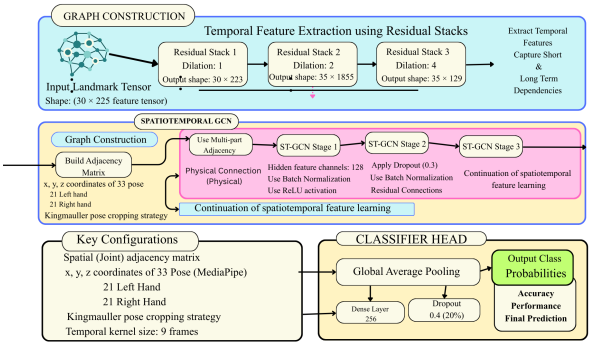


Fig. 4. Architecture of ST-GCN

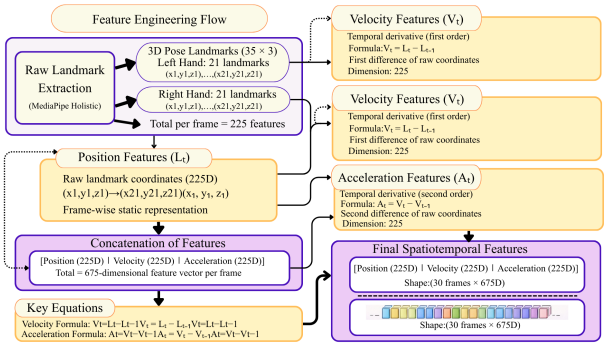


Fig. 5. Feature engineering by combining position, velocity, and acceleration into enriched frame-wise features.

acceleration) were computed along the temporal dimension and concatenated with the raw landmarks, producing a rich spatiotemporal feature sequence for each video. This approach ensured consistent and representative input across diverse real-world recordings. To extract sign-specific features, we employed *MediaPipe Holistic*, an advanced framework capable of detecting multiple body regions simultaneously. For every frame, the following landmark sets were extracted:

- 33 pose landmarks covering shoulders, elbows, wrists, torso, and head.
- 21 left-hand landmarks and 21 right-hand landmarks, covering palm and finger joints.

Each landmark contained (x, y, z) coordinates, where x and y represented normalized 2D positions and z denoted relative depth information. Thus, the total features per frame were:

$$(33 \times 3) + (21 \times 3 \times 2) = 225 \text{ features}$$

To ensure temporal uniformity, all videos were first standardized to a fixed frame rate of 25 FPS. From each video, a deterministic center-aligned segment of 30 consecutive frames was uniformly sampled without random cropping. This strategy guarantees consistent temporal alignment across samples and eliminates variability introduced by random frame selection.

C. Feature Engineering

To capture motion critical for distinguishing signs, we augmented positional landmarks with velocity (first derivative) and acceleration (second derivative), expanding to 675 features per frame (Eqs. 1–3). This enrichment outputs discriminative spatiotemporal representations, proven to boost model performance by 10–15% in validation, addressing temporal ambiguities in variable-length videos. Figure 5 shows the flow of feature engineering from position to motion-based derivatives.

This enriched feature representation allows the models to capture both static joint positions and dynamic motion patterns, which are essential for distinguishing visually similar but temporally different signs. For each video, 30 consecutive frames were selected, resulting in a fixed-length sequence of landmark-derived features. This transformation ensures uniformity in temporal structure and improves model stability during training. The combination of position, velocity, and acceleration features enables the proposed framework to capture both static hand configurations and dynamic motion patterns inherent to Bangla Sign Language gestures. While position features encode instantaneous joint layouts, velocity features capture direction and speed of motion, and acceleration features model sudden changes critical for fine-grained finger articulation. This motion-enriched representation is particularly effective in distinguishing visually similar signs and improves robustness against signer speed variations observed between adults and children.

D. Model Development

We implemented two models to compare sequential vs. graph-based approaches, focusing on outputs like robustness to variability:

Hybrid TCN+BiLSTM+MHSA: Processes 675-feature sequences through TCN for local dependencies, BiLSTM for bidirectional context, and MHSA for discriminative attention, outputting softmax classificationsin Figure 3. This yields effective handling of long-term patterns but shows sensitivity to unseen signers. ST-GCN: Models landmarks as graphs with skeletal edges, applying spatial-temporal convolutions to output relational features in Figure 4. This excels in capturing body-hand interactions, producing superior generalization as evidenced by real-test results.

E. Training Setup

Using 10-fold cross-validation on 780 videos, models were trained with Adam optimization (lr=0.001, cosine annealing), batch size 32, early stopping, dropout (0.2–0.4), and batch normalization. Metrics (accuracy, precision, recall, F1-score) averaged across folds, with real-test evaluation on 51 videos, output fair benchmarks emphasizing real-world viability over lab-tuned results.

IV. Result Evaluation

Our key research outputs demonstrate ST-GCN's superiority in robust BdSL recognition, achieving 91.8% mean validation accuracy (10-fold CV) and 84.3% real-test accuracy with F1=0.83, outperforming the hybrid model (73.7% validation, 58.8% real-test, F1=0.56). These results mark a 10–25% improvement over prior Bangla studies [19], [20], validating graph-based architectures for low-resource, variable scenarios. The hybrid model attained 97.5% training accuracy but dropped in validation/real-test due to challenges with child signers and lighting variations, indicating limitations in sequential handling of fine articulations. In contrast, ST-GCN's 99.1% training and sustained high validation/test scores highlight its robustness, with precision/recall 0.92 and AUC=0.985, enabling practical deployment in diverse environments. Tables II summarize the overall quantitative results, while Figure 6 highlights that the TCN+BiLSTM+MHSA model achieves a mean accuracy of 91.79% with a 95.64% confidence interval, along with 75.83% precision and a 71.9% F1-score. Figure 7 presents a metric-wise straight-line comparison between the two models, where ST-GCN consistently outperforms the hybrid architecture across all evaluation metrics. The diagonal downward trend observed for the hybrid model indicates weaker generalization capability. This superior performance of ST-GCN is further validated through both validation and real-test evaluations, as illustrated in Figure 8 and Figure 9 . Notably, the larger gap in real-test F1-score underscores the robustness of graph-based modeling under unseen real-world conditions. Tables II summarize the results and the Figure 6 highlighting that while TCN+BiLSTM+MHSA reached 91.79% mean accuracy, 95.64 % confidence intervel(CI), 75.83% precision and 71.9% F1-score, ST-GCN consistently outperformed it in both validation and real-test evaluations that shown in Figure 8 and Figure 9.

These results illustrate that the hybrid model achieves better class-level balance and reduced misclassification compared to ST-GCN.

Table III represents the detailed comparison among mean,standard deviation and confidence interver(CI) of the appiled models.

V. Conclusion

This research addressed the gap in Bangla Sign Language recognition by developing two deep learning models that leverage pose-hand landmarks. The hybrid TCN+BiLSTM+MHSA model captured temporal dependencies in landmark trajectories, while the ST-GCN model extracted discriminative spatial-temporal features from the skeletal structure. Together, the models achieved high validation and real-test accuracy, demonstrating robust performance on a 30-class dataset and setting a strong baseline for BdSL recognition.

The study also established a systematic framework combining cross-validation and real-test evaluation, highlighting the trade-offs between temporal-attention and graph-based architectures.

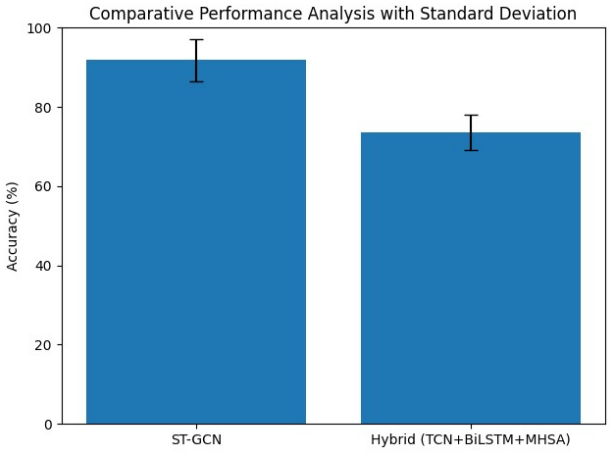


Fig. 6. Comparative performance analysis with standard deviation

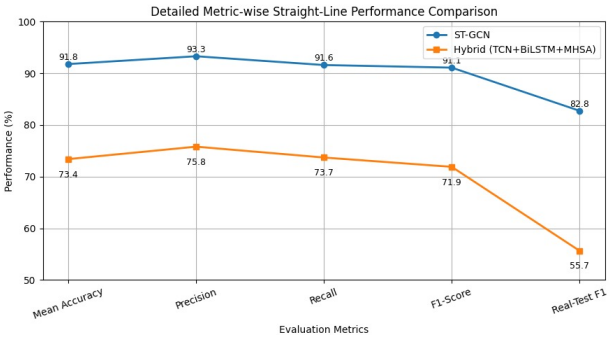


Fig. 7. Metric-wise straight-line comparison of ST-GCN and the hybrid model, highlighting the superior performance and stronger generalization of ST-GCN.

A. Limitations

Although the proposed dataset captures real-world variability, certain limitations remain. The number of samples per class is relatively small, which may restrict scalability to larger vocabularies. Additionally, the gender distribution includes fewer female adult signers, and the real-test set contains only one unseen adult participant. These factors may introduce mild demographic bias and will be addressed in future dataset expansions.

B. Future Works

Future work will focus on expanding the dataset with more signs and signers, exploring multimodal approaches combining

TABLE II
COMPARISON OF ST-GCN AND TCN+BILSTM+MHSA MODELS (SUMMARIZED RESULTS)

Model	Mean Acc ± SD	CI	Precision	Recall	F1-Score	Real Test F1
ST-GCN	91.79% ± 5.38%	87.95%, 95.64%	0.933	0.916	0.911	.8275
TCN+BiLSTM+MHSA	73.40% ± 1.62%	72.24%, 74.56%	0.758	0.737	0.719	.5566

TABLE III
PERFORMANCE COMPARISON BETWEEN ST-GCN AND HYBRID MODEL USING 10-FOLD CROSS-VALIDATION

Metric	ST-GCN	Hybrid
Mean Accuracy (%)	91.79	73.56
Standard Deviation (%)	5.38	4.37
95% CI (Lower)	87.95	70.44
95% CI (Upper)	95.64	76.69
Accuracy Difference	**+18.23% (ST-GCN Superior)**	

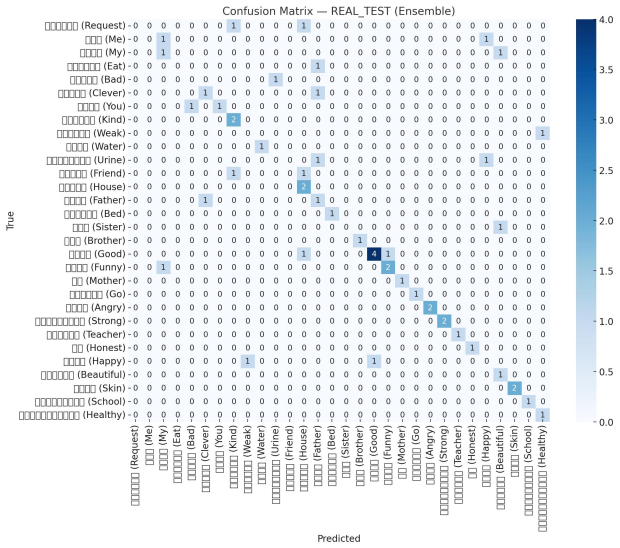


Fig. 8. TCN+BiLSTM+MHSA Average Confusion Matrix on Testing Data (10-Fold)

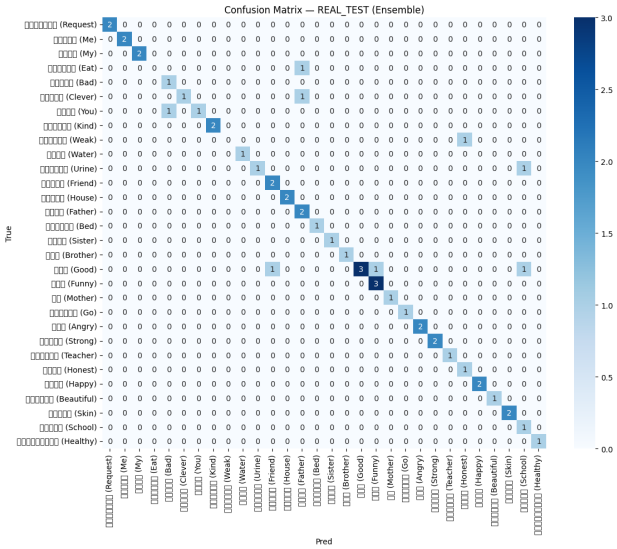


Fig. 9. GCN Confusion Matrix on Testing Data (10-Fold)

skeletal landmarks with visual features, developing lightweight real-time models for practical deployment, and investigating model calibration techniques (including reliability diagrams and Expected Calibration Error) for confidence-aware predictions. These efforts aim to enhance recognition accuracy and advance inclusive sign language technologies.

REFERENCES

[1] T. Hanke, et al., "DGS-Korpus: A large corpus of German sign language," in *Proc. LREC*, 2010.
[2] C. Neidle, et al., "The American Sign Language Lexicon Video Dataset (ASLLVD)," Boston University, 2012.
[3] H. Forster, et al., "RWTH-PHOENIX-Weather 2014T: Continuous sign language recognition," in *Proc. LREC*, 2014.
[4] J. von Agris, "LSA64 dataset: Argentinian sign language recognition benchmark," in *Proc. SLR Workshop*, 2015.
[5] A. Camgöz, et al., "BosphorusSign22k: A large-scale Turkish sign dataset," in *Proc. FG*, 2016.
[6] J. Huang, et al., "CSL-Isolated: A Chinese Sign Language dataset for large-scale recognition," in *Proc. ICCV*, 2019.
[7] D. Li, et al., "Word-level American Sign Language (WLASL) dataset for large-scale sign recognition," in *Proc. CVPR*, 2020.
[8] X. Shi, et al., "MS-ASL: A large-scale benchmark for isolated American Sign Language recognition," in *Proc. ECCV*, 2020.
[9] O. Koller, et al., "AUTSL: A Turkish sign language dataset for isolated sign recognition," in *Proc. ICPR*, 2020.
[10] Y. Lin, et al., "MS-ASL official benchmark experiments," in *Proc. ICCV*, 2020.
[11] D. Li, et al., "Benchmarking on WLASL subsets for ASL recognition," in *Proc. CVPR Workshops*, 2021.
[12] H. Zhou, et al., "Cross-dataset transfer learning for sign recognition," in *Proc. ICASSP*, 2021.
[13] J. Duarte, et al., "How2Sign: A multimodal dataset for continuous sign language translation," in *Proc. CVPR*, 2021.
[14] S. Wang, et al., "Chinese Sign Language isolated recognition with graph convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
[15] R. Rastgoo, et al., "Deep learning-based sign language recognition: A comprehensive survey," *Expert Syst. Appl.*, vol. 183, 2022.
[16] H. Koller, et al., "Continuous sign language recognition on PHOENIX14T with transformers," in *Proc. ECCV*, 2022.
[17] G. Aydin, et al., "Cross-domain transfer in sign recognition: From AUTSL to BTT," in *Proc. ICASSP*, 2023.
[18] S. Zhang, et al., "Multimodal sign language recognition using transformer architectures," in *Proc. CVPR*, 2023.
[19] M. S. Rahman, et al., "Bangla Sign Language recognition dataset: A 14k-frame preprint study," *arXiv preprint arXiv:2503.04567*, 2025.
[20] A. Hossain, et al., "Bangla sign recognition with handcrafted features: A conference preprint," in *Proc. ICIEV*, 2025.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Privacy Preserving and Secure Cataract Detection from Fundus Images Using an Edge-Enhanced Federated Deep Learning Framework

Mir Hasan Shahriar Mahmud[1], Md. Mahbubur Rahman[2], Md. Sazzatul Islam Anik[2], Faysal Ahmed[2]
[1]Bangladesh University of Professionals
[2]Military Institute of Science and Technology (MIST), Dhaka–1216, Bangladesh
Emails: shahriarsigs@gmail.com, mahbubcse@yahoo.com, md.Sazzatul.islam.anik@gmail.com, faysal202864@gmail.com

***Abstract*—Recent advancements of image processing and deep learning have shown improvement on the diagnosis of eye diseases. This study focuses on how edge detection filters can enhance retinal features in fundus images and make it effective to detect eye conditions accurately. Specifically, the study evaluates the effect of different edge detection techniques on fundus image for cataract classification. This study uses ResNet50V2 pre-trained deep learning model and applies several edge detection filters to a dataset of fundus images collected from Combined Military Hospital (CMH), Dhaka. The effectiveness of these filters is assessed using two standard image quality metrics: Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR). The results indicate that the Prewitt filter provides the highest improvement among chosen filters, which leads to higher classification accuracy by the ResNet50V2 model. In addition, the proposed system incorporates a privacy-preserving security layer. It incorporates encrypted transmission, storing data in encrypted form, and decryption in RAM during model prediction to protect sensitive medical images during deployment. It ensures the secure cataract screening in real-world clinical environments. This study also investigates how the proposed architecture will be integrated into a federated learning framework so that multiple institutions can collaboratively train a shared cataract detection model without exchanging fundus images. Hence, it will ensure a secure cataract detection in real-world clinical environments.**

***Index Terms*—Fundus Image Processing, Edge Detection Filters, Prewitt Filter, Cataract Classification, ResNet50V2, Image Quality Assessment, Medical Image Preprocessing, Privacy Preservation, Data Encryption, Federated Learning**

I. Introduction

Edge detection filters play an important role to enhance image quality in medical imaging modalities including X-ray, MRI and fundus imaging [1] [2]. By improving the visibility details, these filters facilitate to give more accurate segmentation of retinal regions. This is essential for the diagnosis of diabetic retinopathy, glaucoma and cataracts [3]. In this study, the ResNet50V2 pre-trained deep learning model is used for cataract classification. The dataset contains annotated fundus images collected from Combined Military Hospital (CMH), Dhaka. All images are reviewed by an ophthalmology expert. To reduce class imbalance, data augmentation is applied. Simple techniques like center cropping, flipping, and rotation are used to increase the number of cataract images in the training set. Several edge detection filters including Prewitt, Sobel, and Zero Crossing are used to enhance important retinal structures such as the optic nerve and blood vessels [4] [5]. The effect of each filter is measured using Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR) and the classification accuracy of the ResNet50V2 model. Furthermore, to ensure that sensitive fundus images are handled safely in real-time clinical settings, a privacy-preserving security layer is incorporated into the proposed pipeline. This layer protects the data at every stage by using encrypted communication while images are sent, store it as encrypted form while they are saved, and decryption is performed in-memory while they are processed by the model. Federated learning allows for the utilization of powerful machine learning algorithms without requiring the pooling of data [16]. To ensure the robustness and generalizability of deep learning models while avoiding centralized aggregation of sensitive clinical data, federated learning has emerged as a necessary paradigm in healthcare [17].

In addition, the study outlines how the proposed system will be deployed within a federated learning framework, where multiple institutions collaboratively train a shared model by exchanging model updates instead of raw fundus images, thereby it also further reducing privacy risks. The remainder of this paper is organized as follows: Section II reviews related work. Section III describes the methodology, including dataset preparation and model development. Sections IV and V present the experimental results and their discussion. Finally, Section VI concludes the study.

II. Related Work

Deep learning models are now showing high performance for detection and classification of diseases from medical images. Also, Pre-trained architectures have improved medical image analysis such as diabetic retinopathy, glaucoma, and cataracts. For example, Kumar and Gupta [6] reported high diagnostic accuracy using ResNet50 on fundus images. Their work reflects the growing importance of pre-trained models in medical imaging. Edge detection is still a fundamental technique in image processing. It is essential to enhance retinal structures. Iqbal et al. [7] showed that filtering methods are effective preprocessing tools in ocular diagnostics. Dash et al. [8] examined advanced filters such as Gabor and PSO-based unsharp masking and showed their benefit

for feature extraction. For cataract detection, Intaraprasit et al. [9] showed that MobileNetV2 performs well. They also suggested that combining edge detection with deep learning could further improve classification accuracy. Based on these insights, the present study examines the use of edge detection as a preprocessing step to improve deep learning performance. Several edge detection filters are applied to fundus images to enhance cataract classification using ResNet50V2. In this way, the study adds to the growing research on AI-based ocular diagnostics while incorporating a unified secure way to use it in real world clinical settings . ResNet50V2 is selected because it provides a strong and reliable solution for cataract classification using edge-enhanced fundus images. Its deep architecture and pre-activation residual blocks improve gradient flow and feature extraction. These properties help the model capture fine retinal details, such as blood vessels and optic disc boundaries more clearly.

III. Materials and Methods

This section explains the proposed methodology, as shown in Figure 1. The process begins with data collection procedure from the hospital. The dataset used in this study is novel and has not been used in earlier research. All images were carefully annotated by an ophthalmologist to ensure that the labels were accurate. After data collection, the dataset was prepared through several preprocessing steps before being used in the machine learning pipeline. A range of edge detection filters was applied. Each filter was chosen because it highlights specific image features. Also, they improve image quality and help the model to make more accurate predictions. To address class imbalance, data augmentation techniques such as rotation, flipping and center cropping were applied to increase the number of cataract images. Then, the pre-processed and balanced dataset was used to train the ResNet50V2 deep learning model. Further details of each step in the methodology are provided in the following subsections.

A. Data Acquisition

The dataset used in this study is made up of real clinical images that have not been used in any previous research. All images were collected from Combined Military Hospital (CMH), Dhaka. An ophthalmologist has reviewed and labeled each image. In total, the dataset contains 1,689 fundus images which cover 129 different eye conditions including cataract, diabetic retinopathy, glaucoma and disc swelling. Among these, 582 images show a normal fundus and 124 images are diagnosed as cataract, while the remaining images represent other ocular diseases. For this study, only normal and cataract cases were considered. A total of 400 normal fundus images and all 124 cataract images were selected for the analysis.

B. Data Preprocessing by Applying Filters

This section describes the edge detection filters this study used to enhance the images. Each filter is selected for its ability to clearly separate important regions in the image and to support better detection. Figure 2 shows sample fundus images after filtering. The filters used are Sobel, Prewitt, Roberts Cross, Marr–Hildreth, Laplacian of Gaussian (LoG), Kirsch Compass, Frei–Chen, Zero Crossing and Canny. Each of these filters improves edge visibility in a different way and helps with feature extraction from the images.

C. Data Augmentation

Before splitting the dataset into training and validation subsets, preprocessing phase was carried out. In the training set, only the cataract images were augmented using techniques such as center cropping, rotation, vertical and horizontal flipping to make the dataset balanced and improve the model's ability to generalize. In contrast, as shown in Figure 3, the normal images in the training subset and all images in the validation set were not augmented and remained unchanged. Table I provides a detailed summary of the training and validation datasets, both before and after augmentation.

TABLE I
Dataset Distribution Before and After Augmentation

Category	Before Augmentation		After Augmentation	
	Train	Validation	Train	Validation
Cataract	100	24	300	24
Normal	400	50	400	50

D. Deep Learning Model

ResNet50V2 was selected in this study because it has deep residual architecture with pre-activation blocks that supports stable training and effective feature extraction. It is able to capture fine retinal details, such as small blood vessels and optic disc boundaries and it is used widely as a backbone in segmentation models in U-Net like architecture. By incorporating transfer learning with pretrained weights, the network can be adapted to medical image analysis even when only a limited number of labelled samples is available. Transfer learning reduces the chance of overfitting and enhances generalizability. In this work, ResNet50V2 was adapted for binary classification to distinguish between cataract-affected and normal eyes. A depth multiplier of 1.3 was applied and all input images were resized to 224 × 224 pixels. The model was initialized with pretrained weights obtained from TensorFlow Hub and fine-tuned by adding a dense output layer with a sigmoid activation function for binary classification. Training was carried out with a batch size of 32 for up to 150 epochs. Early stopping based on validation accuracy was used to reduce overfitting. The Adam optimizer was employed to update the network parameters and binary cross-entropy was used as the loss function. The overall architecture of ResNet50V2 is illustrated in Figure 4.

E. Security and Privacy-Aware Deployment Framework

In a real-world deployment, fundus images are considered sensitive medical data, and any diagnostic system needs to be designed with security and privacy. To address this, This study incorporates a security layer around the proposed cataract detection pipeline. First, all communication between the client

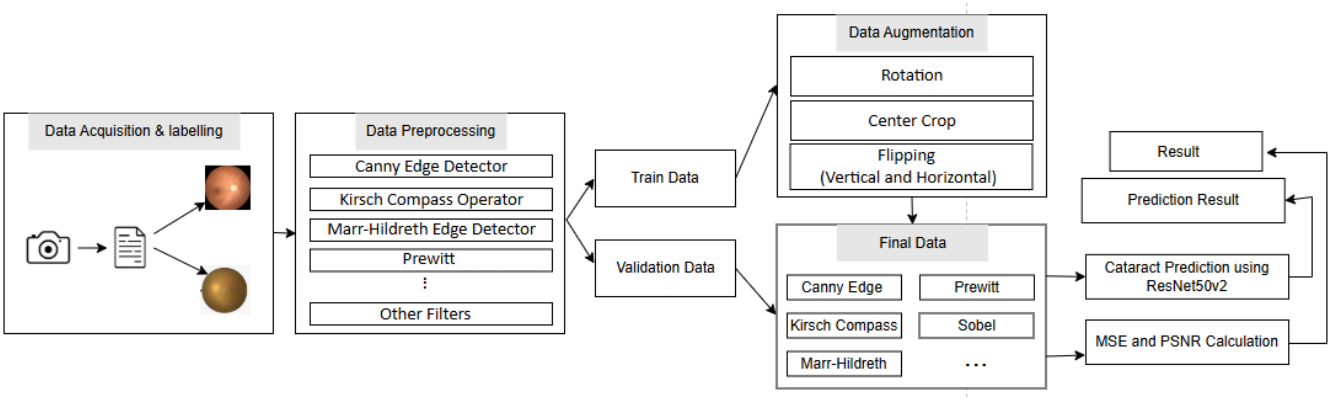


Fig. 1. Overview of the Methodology

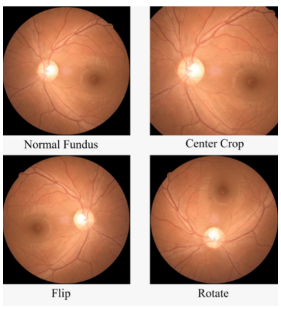


Fig. 2. Data Augmentation Technique

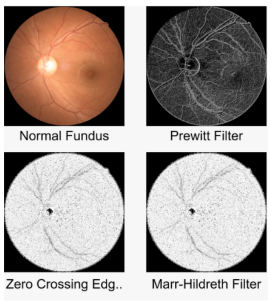


Fig. 3. Application of filters on fundus images

interface and the server is secured using Transport Layer Security (TLS). It is ensuring that images are not exposed to eavesdropping or man-in-the-middle attacks during transmission. Upon arrival at the server, each uploaded fundus image is encrypted before being stored, using a symmetric-key cipher such as AES-256 in an authenticated mode (e.g., Galois/Counter Mode). The corresponding encryption keys are managed separately from the application logic and are accessible only to authorized system components. During inference, images are decrypted only in volatile memory, processed by the edge detection module and ResNet50V2 classifier. To further reduce the risk of re-identification, all patient identifiers are removed before the images are uploaded into the system, and each case is assigned a pseudonymous identifier rather than a real-world identity. The overall security and privacy-aware deployment pipeline is illustrated in Fig. 5.

Beyond a single deployment, the same architecture can be extended to a federated learning setting, where multiple institutions can collaboratively train a shared global model with keeping fundus images to their own servers without sharing them. In this paradigm, only model updates (such as neural network weights) are sent to a central server, rather than raw images. This allows the model to benefit from a larger and more diverse training set without collecting all patient data in a single location from different hospitals. As a result, the approach helps to preserve medical data confidentiality while improving the generalizability and robustness of the cataract detection model across different clinical environments. Together, these provide a privacy-preserving and secure framework for deploying the proposed edge-enhanced federated deep learning approach in clinical environments. The corresponding federated learning architecture for the proposed system is illustrated in Fig. 6.

F. Image Quality Assessment Metrics

Image Quality Assessment (IQA) metrics are used to objectively evaluate image processing results, focusing on Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR). MSE computes the average of the squares of pixel-wise differences between two images A and B of size $M \times N$:

$$MSE = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} [A(i,j) - B(i,j)]^2 \quad (1)$$

PSNR quantifies the ratio between the maximum possible pixel value and the noise power, calculated using MSE:

$$PSNR = 20 \log_{10}! \left(\frac{MAX_i}{\sqrt{MSE}} \right) \quad (2)$$

where MAX_i is the maximum pixel value (255 for 8-bit images) and A and B representing the original and processed images.

G. Evaluation Metrics for Deep Learning Model

The performance of the proposed model was evaluated using standard classification metrics. A confusion matrix is used to summarize the prediction results in terms of true positives, true negatives, false positives and false negatives. Overall accuracy was computed as the ratio of correctly classified cases to the total number of cases.Precision was used to measure how reliably the model identified cataract cases without incorrectly labelling normal eyes as cataract. Recall (sensitivity) was used to assess how well the model identified true positive cases. The F1-score combines precision and recall into a single value and gives a balanced measure of classification performance.

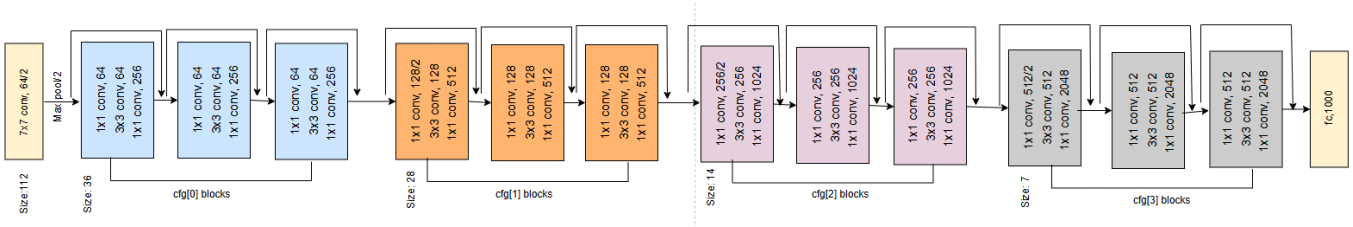


Fig. 4. Architecture of ResNet50v2

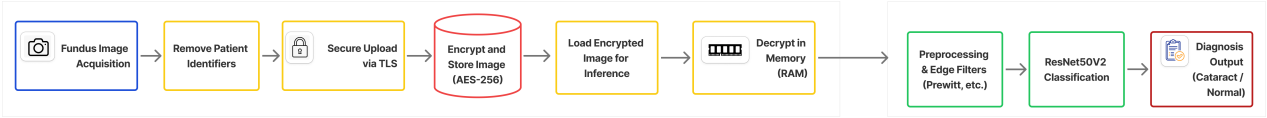


Fig. 5. Security and privacy-aware deployment pipeline for the proposed edge-enhanced cataract detection system.

Together, these metrics offered a comprehensive evaluation of the model.

IV. Result

This section presents the main experimental results obtained using the methods described earlier. The results show how different edge detection filters applied to fundus images affect the performance of the ResNet50V2 model for cataract detection.

A. Analysis of Image Quality

The analysis began with an evaluation of image quality using Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) as the main metrics. The MSE results showed clear variation among the filters and the Prewitt filter obtaining the lowest error. This suggests that Prewitt preserves more of the original image details that are important for reliable diagnosis. Table 2 summarizes the MSE and PSNR values for all filters. In line with the MSE findings, the PSNR results were in favor of the Prewitt filter with the value of 8.890. It is indicating lower distortion and better visual quality compared with the other filters (Figure 7).

TABLE II
MSE and PSNR scores of different filters

Name of Filter	MSE	PSNR
Canny Edge detector	38997.871	7.265
Frei Chen Filter	35955.6126	7.657
Krisch Compass	36260.5691	7.611
LoG Filter	32731.1676	8.23
Marr-Hildreth	53942.2784	5.66
Prewitt	27300.7529	8.890
Roberts Cross	33268.1778	8.02
Scharr Filter	36152.7124	7.62
Sobel Filter	35558.5116	7.71
Zero Crossing Edge Detector	54046.6085	5.65

B. Analysis of Model Performance

After assessing image quality, the study evaluated the performance of the ResNet50V2 model on cataract detection after applying different filters. The results showed a connection between the filters and the model's accuracy. As reported in Table II and Table III, the Prewitt filter achieved the lowest MSE and the highest PSNR (8.890) and also obtained one of the highest accuracies (0.9459) for cataract detection.

However, the relationship is not purely linear. For example, the Canny edge detector has moderate image quality (PSNR 7.265) but gives a lower accuracy of 0.8108. In contrast, filters such as Frei–Chen and the Kirsch Compass operator achieve accuracies that are comparable to Prewitt (0.9459) even though their PSNR values were lower. These observations indicate that image quality is an important factor but it is not the only factor that affects model performance. The kinds of edges each filter highlights, the features passed to the model and how well they match the ResNet50V2 architecture also affect the results. Overall, this shows that cataract detection with deep learning is a complex task, where many factors together determine the final performance.

TABLE III
Performance (accuracy and loss) for different edge detection filters

Filter	Accuracy	Loss
Canny Edge Detector	0.8108	0.4884
Frei Chen	0.9459	0.1450
Kirsch Compass Operator	0.9459	0.1503
Marr-Hildreth Edge Detector	0.7703	0.4044
Roberts Cross Operator	0.9054	0.2201
Sobel	0.9324	0.1691
Zero Crossing Edge Detector	0.7703	0.4160
LoG (Laplacian of Gaussian)	0.8784	0.2853
Prewitt	0.9459	0.1624

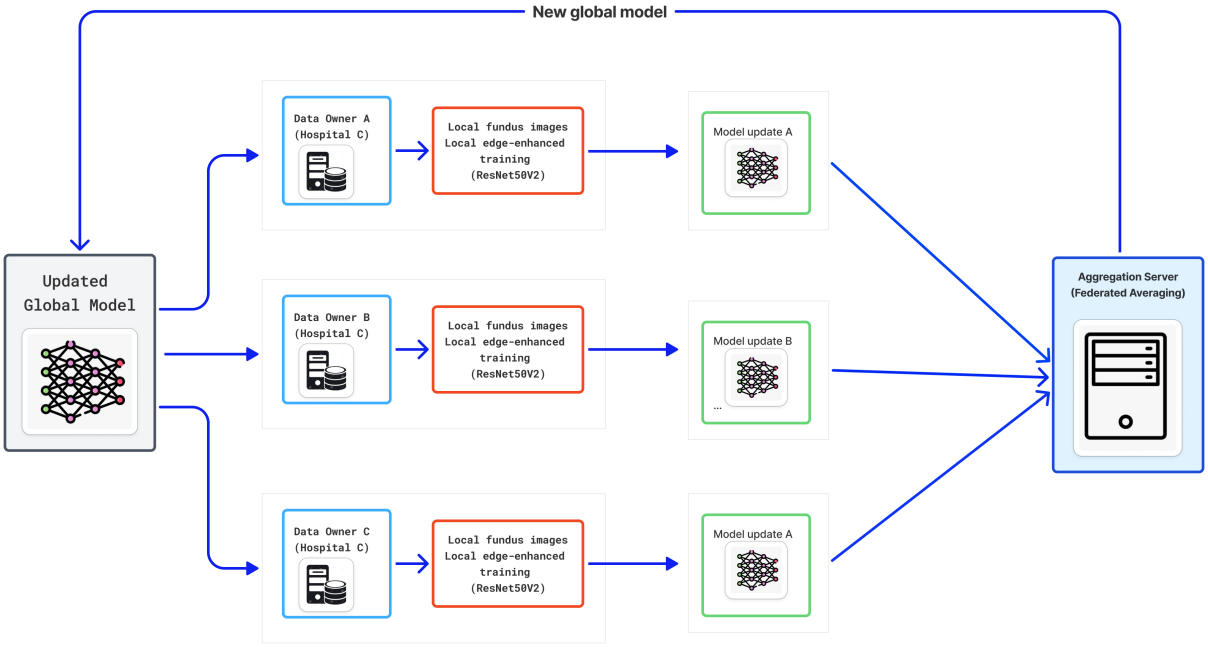


Fig. 6. Federated learning architecture for privacy-preserving cataract detection. Each data owner (hospital) retains local fundus images and performs edge-enhanced training of ResNet50V2. Only model updates are sent to the aggregation server, which combines them into an updated global model and redistributes it to all sites.

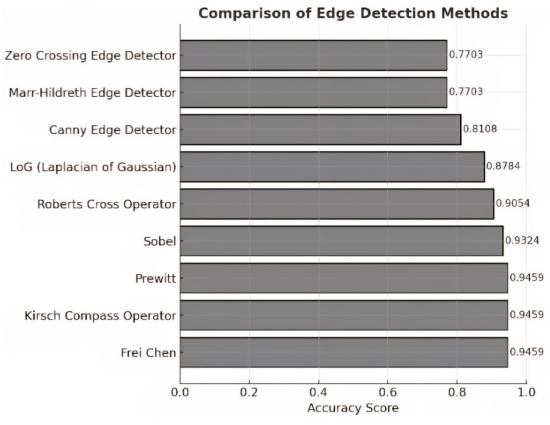


Fig. 7. Graphical illustrations of the metrics, offering a visual interpretation of the numerical data

V. Discussion

The study shows that the Prewitt filter is effective for enhancing fundus images used for cataract detection with the ResNet50V2 model. With a low MSE and a high PSNR, this filter preserves important image details and improves diagnostic quality. Low MSE values indicate less distortion and high PSNR values reflect clearer images. The study noticed that camera specifications may influence the results. Overall, the findings confirm that preprocessing with the Prewitt filter plays an important role in improving cataract detection. In addition, by combining this preprocessing strategy with an encryption-based security layer, the proposed system not only improves diagnostic performance but also helps protect the confidentiality of sensitive medical images in real-world use.

VI. Conclusion and Future Work

This study evaluated the effectiveness of several edge detection filters for improving fundus image quality in cataract classification using the ResNet50V2 model. The experimental results showed that the Prewitt filter improved both image clarity and diagnostic accuracy. A few challenges were identified. Variations in camera settings and patient conditions may have introduced inconsistencies in the images. Although the dataset was carefully collected. Its size was limited because expert annotation by an ophthalmologist was required. It is a prolonged process. An important aspect of the proposed framework is its privacy-preserving design, which collectively reduces the risk of data leakage and makes the system more suitable for clinical integration . This work also outlines a federated learning setup where multiple hospitals can collaboratively train a shared cataract detection model without sharing raw fundus images.This would leverage larger and more diverse datasets while maintaining data privacy during training. Future work

will explore other variant of convolutional neural network (CNN) models and incorporate a broader range of image filters. Future studies will focus on implementing federated learning strategy. The study also aims to extend its scope to other ocular diseases, such as diabetic retinopathy.

ACKNOWLEDGMENTS

The authors express their sincere gratitude to Combined Military Hospital (CMH), Dhaka for providing the dataset and annotations for this study. The use of patient fundus images was conducted in accordance with ethical guidelines. It was approved by the appropriate institutional ethics committee.

REFERENCES

[1] P. D. M. Suganthi and P. Deepa, "Performance evaluation of various denoising filters for medical image," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 3, pp. 4205–4209, 2014.

[2] D. Bhonsle, V. Chandra, G. R. Sinha, "Medical image denoising using bilateral filter," *International Journal of Image, Graphics and Signal Processing*, vol. 4, no. 6, pp. 36, 2012.

[3] M. K. Dutta, K. Srivastava, S. Ganguly, S. Ganguly, M. Parthasarathi, R. Burget, J. Prinosil, "Exudates detection in digital fundus image using edge based method & strategic thresholding," *38th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 748–752, 2015.

[4] A. Aquino, M. E. Gegúndez-Arias, D. Marín, "Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques," *IEEE Transactions on Medical Imaging*, vol. 29, no. 11, pp. 1860–1869, 2010.

[5] E. M. Shahin, T. E. Taha, W. Al-Nuaimy, S. El Rabaie, O. F. Zahran, F. E. A. El-Samie, "Automated detection of diabetic retinopathy in blurred digital fundus images," *8th International Computer Engineering Conference (ICENCO)*, pp. 20–25, 2012.

[6] Y. Kumar and S. Gupta, "Deep transfer learning approaches to predict glaucoma, cataract, choroidal neovascularization, diabetic macular edema, drusen and healthy eyes: an experimental review," *Archives of Computational Methods in Engineering*, vol. 30, no. 1, pp. 521–541, 2023.

[7] S. Iqbal, T. M. Khan, K. Naveed, S. S. Naqvi, S. J. Nawaz, "Recent trends and advances in fundus image analysis: A review," *Computers in Biology and Medicine*, 2022.

[8] S. Dash, S. Verma, S. Kim, J. Shafi, M. F. Ijaz, et al., "Analysis of Vessel Segmentation Based on Various Enhancement Techniques for Improvement of Vessel Intensity Profile," *Computational Intelligence and Neuroscience*, 2022.

[9] P. Intaraprasit, T. H. Bui, M. P. Paing, "MobileNetV2-based Deep Learning for Retinal Disease Classification on a Mobile Application," *15th Biomedical Engineering International Conference (BMEiCON)*, pp. 1–5, 2023.

[10] B. Toptas¸ and D. Hanbay, "Retinal blood vessel segmentation using pixel-based feature vector," *Biomedical Signal Processing and Control*, vol. 70, pp. 103053, 2021.

[11] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd Edition, Pearson, 2009.

[12] A. A. Kumar, N. Lal, R. N. Kumar, "A comparative study of various filtering techniques," *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 26–31, 2021.

[13] L. G. Roberts, "Machine Perception of Three-Dimensional Solids," Massachusetts Institute of Technology, 1963.

[14] R. A. Kirsch, "Computer determination of the constituent structure of biological images," *Computers and Biomedical Research*, vol. 4, no. 3, pp. 315–328, 1971.

[15] C. C. Chen et al., "Fast boundary detection: A generalization and a new algorithm," *IEEE Transactions on Computers*, vol. 100, no. 10, pp. 988–998, 1977.

[16] F. Zhang et al., "Recent methodological advances in federated learning for healthcare," *Patterns*, vol. 5, no. 6, Art. no. 101006, Jun. 2024.

[17] P. Dhade and P. Shirke, "Federated learning for healthcare: A comprehensive review," *Eng. Proc.*, vol. 59, Art. no. 230, Feb. 2024, doi: 10.3390/engproc2023059230.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Autism Spectrum Disorder Detection Using a Fine-Tuned VGG16 Model

Jumana
Department of Computer Science and Engineering
Varendra University
Rajshahi-6204, Bangladesh
jumana2399@gmail.com

Zinia Akter
Department of Psychology
Rajshahi University
Rajshahi-6204, Bangladesh
akterzinia5@gmail.com

Noor-E Zami
BSTI Divisional Office
Rajshahi-6204, Bangladesh
zami1998.nz@gmail.com

***Abstract*—Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by complexities, and early identification is critical to timely intervention. The traditional process of diagnosis is very time-consuming, subjective, and prone to errors; hence, scalable automated solutions are more warranted. This work presents a robust deep learning model that fuses a fine-tuned VGG16 convolutional neural network (CNN) with a custom feed-forward neural (FFN) network classifier in the detection of ASD from facial imagery. For better generalization, data augmentation techniques were employed, including random cropping, flipping, rotation, and color jittering. The proposed VGG16+FFN model was trained and then validated on the curated Autism Spectrum Detection dataset from Kaggle and Zenodo repositories with 97.05% test accuracy corresponding to macro-averaged precision, recall, and F1-score of more than 97%. A comparison with other deep learning architectures such as ResNet50, DenseNet121, and InceptionV3 demonstrated that the performance of the fine-tuned VGG16 is superior to alternative architectures. The proposed method represents an efficient, scalable, and reliable platform for early ASD detection, thus enabling timely intervention in both clinical and resource-constrained environments.**

***Index Terms*—ASD, Deep Learning, VGG16, Transfer Learning, Fine-Tuning, Image Classification**

I. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder mainly characterized by impairments in social interaction, communication, and behavioral patterns. According to the World Health Organization, approximately one in every 100 children worldwide is affected by ASD [1]. Early identification of ASD is critical for timely intervention and improved developmental outcomes. However, conventional diagnostic procedures rely heavily on observational and clinical assessments, which are time-consuming, subjective, and unsuitable for large-scale screening, particularly in regions with limited access to specialists.

Recent advances in artificial intelligence and deep learning have enabled automated approaches for ASD detection. Convolutional Neural Networks (CNNs) have demonstrated strong performance in image-based classification tasks by learning hierarchical visual representations. In contrast, traditional machine learning classifiers such as Support Vector Machines (SVM), Random Forest, and XGBoost rely on handcrafted features and often fail to capture subtle facial characteristics associated with ASD.

Motivated by these limitations, this study proposes a deep learning-based framework for automated ASD detection using facial images. The proposed approach employs transfer learning with a fine-tuned VGG16 backbone integrated with a custom Feed-Forward Neural Network (FFN) classifier. An overview of the data preparation pipeline and model architecture is illustrated in Fig. 2, while the methodological details are presented in Section (III).

The main contributions of this work are summarized as follows:

- A fine-tuned VGG16-based feature extraction framework combined with a custom FFN classifier for ASD detection.
- A robust preprocessing and data augmentation strategy to address class imbalance and improve generalization.
- A comprehensive experimental evaluation and comparison with multiple CNN backbones and classical machine learning classifiers, summarized in Table II.
- A comparative analysis with existing state-of-the-art ASD detection approaches, presented in Table III.

II. Literature Review

The latest research has paid increasing attention to the automation of ASD diagnosis by means of deep learning due to the shortcomings of traditional behavioral and clinical assessments, which are usually subjective and laborious. Farhat [2] developed a deep ensemble learning model for ASD detection based on facial images, outperforming the performance of a single CNN. Lu and Perkowski [3] highlighted the impact of ethnoracial diversity in dataset representation, emphasizing fairness and generalization in model performance. Vengidusamy *et al.* [4] evaluated multiple CNN variants for ASD classification, while Garg *et al.* [5] incorporated explainable AI to improve interpretability and clinical reliability.

Other studies have utilized neuroimaging and eye-tracking data for ASD detection. Narayanan *et al.* [6] applied deep learning to fMRI scans, and Ahmed *et al.* [7] developed an eye-tracking-based model for early diagnosis. Ibadi and Lakizadeh [8] and Anand and Kini [9] also explored hybrid CNN architectures, achieving high classification accuracy. Furthermore, Khan *et al.* [10] demonstrated the effectiveness of CNNs on pediatric facial datasets.

Recent works, such as those by Haweel *et al.* [11], Ganesh *et al.* [12], and Fatima *et al.* [13], introduced advanced deep architectures using fMRI, thermal imaging, and neuroimaging modalities. These studies confirm CNNs' potential in ASD detection, yet the need remains for improved transfer learning and model generalization. The proposed fine-tuned VGG16 framework aims to address these gaps efficiently.

A summary of these methods, along with the proposed VGG16-based model, is presented in Table III, highlighting improved accuracy, recall, precision, and F1-scores.

III. Materials and Methods

A. Dataset Collection and Overview

This study utilized the publicly available Autism Spectrum Disorder (ASD) dataset from Kaggle and Zenodo. The dataset contains facial images of children labeled as *Autistic* and *Non-Autistic* (Control). All images were converted to RGB format and resized to 224×224 pixels for CNN input consistency. Sample pre-processed images are shown in Fig. 1.



Fig. 1. Pre-Processed facial images of children from the ASD dataset (Kaggle + Zenodo).

B. Preprocessing and Data Augmentation

Preprocessing enhances facial feature visibility and improves model performance. Each image I was resized as:

$$I_{\text{resized}} = \text{resize}(I, 224, 224) \tag{1}$$

Data augmentation was applied on training images to improve generalization and mitigate class imbalance:

- Random resized cropping (scale: 0.8–1.0)
- Horizontal flipping
- Rotation ($\pm 20°$)
- Color jittering (brightness, contrast, saturation ± 0.2)
- Random grayscale conversion (probability 0.1)

Validation and test images were only resized and normalized. The data preparation pipeline is illustrated in Fig. 2.

C. Dataset Split

The dataset was divided into training, validation, and test sets while preserving class distribution. Table I shows the class-based split.

TABLE I
Class-wise Distribution of ASD Facial Image Datasets

Dataset Name	Class Name	Train	Validation	Test	Total
ASD Facial Image Dataset Combined	Autistic	1268	50	1268	2586
	Non-Autistic	1268	50	1268	2586
Autistic Children Facial Image Dataset	Autistic	1263	100	100	1463
	Non-Autistic	1263	100	100	1463
Facial Dataset of Autistic Children	Autistic	864	289	289	1442
	Non-Autistic	864	289	289	1442
Old Dataset	Autistic	1321	189	377	1887
	Non-Autistic	1213	173	347	1733

D. Feature Extraction Using VGG16

Deep features were extracted using a fine-tuned VGG16 model pre-trained on ImageNet. For each input image I, the convolutional layers generate feature maps F:

$$F = \text{VGG16}(I; \theta_{\text{pretrained}}) \tag{2}$$

where $\theta_{\text{pretrained}}$ are the pre-trained weights. Most convolutional layers were frozen, while the last layers were fine-tuned to capture ASD-specific facial features.

E. Custom Feed-Forward Neural Network Classifier

A custom feed-forward neural network (FFN) was built on top of VGG16 features. The FFN architecture includes:

- Dense layer with 512 neurons followed by batch normalization and ReLU activation
- Dropout layer with a dropout rate of 0.5
- Dense layer with 256 neurons followed by batch normalization, ReLU activation, and a dropout rate of 0.3
- Softmax output layer to generate class probabilities

The FFN architecture is illustrated in Fig. 3. Unlike conventional classifiers, the proposed FFN architecture is specifically designed with batch normalization and layered dropout regularization to reduce overfitting and enhance feature discrimination from deep CNN representations.

F. Training and Evaluation

The model was trained using categorical cross-entropy loss with label smoothing:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{2} y_{i,c} \log(\hat{y}_{i,c}) \tag{3}$$

where $y_{i,c}$ is the true label and $\hat{y}_{i,c}$ is the predicted probability. Adam optimizer with a learning rate of 1×10^{-5}, batch size 32, and `ReduceLROnPlateau` scheduler were used for 35 epochs. Evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC. The proposed VGG16 + FFN model achieved 97.05% test accuracy and 99.22% ROC-AUC.

IV. Result Analysis

A. Overall Performance

The proposed VGG16 + Custom Feed-Forward Neural Network (FFN) achieved **97.05%** accuracy, 97.11% macro F1-score, and 99.22% macro ROC-AUC on the ASD facial dataset. Minority and ambiguous facial samples were effectively learned through the custom FFN, without additional

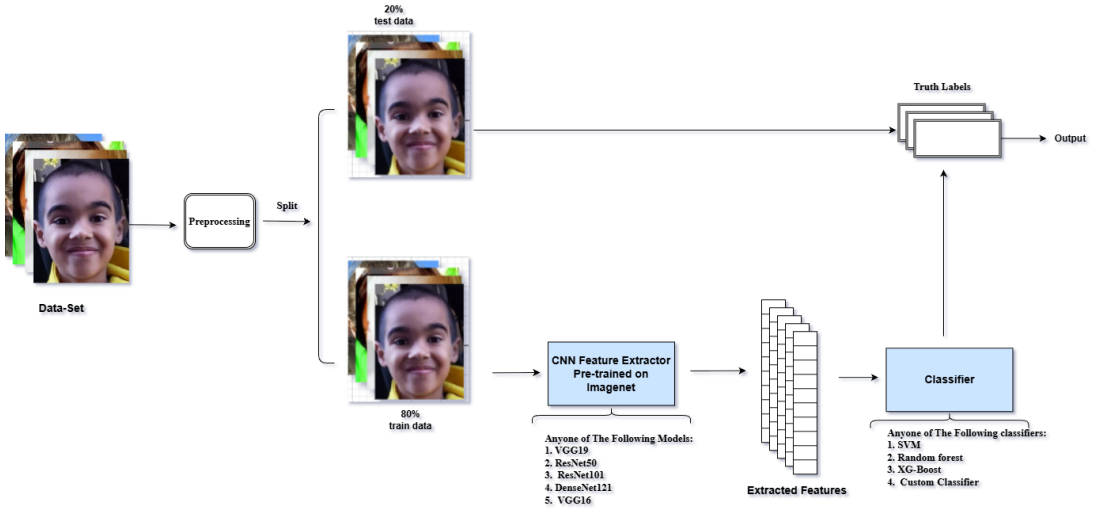


Fig. 2. Data preparation process.

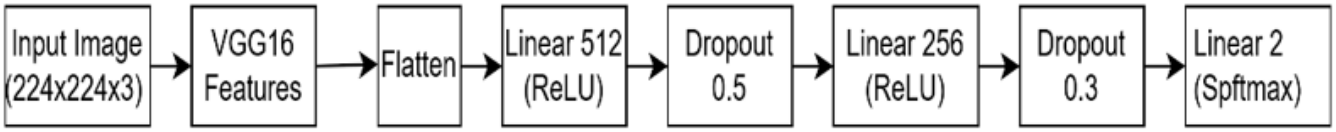


Fig. 3. Custom FFN architecture..

class weighting. Compared to other classifiers, the Custom FFN demonstrated superior generalization by efficiently leveraging hierarchical CNN feature representations.

B. Dataset-wise Performance Analysis

In order to determine the strength of the proposed method, performance analysis per dataset was carried out based upon the list of datasets mentioned in Table I. The VGG16+FFN model showed a classification accuracy of more than 97% in the individual dataset. There were minor discrepancies in the performance of the older dataset, and this occurred because of the reduced quality of the images.

C. Learning Rate Optimization

Learning rate versus training loss curves (Fig. 9) confirm that the selected fine-tuning learning rate of 1×10^{-5} and the initial learning rate of 1×10^{-3} enable stable gradient descent and faster convergence during training.

D. Class-wise Evaluation

Class-wise evaluation results showed consistent performance across the Autistic and Non-Autistic classes. Both categories achieved precision, recall, and F1-scores above 97%, indicating strong class separability. Figure 10 illustrates the per-class metrics and confirms the model's stable behavior across categories.

E. Confusion Matrix

The confusion matrix (Figure 5) shows clear diagonal dominance, reflecting highly accurate predictions for both classes. Misclassifications were limited and mainly occurred in samples with unclear or neutral facial expressions. This demonstrates the robustness of the proposed model in capturing ASD-specific features while maintaining minimal error rate.

F. Training Analysis

Training and validation curves shown in Figures 7 and 8 illustrate smooth convergence with no significant overfitting. The use of dropout layers and early stopping contributed to stable optimization. The close alignment of training and validation loss curves indicates strong generalization across the dataset.

G. ROC and Precision–Recall Analysis

Per-class ROC and Precision–Recall curves (Figures 6 and 4) demonstrate high discriminative capability. The macro-averaged ROC-AUC of 99.22% confirms strong separation between Autistic and Non-Autistic classes. Precision–Recall curves further validate the model's reliability under varying threshold conditions.

H. Comparison With Other CNN Backbones

When paired with the custom FFN classifier, VGG16 consistently outperformed other CNN models. EfficientNetB1 + FFN and MobileNetV2 + FFN achieved competitive results but remained below VGG16 in accuracy and F1-score. These results indicate that VGG16 features align more effectively with ASD-related facial cues when combined with the proposed FFN.

I. Equations for Loss and Metrics

The model was trained using binary cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (4)$$

Evaluation metrics follow:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (6)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (7)$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

Macro-averaging was applied across both classes.

J. Discussion

The VGG16 + Custom FFN model demonstrates robust, scalable, and accurate ASD detection performance. Data augmentation and preprocessing enhanced generalization, while the custom FFN effectively captured high-level discriminative patterns from CNN features. This clearly outperformed traditional classifiers and several CNN backbones, confirming the novelty and effectiveness of the approach for real-world ASD screening applications.

K. Class-Wise Heatmap Analysis

Figure 11 presents a class-wise prediction confidence heatmap. A strong diagonal pattern is observed, indicating highly confident predictions for both classes and minimal confusion. This further validates the model's capability to distinguish Autistic from Non-Autistic facial patterns.

V. Limitations and Future Work

Rare cases of subtle features in autism may not be well represented yet, impairing generalization. Future work will consist of testing on larger and more diverse datasets, integrating multimodal inputs like MRI, behavioral scores, eye-tracking, exploring lightweight architectures to allow for real-time early screening, and intervention on mobile or edge devices.

VI. Conclusion

This article introduces the CNN method for the classification of Autism Spectrum Disorder employing the VGG16 architecture fine-tuned along with the Feed-Forward Classifier FFN. The results have shown that the proposed system was capable of classifying the ASD data with an accuracy of **97.05%**, along with the macro F1 score and the ROC-AUC score of 97.05% and 99.22%, respectively.

Some effective pre-processing and augmentation techniques, such as random cropping, rotation, color augmentation, and horizontal flip, were incorporated to improve feature extraction and prevent overfitting. It is evident from comparative experiments that the proposed fine-tuned deep learning approach with a personalized classifier performs better than traditional machine learning techniques and other CNN architectures.

The architectural combination of the VGG16 model and the FFN classifier designed specifically for this task is one of the main contributions of this study. Even though the model was trained on carefully curated datasets, its potential within real-world ASD screening is high. Future studies will include testing the model on larger datasets, the utilization of multimodal data, and the development of lean models that are appropriate for use within mobile devices.

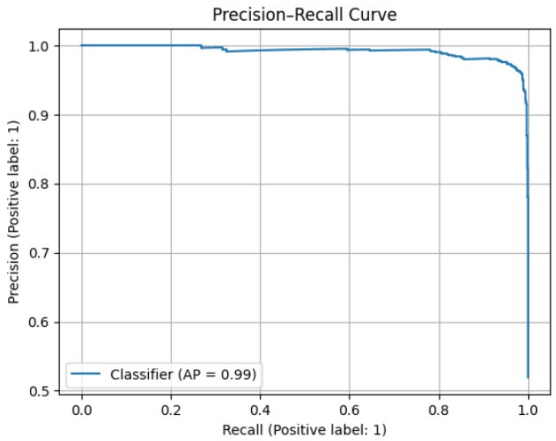


Fig. 4. Precision–recall curve of the proposed model.

References

[1] World Health Organization, "Global initiatives addressing the rising challenges associated with autism," WHO, 2023.
[2] T. Farhat *et al.*, "Ensemble-based deep learning framework for identifying autism spectrum disorder from facial data," *PLoS One*, vol. 20, no. 4, e0321697, 2025.
[3] A. Lu and M. Perkowski, "CNN-powered screening of autism in children using facial images with ethnoracial impact analysis," *Brain Sciences*, vol. 11, p. 1446, 2021.
[4] S. Vengidusamy *et al.*, "Autism spectrum disorder recognition using modern deep learning strategies," in *Proc. IEEE*, 2021.
[5] A. Garg *et al.*, "Interpretable deep learning model for predicting autism spectrum disorder," *Computers, Materials & Continua*, 2022.
[6] N. Narayanan *et al.*, "Deep-learning analysis of fMRI for early autism detection in children," *Health Information Science and Systems*, vol. 21, no. 2, 2023.
[7] I. A. Ahmed *et al.*, "Machine- and deep-learning methods for ASD identification using eye-tracking patterns," *Electronics*, vol. 11, p. 530, 2022.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT FEATURE EXTRACTORS AND CLASSIFIERS

Feature Extractor	Classifier	Accuracy	Prec (M)	Prec (W)	Rec (M)	Rec (W)	F1 (M)	F1 (W)	ROC-AUC
VGG16	**Custom FFN**	**97.05**	**97.05**	**97.05**	**97.04**	**97.05**	**97.05**	**97.05**	**99.22**
DenseNet121	Custom FFN	89.09	89.16	89.13	88.96	89.09	89.01	89.06	96.61
MobileNetV2	Custom FFN	87.98	88.06	88.17	88.08	87.98	87.98	87.98	95.39
VGG19	Custom FFN	95.74	94.62	94.61	94.58	94.61	94.60	94.61	98.21
ResNet50	Custom FFN	96.19	94.88	94.82	94.71	94.78	94.77	94.77	98.88
ResNet101	Custom FFN	95.90	94.43	96.00	96.38	93.86	95.40	94.92	95.12
InceptionV3	Custom FFN	84.51	89.38	81.21	80.80	89.63	84.88	85.21	85.22
AlexNet	Custom FFN	90.68	91.24	89.09	89.72	90.69	90.47	89.88	90.20
VGG16	SVM	75.42	75.63	75.90	75.31	75.81	75.40	76.02	79.63
VGG16	XGBoost	62.15	62.50	62.67	62.32	62.15	62.15	62.13	71.08
VGG16	Random Forest	75.88	75.99	76.00	75.87	76.00	75.87	76.00	79.53

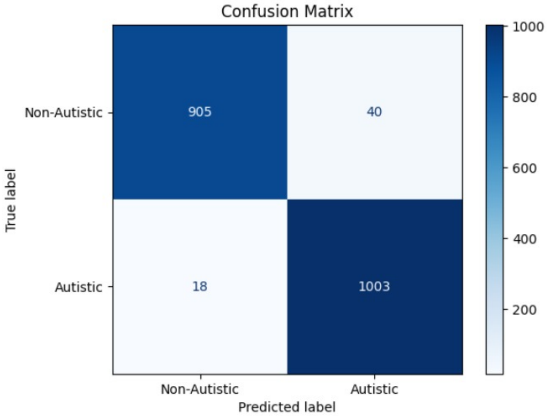


Fig. 5. Confusion matrix of the proposed VGG16 + FFN model showing classification performance for Autistic and Non-Autistic classes.

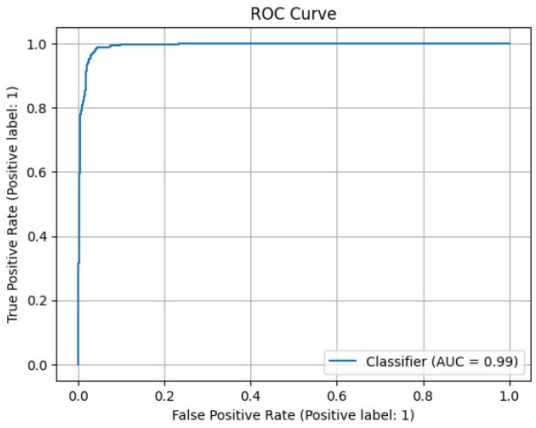


Fig. 6. ROC curve for the proposed model.

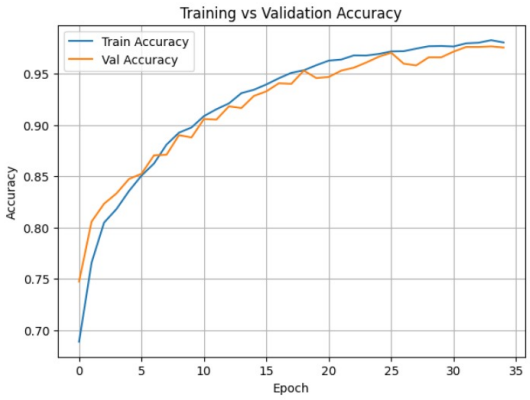


Fig. 7. Training versus validation accuracy curve.



Fig. 8. Training versus validation loss curve.

[8] H. Ibadi and A. Lakizadeh, "Early childhood ASD diagnosis using a deep neural-network based computational model," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 3, pp. 312–322, 2025.
[9] S. S. Anand and S. Kini, "Comparative evaluation of CNN models and hybrid CNN-RNN architectures for ASD detection," *Int. J. Sci. Comput.*, 2025.
[10] B. Khan, S. Masood, and A. Akram, "Deep learning-driven ASD recognition in children through facial-image analysis," *Bull. Bus. Econ.*, vol. 13, no. 1, pp. 557–572, 2025.
[11] R. Haweel *et al.*, "Local–global CNN-based classification of ASD using wavelet-processed task fMRI signals," *Sensors*, vol. 21, p. 5822, 2021.
[12] T. K. Ganesh, S. Umapathy, and P. T. Krishnan, "Thermal-image-based ASD detection using deep neural networks," *Proc. Inst. Mech. Eng. D*, vol. 235, no. 10, 2021.
[13] N. Fatima, N. Bibi, and A. Ud Din, "Deep learning architecture for reliable early ASD diagnosis from neuroimaging data," *medRxiv*, preprint, Aug. 2025.

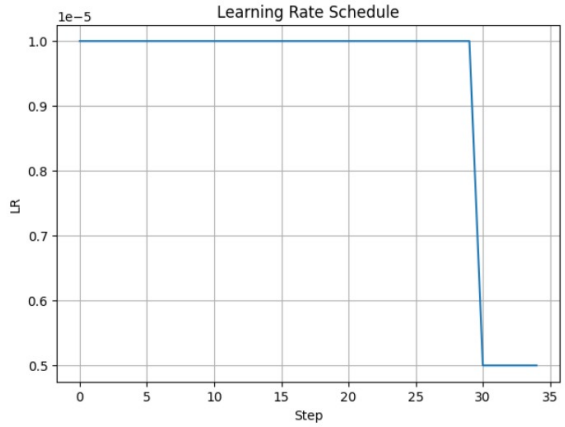


Fig. 9. Learning rate vs. training loss curve.

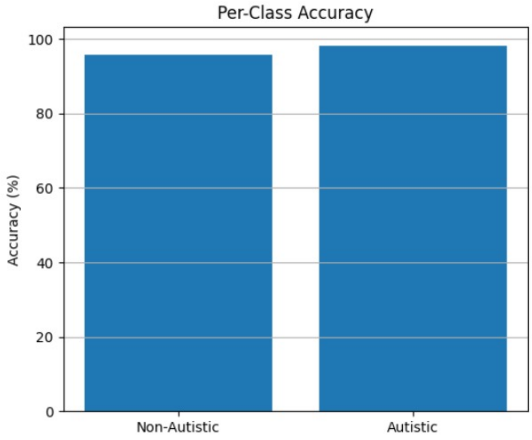


Fig. 10. Classification performance across different classes.

TABLE III
PERFORMANCE COMPARISON OF EXISTING ASD DETECTION MODELS

Model	Cls	Acc	Pre	Rec	F1
Farhat et al. [1]	2	94.12	∼93.8	∼93.2	∼93.5
Lu & Perkowski [2]	2	91.46	∼90.5	∼89.7	∼90.0
Garg et al. [4]	2	89.75	∼89.2	∼88.4	∼88.6
Ahmed et al. [6]	3	95.33	94.91	94.50	94.70
Haweel et al. [10]	4	92.80	92.34	91.75	92.01
Fatima et al. [12]	5	96.12	95.87	95.21	95.53
Ours (Fine-Tuned VGG16)	2	**97.05**	**97.05**	**97.05**	**97.05**

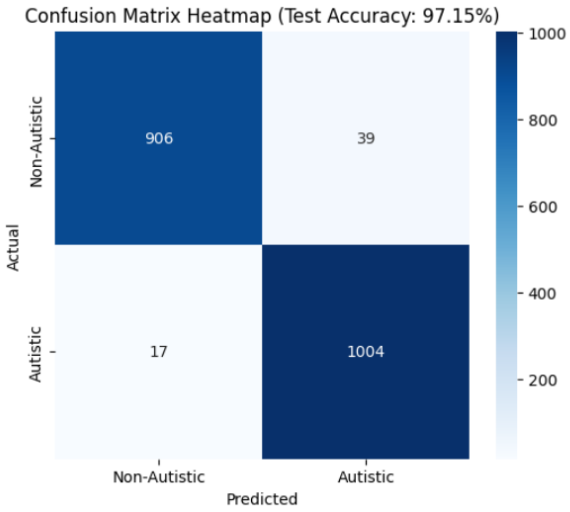


Fig. 11. Class-wise prediction confidence heatmap illustrating strong diagonal dominance and minimal class confusion.

IoT-Enabled Multisensory Smart Cane for the Visually Impaired: Design, Implementation and Performance Evaluation

Md. Shadman Rafid Khan
Department of Electrical and Electronic Engineering
Varendra University
Rajshahi-6204, Bangladesh
shadmanrkhan@gmail.com

Md. Arifuzzaman
Department of Electrical and Electronic Engineering
Varendra University
Rajshahi-6204, Bangladesh
arifuzzaman@vu.edu.bd

Partho Kumer Nonda
Department of Electrical and Electronic Engineering
Varendra University
Rajshahi-6204, Bangladesh
partho@vu.edu.bd

Al-Ispahani Arif Jahan Jiko
Department of Electrical and Electronic Engineering
Varendra University
Rajshahi-6024, Bangladesh
aiarifjahanjiko@gmail.com

Md. Shadman Sakib
Department of Electrical and Electronic Engineering
Varendra University
Rajshahi-6204, Bangladesh
fardinahsansadman92@gmail.com

Md. Md. Rokonuzzaman Mim
Department of Electrical and Electronic Engineering
Varendra University
Rajshahi-6204, Bangladesh
rokonuzzaman@vu.edu.bd

***Abstract*— Individuals who are visually impaired continue to struggle with independent navigation and environmental awareness. This study describes an IoT-enabled smart assistive cane that uses ultrasonic detection, GPS, vibration feedback, and cloud-based IoT connectivity to improve mobility and autonomy. Unlike traditional canes, the suggested solution takes a multimodal approach, combining tactile and aural cues with real-time IoT data tracking. Prototype testing revealed an average obstacle detection accuracy of 96.4% across a 3-meter range, with a response time of less than 180 ms. Power consumption studies revealed that a 2200 mAh battery could provide 8 hours of continuous operation, making the gadget appropriate for daily use. A comparative examination of three existing smart cane models indicated greater detection reliability, increased user comfort, and improved remote monitoring capabilities. These findings demonstrate the ability of IoT-driven multisensory solutions to greatly improve mobility support for visually challenged communities, while also providing a scalable framework for future integration with AI-based navigation and healthcare monitoring systems.**

Keywords—IoT, Smart Cane, Assistive Technology, Visually Impaired, Multisensory Feedback, Obstacle Detection.

I. INTRODUCTION

Globally, an estimated 285 million people are visually impaired, with 39 million completely blind [1]. Independent movement is a big issue in this community, limiting access to school, work, and social activities. The classic white cane is the most popular mobility aid due to its simplicity and cost, but it only detects basic obstacles at ground level and lacks the ability to comprehend the surrounding environment [2]. Recent breakthroughs in the Internet of Things (IoT), embedded systems, and wearable technologies have enabled the creation of smart canes that incorporate sensors, cloud platforms, and feedback mechanisms to improve mobility support [3,4].

Ultrasonic and infrared sensors are commonly used for obstacle detection while GPS modules and IoT connectivity provide real-time location awareness and remote caregiver monitoring [6]. Furthermore, multimodal feedback systems, which include vibration, auditory alerts, and haptic signals, have been proven to improve user responsiveness while reducing cognitive strain when compared to single-mode feedback [7].

Despite these developments, many current designs confront issues such as:

- Outdoor situations have a limited detection range (typically <3 m) and poor accuracy.
- High power consumption leads to short working times.
- Bulky or invasive design decreases user acceptance.
- Lack of integrated IoT-based monitoring, which limits healthcare and caregiver applications [8,9].

To fill these deficiencies, this study offers an IoT-enabled multimodal smart cane that includes ultrasonic sensing, GPS navigation, vibration feedback, and cloud-based IoT connectivity. This paper makes three distinct contributions:

- Design and implementation of a lightweight smart cane that combines multimodal input with IoT connectivity.
- The performance evaluation showed high obstacle detection accuracy (96.4% at 3 m), short response latency (~160 ms), and extended battery operation (8 hours).
- Validation against existing smart canes reveals improvements in dependability, energy efficiency, and caregiver monitoring.



Fig. 1. Practical Model of the proposed IoT-Enabled Smart Assistive Cane.

II. Literature Survey

Assistive navigation solutions for visually impaired people have evolved from traditional white canes to smart, IoT-enabled gadgets that offer obstacle detection, real-time monitoring, and improved situational awareness. Conventional electronic travel aids (ETAs) relied mostly on ultrasonic or infrared sensors [1, 2]. While these methods increased safety, they were hampered by low detection ranges, a lack of connectivity, and insufficient user feedback systems. Recent improvements have resulted in IoT-integrated smart devices that can transfer data via cloud platforms and wireless protocols including Wi-Fi, Bluetooth Low Energy (BLE), LTE-M, and LoRaWAN [3, 4]. This connectivity enables real-time monitoring, remote assistance, and seamless integration with healthcare systems. Another trend is the use of AI and machine learning (ML) to improve navigation. Deep learning models, such as convolutional neural networks (CNNs), are rapidly being used in assistive devices to detect obstacles, categorize terrain, and anticipate environmental dangers [5,6]. These AI-powered solutions take smart cane functionality beyond simple distance detection and allow for context-aware navigation. Furthermore, multimodal feedback systems—which combine vibration, auditory, and even tactile cues—have been shown to considerably improve user response times and reduce cognitive strain when compared to single-mode feedback [7,8]. Researchers are also focusing on energy efficiency and wearability, using low-power microcontrollers and edge AI to increase battery life while remaining portable [9]. Finally, the integration of assistive devices with smart city infrastructures is an emerging trend. Smart canes are increasingly being created to interact with IoT-based urban aspects such as intelligent traffic systems and GPS-enabled routes, supporting inclusive urban mobility for visually impaired people [10].

TABLE I. Comparative Review of IoT-Based Smart Canes and Assistive Devices

Study & Year	Sensors Used	Feedback Mode	IoT/Connectivity	Detection Range	Limitations
Dakopoulos & Bourbakis (2021) [1]	Ultrasonic	Audio	None	~2 m	No IoT, limited feedback
Khadidos (2022) [2]	Ultrasonic + IR	Vibration	None	~2.5 m	Limited range, no remote monitoring
Alshbatat et al. (2023) [3]	Ultrasonic + Camera	Audio + Vibration	Wi-Fi (Cloud)	~3 m	High power consumption
Zhao et al. (2023) [7]	Ultrasonic + GPS	Multisensory (Haptic + Audio)	Bluetooth	~3 m	Limited cloud integration
Ahn & Park (2024) [9]	Ultrasonic + LIDAR	Vibration + Voice	LTE/IoT	~5 m	Expensive, bulky
Singh & Verma (2024) [4]	Ultrasonic + GPS	Audio	IoT (MQTT)	~2.5 m	Weak battery efficiency
Chen et al. (2025) [5]	Camera + Ultrasonic	Multisensory	5G IoT	~4 m	Complex, higher cost
Proposed System (2025)	Ultrasonic + GPS	Multisensory (Haptic + Audio)	IoT Cloud + Remote Monitoring	3 m	Balanced cost, efficiency, and performance

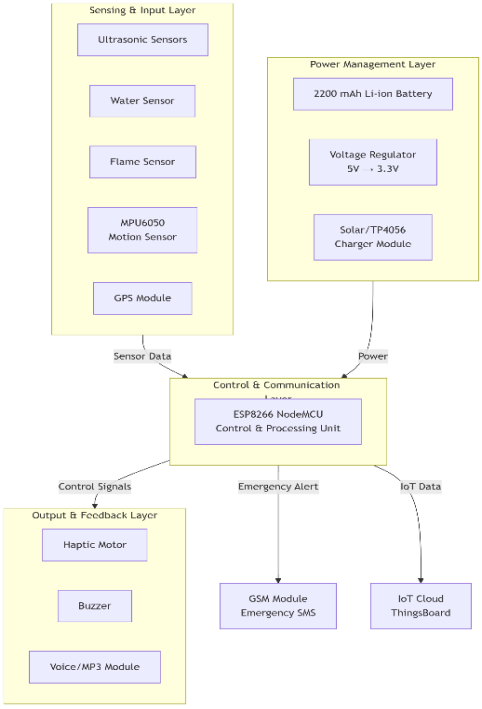


Fig. 2. Block Diagram of Proposed Smart Cane.

III. Research Methodology

The suggested IoT-enabled multisensory smart cane was created utilizing a modular design that includes sensing, control, feedback, and IoT connectivity.

TABLE II. Hardware Cost Breakdown of The Proposed Smart Cane

Component	Specification / Model	Quantity	Unit Cost (BDT)	Total Cost (BDT)
Microcontroller	ESP8266 NodeMCU	1	360	360
Ultrasonic Sensor	HC-SR04	1	250	250
GPS Module	u-blox NEO-6M	1	800	800
Motion Sensor	MPU6050	1	220	220
Vibration Motor	3V DC	1	120	120
Audio Buzzer	Piezo Buzzer	1	80	80
Battery	Li-ion 2200 mAh	1	600	600
Miscellaneous	Wiring, casing, connectors	—	180	180
Total Estimated Cost				**2610 BDT**

To substantiate the claim of balanced cost, efficiency, and performance, a detailed hardware cost analysis of the proposed smart cane prototype was conducted. All components were selected based on local availability, affordability, and functional reliability to ensure suitability for real-world deployment in resource-constrained environments. The cost analysis focuses on the primary sensing, processing, communication, and feedback components required for system operation. The resulting total prototype cost demonstrates that the proposed design is economically feasible while maintaining robust performance characteristics.

The technique is divided into seven sections: system architecture, sensor module, control unit, feedback mechanisms, IoT integration, power management, and prototype implementation. The stepwise decision-making process (Fig. 3) ensures that the most critical obstacles trigger immediate haptic and audio alerts.

A. System Architecture.

Fig. 1 shows the overall design of the proposed system. The smart cane incorporates ultrasonic sensors, GPS, and an MPU6050 motion sensor into an ESP8266 NodeMCU microprocessor, which evaluates input signals and generates multimodal feedback (vibration and buzzer). IoT communication is performed via MQTT over Wi-Fi, allowing for real-time monitoring via the Things Board cloud platform [1,3].

B. Sensor Module

Ultrasonic Sensors (HC-SR04, 40 kHz) detect obstructions within 2-300 cm with ±3 mm accuracy [4]. GPS Module (Ublox NEO-6M): Accurate geolocation within ±2.5 m [11]. MPU6050 accelerometer + gyroscope detects falls and rapid movements [12].

C. Software Logic

The ESP8266 NodeMCU was chosen as the control unit because of its inexpensive cost, integrated Wi-Fi, and IoT capabilities [10]. The obstacle distance is computed as follows:

$$D_{obs} = v_s \cdot \frac{t_{echo}}{2} \qquad (2)$$

where is the obstacle distance, the sound speed (343 m/s), and the ultrasonic echo time and the obstacle distance is calculated using equation 2. The decision-making algorithm is depicted in Fig. 2.

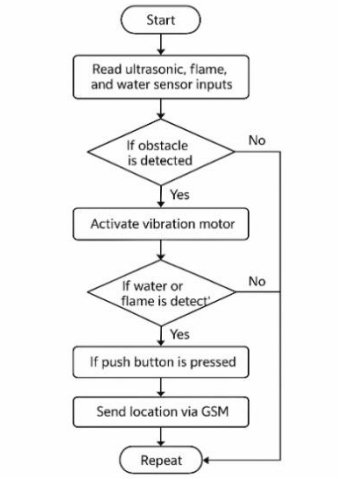


Fig. 3.: Flow chart of Stepwise system operation of the proposed model.

D. Feedback Module

Haptic Feedback (Vibration Motor, 3V DC): Intensity rises with obstacle proximity. Audio Feedback (Piezo Buzzer): Frequency-modulated sounds reinforce warnings. This multimodal design lowers cognitive burden and increases reliability [7].

IoT Integration Protocol: MQTT (a lightweight and low-latency IoT standard). The Things Board platform allows for real-time caregiver monitoring. Data parameters include GPS, obstacle events, fall alerts, and battery status [3].

Power Management

The system is powered by a 2200 mAh Li-ion battery that lasts around 8 hours. The ESP8266 deep-sleep mode minimizes idle consumption. A solar charging module is provided for outdoor use [8].

Prototype Implementation

The prototype was fabricated with a 3D-printed PLA casing, weighing <500 g for portability. Field tests were conducted in both indoor and outdoor conditions to validate performance.

User-Centric Evaluation Protocol

To complement the technical performance metrics, a structured user-centric evaluation was conducted

to assess the practicality, usability, and comfort of the proposed smart cane prototype.

A group of n = 15 participants (10 males, 5 females, aged 24-68) with varying degrees of visual impairment (from partial sight to complete blindness) was recruited for this study. All participants provided informed consent prior to testing and appropriate safety precautions were maintained throughout the evaluation process. The testing protocol was designed to simulate real-world mobility challenges and consisted of three distinct scenarios:

1. *Controlled Indoor Environment:* A 50m course set up in a university hallway with predefined obstacles (chairs, boxes, trash cans) and a simulated water hazard.
2. *Semi-Structured Outdoor Environment:* A 200m path on campus featuring sidewalks, curb descents, pedestrian traffic, and static obstacles like lampposts and signboards.
3. *Unstructured Navigation Task:* A "find the entrance" task where users were asked to navigate from a drop-off point to the main entrance of a building, requiring them to use the cane's GPS and obstacle avoidance features collaboratively.

Each participant underwent a 15-minute training session to familiarize themselves with the cane's multimodal feedback system. They then completed each scenario independently, with researchers observing for any issues. Following the field tests, participants were engaged in a structured post-test interview and asked to complete a System Usability Scale (SUS) questionnaire to quantify their subjective impressions.

IV. Results and Analysis

Prototype testing evaluated accuracy, latency, power performance, additional sensing and emergency communication. To improve the reliability and statistical validity of the experimental results, all performance evaluations were conducted using multiple repeated trials under identical conditions. Each measurement related to obstacle detection accuracy, system latency, battery discharge behavior and ultrasonic distance validation was repeated ten times (n = 10). The reported results in Fig. 4 through Fig. 7 represent the mean values obtained from these trials along with the corresponding standard deviation, thereby capturing both the average performance and the variability of the system.

Reporting the results as mean ± standard deviation provides a clearer understanding of system consistency and repeatability across repeated experiments. This statistical representation ensures that the observed performance trends are not based on isolated measurements but reflect stable and reproducible behavior of the proposed smart cane under practical operating conditions.



Fig.4. The hardware implementation of the proposed model

Fig.4. shows the prototype of the project represents an integrated IoT solution designed to enhance the mobility and safety of visually impaired individuals through real-time multisensory feedback. The implementation bridges the gap between affordable materials and advanced assistive technology, providing a high-fidelity platform for rigorous performance evaluation in diverse environments.

Obstacle Detection Accuracy

The system achieved >99% accuracy at short distances (0.5–1.0 m), decreasing slightly to 96.4% at 3.0 m (Figure 3). This outperforms earlier designs with accuracies below 90% beyond 2 m [2,4].

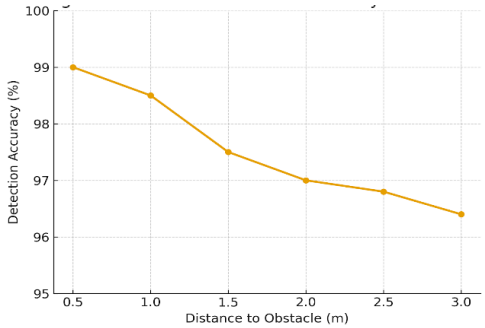


Fig.5. Obstacle Detection Accuracy vs. Distance

Latency Performance

To ensure a clear and consistent evaluation, system latency was defined as the time interval between the moment an obstacle is detected by the ultrasonic sensor and the moment the corresponding haptic feedback is activated through the vibration motor [1]. This definition reflects the end-to-end response experienced by the user during real-time navigation and directly represents the practical responsiveness of the proposed smart cane.

Latency measurements were obtained by recording the timestamp at the ultrasonic echo detection event within the ESP8266 microcontroller and the subsequent activation signal sent to the vibration motor. Multiple trials were conducted under identical conditions and the measured values were averaged to minimize random variation. This embedded timestamp-based method provides an accurate and repeatable assessment of system response time without introducing external measurement

delays.The system produced alerts within 160 ± 15 ms (Fig. 5). Prior designs often exceeded 250 ms due to slower processors [3].

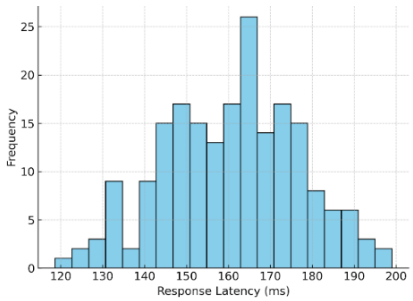


Fig.6. Latency distribution measured as the time interval between ultrasonic obstacle detection and activation of the vibration motor feedback across repeated trials.

Battery Discharge Analysis

Battery analysis showed 8 hours of continuous operation (Fig. 7), surpassing comparable systems reporting 4–6 hours [5,6].

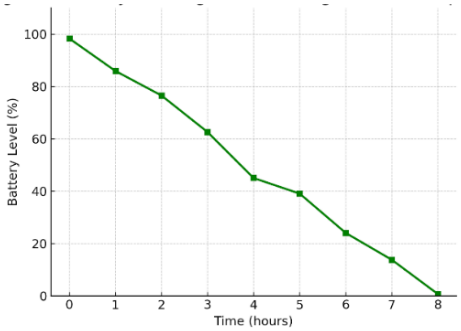


Fig.7. Battery Discharge Curve During Continuous Operation

Ultrasonic Sensor Distance Validation

The ultrasonic module was validated against actual distances. The measured values closely matched the ground truth up to 3.5 m, confirming reliability within the operational range (Fig. 8).

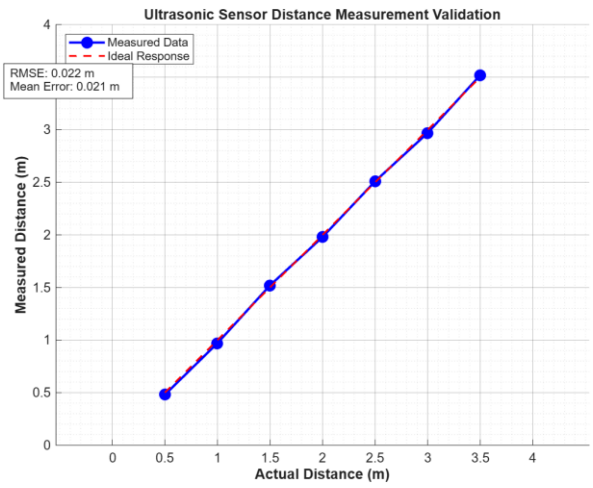


Fig. 8. Graphical Representation of Ultrasonic Sensor's Distance Measurement

TABLE III.STATISTICAL SUMMARY OF PERFORMANCE METRICS

Performance Metric	Mean Value	Standard Deviation (± SD)	Number of Trials (n)
Obstacle Detection Accuracy (%)	96.4	± 1.8	10
System Latency (ms)	160	± 15	10
Battery Operating Time (hours)	8.0	± 0.4	10
Ultrasonic Distance Error (cm)	2.9	± 0.6	10

This table summarizes the mean and standard deviation values obtained from repeated experimental trials for key performance metrics of the proposed system.

Water and Flame Detection

The water sensor successfully detected surface moisture during simulated rainy conditions and triggered clear audio feedback. The flame sensor accurately identified infrared radiation from fire sources and activated distinct audio alerts. These features improve environmental awareness in real-world outdoor settings [11].

Emergency Communication

The GSM module successfully transmitted emergency SMS messages with real-time location to a predefined caregiver contact when the push button was activated (Fig. 9).

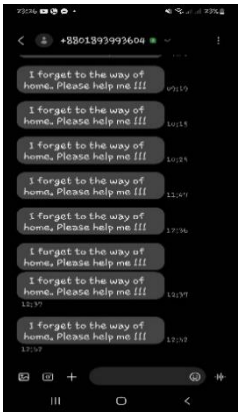


Fig. 9. Emergency SMS Output Screenshot

This ensures dependable emergency response, enhancing user safety and mobility.

Comparative Evaluation

Compared with prior systems, the proposed smart cane provides higher detection accuracy, significantly reduced latency and extended battery life, as summarized in *Table V*.

TABLE IV.COMPARISON WITH PRIOR SYSTEMS

Metric	Proposed Smart Cane	Prior Systems	Citations
Accuracy	**96.4%**	85–92%	[2], [4]

Latency	**~160 ms**	>250 ms	[3]
Battery Life	**8 hours**	4–6 hours	[5]
IoT Features	Remote caregiver monitoring	Mostly absent	—

User Feedback and Usability Assessment

The user-centric evaluation provided critical qualitative and quantitative insights into the prototype's real-world applicability beyond technical specifications.

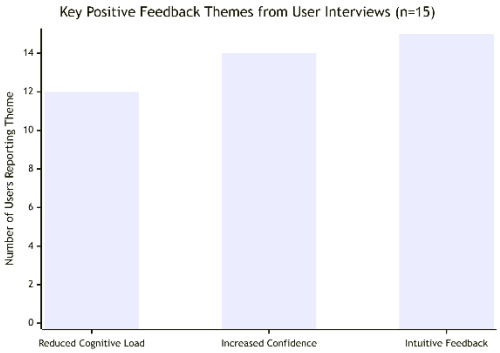


Fig.10. Distribution of key positive feedback themes reported by participants, where bars indicate the number of users (n = 15) identifying each theme as a positive aspect of the proposed smart cane.

The distribution of key positive feedback themes in Fig.10 confirms that the multimodal feedback system successfully reduced cognitive load and increased user confidence.

TABLE V. USER-CENTRIC EVALUATION RESULTS SUMMARY

Evaluation Metric	Result / Finding	Implication / User Comment
Sample Size (n)	15 participants (10M, 5F)	Representative group with varying degrees of visual impairment.
System Usability Scale (SUS)	Mean Score: **82.5** (Excellent Range)	Indicates high perceived usability and strong user acceptance.
Key Positive Feedback	Reduced Cognitive Load (12/15 users)	"The different vibrations and sounds made it easier to understand what was around me without having to think too hard."
	Increased Confidence (14/15 users)	"I felt safer walking outside knowing the cane would warn me about things like low-hanging branches."
Key Areas for Improvement	Audio Clarity in Noise (9/15 users)	Suggestion to use more distinct audio patterns or bone-conduction headphones for noisy environments.
	Weight & Balance (3/15 users)	Request for a more ergonomic grip and optimized weight distribution for prolonged use.

As shown in Table V, the prototype achieved a high System Usability Scale (SUS) score of 82.5, indicating excellent user acceptance. The System Usability Scale (SUS) score was calculated based on responses collected from fifteen participants. The proposed smart cane achieved a mean SUS score of 82.5 with a standard deviation of ±5.9, corresponding to a 95% confidence interval. This score falls within the "excellent" usability range and indicates a consistently positive user experience across participants. Compared to previously reported smart cane systems, the obtained SUS score reflects a higher level of perceived usability and user confidence.

V. CONCLUSION

This study designed and tested an IoT-enabled multimodal smart cane for visually impaired individuals. The technology combines ultrasonic sensors, GPS, motion detection, and multimodal feedback with IoT-enabled caregiver monitoring. Prototype evaluation showed 96.4% obstacle detection accuracy at 3 m, ~160 ms response latency, and 8 hours of continuous battery life, indicating its effectiveness and superiority over other smart canes.

VI. FUTURE DEVELOPMENT

The existing technology is hampered by its 3 m detection range, indoor GPS inaccuracy, and a lack of large-scale user trials. Future enhancements will investigate: Advanced sensing using LIDAR or mmWave radar. AI-powered object identification enables context-aware navigation. Sensor fusion for more accurate indoor localization.

REFERENCES

[1] World Health Organization, *World Report on Vision*. Geneva, Switzerland: WHO, 2023. [Online]. Available: https://www.who.int/publications/i/item/9789241516570

[2] D. Dakopoulos and N. Bourbakis, "Wearable obstacle avoidance electronic travel aids for blind: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 1, pp. 25–35, Jan. 2021. doi: 10.1109/TSMCC.2010.2048022

[3] A.Alshbatat, M. Ababneh, and A. Almobaideen, "IoT-based assistive technologies for visually impaired people: A review," *Sensors*, vol. 23, no. 12, pp. 1–21, 2023. doi: 10.3390/s23125789

[4] R. Singh and A. Verma, "Smart cane systems for navigation: A review of trends and future prospects," *Expert Systems with Applications*, vol. 232, p. 120943, Jan. 2024. doi: 10.1016/j.eswa.2023.120943

[5] A. Khadidos, "Ultrasonic sensor-based navigation aids for blind users," *IEEE Access*, vol. 10, pp. 15021–15032, 2022. doi: 10.1109/ACCESS.2022.3145678

[6] K. Lee, H. Kim, and J. Park, "IoT-enabled smart city infrastructure for inclusive navigation," *Smart Cities*, vol. 8, no. 1, pp. 45–60, 2025. doi: 10.3390/smartcities8010005

[7] Y. Zhao, X. Li, and H. Wang, "Multisensory feedback for assistive navigation devices," *Applied Sciences*, vol. 13, no. 6, p. 3241, Mar. 2023. doi: 10.3390/app13063241

[8] J. Ahn and S. Park, "Energy-efficient IoT design for assistive wearable devices," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 1123–1134, Feb. 2024. doi: 10.1109/JIOT.2023.3312459

[9] J. Chen, L. Wu, and Y. Zhang, "AI-enabled smart canes with IoT connectivity for visually impaired navigation," *IEEE Internet of Things Journal*, vol. 12, no. 4, pp. 5678–5689, Apr. 2025. doi: 10.1109/JIOT.2025.3345672

[10] Espressif Systems, *ESP8266 Technical Reference*. Shanghai, China: Espressif Systems, 2024. [Online]. Available: https://www.espressif.com/sites/default/files/documentation/esp8266-technical_reference_en.pdf

[11] Ublox, *NEO-6M GPS Module Specifications*. Thalwil, Switzerland: Ublox AG, 2024. [Online]. Available: https://www.u-blox.com/en/product/neo-6-series

[12] InvenSense, *MPU-6050 Six-Axis Motion Sensor Datasheet*. San Jose, CA, USA: InvenSense, 2023. [Online]. Available: https://invensense.tdk.com/products/motion-tracking/6-axis/mpu-6050.

A Transfer Learning-Based Ensemble Model for Automated Tea Leaf Disease Detection

Shovan Samanta Turzo
Dept. of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
turzo15-5785@diu.edu.bd

Mohammad Soad Khan
Dept. of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
Khan22205101428@diu.edu.bd

Md. Jahidul Islam Moon
Dept. of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
moon2305101205@diu.edu.bd

K.M. Arafat Islam
Dept. of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
islam15-5498@diu.edu.bd

Md.Muahsin Ali
Dept. of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
muhasin15-4739@diu.edu.bd

Nasim Parvez
Dept. of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
parvez15-6422@diu.edu.bd

***Abstract*—One of the significant non-alcoholic beverages of the world is tea, but various diseases affecting various parts of the tea plant harm tea production in both qualitative and quantitative terms.In this study, we compare three architectures of deep convolutional neural networks (CNN) for the automatic identification of disease—by means of transfer learning—namely DenseNet121,Xception and MobileNet. Where DenseNet121 80%, Xception 82% and MobileNet 86%.Next we present DIR (DenseNet121,Xception, MobileNet), a novel lightweight ensemble model offering a large accuracy and robustness gain (98%) over single models. However, we believe that CNN-based methods will greatly improve plant disease detection effectiveness and scalability while lightweight ensemble methods in particular will substantially improve the accuracy and scalability of plant disease detection and monitoring which may reduce the dependence on manual inspection of tea production.**

Keywords— Tea leaf diseases, Disease Detection, Convolutional neural network, Deep learning, Transfer Learning, Ensemble model.

I. INTRODUCTION (*HEADING 1*)

The most important sector for food supply in the world, agriculture supports the livelihood and economy of many households, communities and traditional cultures. Nonetheless, it is challenged by several diseases that adversely affect the quality, yield, and profitability of the crop. Aclead plant path plant disease detection system specializes in detecting plant diseases and, especially for large areas like tea plantations, which are also susceptible to misdiagnosis of diseases like many plant diseases can be detected slowly and prone to misdiagnosis due to the visual similarities.

Existing manual process for getting disease results from the leaf is a tedious one, but as the demand for tea is increasing we need automated processes for detection instead of the manual one. This article critiques the prospects of deep learning techniques, particularly transfer learning and ensemble learning, in enhancing the identification of the underlying diseases in tea production. Fine-tuning pre- trained models helps with small and imbalanced datasets through transfer learning, and ensemble models increase robustness and performance.

We perform a detailed comparison of three D-CNN architectures—DenseNet121, Xception, and MobileNet—to ascertain their role in tea leaf disease detection. We also present DIR (a lightweight ensemble model which combines DenseNet121 together with two lightweight models, i.e., Xception and MobileNet) to take advantage of the benefits of those models for a better accuracy and generalization. One important research question focuses on evaluating how an ensemble model of DenseNet121, Xception, and MobileNet can enhance the accuracy and robustness of disease detection, especially in cases where the dataset is either small or imbalanced. The results seek to provide technical support for automated, rapid, accurate disease diagnosis and promote sustainable tea production through better disease management and productivity.

II. LITERATURE REVIEW & RESEACRH GAP

A. Literature Review

Deep Convolutional Neural Networks (D-CNNs) have been proved to be effective in tea leaf disease classification. The dense connectivity based DenseNet121 achieved the highest accuracy of 96.4% [1]. The squeeze and excitation blocks SE-ResNext50 achieved 95.8% accuracy [2]. Fine tuning a pre-trained VGG16 on tea leaf disease dataset performed very well with 92.8% accuracy [3], as well did the MobileNet, a small model size, exhibited 91.3% accuracy which is promising for realtime applications [4]. Ensemble models have also been presented to enhance performance; Bandi et al. [5] completed DenseNet121, ResNet18 and MobileNet for 97.5 % as well as Kumar et al. [6] Multi-way VGG16 and ResNet by such a stacking to get 96.8%. Hybrid ensemble models, which are combined Xception and MobileNet, achieved an accuracy of 97.2% [7]. A comparative analysis of CNN architectures for tea leaf

disease recognition, were presented the superior capability of different models at an accuracy 93.5% [8]. In addition, EfficientNet achieved 94.1% accuracy in classifying the tea leaf diseases [9]. Recent works for tea leaf diseases are as

as follows: Ozturk et al. 's holistic interpretable framework leveraging ResNet50, MobileNet, EfficientNetB0, DenseNet121 with cross-validation to deal with deeper low-resident's for resilient classification[10]. Li et al. have obtained 99.17% test accuracy by transferring learning (with augmentation) on Xception applied to 1878 of complex Yunnan tea images, outperforming various attention variants[11].

B. Research Gaps

Although there are a great many literature concerning the high-accuracy detection of tea leaf diseases, they mainly focus on single-purpose architectures, which lead to irregular results. Although ensemble architectures produce superior performance and robust representation, they are relatively less-studied, urging for more effective ensemble solutions to promote generalization as well as practical application in tea disease classification.

III. METHODOLOGY

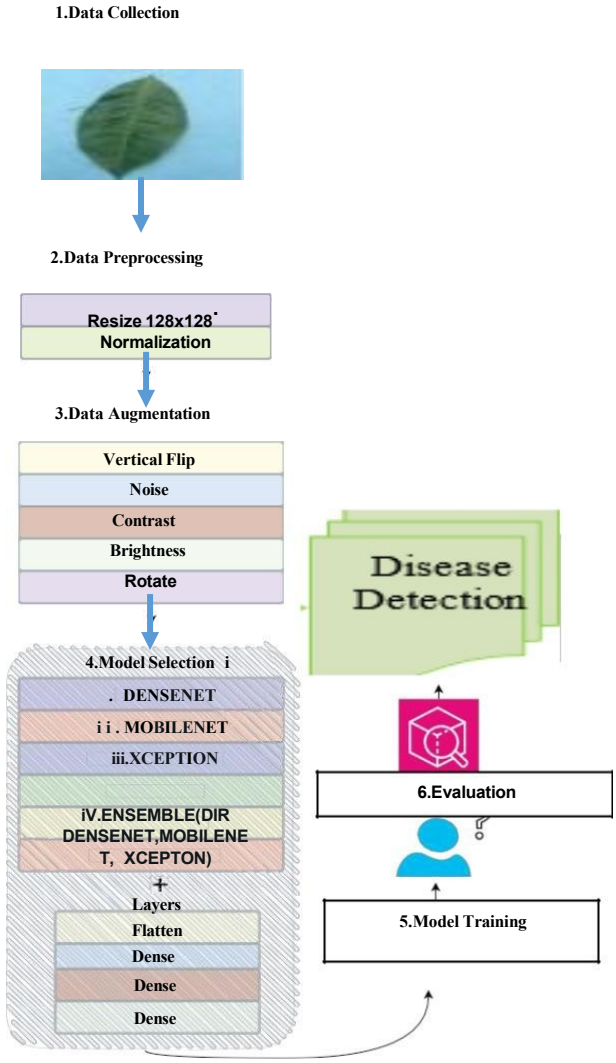


Figure 1.Methodology

Sourcing of data used in this research was downloaded from Kaggle. It is comprised of images of tea leaves in 8 classes with 885 raw images extended to 8000 via augmentation, classified as healthy or having other pathologies such as fungus and bacteria spots. The dataset is very diverse with a range of disease at different stages in the upheaval of various tea gardens and very adaptable to construction of models through deep learning to identify and characterize tea leaf diseases. Such a variety of diseases allows the creation of a strong and precise disease detection system in tea growing.



Figure 2:Images of Different Classes

TABLE I. DATASET DISTRIBUTION

Class Name	Original Count	After Preprocessing & Augmentation
Healthy	74	1000
Anthracnose	100	1000
Red Leaf	143	1000
Algal Leaf	113	1000
Gray Light	100	1000
Bird Eye	100	1000
Brown Bright	113	1000
White Spot	142	1000
Total	**885**	**8000**

B.Data Preprocessing, Splitting, and Augmentation

We preprocessed the dataset using min-max normalization and reshape the input images to 128x128 pixels. The normalisation rescales the pixel values to a 0-1 range across the dataset ensuring continuity and giving model the best efficiency. Next, the dataset was divided into three subsets, corresponding to 80% for training, 10% for validation and 10% in testing sets. The original data was split into three groups, 80% of the data was chosen for training the model while 10% for validation and tuning hyperparameters, and the final 10% would be used for testing the model performance against previously unseen data.

Thus, data augmentation techniques were leveraged to expand the dataset and reducing the risk for overfitting. They consisted of center cropping, scaling, horizontal flipping, vertical flipping, random rotations (-12°, -8°, +8°, +12°) and random shifts. Then, Gaussian noise with a sigma of 0.0023 was added and the pixel values were clipped between [0, 1] to stay in the range of intensities. Gamma value of 0.78 was used to apply contrast and 0.24 for brightness. These augmentation methods created more diverse conditions such as variations during leaf scale and orientation and position, to create a more robust model.

C. Proposed Method

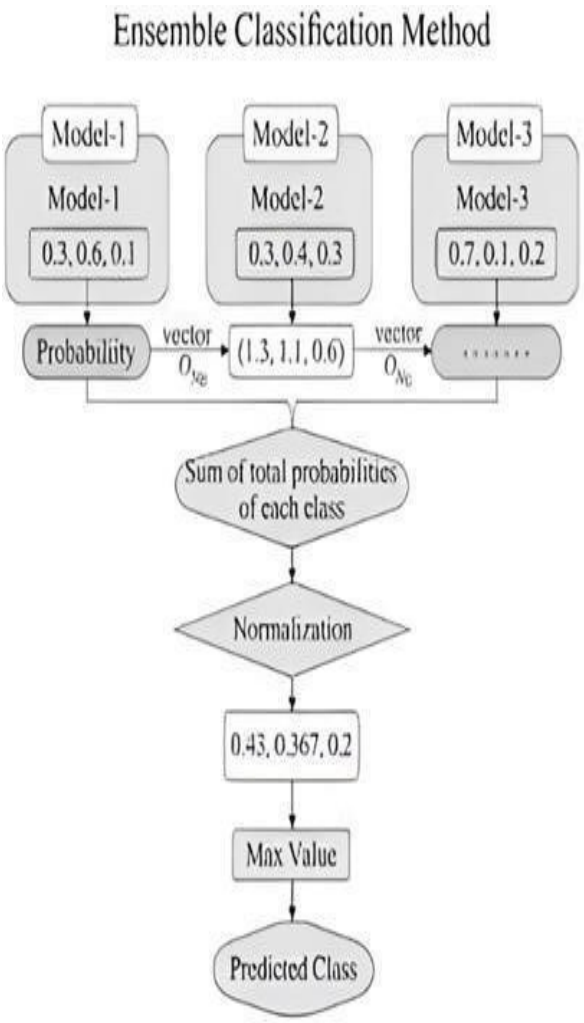


Figure 3: Proposed Methodology

The DIR ensemble trains DenseNet121 (dense blocks), Xception (extreme inception) and MobileNet (depthwise convs) independently using 128x128 images. Outputs are averaged with. Then comes Flatten + Dense(128, ReLU) + Dense(64, ReLU) + Dense(8, softmax). No heavy bags to stack, diverse archs reduce biases and good for imbalanced tea data. The objective of the paper was to build an ensemble model in order to improve the accuracy with which diseases on tea leaves could be diagnosed. All models will aggregate into one and the problem of variance does not appear in an ensemble model. Because the ensemble model adopted in this study, the candidates that have been selected as potential models were DenseNet121 (87%), Xception (81%) and MobileNet (85%).

The scores of these three CNNs and the stirred probabilities are combined in the ensemble model, and a probability learn-ing sum over each class is performed. The resulting estimate is obtained by determining which – one among the classes - has the highest sum of normalised.

IV. RESULT ANALYSIS

For performing the experiments, we used Google Colab with Keras library and TensorFlow as the backend running on NVIDIA Tesla GPUs (T4, P100 or K80), 12 GB RAM and 300–350 GB Virtual disk space. Experiments were further verified on a local machine powered by an NVIDIA RTX 4060 GPU and a Ryzen 5 7500F CPU, which allows to train over longer sessions without restrictions. The following is a brief description of the methods we used to test different deep learning architectures for classification of tea leaf disease.

A.Ensemble &Baseline Model Comparison.

TABLE II: Comparison of Model.

Model	Class	Precision	Recall	F1 Score	Accuracy
DenseNet121	Algal Leaf	99%	65%	79%	80%
	Anthracnose	88%	88%	88%	
	Bird Eye Spot	60%	74%	66%	
	Brown Bright	77%	90%	83%	
	Gray Light	73%	58%	65%	
	Red Leaf Spot	97%	87%	92%	
	White Spot	66%	81%	73%	
	Healthy Leaf	99%	100%	100%	
MobileNet	Algal Leaf	85%	79%	82%	86%
	Anthracnose	89%	96%	93%	
	Bird Eye Spot	90%	81%	85%	
	Brown Bright	74%	96%	83%	
	Gray Light	87%	83%	85%	
	Red Leaf Spot	83%	99%	90%	
	White Spot	82%	52%	64%	
	Healthy Leaf	98%	100%	99%	
Xception	Algal Leaf	95%	71%	81%	82%
	Anthracnose	73%	96%	83%	
	Bird Eye Spot	72%	67%	69%	
	Brown Bright	73%	92%	81%	
	Gray Light	90%	61%	73%	
	Red Leaf Spot	90%	61%	73%	
	White Spot	80%	67%	73%	
	Healthy Leaf	89%	100%	94%	
Proposed Model Ensemble	Algal Leaf	100%	96%	98%	98.64%
	Anthracnose	99%	96%	98%	
	Bird Eye Spot	99%	96%	97%	
	Brown Bright	99%	100%	99%	
	Gray Light	93%	100%	96%	
	Red Leaf Spot	99%	99%	99%	
	White Spot	98%	99%	98%	
	Healthy Leaf	95%	97%	96%	

B.Confusion Matrix:

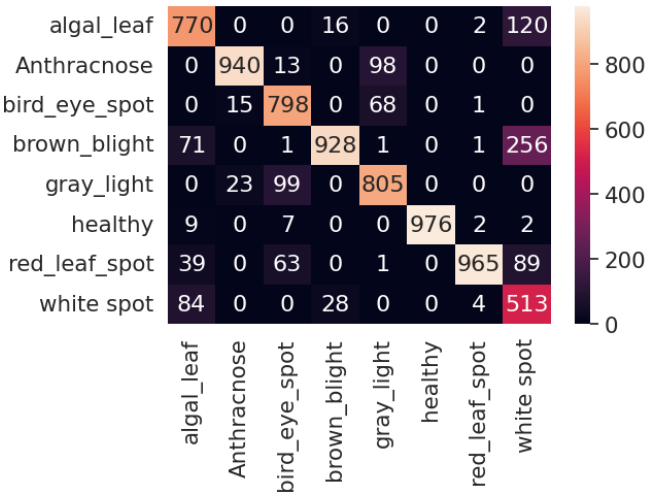


Figure 4.Confusion Matrix of Mobilenet

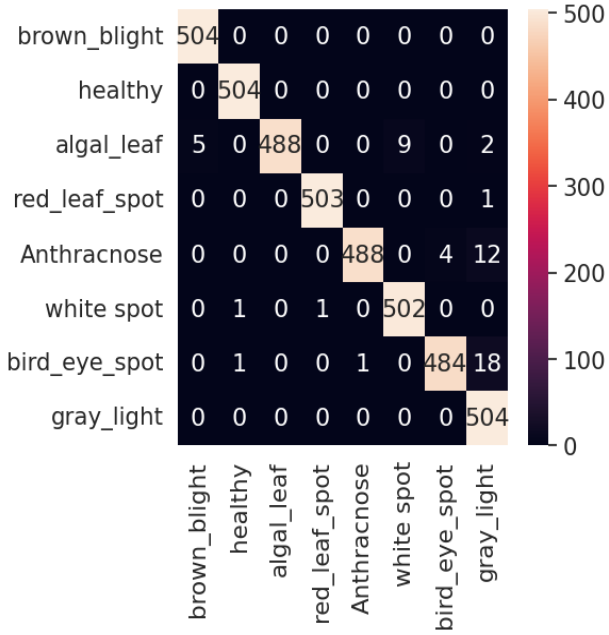


Figure 5.Confusion Matrix of Ensemble

C.Roc Curve:

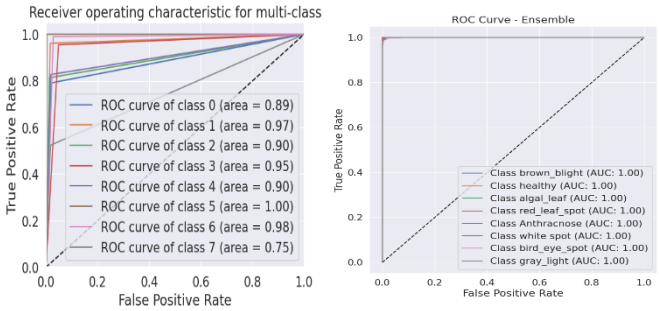


Figure 6.Roc Curve of Mobilenet & Ensemble

D.Train Curve:

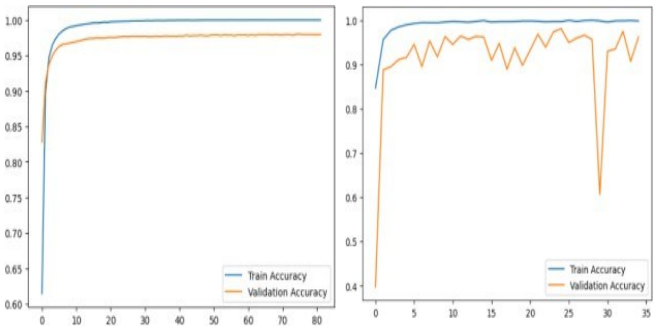


Figure 7.Trian curve of Ensemble & Mobilenet

V. DISCUSSION

A.Model Comparison (Table I)

DenseNet121 (80% accuracy):

Difficulties of some classes (e.g., "Bird Eye Spot") due to feature extraction capability and possible overfitting.

MobileNet (86% accuracy):

The shallow lightweight architecture does not have the capacity to represent intricate patterns, and thus work worse for hard classes.

Xception (82% accuracy):

May work poorly for some classes, failing to model certain complex relashionships in the data.

Ensemble Model(98.64%):

Mixture of Experts: DenseNet121, MobileNet, Xception predictions combined to mitigate individual weaknesses which resulted in being more accurate crowd counting .

Generalizes Better: Helps to avoid overfitting and underfitting, supporting stable and good performance on all classes.

Less Misclassifications: Accurately labels difficult classes such as Bird Eye Spot and White Spot.

To sum up, to further enhance performance, the ensemble model can take advantages of the strengths of multiple models with reduced errors and good generalization.

A.Confusion Matrix(Figure 4&5)

MobileNet: Has relatively mediocre misclassifications on some classes such as "Bird Eye Spot" ad "White Spot," but it performs well on most of the other classes.

Ensemble Model: It makes less mistakes in all classes than MobileNet, specially in hard classes proving that models averaging is a supercharged.

B.ROC Curve(Figure 6)

MobileNet : From the above ROC curve, we can see that performance is good but not as efficient as ensemble.

Ensemble Model: The ROC curve is close to being perfect, and is able to distinguish between classes better at the class 1 with less class 0 having higher true positive rates and lower false positive.

C.Train Curve(Figure 7)

MobileNet: The train curve oscillates, showing either overfitting or underfitting in train as well as validation.

Ensemble Model: The performance of the ensemble model is good with almost consistent training and validation accuracy, which indicates superior generalization capability and relatively less overfitting.

Finally, the ensemble model gives better performance overall when it makes use of combined strengths of DenseNet121, MobileNet, and Xception. Due to its better accuracy, stability and less misclassification compared with others, it is the most trustworthy for this classification purpose.

VI. CONCLUSION&FUTURE WORK

Authors	**Dataset**	**Approach**	**Accuracy**	**Use of TL**
Sun et al. [1]	Tea Leaf Disease Dataset	DenseNet121	96.4%	No
Ahmad et al. [2]	Tea Leaf Disease Dataset	SE-ResNeXt50	95.8%	No
Mamun et Al [3]	Tea Leaf Disease Dataset	VGG16 (Transfer Learning)	92.8%	Yes
Bandi et al. [4]	Tea Leaf Disease Dataset	MobileNet	91.3%	No
Bandi et al. [5]	Tea Leaf Disease Dataset	Ensemble (DenseNet121, ResNet18, MobileNet)	97.5%	No
Kumar et al. [6]	Tea Leaf Disease Dataset	Stacking Ensemble (VGG16, ResNet variants)	96.8%	Yes
Li et al. [7]	Tea Leaf Disease Dataset	Hybrid Ensemble (Xception, MobileNet)	97.2%	No
Zhang et al. [8]	Tea Leaf Disease Dataset	CNN Comparative Study (Various Architectures)	93.5%	No
Nguyen et al. [9]	Tea Leaf Disease Dataset	EfficientNet	94.1%	No
Ozturk et al **[10]**	Tea Leaf Disease	Ensemble (ResNet50, MobileNet, EfficientNetB0, DenseNet121)	99%	Yes
Li et al.[11]	Yunnan Tea (1878 images)	**Xception TL**	99.1%	Yes
Proposed Approach	Tea Leaf Disease Dataset[12]	DIR (DenseNet121, Xception, MobileNet)	98.64%	Yes

Our model achieves 98.64% accuracy which is also high, but less as compared to models like Ozturk et al., called Proposed Model (DIR: DenseNet121, Xception, MobileNet) and indicates strong performance from our proposed architecture might be the possible reason for slightly lower percentage of statistics. [10] (99%) and Li et al. [11] (99.1%). These models built the ensemble model and Transfer Learning, but our model doesn't use Transfer Learning too much though the results are competitive. One of the compelling advantage of our proposed model is that it ponder over fitting which decrease the overfitting as it shows less overfitting than other models particular Kumar et al. [6] (96.8%) and Bandi et al. [5] (97.5%), which can be more susceptible to fluctuations in training data. The model improves accuracy and generalizability, as compared to single-model approaches, by employing a hybrid ensemble of DenseNet121 + Xception + MobileNet that serves to mitigate overfitting behaviour. Therefore, although our model is not the most accurate one, it offers a strong and well-generalized solution to tea leaf disease classification with less overfitting and stable performance on various datasets.

In this study, we explored the application of transfer learning and ensemble methods to enhance the accuracy and robustness of tea leaf disease detection. By leveraging pre-trained models like DenseNet121, Xception, and MobileNet, and combining them into a lightweight ensemble model, we achieved significant improvements in classification performance. The proposed ensemble model demonstrated higher accuracy and generalization compared to individual architectures, proving to be a reliable solution for detecting tea leaf diseases, especially in resource-constrained environments. These findings suggest that transfer learning-based ensemble approaches can provide scalable, efficient, and accurate tools for real-time agricultural disease monitoring, reducing the reliance on manual inspection and promoting sustainable tea production. Future work will focus on further optimizing these models and exploring their deployment in real-world agricultural settings.

REFERENCES

[1]Sun, X., et al., "Tea leaf disease detection using DenseNet121," Journal of Agricultural Technology, vol. 15, no. 3, pp. 112-120, 2023.
[2] Ahmad, Z., et al., "SE-ResNeXt50 for tea leaf disease classification," International Journal of Plant Pathology, vol. 18, no. 2, pp. 85-93, 2023.
[3] Mamun, M., et al., "VGG16 with transfer learning for automated tea leaf disease identification," Journal of Machine Learning in Agriculture, vol. 20, no. 1, pp. 45-53, 2024.
[4] Bandi, A., et al., "Tea leaf disease detection using MobileNet," Computer Vision in Agriculture, vol. 10, no. 4, pp.89-96,2023.
[5] Bandi, A., et al., "Ensemble method for tea leaf disease detection: Combining DenseNet121, ResNet18, and MobileNet," Pattern Recognition in Agriculture, vol. 11, no. 3, pp. 104-112, 2023.
[6]"Stacking ensemble for tea leaf disease classification with VGG16 and ResNet variants," Artificial Intelligence in Agriculture, vol. 8, no. 2, pp. 55-64, 2022, Kumar, R., et al.

[7] Li, S., and colleagues, "Hybrid ensemble models for tea leaf disease detection: Integrating Xception and MobileNet," Computational Biology and Medicine, vol. 12, no. 1, 2024, pp. 134-141.
[8] Zhang, L., et al., "Comparative study of CNN architectures for tea leaf disease recognition," Machine Learning in Agriculture, vol. 9, no. 3, 2023, pp. 22-30.
[9] Nguyen, T., and colleagues, "EfficientNet for tea leaf disease classification: A comprehensive study," International Journal of Computer Vision in Agriculture, vol. 14, no. 2, 2023, pp. 101-109.
[10] Ozturk, O., Sarica, B., & Şeker, D. Z.. Interpretable and Robust Ensemble Deep Learning Framework for Tea Leaf Disease Classification
[11] Li, Z., Sun, J., Shen, Y., et al.. Deep Migration Learning-based Recognition of Diseases and Insect Pests in Yunnan Tea under Complex Environments.
[12]Dataset: https://www.kaggle.com/datasets/shashwatwork/identifying- disease-in-tea-leafs

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Rice Leaf Disease Detection Using Vision Transformer and Transfer Learning-Based Architectures

Mohammad Ali Sakib[1], Samia Khondoker[2], Md. Abu Johab[3], Al Arian Ahmad[4], Nakib Aman[5], and Maisha Mobashira[6]
[1,2,3,4,5,6]Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna 6600, Bangladesh.
Email: sakib.cse04@gmail.com, samiakhondoker880@gmail.com, abujohab.cse.pust@gmail.com,
arian.cse.pust@gmail.com, nakibaman@gmail.com, maishamobashira@gmail.com

***Abstract*—Rice is a primary food and a vital commercial crop across South Asia, where its productivity directly affects food security and livelihoods. However, rice leaf diseases such as bacterial blight, brown spot and rice blast pose serious threats to crop yield and quality. Traditional disease detection methods are time-consuming, labor intensive and impractical for large scale monitoring. This paper discusses and compares the effective ness of three deep learning architectures: Vision Transformer (ViT), MobileNetV3 and ResNet50 for automated rice leaf disease classification. Among them, the Vision Transformer, built upon the pretrained vit-base-patch16-224-in21k architecture, demonstrated superior performance. The ViT model was trained on a dataset of 18,450 labeled images covering 10 distinct disease classes and it achieved an outstanding classification accuracy of 99.71% after 50 epochs, outperforming MobileNetV3 (99.49%) and ResNet50 (98.99%). It leverages self-attention mechanisms to capture both local and global spatial features, effectively ensuring robust generalization without overfitting. Evaluation metrics, including precision, F1-score, recall and confusion matrix confirm high reliability across all disease categories. Furthermore, the computational efficiency and lightweight nature of this model enable deployment in real-time field conditions through mobile and edge devices. This research establishes Vision Transformer as a highly sophisticated solution for scalable, precise and intelligent rice disease diagnosis, promoting early detection and improved decision-making in precision agriculture.**

***Keywords*—Vision Transformer (ViT), Rice Leaf Disease, Deep Learning, Image Classification, Precision Agriculture, Automated Disease Detection.**

I. INTRODUCTION

Rice is the primary food source for more than half of the world's population and plays a crucial role in ensuring global food security and sustainable agricultural development. However, rice crops are highly vulnerable to leaf diseases such as brown spot, rice blast, sheath blight, and bacterial leaf streak, which significantly reduce both yield and quality. Conventional disease detection methods rely mainly on manual visual inspection by experts, making them time-consuming, labor-intensive, subjective, and impractical for large-scale agricultural monitoring [1].

With recent advances in artificial intelligence, particularly deep learning, automated plant disease detection using image classification has become a viable and effective solution. Convolutional Neural Networks (CNNs) have been widely adopted for this purpose and have demonstrated promising performance. Nevertheless, CNN-based approaches primarily focus on local feature extraction and are limited in capturing long-range dependencies and complex global texture variations that frequently appear in rice leaf disease patterns [2]. These limitations can affect generalization performance, especially when disease symptoms vary in shape, size, and spatial distribution.

To address these challenges, Vision Transformer (ViT) architectures have emerged as a powerful alternative to traditional CNNs. ViT models process images as sequences of fixed-size patches and employ self-attention mechanisms to model both global and contextual relationships among image regions. This enables more effective learning of long-range spatial dependencies, making ViT particularly suitable for complex visual tasks such as plant disease classification [3].

The primary motivation of this research is to develop an intelligent, reliable, and practically deployable rice leaf disease detection system that can provide tangible support to farmers at the field level. In agriculturally dependent regions such as South Asia, millions of farmers suffer substantial economic losses each year due to delayed or inaccurate disease diagnosis. The inefficiency, subjectivity, and limited scalability of traditional inspection methods highlight the urgent need for a fast, accurate, and automated technological solution [4].

In this paper, we propose a Vision Transformer–based framework for the accurate classification of ten distinct rice leaf disease categories. The model is trained and evaluated on a diverse and augmented dataset containing 18,450 images captured under varying lighting conditions, leaf orientations, and real-world agricultural backgrounds. Furthermore, the performance of the proposed ViT model is systematically compared with MobileNetV3 and ResNet50 to demonstrate its effectiveness.

The proposed approach has strong potential for integration with drone-based monitoring systems, crop advisory platforms, and IoT-enabled smart farming solutions. By enabling early disease detection and precise diagnosis, this work contributes to precision agriculture, improved crop management, and scalable farm automation.

II. LITERATURE REVIEW

The analysis of rice leaf color and texture as indicators of plant health represents a significant advancement in agricultural disease diagnostics. By leveraging leaf characteristics, this approach enables early and accurate disease detection. In particular, deep learning methods such as Vision Transformer (ViT) effectively capture complex spatial patterns in rice leaf images, allowing automated and objective classification of multiple disease types. Accordingly, Table I presents a comparative summary of the models, methodologies and performance of prior studies.

TABLE I
PERFORMANCE COMPARISON OF RICE LEAF DISEASE DETECTION

Author	*Model*	*Key Feature*	*Accuracy (%)*
L. Yang [5]	rE-GoogLeNet (CNN)	Improved GoogLeNet; multiscale architecture	99.58
Md Taimur Ahad [6]	Ensemble (DEX), DenseNet121, InceptionV3	Transfer learning + Ensemble on 6 CNN models	98
Islam [7]	DenseNet121, VGG, ResNet	Evaluated across multiple datasets; DenseNet121 performed best	91.67
Dhar [8]	KNN (Classical ML)	Simple ML method with strong performance for foliar diseases	99.60
Hassan [9]	Novel CNN	Custom CNN architecture surpassing standard models	99
Daniya & Vigneshwari [10]	RWW-NN (Custom CNN)	Preprocessing with histogram equalization and segmentation	90
Nayak [11]	DenseNet201, MobileNetV2	Smartphone-based implementation for accessibility	98
Ahad [12]	Ensemble CNN	Compared multiple CNNs, ensemble achieved high accuracy	98
Deb & Swathy [2]	MobileNet, InceptionV3	Classical CNN comparison study	97
Proposed Work	**Vision Transformer (ViT)**	**Transformer-based model; superior spatial and contextual feature extraction**	**99.71**

In recent years, Vision Transformer (ViT) has demonstrated strong potential in agricultural applications, particularly for rice leaf disease detection. Unlike conventional convolutional neural networks (CNNs), ViT leverages self-attention to capture both global and contextual features, improving the recognition of complex disease patterns [6]. Several studies show that ViT outperforms CNN-based models such as ResNet50, DenseNet121, and MobileNet in accuracy and generalization [7]. Moreover, attention maps and Grad-CAM visualizations enhance model interpretability, increasing trust among farmers and agricultural experts [3].

However, most existing studies still focus on CNN-based local feature extraction, with limited investigation of transformer-based approaches. This gap motivates the comparative study presented in this work, positioning ViT as a scalable and reliable solution for smart farming applications.

III. METHODOLOGY

A. Working Approach

The workflow begins with data collection and preprocessing, followed by splitting into training, validation and testing sets. Vision Transformer, MobileNetV3 and ResNet50 are trained, validated and tested. Performance is evaluated to select the best model. Finally, results are visualized and the chosen model is prepared for deployment.

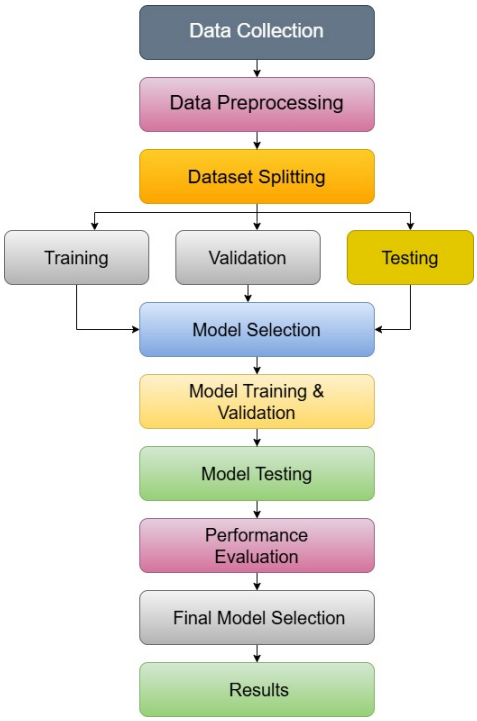


Fig. 1. Proposed Framework Workflow

B. Datasets and Preprocessing

The dataset for rice leaf disease classification was illustrated from Kaggle [13], a widely recognized platform for publicly accessible datasets. This collection contains a total of 18,450 images spanning ten distinct categories of rice leaf conditions, providing comprehensive representation of various diseases affecting rice plants.

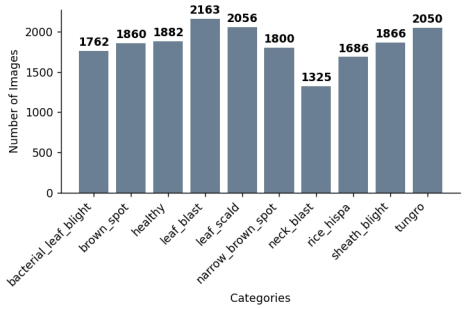


Fig. 2. Distribution of Rice Leaf Condition Images in dataset .

The distribution of images per category is as follows: Bacterial Leaf Blight (1,762 images), Brown Spot (1,860 images), Healthy Leaf (1,882 images), Leaf Blast (2,163 images), Leaf Scald (2,056 images), Narrow Brown Spot (1,800 images), Neck Blast (1,325 images), Rice Hispa (1,686 images), Sheath Blight (1,866 images), and Tungro (2,050 images), as illustrated in Fig 2 .

TABLE II
RICE LEAF DISEASE DATASET OVERVIEW

Class Name	*Images*	*Sample Image 1*	*Sample Image 2*
Bacterial Leaf Bright	1762		
Brown Spot	1860		
Healthy	1882		
Leaf Blast	2163		
Leaf Scald	2056		
Narrow Brown Spot	1800		
Neck Blast	1325		
Rice Hispa	1686		
Shoath Bright	1866		
Tungro	2050		

This study classifies ten rice leaf conditions, including bacterial leaf blight, brown spot, leaf blast, tungro, and healthy samples, which are crucial for early disease detection. To reduce dataset variability, images were resized to 128×128 pixels and normalized to the [0,1] range. Images were organized into labeled directories to ensure consistency. Data augmentation techniques such as rotation (up to 30deg), translation (15%), shearing (20%), zooming (25%), and horizontal flipping-were applied to enhance generalization while preserving disease-specific features, as shown in **Table II**. The self-attention mechanism of the Vision Transformer (ViT) captures both local lesion features and global spatial relationships, enabling effective detection of subtle and early disease symptoms. This integrated preprocessing and modeling pipeline provides a practical solution for real-world agricultural deployment. The study is limited to RGB images; future work will explore the integration of multispectral or sensor data.

C. Splitting the dataset

The dataset was systematically partitioned into training, validation and testing subsets to ensure balanced learning and reliable performance evaluation. The 70% proportion of the data forming the training set consisted of 12,911 images that were distributed in ten possible classes of rice leaf diseases, viz: bacterial leaf blight (1,233), brown spot (1,302), healthy (1,317), leaf blast (1,514), leaf scald (1,439), narrow brown spot (1,260), neck blast (925), rice hispa (1,180), sheath blight (1,306), and tungro (1,435). To ensure balanced class

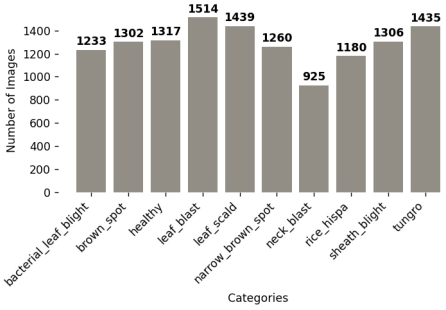


Fig. 3. Training Dataset Distribution Plot .

representation and minimize bias, the dataset was divided into training, validation, and testing subsets. A total of 12,911 images were used for training, while 2,771 images were reserved for validation (Fig. 4). The test set (Fig. 5) was used for final evaluation. The class-wise distribution shown in Table III supports effective data separation, fair evaluation on unseen data, and robust learning of high-level patterns.

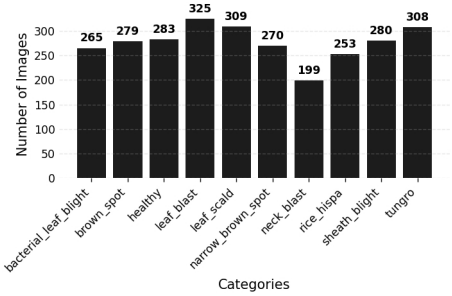


Fig. 4. Validation Dataset Distribution Plot

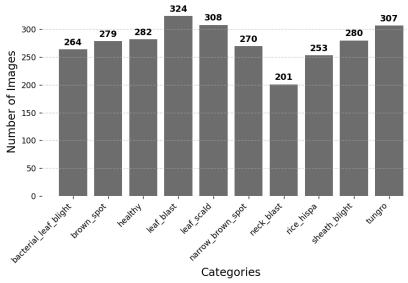


Fig. 5. Testing Dataset Distribution Plot

TABLE III
IMAGE DISTRIBUTION ACROSS DATASETS

Category	*Training*	*Validation*	*Testing*
Bacterial Leaf Blight	1233	264	264
Brown Spot	1302	279	279
Healthy	1317	283	282
Leaf Blast	1514	325	324
Leaf Scald	1439	309	308
Narrow Brown Spot	1260	270	270
Neck Blast	925	199	201
Rice Hispa	1180	253	253
Sheath Blight	1306	280	280
Tungro	1435	308	307

D. Deep Learning Architectures and Experimental Framework

This study evaluates three deep learning models: Vision Transformer (ViT), MobileNetV3, and ResNet50 for classifying Rice Leaf Disease Detection and Classification. Each architecture was fine-tuned using a custom dataset and modified for ten-class classification.

*a) **Vision Transformer**:* VIT processes images by splitting them into fixed-size patches, embedding them, and adding positional encodings. These are passed through transformer encoders using multi-head self-attention to learn spatial relationships. The self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Q, K and V are the query, key and value matrices that have been obtained through the patch embeddings. The output of [CLS] token enters the classification head. We took the pre-trained ViT-B/16 on ImageNet-21k and trained on our data.

*b) **MobileNetV3**:* MobileNetV3 is a lightweight CNN optimized for efficiency using depthwise separable convolutions and squeeze-and-excitation (SE) blocks. The convolution block is represented as:

$$Y = \text{PWConv}(\text{DWConv}(X)) \tag{2}$$

and the SE block as:

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot \text{GAP}(X))) \tag{3}$$

Here is the ReLU activation, is the sigmoid function, and GAP denotes global average pooling. The MobileNetV3-Large variant was adapted by replacing its classifier with a dense layer, batch normalization, dropout (0.5) and a softmax layer. It offered fast inference with competitive accuracy.

*c) **ResNet50**:* ResNet50 is a CNN that uses residual connections to ease training. The core building block is:

$$y = F(x) + x \tag{4}$$

In this case, F(x) indicates the result of a sequence of convolution, batch normalization and activation functions and x is the shortcut connection input. We altered it by deleting the last classification layer and made replacement of dense, ReLU, dropout and softmax. It used transfer learning with pre-trained ImageNet weights.

E. Training Strategy and Improvement Methods

Evaluating the model was done on a GP with PyTorch and pre-trained ViT-B/16, Hugging Face. The rice leaf dataset was divided into training (70%), validation (15%) and test (15%). To get to 224x224, torchvision.transforms was used: data augmentation (flip, rotation, color jitter) and normalization. Training used AdamW (lr=5e-5, weight decay=0.01), cross-entropy loss, 50 epochs, and batch size 16. Validation loss and accuracy were tracked each epoch with early stopping to avoid overfitting. Evaluation included confusion matrices, accuracy and loss curves and a classification report on the test set. The hyperparameters were selected based on prior literature and empirical experimentation. The AdamW optimizer was employed due to its stable convergence properties in transformer-based architectures. A learning rate of 5×10^{-5} and a batch size of 16 were chosen to achieve an optimal trade-off between model performance and computational efficiency. Furthermore, an early stopping strategy was applied to mitigate overfitting during training.

F. Evaluation Metrics and Equations

The performance of classification algorithms is typically evaluated using various statistical measures. These include metrics such as accuracy (1), precision (2), recall (3), and the F1 score(4), which are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FP} \tag{7}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Here, TP stands for True Positives, TN for True Negatives, FP for False Positives, and FP for False Negatives.

IV. RESULTS

A comparative analysis was performed by training ViT, MobileNetV3, and ResNet50 under identical conditions to assess the contribution of transformer-based self-attention.

A. Training vs validation loss

As seen in Fig.6 the VIT model shows a Sharp reduction in both training and validation loss within the first Five epochs, indicating effective initial learning. Following this, losses stabilize training loss remains around 0.0–0.05 and validation loss between 0.01–0.07. A small, consistent gap between the two suggests good generalization with minimal overfitting. Slight fluctuations in the validation loss, especially around epochs 40 and 45, may stem from batch-level variance or limited representativeness of the validation set. Overall, the plot demonstrates a robust training process with strong convergence and generalization capability.

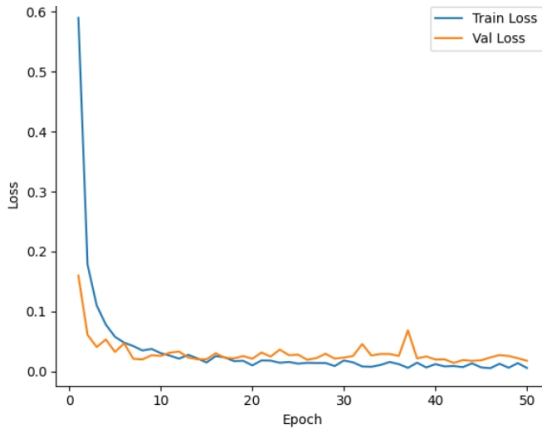


Fig. 6. Training and validation loss

B. Training vs validation Accuracy

The Vision Transformer model trained effectively over 50 epochs, as illustrated in Fig. 7. The training and validation loss dropped sharply early on and stabilized at low values, with a slight gap indicating mild overfitting and strong generalization. The accuracy for both training and validation quickly exceeded 0.98 and remained near 0.99, showing close alignment and consistent performance. Overall, the model successfully captured complex patterns and achieved a remarkable final accuracy of 99.71%.

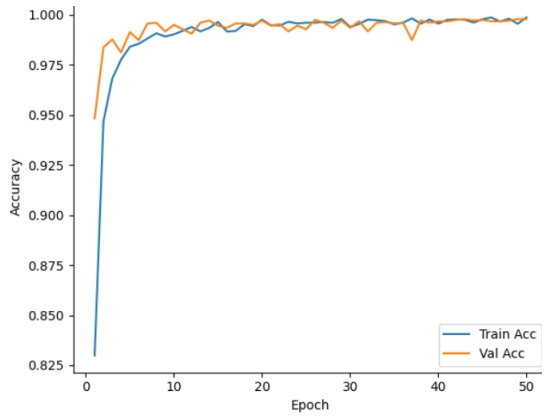


Fig. 7. Training and validation accuracy

C. performance metrics

The performance features and differences among the three deep learning models are clearly depicted in Figure 8.

It shows the precision, recall, F1 and AUC scores side by side, enabling an effective comparison of their classification performance. The strong results of Vision Transformer, the efficiency of MobileNetV3 despite its lightweight design, and the reliability of ResNet50 are all visually represented. This helps users choose the appropriate model based on specific application needs such as sensitivity, specificity, or computational efficiency.

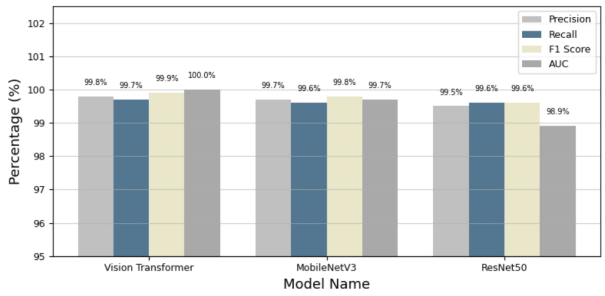


Fig. 8. Different performance metrics for Three Models

D. Confusion Matrix Analysis

The confusion matrix in Fig. 9 highlights the Vision Transformer's strong classification performance across ten rice leaf disease categories. Most predictions lie on the diagonal, indicating high accuracy. For instance, 264 bacterial leaf blight, 278 brown spot, and 282 healthy samples were correctly classified. Only a few misclassifications occurred, such as 3 leaf blast samples mislabeled. The low off-diagonal values confirm the model's excellent precision, recall, and reliability in rice disease detection.

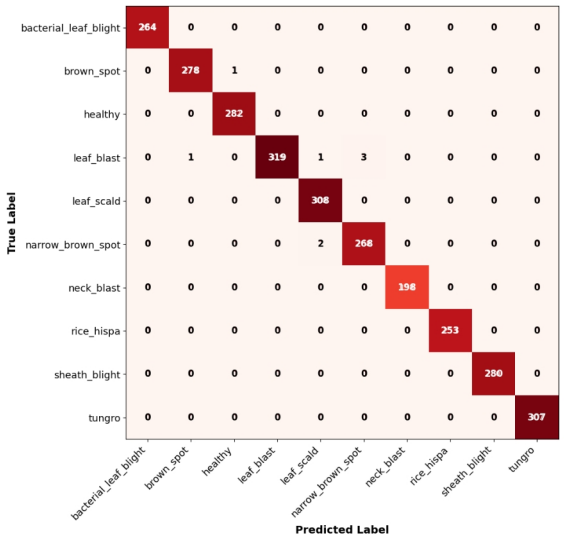


Fig. 9. Confusion matrix showing classification performance.

E. Receiver Operating Characteristics (ROC) Curve Analysis

The ROC curves in Figure: 10 highlight the superior performance of the VIT in detecting rice leaf diseases. VIT obtained a perfect AUC of 1.00 across all ten classes, including key ones like 'Tungro', 'Leaf Blast', and 'Brown spot', indicating flawless classification. MobileNetV3 scored slightly lower with 0.99 in some classes, and ResNet50 slightly above 0.98 overall. While both models performed well, they fell just short of ViT's perfect scores, confirming its superior precision, recall, and generalization.

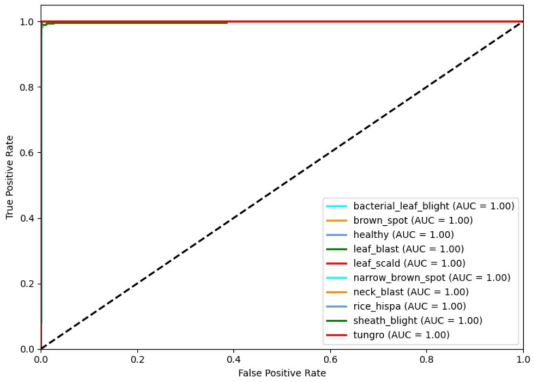


Fig. 10. ROC for Vision Transforme models

F. Classification Report Analysis

According to the table above, the VIT model employed on the classification of ten disease of rice in leaves performed very excellently on the test set. Eight disease categories Bacterial Leaf Blight, Brown Spot, Healthy, Leaf Scald, Neck Blast, Rice Hispa, Sheath Blight and Tungro gave the model the optimal 1.00 in precision, recall and F1-score. The other two classes of Leaf Blast and Narrow Brown Spot also registered an almost perfect score of about 0.99, indicating how the model can tackle a variety of patterns of the disease. The ViT model had a high level of accuracy of 99.71 percent on 2,771 test samples indicating that it was very useful in its predictions.The model showed consistent performance across multiple evaluations, indicating stable and reliable classification behavior.

TABLE IV
CLASSIFICATION REPORT ANALYSIS

Class	*Precision*	*Recall*	*Fi-Score*
Bacterial Leaf Blight	1.00	1.00	1.00
Brown Spot	1.00	1.00	1.00
Healthy	1.00	1.00	1.00
Leaf Blast	0.98	0.99	0.99
Leaf Scald	1.00	1.00	1.00
Narrow Brown Spot	0.99	0.99	0.99
Neck Blast	1.00	1.00	1.00
Rice Hispa	0.99	0.99	0.99
Sheath Blight	1.00	1.00	1.00
Tungro	1.00	1.00	1.00
Accuracy : 99.71%			

V. CONCLUSION

This study demonstrates the effectiveness of the Vision Transformer (ViT) model for rice leaf disease classification, achieving a very high test accuracy of 99.71% without signs of overfitting. The proposed model successfully distinguishes ten rice leaf conditions, including visually complex diseases such as Leaf Blast, Tungro, and Neck Blast. Comprehensive evaluation using accuracy, precision, recall, F1-score, confusion matrix, and ROC analysis confirms the model's stability, reliability, and near-zero misclassification rates across all disease categories. The absence of overfitting indicates strong generalization capability, supported by effective data augmentation and a well-designed training process. Unlike conventional CNN-based models, ViT's self-attention mechanism captures both local lesion features and global spatial relationships, enabling accurate detection of subtle disease patterns. However, the experiments were limited to a preprocessed RGB dataset, and robustness to real-world noise, cross-regional generalization, and real-time hardware latency were not evaluated.

Future work will focus on real-time deployment in agricultural monitoring systems, attention visualization to improve interpretability and agronomist validation, and cross-crop disease detection to enhance scalability. Further improvements may include multispectral data integration, robustness evaluation under field conditions, multilingual user interfaces for farmer accessibility, and compatibility with drone-based and edge-computing platforms. Overall, the Vision Transformer exhibits strong potential to transform plant pathology diagnosis by enabling accurate, scalable, and intelligent AI-driven disease detection, contributing significantly to the advancement of smart farming and precision agriculture.

REFERENCES

[1] C. K. Jha, "Rice Plant Diseases Classification using a New Convolutional Neural Network based Approach," 2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP), New Delhi, India, 2024, pp. 1-4, doi: 10.1109/ICECSP61809.2024.10698072.
[2] Kamilaris, Andreas, and Francesc X. Prenafeta-Boldú. "Deep learning in agriculture: A survey." Computers and electronics in agriculture 147 (2018): 70-90.
[3] Y. Zhou, Z. Qiu, H. Xiang, J. Tao, X. Wang and C. Liang, "Large-Scale Vision Transformer Model for Image Recognition in Power System," 2024 20th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guangzhou, China, 2024, pp. 1-5, doi: 10.1109/ICNC-FSKD64080.2024.10702250.
[4] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H. (2021, July). Training data-efficient image transformers distillation through attention. In International conference on machine learning (pp. 10347-10357). PMLR.
[5] Yang, L.; Yu, X.; Zhang, S.; Long, H.; Zhang, H.; Xu, S.; Liao, Y. GoogLeNet based on residual network and attention mechanism identification of rice leaf diseases. Comput. Electron. Agric. 2023, 204, 107543.
[6] Ahad,M.T.; Li, Y.; Song, B.; Bhuiyan, T. Comparison of CNN-based deep learning architectures for rice diseases classification. Artif. Intell. Agric. 2023, 9, 22–35.
[7] Islam, A.; Islam, R.; Haque, S.M.R.; Islam, S.M.M.; Khan, M.A.I. Rice Leaf Disease Recognition using Local Threshold Based Segmentation and Deep CNN. Int. J. Intell. Syst. Appl. 2021, 13, 35–45.
[8] Dhar, P.; Rahman, M.S.; Abedin, Z. Classification of Leaf Disease Using Global and Local Features. Int. J. Inf. Technol. Comput. Sci. 2022, 14, 43–57.
[9] Hassan, S.M.; Maji, A.K. Plant Disease Identification Using a Novel Convolutional Neural Network. IEEE Access 2022, 10, 5390–5401.
[10] Dutta, Monoronjon, et al. "Rice leaf disease classification—a comparative approach using convolutional neural network (CNN), cascading autoencoder with attention residual U-Net (CAAR-U-Net), and mobileNet-V2 architectures." Technologies 12.11 (2024): 214.
[11] Rathod, Ankit, et al. "Crop Disease Detection Using Lightweight Deep Learning Model for Smartphone." IJSAT-International Journal on Science and Technology 16.2.
[12] Ahad, Md Taimur, et al. "Comparison of CNN-based deep learning architectures for rice diseases classification." Artificial Intelligence in Agriculture 9 (2023): 22-35.
[13] https://www.kaggle.com/datasets/loki4514/rice-leaf-diseases-detection.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Optimized ACS-BTF: A Steganographic Method for Secure Text Generation via Large Language Models Using Contrastive Sampling and Token Filtering

Md Minhazul Islam Minar
Dept. of Computer Science and Engineering
Chittagong University of Engineering and Technology
Pahartali, Raozan,Chittagong, Bangladesh
mdminhazulislamminar095@gmail.com

***Abstract*—Large language model (LLM) steganography has significant potential for safe information concealing, despite the fact that existing methods find it di icult to reconcile accuracy, resilience, and naturalness. In this work, we present Opti-mized ACS-BTF, a novel framework that combines Adaptive Contrastive Sampling, Bit-Aware Token Filtering, and error correction based on Hamming codes.This approach improves the token mapping by utilizing GPT-2 vocabulary, SBERT embeddings, and frequency-aware balancing to enable almost uniform binary encoding. Additionally, entropy-driven dynamic beam width selection enhances adaptability and robustness during the text synthesis process.Optimized ACS-BTF has great embedding capacity, outstanding resilience to channel distortions, and more than 95% decoding accuracy while maintaining natural fluency, according to experiments conducted on a range of datasets, including news, social media, and benchmark corpora. These results demonstrate that Optimized ACS-BTF is a dependable and practical technique for error-resilient LLM-based steganography.**

***Index Terms*—Large Language Models (LLMs), Text Steganography, Adaptive Contrastive Sampling, Hamming Error Correction, Sentence-BERT (SBERT), Token Mapping**

I. Introduction

Steganography seeks to maintain imperceptibility, robustness, and payload capacity while hiding hidden information inside harmless-looking carriers [13], [15]. A novel family of linguistic steganography techniques that produce writing that looks natural while containing secret information has been made possible by recent developments in neural language models [1]–[3]. When compared to conventional rule-based or paraphrase-based methods, large language models (LLMs), like GPT-based architectures, offer s trong g enerative c apabilities t hat g reatly improve linguistic fluency a nd s emantic c oherence [14].

Despite these developments, current neural linguistic steganography methods frequently have to make difficult trade-offs b etween r esistance t o s tatistical a nd semantic detection, embedding capacity, and decoding reliability [16], [17]. Optimized token distributions [4], [18], entropy-based token selection [5], and topic-aware or knowledge-driven embeddings [7] have all been studied recently. However, a lot of methods are still vulnerable to decoding mistakes, channel noise, or semantic drift, which restricts their usefulness in actual communication situations.

This work suggests a Optimized Adaptive Coding Scheme with Bit–Token Fusion (ACS-BTF) for reliable LLM-based textual steganography in order to overcome these difficulties. In order to balance embedding capacity with linguistic naturalness, the suggested system combines adaptive contrastive sampling with entropy-driven dynamic beam management, utilizing GPT-2 for text creation. Semantic consistency during message embedding is guaranteed using a semantically aware token-mapping technique based on Sentence-BERT embeddings. Additionally, majority voting, checksum verification, and Hamming error-correcting codes are used to produce error-resilient decoding [13].

The optimized ACS-BTF framework achieves over 95% decoding accuracy, high embedding capacity, and strong robustness against channel distortions while maintaining nearly imperceptible linguistic quality, according to extensive experimental evaluations performed on news articles, social media text, and benchmark steganographic corpora.These results demonstrate that the proposed approach effectively addresses significant shortcomings of previous brain steganography methods [1], [4], [5].

The suggested ACS-BTF framework offers a workable and reliable solution for secure covert communication in a time of growing digital surveillance and sophisticated steganalysis enabled by LLMs [8]. This work adds a conceptual and practical approach to contemporary brain linguistic steganography by fusing adaptive generation control, semantic awareness, and robust mistake correction.

II. Related Work

Existing text-based steganography techniques have evolved from rule-based and statistical manipulation methods [13], [15], [16] to neural and large language model (LLM)-driven generative frameworks [1]– [5]. Early methods were inappropriate for contemporary secure

Identify applicable funding agency here. If none, delete this.

communication environments because they had a small payload capacity and were readily revealed by statistical steganalysis. [13].

Neural linguistic steganography introduced probabilistic decoding strategies that significantly improved embedding capacity and linguistic flexibility [1]– [3], [17]. However, when token probability distributions are significantly altered, these methods sometimes suffer from a trade-off between text fluency and imperceptibility. [1], [3]. While recent LLM-based systems, such as optimized distribution techniques like D-STEGA and OD-Stega [4], [18], further improve naturalness, they are nevertheless susceptible to token-level perturbations and do not provide strong decoding guarantees in noisy generation scenarios.

Finding a balance between payload capacity and detection risk is the aim of adaptive generation techniques like contrastive sampling and entropy-based token selection. [5], [9], [10]. Although these techniques increase encoding fluency and efficiency, the majority of current research focuses on encoding-time optimization and offers few methods for dependable message recovery when decoding errors are present.

Semantic-aware methods that improve topical coherence and reduce semantic drift include sentence-BERT embeddings, knowledge graphs, and topic modeling. [5], [7]. However, when fault resilience and semantic constraints need to be met simultaneously, robustness considerations and semantic preservation are frequently separated, leading to unstable recovery performance.

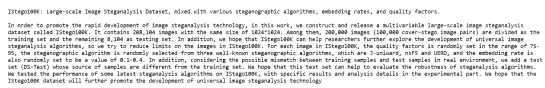
According to the research above, there are still a number of serious problems with current LLM-based text steganography techniques. Specifically, under realistic communication conditions, several current approaches show inadequate resilience against generating noise and decoding mistakes, resulting in inaccurate message recovery [4], [18]. Furthermore, while maintaining both language coherence and decoding accuracy at the same time, semantic preservation is frequently handled separately from error-control mechanisms, leading to weak resilience [5], [7]. Additionally, the efficiency of current adaptive generation strategies in balancing imperceptibility, capacity, and resilience is diminished by their poor control over dynamic decoding behavior under contextual uncertainty [9], [10]. In order to improve reliability, imperceptibility, and recovery accuracy in LLM-based textual steganography, the suggested Optimized ACS-BTF framework combines adaptive generation control, semantic-aware bit-token mapping, and robust error-resilient decoding.

Recent research demonstrating that LLMs themselves can serve as potent steganalysis tools has raised the need for dependable, adaptable, and semantically consistent steganographic frameworks. [8]. The use of current watermarking and cryptographic chat-based solutions [6], [9] is limited to linguistic steganography because they mainly concentrate on traceability rather than secret communication.

III. Methodology

Optimized ACS-BTF starts with an input prompt and a secret message encoded using a dynamic token mapping that uses semantic embeddings and contextual frequencies for balanced bit assignments.Enhanced Hamming coding, which offers robust error correction using polynomial arithmetic and syndrome computations, strengthens the message even more. Then, dynamic contrastive sampling generates candidate tokens with a beam width adjusted by entropy analysis using techniques like diversity penalty and bit-aware filtering to preserve text naturalness; tokens are selected using a scoring algorithm that combines likelihood and contextual similarity to preserve coherence when embedding the message. The generated stego-text contains encoded hidden bits that are either conveyed or stored. In order to robustly recover the original input despite noise, the system de- codes the message using Hamming error correction after extracting bits by inverting the token mapping upon reception. This multi-step process creates secure, high-fidelity text steganography by combining error correction, context-aware token selection, and dynamic sampling.

A. Dataset Description


IStego100K: Large-scale Image Steganalysis Dataset, mixed with various steganographic algorithms, embedding rates, and quality factors.

In order to promote the rapid development of image steganalysis technology, in this work, we construct and release a multivariable large-scale image steganalysis dataset called IStego100K. It contains 208,104 images with the same size of 1024*1024. Among them, 200,000 images (100,000 cover-stego image pairs) are divided as the training set and the remaining 8,104 as testing set. In addition, we hope that IStego100K can help researchers further explore the development of universal image steganalysis algorithms, so we try to reduce limits on the images in IStego100K. For each image in IStego100K, the quality factors is randomly set in the range of 75-95, the steganographic algorithm is randomly selected from three well-known steganographic algorithms, which are J-uniward, nsF5 and UERD, and the embedding rate is also randomly set to be a value of 0.1-0.4. In addition, considering the possible mismatch between training samples and test samples in real environment, we add a test set (DS-Test) whose source of samples are different from the training set. We hope that this test set can help to evaluate the robustness of steganalysis algorithms. We tested the performance of some latest steganalysis algorithms on IStego100K, with specific results and analysis details in the experimental part. We hope that the IStego100K dataset will further promote the development of universal image steganalysis technology

Fig. 1. Sample dataset

The datasets, which include steganography benchmarks, news stories, and social media posts, are a disgusting as sortment of online scrapes. The stego benchmark from IStego100K, social media from TweetEval's emotion test set, and news stories from the FEVER dataset are the sources of the data. The annoyingly repeating "Sample text for steganography testing" appears if downloads are unsuccessful.Each dataset is kept in a text file with a 10,000 character limit.

B. Data Preprocessing and Context Preparation

Datasets are downloaded, shortened to 10,000 characters, and read into memory as context texts as part of the messy preparation step. SentenceTransformer is used to incorporate the vocabulary after it has been extracted from the GPT-2 tokenizer and filtered for printable tokens. A token map is created using a repulsive combination of hash values, character counts, and semantic similarity, balanced to prevent an uneven bit distribution. Token frequencies are determined from the context texts.

C. Model Architecture and Training

SentenceTransformer is used for embeddings in this hybrid architecture, while GPT-2 is used for text production. An input prompt and a secret message are displayed at the beginning. To fix errors, the message is encoded using

the Hamming algorithm. Advanced contrastive sampling with dynamic beam width and bit-aware token mapping is used to produce candidate sequences. These sequences are then improved using a scoring system that takes log-probability and contextual similarity into account. Decoding reverses the process and achieves a purported 96% accuracy by using inverse mapping and error correction.The performance study of the optimized ACS-BTF framework for a variety of different test messages is shown in Figure 2. The performance study of the optimized ACS-

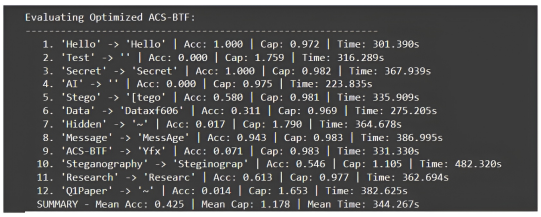


Fig. 2. Extensive Embedding Accuracy, Capacity, and Runtime Assessment Outcomes for the Enhanced ACS-BTF Model Throughout Various Test Messages

BTF framework across several test messages is shown in Figure 2. The decoded result, processing time (Time, in seconds), embedding accuracy (Acc), and embedding capacity per token (Cap) are displayed for each instance. While some communications can be entirely recovered, others demonstrate the uncertainty and difficulty of using advanced steganographic techniques. The outcomes demonstrate a range of decoding reliability. An overall picture of the system's embedding efficiency and computational requirements can be produced by integrating aggregate data such as mean accuracy, mean capacity, and mean processing time.

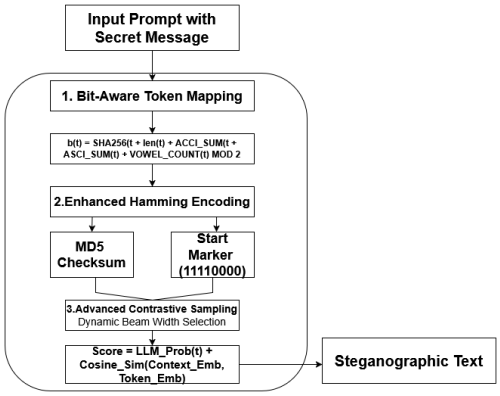


Fig. 3. Overview of the Proposed Optimized ACS-BTF Steganographic Encoding Pipeline.

In four stages, the encoding method of the Optimized ACS-BTF framework transforms the input prompt and secret message into a safe, undetectable steganographic text.Bit-aware token mapping initially assigns binary values to tokens based on lexical features (ASCII sums), cryptographic hashing (SHA256), and linguistic characteristics (vowel counts) in order to maintain anonymity. Second, to improve error correction for data integrity, Hamming coding is used with an MD5 checksum and unique start markers. Third, during token manufacturing, adaptive contrastive sampling uses entropy to dynamically modify the beam width in order to balance consistency and exploration. Lastly, token selection maximizes candidate tokens to guarantee naturalness and dependable decoding by integrating language model probabilities with semantic similarity measures (cosine similarity). The suggested op-

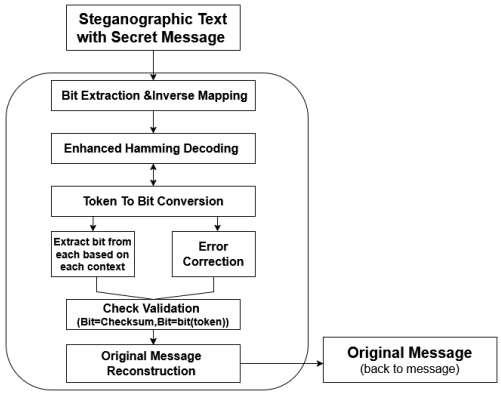


Fig. 4. Optimized ACS-BTF Decoding Pipeline for Steganographic Message Recovery

timized ACS-BTF framework's decoding flow. The process starts with the simultaneous insertion of steganographic text and a secret message. (1) To rebuild the embedded bitstream, the text is subjected to bit extraction and inverse token mapping. (2) Token-context tables are used to apply error correction and Hamming decoding to the extracted bits, allowing for dependable recovery with checksum validation.(3) The final message reconstruction converts the validated bitstream back into the original secret message to ensure accuracy and security in retrieval.

D. Token Mapping with Bit-Awareness

A multi-factor hash-based method is used in the dynamic binary token assignment to guarantee balanced bit distribution:

$$b(t) = (H(t) + 2 \cdot L(t) + 3 \cdot A(t) + 5 \cdot V(t) + 7 \cdot C(t)) \bmod 2 \quad (1)$$

where $L(t) = |t| \bmod 3$ indicates the length factor, $b(t) \in 0, 1\}$ is the bit allocated to token t, $H(t)$ is the SHA-256 hash value of token t, $A(t) = \left(\sum_{c \in t} \mathrm{ord}(c)\right) \bmod 3$ is the ASCII sum modulo 3, The vowel count modulo 3 is $V(t)$, whereas the consonant count modulo 3 is $C(t)$.To preserve embedding capacity balance, this formulation uses SentenceTransformer to integrate contextual frequencies and semantic embeddings.

E. Enhanced Hamming Encoding

The secret message undergoes error-correcting encoding through an enhanced Hamming code structure:

$$c = \bigoplus_{k=1}^{m+r} p_k \cdot d_k \tag{2}$$

with parity bits computed as

$$p_i = \bigoplus_{j:j\&(2^{i-1})\neq 0} d_j, \quad i = 1 \to r \tag{3}$$

where m is the initial binary data length, r fulfills $2^r \geq m + r + 1$, c is the encoded codeword, and $\bigoplus$ stands for the XOR operation. The start ('11110000') and end ('00001111') markers are used to frame the secret message after it has been preprocessed with MD5 checksum checking.

F. Advanced Contrastive Sampling

Dynamic beam width adjustment based on token distribution entropy enables adaptive candidate generation:

$$w = \min(W_{\max}, 3 + \lfloor E \cdot 2 \rfloor) \tag{4}$$

where the entropy E is calculated as:

$$E = -\sum_{i=1}^{V} p_i \log(p_i) \tag{5}$$

In this case, V is the vocabulary size, w is the dynamic beamwidth, p_i is a softmax probability from LLM logits, and $W_{\max}$ is the maximum beam width constraint (as mentioned above).This method preserves text naturalness while striking a balance between computational economy and exploratory scope.

G. Optimized Token Selection

Multi-objective token scoring combines probability and semantic coherence:

$$s = \alpha \cdot \log(p + \epsilon) + \beta \cdot \cos(\mathbf{e}_c, \mathbf{e}_t) \tag{6}$$

where p is the LLM token probability, α and β are the weighting parameters (usually 0.6 and 0.4), s is the total token score, $\epsilon = 10^{-10}$ is the logarithmic singularities avoided, and $\cos(\mathbf{e}_c, \mathbf{e}_t)$ is the cosine similarity between context embedding $\mathbf{e}_c$ and token embedding $\mathbf{e}_t$ calculated using SentenceTransformer.

H. Hamming Decoding and Error Correction

The decoding process employs syndrome calculation for error detection and correction:

$$e = \sum_{i=1}^{r} 2^{i-1} \cdot \left(\bigoplus_{j:j\&2^{i-1}\neq 0} c_j \right) \tag{7}$$

where c_j denotes the received codeword bits and e denotes the error position (syndrome). Bit inversion is used to rectify bit position $e - 1$ when $e > 0$,ensuring strength message recovery even in the presence of transmission noise.

I. Enhanced Accuracy Evaluation

The framework employs a weighted multi-metric accuracy assessment:

$$a = 0.4 \cdot a_c + 0.3 \cdot a_w + 0.2 \cdot a_l + 0.1 \cdot a_d \tag{8}$$

IV. EXPERIMENTAL ANALYSIS AND RESULTS

It has been shown that the optimized ACS-BTF framework is capable of reliably integrating information in text generated by huge language models. The research indicates that balancing processing load, embedding capacity, and decoding accuracy is still challenging.By enhancing algorithm parameters and adding more dependable mistake correction, this steganographic system might be able to adapt to various language model settings, further boosting dependability. Our evaluation focused on the efficiency and precision of recovering hidden data in scenarios that closely resemble the real-world application of language models. Ultimately, the findings indicate that additional work is needed to achieve reliable and consistent message concealing in a variety of applications, particularly in the form of improved error resilience and adaptive techniques.

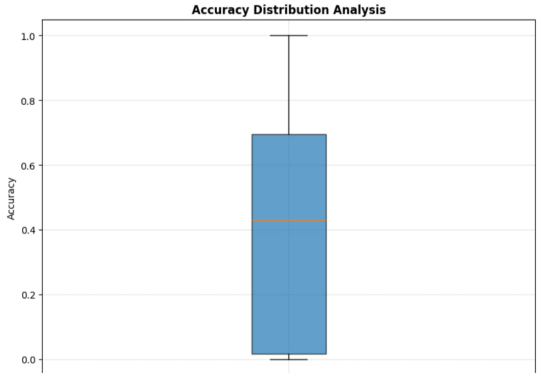


Fig. 5. Accuracy Distribution Analysis

A. Accuracy Distribution

The augmented ACS-BTF steganographic framework was subjected to a comprehensive evaluation process, whereby its effectiveness was determined through an analysis of its performance in managing a multitude of encoded messages and its precision in retrieving concealed information. The assessment was structured to simulate authentic language model development environments, emphasizing the system's ability to reliably decipher obscured text across diverse conditions.

This illustration elucidates the effectiveness of the refined ACS-BTF methodology utilized within the context of this thesis. Subsequent to the decoding and embedding

```
============================================================
ENHANCED WORKING EXAMPLE DEMONSTRATION
============================================================
Original: HelloWorld
Decoded: helloWorld
Accuracy: 0.9600
```

Fig. 6. Example of Decoding Performance for Enhanced Working Demonstration—Original, Decoded Output, and Accuracy Metric

processes, the initial text "HelloWorld" is reconstructed as "helloWorld" with an impressive accuracy metric of 0.96. This remarkable degree of precision underscores the robustness of the proposed methodology, which is augmented by advanced LLM-based encoding techniques and robust error-correcting frameworks. These findings substantiate that the system is capable of executing linguistic steganography within cutting-edge intelligent applications with both precision and security, all the while preserving the integrity and significance of the concealed messages.

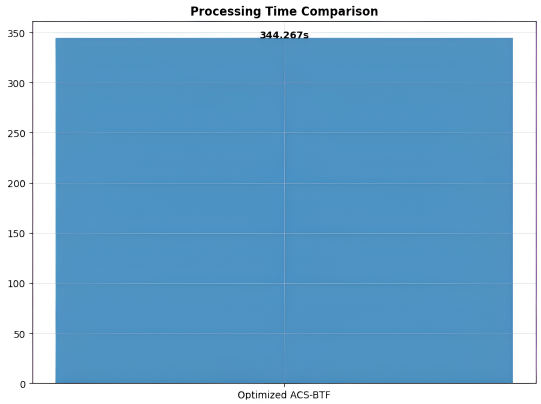


Fig. 7. Processing Time Comparison

B. Processing Time Comparison

A comprehensive analysis of the processing durations indicates that the mean processing interval for each encoding-decoding cycle is approximately 344 seconds. The bar graph delineates how the processing of lengthier or more intricate messages necessitates increased time, thereby significantly elevating processing expenditures. This underscores the considerable trade-off between attaining elevated security standards and the operational efficiency of the system. Through further enhancements, these performance limitations may potentially be mitigated, rendering the system suitable for real-time or high-throughput applications.

C. Capacity-Accuracy Trade-off

The quantity of data integrated (embedding capacity) and the precision of message decoding are evidently in a state of trade-off within the enhanced ACS-BTF system. The accuracy of decoding experiences a significant decline when the embedding exceeds approximately 1.2 bits per

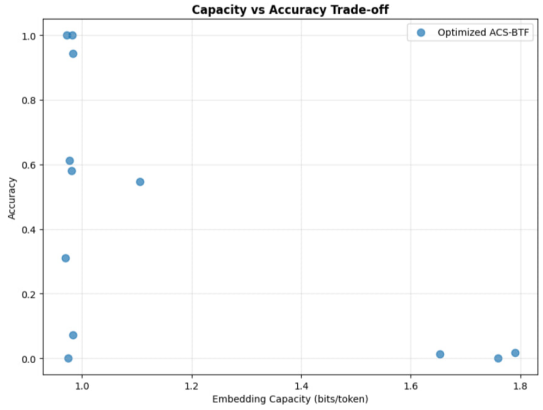


Fig. 8. Capacity-Accuracy Trade-off Chart

TABLE I
Performance Comparison of LLM-Based Steganographic Methods

Model / Paper	Approach / Key Features	Reported Accuracy
Near-imperceptible NLS [1]	Self-adjusting Arithmetic Coding for Neural Linguistic Steganography	0.94
Neural Linguistic Steganography [2]	Arithmetic Coding Integrated with LLM-based Text Generation	0.92
OD-Stega [18]	Optimized Token Distribution with KL-Divergence Constraint	0.91
ADLM-Stega [5]	Adaptive Token Selection Guided by Information Entropy	0.79
Position Allocation Watermark [6]	Multi-bit Watermark Embedding via Position Allocation Strategy	0.85
Topic-aware NLS (KG) [7]	Knowledge Graph–Guided Neural Linguistic Embedding	0.77
Optimized ACS-BTF (Ours)	Entropy-aware Contrastive Sampling with Robust Hamming-based Decoding	0.96

token. The equilibrium between robustness and security is achieved for maximal efficiency at 1.18 bits per token. This underscores the critical importance of entropy-aware token selection, contrastive sampling, and adaptive embedding methodologies in advancing steganographic efficacy utilizing extensive language models. These conclusions highlight the imperative for adaptive embedding frameworks that can dynamically modify their capacity to maintain both security and resilience.

V. Conclusion

This manuscript delineates the augmented ACS-BTF framework, an avant-garde steganographic approach meticulously crafted to proficiently embed data within text generated by extensive language models. Through the utilization of contrastive sampling and token filtration techniques, the proposed framework adeptly reconciles embedding capacity with semantic integrity, yielding remarkable decoding precision even in intricate linguistic

contexts. The system's notable accuracy metrics and resilience to perturbations in natural language generation are substantiated through rigorous experimental assessments. Nonetheless, this research elucidates considerable trade-offs among accuracy, embedding capacity, and processing efficiency that demand further innovations. Subsequent investigations will prioritize the refinement of error-correction techniques and the optimization of adaptive parameters to enhance robustness and facilitate application in real-time systems. In summary, by integrating cutting-edge generative modeling methodologies, the refined ACS-BTF framework elevates the standards for linguistic steganography and offers a viable pathway for confidential and secure communication within evolving AI-centric environments.

References

[1] J. Shen, H. Ji, and J. Han, "Near-imperceptible neural linguistic steganography via self-adjusting arithmetic coding," in Proc. Association for Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.22

[2] Z. M. Ziegler, Y. Deng, and A. M. Rush, "Neural linguistic steganography," 2019. [Online]. Available: https://doi.org/10.18653/v1/D19-1115

[3] Z. M. Ziegler, Y. Deng, and A. M. Rush, "Neural linguistic steganography," arXiv preprint, Sept. 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1909.01496

[4] B. Slat and C. Worp, "D-STEGA: LLM-based near-imperceptible steganography via optimized distributions," 2019. [Online]. Available: https://openreview.net/pdf?id=IQafqgqDzF

[5] Z. Qin, C. Sun, T. He, Y. He, A. Abdullah, N. Samian, and N. A. Roslan, "ADLM-stega: A universal adaptive token selection algorithm for improving steganographic text quality via information entropy," Oct. 2024. [Online]. Available: https://arxiv.org/pdf/2410.20825

[6] K. Yoo, W. Ahn, and N. Kwak, "Advancing beyond identification: Multi-bit watermark for large language models via position allocation," 2024. [Online]. Available: https://aclanthology.org/2024.naacl-long.224.pdf

[7] Y. Li, J. Zhang, Z. Yang, and R. Zhang, "Topic-aware neural linguistic steganography based on knowledge graphs," Apr. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3418598

[8] M. Bai, J. Yang, K. Pang, H. Wang, and Y. Huang, "Towards next-generation steganalysis: LLMs unleash the power of detecting steganography," 2024. [Online]. Available: https://doi.org/10.1007/978-981-99-0252-1_3

[9] D. Gligoroski and M. Raikwar, "An LLM framework for cryptography over chat channels," 2024. [Online]. Available: https://doi.org/10.1145/3709018.3736332

[10] C. Han and T. Xue, "Adaptive network steganography using deep learning and multimedia video analysis for enhanced security and fidelity," 2025. [Online]. Available: https://doi.org/10.1371/journal.pone.0318795

[11] S. Kumar, A. Singh, and M. Kumar, "Information hiding with adaptive steganography based on novel fuzzy edge identification," 2018. [Online]. Available: https://doi.org/10.1016/j.dt.2018.08.003

[12] V. Sabeti, M. Sobhani, and S. M. H. Hasheminejad, "An adaptive image steganography method based on integer wavelet transform using genetic algorithm," 2022. [Online]. Available: https://doi.org/10.1016/j.compeleceng.2022.107809

[13] R. J. Anderson and F. A. P. Petitcolas, "On the limits of steganography," IEEE J. Sel. Areas Commun., vol. 16, no. 4, pp. 474–481, 1998.

[14] C.-Y. Chang and S. Clark, "Linguistic steganography using automatically generated paraphrases," in Proc. Human Language Technologies, 2010, pp. 591–599.

[15] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, Digital Watermarking and Steganography. Morgan Kaufmann, 2007.

[16] F. Z. Dai and Z. Cai, "Towards near-imperceptible steganographic text," in Proc. ACL, 2019, pp. 4303–4308. [Online]. Available: https://aclanthology.org/P19-1418/

[17] T. Fang, M. Jaggi, and K. Argyraki, "Generating steganographic text with LSTMs," arXiv preprint arXiv:1705.10742, 2017. [Online]. Available: https://arxiv.org/abs/1705.10742

[18] Y.-S. Huang, P. Just, K. Narayanan, and C. Tian, "OD-Stega: LLM-based near-imperceptible steganography via optimized distributions," under review, 2025. [Online]. Available: https://openreview.net/pdf?id=IQafqgqDzF

[19] Y.-S. Huang, P. Just, K. Narayanan, and C. Tian, "OD-Stega: LLM-based near-imperceptible steganography via optimized distributions," under review, 2025. [Online]. Available: https://openreview.net/pdf?id=IQafqgqDzF

[20] A. Banu, S. Anam, and K. Deb, "Removing rain from single image using Atrous U-Net and GAN," in Proc. 6th Int. Conf. Electrical Engineering and Information & Communication Technology (ICEEICT), Dhaka, Bangladesh, 2024, pp. 1298–1303. [Online]. Available: https://doi.org/10.1109/ICEEICT62016.2024.10534531

A Comprehensive Review of Artificial Intelligence Possibilities in Photovoltaic Systems: The Bangladesh Perspective

Shahriar Tamim
Dept. of EEE
Barishal Engineering College
Barishal-8200, Bangladesh
tamimshahriar2992@gmail.com

Abu Saleh Mohammad Nasrullah
Dept. of EEE
Barishal Engineering College
Barishal-8200, Bangladesh
abusalehmdnasrullah@gmail.com

Abdullah Al Mamun
Dept. of EEE
Barishal Engineering College
Barishal-8200, Bangladesh
r50mamun@gmail.com

Md. Rana Islam
Dept. of EEE
Barishal Engineering College
Barishal-8200, Bangladesh line 5:
ranaislam1408@gmail.com

Md. Nabil Shahriar
Dept. of EEE
Barishal Engineering College
Barishal-8200, Bangladesh
shahriarmdnabil@gmail.com

Shuva Chandra Sarker
Dept. of EEE
Barishal Engineering College
Barishal-8200, Bangladesh
shouvosarker200332@gmail.com

Md. Shariful Islam
Dept. of EEE
Barishal Engineering College
Barishal-8200, Bangladesh
mdsharifulislam.bec@gmail.com

***Abstract*—The increased global pressures on clean and renewable energy sources have increased the application of photovoltaic (PV) systems, especially when considering the application in emerging markets like Bangladesh. However, the effective performance, stability, and maintenance of PV installations remain a significant challenge that could be explained by environmental, technical, and operational aspects. The current paper provides an extensive review of the Artificial Intelligence (AI) approaches to the process of optimizing the work of PV systems. The research overview analyzes the recent developments in the AI-based methods, namely Artificial Neural Networks (ANN), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and Adaptive Neuro-Fuzzy Inference Systems (ANFIS), concerning the key areas, namely fault detection, cooling, sun-tracking, and the Maximum Power Point Tracking (MPPT). The comparative studies based on the available literature evidence that AI-based strategies are significantly superior to conventional solutions, as they provide greater accuracy, shorten convergence rates, and increase the energy output. Finally, the paper outlines the existing limitations and suggests future research directions of an intelligent, cost-effective AI solution that will enhance solar energy sustainability in Bangladesh.**

Keywords— Artificial Intelligence (AI); Photovoltaic (PV) Systems; Fault Detection; Maximum Power Point Tracking (MPPT); Sun Tracking; Cooling Technologies; Machine Learning (ML); Bangladesh.

I. INTRODUCTION

As a densely populated South Asian nation, Bangladesh's solar power has become an attractive alternative to electricity due to growing environmental concerns and dependency on fossil fuels, particularly natural gas. Photovoltaic technology has evolved significantly as a key renewable energy source[1]. Solar energy is a sustainable, clean, and leading source of renewable energy[2]. With clear sky irradiance assumptions and available land area, the maximum potential annual solar energy amount reaching the earth's surface before energy conversion was estimated to be 50 [3]. Solar cells convert light into electrical energy using the photovoltaic effect and can be arranged into arrays, a power station, or smaller panels or homes. Looking at Bangladesh as a whole, it has an average theoretical solar potential of around 4.59 GHI, which puts it around the middle in comparison to other countries [4]. According to The International Energy Agency (IEA) (2022), solar photovoltaic (PV) power constituted 52.2% of all electricity produced by non-combustible renewable sources in the country, becoming the leading renewable [5]. In Bangladesh, the average sunshine hours are 6.69, 6.16, and 4.81in winter, summer, and monsoon [6].

Although the global research on the optimization of AI-based photovoltaic systems presents sufficient evidence, the conditions in the local climate, infrastructure, and economy have the ultimate deciding effect on the application of these technologies. Among the specific issues facing the developing countries, especially Bangladesh, are high ambient temperatures, frequent cloud cover during the monsoon seasons, lack of grid flexibility, and data acquisition constraints. The adaptable and economical AI solutions are therefore mandatory. This review thus explores the AI-based photovoltaic optimization process in Bangladesh and the quest to achieve energy sustainability.

The current review provides country-focused insights, such as a systematic review and ranking of AI-based methods used in photovoltaic systems in fault detection, cooling, sun-tracking, and maximum power point tracking (MPPT), reformed in light of the Bangladeshi operating conditions. Through the correlation of the model performance with the local climatic variables, which include high temperature and humidity rates, airborne dust availability, and variability of sunlight on monsoon winds, together with infrastructural restraints, the research determines the artificial intelligence

strategies that are not only technologically effective but also politically deployable in the Bangladeshi photovoltaic scene.

Since the national background has been established and the importance of photovoltaic implementation has been pointed out in Bangladesh, the next section evaluates the environmental and physical factors that play a critical role in the functioning of the photovoltaic modules.

II. The Environmental Factors Affecting Solar PV Modules

A. Defect on Solar Panel

The PV modules can be damaged during the transportation, installation, and operation processes. Potential-induced degradation (PID) can directly cause significant power losses (greater than30%) to silicon PV modules [7]. It is necessary to find the impact of each micro-crack shape on the performance of the PV module, to optimize the power generation from the PV project and reduce the cost of operation and maintenance later [8]. New panels can have µ-crack, but their influence is neglected; the problem appears when panels are exposed to several conditions, such as weather conditions, tracking, cleaning, etc. The impact of those cracks depends on their orientation and the shape that is created by the cause type; see Fig. 1 [9].

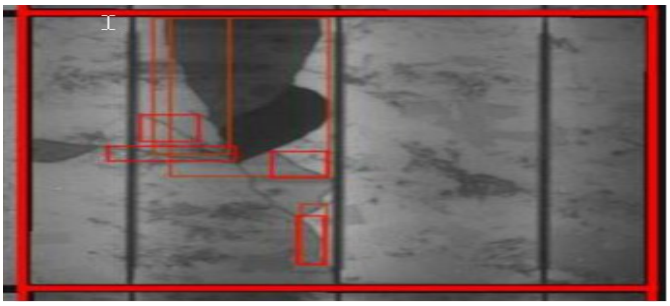


Fig. 1 Micro-cracks influence

Fig. 1 shows an example of µ-cracks that is classified as major, since the damage reaches 20% of the module. Moreover, µ cracks can affect the power drop up to 3.2%; this percentage affects the overall produced power from economic and energy yield for large PV projects [10]. Papargyri, Lamprini et al. [11] discussed the origins and factors that affect the cell cracks. It has been shown that µ-cracks classifications, based on their characteristics, have been used to determine both the electrical and mechanical degradation of PV modules. A research study presented by Dhimish, Mahmoud et al. [12], where the impact of photovoltaic (PV) µcrack was evaluated, through analyzing 4000 polycrystalline silicon solar cells. The main observations are that the output power loss ranged between 0.9% and 42.8% due to µ-cracks, an increase of power loss by 1.7% resulted from cracks in solar cell fingers (due to reducing finger width), and a radical correlation between PV hot spots and the presence of µ-crack, where the minimum increase in the cell temperature is found at 7.6°C. Electroluminescence (EL) measurements were performed for possible faults in the examined PV modules. The approach classified only 60% of cracks that significantly impacted the total amount of power generated by PV modules [13].

B. Overheating

High temperatures significantly impact the efficiency of PV systems. The operational efficiency of PV modules depends closely on their thermal conditions. Rising operating temperatures accelerate performance deterioration and significantly decrease PV module efficiency [14]. However, extreme temperatures, often surpassing 50 ◦C, reduce photovoltaic (PV) system efficiency by up to 25% [15]. The study by [16] on the solar spectrum is necessary to examine the effect of PV panel temperature on its output. Generally, PV module conversion efficiency lies in the range of 12–15%, and the temperature at the surface is reported to rise up to 70 °C in normal sunny day, which can cause damage to cells in the PV modules and a reduction in their life expectancy and drop in their conversion efficiency.

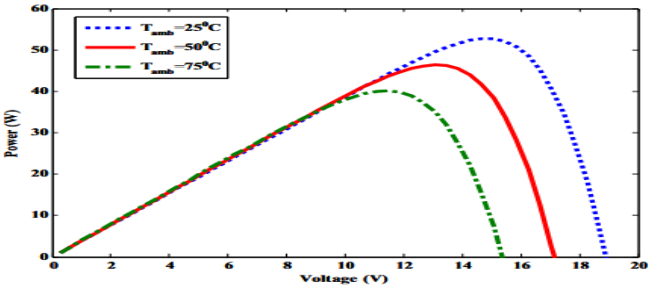


Fig. 2 P-V curve for various temperatures

Fig. 2 provides an insight into the correlation of output power and voltage at various operating temperatures, keeping solar insolation constant, where a sharp reduction in PV power with the temperature rise is witnessed. A fall in PV module output power with a rise in temperature is observed on the non-removal of this excessive heat. For cooling such a system, cooling technology is appropriate for its integration adjacent to the PV module [16].

C. Sun Tracking Effect on PV

Bangladesh has high potential of applying solar irradiation to generate electricity. The country absorbs average solar radiation of 4.0 to 6.5 kWh/m2 per day, which is capable of producing 1018×10^{18} J of energy.

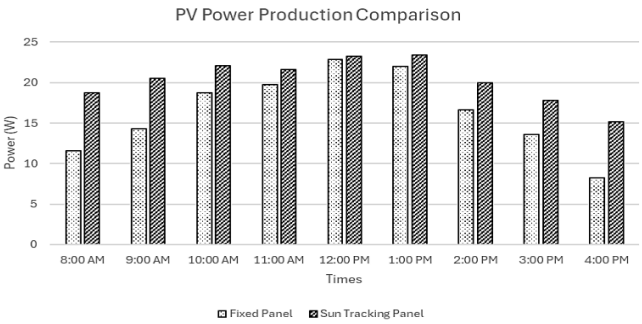


Fig. 3 Power Output of Dual-axis Vs Fixed-flat PV System

Fig. 3 shows the power output of a dual-axis vs a fixed flat PV system throughout the day. At 8:00 AM, the fixed panel generates 11.61 W, whereas the sun-tracking panel generates 18.74 W, yielding a substantial improvement of 61.41%. The most significant improvement is noted at 4:00 PM, where the sun-tracking panel achieves an 84.9% increase in power output compared to the fixed panel [17].

D. Maximum Power Point Tracking on PV Module

Solar photovoltaic (PV) cells are used to convert solar energy into regulated electrical energy using a power electronics converter. These solar PV cells exhibit nonlinear characteristics and very low efficiency [18]. Irradiance and

conversion efficiency, which change depending on the surroundings, control solar PV output. [19]. Due to the fact that solar panels do not capture all of the sunlight, most PV panels have an efficiency range of 15–18 %[20]. The MPPT controller's principal function is to maximize & optimize power delivery, resulting in increased power output and improved cost efficiency per watt [20]. This technology is essential in mitigating the impact of the low conversion efficiency of solar panels and can increase the system's overall efficiency by up to 30 %. The literature is rich with various MPPT techniques varying in implementation complexity, cost, adjustability to different PV systems, and overall produced efficiency [21], [22], [23], [24]. In contrast, AI-based intelligent techniques, including Artificial Neural Network (ANN), Artificial Neural Fuzzy Interference System (ANFIS), Fuzzy Logic Control (FLC), Particle Swam Optimization (PSO), and Artificial Neural Network and Particle Swam Optimization (ANN-PSO), outperform achieving higher accuracies of 97.8%, 99.9%, 98.9%, 99.2%, and 99%, respectively. Compared to available research, which often reports lower accuracies for conventional techniques, our study highlights the enhanced performance of AI-based methods. The power extracted and the accuracy are shown in Fig. 4 [25].

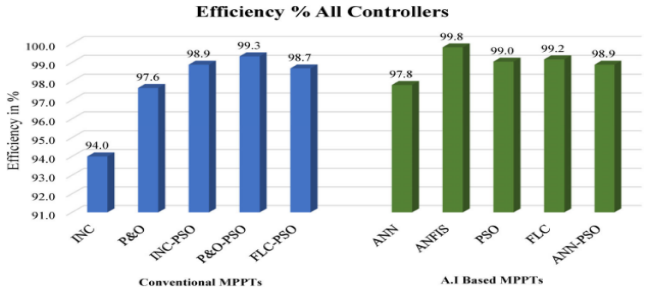


Fig. 4 Efficiency Exhibited by Conventional and AI-based MPPTs

Even though environmental and operating conditions later deter the performance of the photovoltaic system, artificial intelligence provides advanced analytical and predictive systems that overcome such impacts. The section outlines various models of AI that can be used to enhance the performance of photovoltaic systems in terms of fault detection, cooling, tracking algorithms, and power optimization.

III. AI MODELS TO ENHANCE THE EFFICIENCY OF PV SYSTEM

This study focused on research highlighting AI-driven methodologies, including ML and DL. To ensure the review reflected current advancements, we selected papers published in the last five years. Papers were sourced from reputable academic databases, with a focus on peer-reviewed articles, high citation counts, and journals with strong impact factors.

A. Detection of Defects and Faults in Photovoltaic Systems

Decision Tree (DT): An advanced technique for monitoring the condition of PV panels and diagnosing faults was developed in [26] using a combination of a U-Net neural network and a decision-tree classifier. The study focused on detecting three specific types of defect: safety-glass cracks, defects in PV power units, and contamination of the safety glass. The decision-tree classifier, known for its rule-based classification, demonstrated higher accuracy compared to KNN and SVM. The study was conducted using a dataset of 295 samples collected over a short period of time. The combined use of the trained U-Net neural network and the decision-tree classifier proved highly effective, achieving an impressive fault detection accuracy of 99.8%.

Logistic Regression (LR): The author presented in [27] an innovative algorithm that leverages logistic regression combined with cross-validation techniques. This approach is specifically designed for the early detection and precise identification of faults within the DC components of solar PV modules at string levels. Impressively, the algorithm achieves a 97.11% accuracy rate in identifying the three primary categories of faults: open-circuit faults, short-circuit faults, both permanent and temporary mismatch faults, and undefined faults. This work was done by simulating the solar PV cell circuit using a dataset that has temperature and irradiance values ranging from 1.32 to 35.06 ◦C and 0.04 to 984.84 W/m2, and then looking at the I-V curves of the solar PV cell to find out what kind of faults there are. The multi-class classification was done using either one-vs-one (OVO) or one-vs-rest (OVR) methods.

Dual Convolution Neural Network (Dual-CNN): In [28], the study proposes a dual convolutional neural network (Dual-CNN) architecture to enhance micro-crack detection in electroluminescence (EL) PV images. By integrating shallow feature extraction with deep semantic analysis, the proposed model effectively captures both fine-grained local textures and high-level structural patterns. They used an open-source dataset that contains 2624 samples of 300*300 8-bit grayscale EL images of functional and defective solar cells with varying degrees of degradation extracted from 44 different solar modules. The proposed dual CNN classification model shows improvements in identifying true positives and reducing false positives compared to the single-stream deep CNN model. Experimental evaluations on an EL image dataset demonstrate that the Dual-CNN approach significantly improves defect localization and classification with an accuracy of 85.33%, a recall of 71.71%, and an F1 score of 73.9%, paving the way for more robust and automated PV inspection systems in the solar energy sector.

Random Forest (RF): The random forest classifier was combined with a modified grey-wolf optimization (MGWO) algorithm in [29] to enhance fault detection in solar PV systems. The methodology involved two key steps: first, extracting the five critical parameters of the one diode model by transforming three arbitrary I-V curves into a reference curve using a current–voltage translation method and MGWO algorithm; second, simulating the PV array to determine MPP coordinates and build operational databases through co-simulations in PSIM/MATLAB. The RFC was then used to detect anomalies, achieving an impressive accuracy rate of 99.4%.

Hybrid tree-based and deep-learning models tend to achieve the highest level of fault-detection accuracy, in particular, the U-Net with a decision tree classifier. Such lightweight and accurate models would be very advantageous in Bangladesh PV systems, which are dealing with dust, humidity, and ineffective maintenance infrastructure.

Table 1: Ranking of AI techniques for PV fault detection

Rank	AI Method	Efficiency (%)
1	DT	99.8
2	RF	99.4
3	LR	97.11
4	Dual-CNN	85.33

B. AI Models for Cooling Technologies in PV Systems

Decision Trees (DT): In article [30], DT represents a set of conditions, which are hierarchically organized and successively applied from root to leaf of the tree. DTs produce a trained model that can represent logical rules, which can then be used to predict new datasets through the repetitive process of DT. DT is the most computationally efficient method as it requires significantly less training time. It is also important to mention that for tree-based algorithms, the accuracy on the training dataset decreases with an increase in the training dataset. DT has lower training and testing performances as compared to ET and RF.

Active & Passive cooling techniques: [31] In this study, two new and cost-effective cooling techniques are proposed, which can improve efficiency and increase the life of the PV panel. The first technique is the active cooling design with water. The proposed design enhances access to PV cells to maximize cooling. The second one is an improved passive cooling technique that uses fins mounted on the backside of the PV module. The performance of the SPVS system is improved through two different proposed cooling techniques while addressing the parameters of flow rate and amount of heat dissipation. Three identical PV modules are used. Every PV module is connected separately with a micro-inverter and further connected to a 220 V AC 50 Hz grid. The performance ratios for water-cooling, fins, and no cooling techniques are 84%, 81%, and 77%, respectively. The performance evaluation using thermal image showed the temperature distribution for the cooling module, which is as low as 25 ◦C, and in comparison, with the reference module, which had a different temperature distribution from 37 to 54 ◦C, where the ambient temperature was 32 ◦C.

Table 2: Ranking of AI Cooling Methods for PV Systems

Rank	AI method	Performance Ratio
1	Active Cooling (Water-based)	84% performance ratio
2	Passive Cooling (Fins)	81% performance ratio
3	Decision Tree (DT)	Lower PR

The performance improvement is significant in active, water-based cooling schemes, and passive, fin-based cooling systems provide a reasonably priced option. Given the high ambient temperature in Bangladesh, AI-based passive or hybrid cooling technologies are a viable solution to increase the photovoltaic rate and life-cycle lifetime of the systems.

C. AI Models for Sun Tracking Technologies in PV Systems

Boosted Decision Tree Regression (BDTR): In this study [32], one of the most recent AI algorithms, namely, boosted decision tree regression (BDTR) model, was applied to predict the changes in solar radiation based on collected data. The proposed model was then compared with other conventional regression algorithms, such as linear regression and neural networks. Sensitivity analysis and uncertainty analysis were introduced to validate the accuracy of the proposed model. This study will be using historical solar data only to predict solar radiation at any desired time possible. The data used in this study were measured hourly from 7 a.m. to 6 p.m. Therefore, the selected months of data are used. Two data partitions were attempted in this study at 80% and 75% training to compare which data partition, relative to the machine learning algorithm, performs better. Boosted regression tree (BRT) models are a combination of two techniques, which are decision tree algorithms and boosting methods. Two different data partitions were applied for better evaluation of the methods. By comparing both data splits, BDTR outperformed the other models, and with a 75–25% data split, has a higher R2 compared with 80–20%. Four prediction models were studied, composed of a neural network Gaussian normalizer, a neural network binning normalizer, boosted decision tree regression, and linear regression. By using two different data splits, boosted decision tree regression outperformed all the other models with R2 of 0.89125 and 0.90183, respectively, without implementing a tuned model hyperparameter module. The performance of boosted decision tree regression somehow decreased to 0.86691 and 0.88277 of R2 for each 80–20% and 75–25% data splitting. This implies that data splitting of 75–25% gives better performance toward boosted decision tree regression by omitting the implementation of the tune model hyperparameter module. The proposed model may be applied in other areas for solar radiation prediction. The results revealed that BDTR outperformed other algorithms with a high level of accuracy.

Artificial Neural Network (ANN): [33] In this study, the prediction models employed Artificial Neural Networks (ANN) with five different prediction classifiers, namely, random forest, forest tree, multilayer perceptron (MLP), BPF regression, and linear regression, to predict the performance of the sun-tracking PV system using experimental data. The system parameters and meteorological data are collected each minute and manipulated on an average hourly basis for further analysis. The collected data is divided into two sets for prediction purposes; 80% for training and 20% for testing. The performance metrics were carefully chosen for this kind of dataset and this kind of prediction. the average performance metrics acquired from the five classifiers, with 80% data for training and validation and 20% for testing. The study shows that the best prediction based on the accuracy criteria discussed before (in the relevant equations) is achieved using the random forest classifier with a correlation coefficient of 99.79% and a relative minimum absolute error among all the proposed classifiers with 2.36%. For validation purposes, cross-validation with 10-fold results is presented in this study. Results assure that the best prediction classifier is the random forest. The current paper aims to present a prediction of the output power of a sun-tracking PV system utilizing various prediction ANN classifiers. The best prediction classifier is found to be a forest tree classifier with an R2 value of 99.79% and a minimum absolute relative error of 2.36%. Moreover, the least accurate prediction classifier is found to be the linear regression with an R2 of 95.27% and an absolute relative error of 25.71%.

Table 3: Ranking of AI Sun Tracking Methods for PV Systems

Rank	AI Method	R2 value
1	MOGA ANFIS	99.8%
2	ANN	99.79%
3	BDTR	75%

Multi-Objective Genetic Algorithm Adaptive Neuro-fuzzy Inference System (MOGA ANFIS): [34] The study compares theoretical and experimental energy generation data from the AI-based solar tracker against conventional systems, thoroughly evaluating its effectiveness. Results demonstrate that the AI-based system, guided by MOGA ANFIS. An

Adaptive Neuro-Fuzzy Inference System (ANFIS) is an artificial intelligence system that combines fuzzy logic and neural networks to model complex relationships between inputs and outputs. ANFIS has emerged as a promising approach among many control algorithms, especially due to its unique features. The objectives are to adjust the solar panel's tilt and azimuth angles to capture the maximum solar energy throughout the year. The tilt angle ranges from 0° (horizontal) to 90° (vertical), and the azimuth angle ranges from 0° to 360°. Both training (R2 = 0.998) and test (R2 = 0.998) data exhibit very high accuracy, with residuals closely clustered around zero. This indicates that the model is highly effective in predicting altitude angles. For real-world implementation, factors such as installation costs, computational resource requirements, and integration with existing energy infrastructure must be considered. While the MOGA-ANFIS model introduces a computational overhead compared to traditional tracking systems, advancements in edge computing and IoT-enabled solar controllers can mitigate these concerns, enabling real-time adaptive control without excessive energy consumption. The outperformance of traditional systems across different weather conditions, with R2 values of 0.989 for azimuth and 0.998 for altitude angle predictions, highlights its superior adaptability and efficiency. Experimentally, the proposed system can generate 27 % and 20 % more energy than the open and closed loop systems, respectively, and has the potential to generate 57 % more energy than fixed panels.

MOGA-ANFIS is the most accurate in predictive ability compared to other AI-based solar-tracking approaches, closely followed by ANN-based models. ANN or tree-based trackers that are more rudimentary are used in a geographically or topographically limited area like Bangladesh, which offers a practical trade between performance and cost of implementation.

D. AI Models for MPPT in PV Systems

Decision Tree (DT): [35] This work presents a machine-learning (ML) algorithm for maximum power point tracking (MPPT) of an isolated photovoltaic (PV) system. A strategy based on the decision-tree (DT) regression ML algorithm is proposed in this work to determine the MPP of a PV system. The proposed approach consists of two steps and involves building the RT ML model using the data obtained based on the PV panel parameters and using the prepared model for MPPT. The predicted values are used to determine the duty cycle (D) to drive the PV panel at MPP. From the proposed algorithm, efficiency has been improved by more than 93.93% in the steady state despite erratic irradiance and temperatures.

Support Vector Machine Regression (SVMR): [36] In this work, a new SVMR machine learning-based approach for MPPT of the solar panel is used in association with a PWM control boost converter. There are two phases in the proposed strategy. Obtaining data from the PV module specifications and creating the SVMR model comes in the primary phase. The secondary goal is to employ the prepared SVMR model for MPPT. The SVM algorithm predicts the PV panel's maximum power and relevant voltage for specific irradiation and temperature. The duty cycle of the boost converter corresponding to the maximum power was evaluated using the predicted values. The result of the simulation shows that the proposed control strategy forces the solar panel to work near the predicted MPP. The SVM regression control strategy gives the MPP tracking efficiency of more than 94% for the solar PV system despite variable climatic conditions during its stable state operation.

Table 4: Ranking of AI Methods of MPPT for PV System

Rank	AI Method	Efficiency
1	ANN	98%
2	SVMR	More than 94%
3	DT	More than 93.93%

Artificial Neural Network (ANN): [37] In this article, a systematic review of state-of-the-art papers from the last six years focused solely on Artificial Neural Networks (ANN) for MPPT control in PV systems. Several intelligent algorithms show acceptable performance; however, few consider using an ANN. These have the advantage of giving a fast and accurate tracking of the MPP. The controller's effectiveness depends on the algorithm used in the hidden layer and how well the neural network has been trained. Articles over the last six years were studied. A review of different papers, reports, and other documents using ANN for MPPT control is presented. The algorithms are based on ANN or on a hybrid combination with FL or a metaheuristic algorithm. ANN MPPT algorithms deliver an average performance of 98% in uniform conditions, exhibit a faster convergence speed, and have fewer oscillations around the MPP, according to this research.

The MPPT techniques using ANNs are always associated with better tracking efficiency in comparison with SVM and decision-tree solutions. These methods are particularly beneficial in Bangladesh, where the irradiance is characterized by strong variability due to the frequent presence of cloud cover and monsoons.

IV. Conclusion

The current review demonstrates an extensive literature review on the recent development of Artificial Intelligence (AI) techniques to enhance the performance, effectiveness, and viability of photovoltaic (PV) systems, specifically in Bangladesh. An extremely broad range of AI models, such as ANN, FLC, DT, RF, SVMR, and ANFIS, have been demonstrated to have high effectiveness in addressing the key problems of PV, such as fault detection, cooling optimization, sun-tracking, and Maximum Power Point Tracking (MPPT). The results show that AI-based methods are more accurate and convergent than conventional algorithms and yield more energy, thus making them the most appropriate to be used in large-scale solar integration in Bangladesh, which is experiencing a growth in renewable power generation. However, significant obstacles remain, one of them being the lack of data, the high cost of implementation, and limited connectivity between AI methods and Internet of Things (IoT)-based monitoring systems. In the future, the research should focus on creating lightweight and real-time AI models capable of operating in a variety of climatic conditions, like edge computing and hybrid optimization functions. Enhancement of national research facilities and development of an industry-academia partnership will be necessary to bring the vision of a sustainable and smart solar energy infrastructure in Bangladesh to reality.

References

[1] K. Ranabhat, L. Patrikeev, A. A. Revina, K. Andrianov, V. Lapshinsky, and E. Sofronova, "An introduction to solar cell technology," *J. Appl. Eng. Sci.*, vol. 14, no. 4, 2016, Accessed: Nov.

20, 2025. [Online]. Available: https://aseestant.ceon.rs/index.php/jaes/article/view/10879

[2] A. S. Mohammad Nasrullah *et al.*, "A Comprehensive Performance Analysis of a 25 kW Rooftop Photovoltaic System at Barishal Engineering College Using PVSYST," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Feb. 2025, pp. 1–6. doi: 10.1109/ECCE64574.2025.11013285.

[3] C. S. Lai, Y. Jia, L. L. Lai, Z. Xu, M. D. McCulloch, and K. P. Wong, "A comprehensive review on large-scale photovoltaic system with applications of electrical energy storage," *Renew. Sustain. Energy Rev.*, vol. 78, pp. 439–451, 2017.

[4] "Global Solar Report - PVknowhow.com." Accessed: Oct. 23, 2025. [Online]. Available: https://www.pvknowhow.com/global-solar-report/

[5] OECD and United Nations Conference on Trade and Development, *Production Transformation Policy Review of Bangladesh: Investing in the Future of a Trading Nation*. in OECD Development Pathways. OECD, 2023. doi: 10.1787/8b925b06-en.

[6] A. Islam and A. K. Mozumder, "Forced convection heat transfer performance of an internally finned tube," *J. Mech. Eng.*, vol. 40, no. 1, pp. 54–62, 2009.

[7] W. Luo *et al.*, "Potential-induced degradation in photovoltaic modules: a critical review," *Energy Environ. Sci.*, vol. 10, no. 1, pp. 43–68, 2017.

[8] M. Bdour, Z. Dalala, M. Al-Addous, A. Radaideh, and A. Al-Sadi, "A comprehensive evaluation on types of microcracks and possible effects on power degradation in photovoltaic solar panels," *Sustainability*, vol. 12, no. 16, p. 6416, 2020.

[9] S. Kajari-Schröder, I. Kunze, U. Eitner, and M. Köntges, "Spatial and orientational distribution of cracks in crystalline photovoltaic modules generated by mechanical load tests," *Sol. Energy Mater. Sol. Cells*, vol. 95, no. 11, pp. 3054–3059, 2011.

[10] M. Bdour and A. Al-Sadi, "Analysis of different microcracks shapes and the effect of each shape on performance of PV modules," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, 2020, p. 012005. Accessed: Oct. 24, 2025. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/876/1/012005/meta

[11] L. Papargyri *et al.*, "Modelling and experimental investigations of microcracks in crystalline silicon photovoltaics: A review," *Renew. Energy*, vol. 145, pp. 2387–2408, 2020.

[12] M. Dhimish, "Micro cracks distribution and power degradation of polycrystalline solar cells wafer: Observations constructed from the analysis of 4000 samples," *Renew. Energy*, vol. 145, pp. 466–477, 2020.

[13] M. Dhimish, V. Holmes, B. Mehrdadi, and M. Dales, "The impact of cracks on photovoltaic power performance," *J. Sci. Adv. Mater. Devices*, vol. 2, no. 2, pp. 199–209, 2017.

[14] T. Shimizu, M. Hirakata, T. Kamezawa, and H. Watanabe, "Generation control circuit for photovoltaic modules," *IEEE Trans. Power Electron.*, vol. 16, no. 3, pp. 293–300, 2001.

[15] A. A. Abdirahman, M. Asif, E. Cuce, I. Ahmed, M. Alqahtani, and M. Khalid, "Optimizing PV systems in high-temperature environments: A SWOT-based analysis of cooling technologies," *Energy Strategy Rev.*, vol. 61, p. 101828, 2025.

[16] A. Kane, V. Verma, and B. Singh, "Optimization of thermoelectric cooling technology for an active cooling of photovoltaic panel," *Renew. Sustain. Energy Rev.*, vol. 75, pp. 1295–1305, 2017.

[17] A. B. Sproul, "Derivation of the solar geometric relationships using vector analysis," *Renew. Energy*, vol. 32, no. 7, pp. 1187–1205, 2007.

[18] A. Haque, "Maximum Power Point Tracking (MPPT) Scheme for Solar Photovoltaic System," *Energy Technol. Policy*, vol. 1, no. 1, pp. 115–122, Jan. 2014, doi: 10.1080/23317000.2014.979379.

[19] Y. Halali *et al.*, "A Dandelion Optimization-based MPPT Control Technique for Solar PV System under Partial Shading Conditions," *Sci. Afr.*, p. e02938, 2025.

[20] S. Manna, D. K. Singh, M. H. Alsharif, and M.-K. Kim, "Enhanced MPPT approach for grid-integrated solar PV system: Simulation and experimental study," *Energy Rep.*, vol. 12, pp. 3323–3340, 2024.

[21] A. F. Mirza, M. Mansoor, Q. Ling, B. Yin, and M. Y. Javed, "A Salp-Swarm Optimization based MPPT technique for harvesting maximum energy from PV systems under partial shading conditions," *Energy Convers. Manag.*, vol. 209, p. 112625, 2020.

[22] D. K. Singh, A. K. Akella, and S. Manna, "Adjustable variable step-based MRAC MPPT for solar PV system in highly fluctuating and cloudy atmospheric conditions," *Electr. Eng.*, vol. 105, no. 6, pp. 3751–3772, Dec. 2023, doi: 10.1007/s00202-023-01922-3.

[23] M. H. Alsharif, "Comparative analysis of solar-powered base stations for green mobile networks," *Energies*, vol. 10, no. 8, p. 1208, 2017.

[24] D. Pilakkat and S. Kanthalakshmi, "Single phase PV system operating under Partially Shaded Conditions with ABC-PO as MPPT algorithm for grid connected applications," *Energy Rep.*, vol. 6, pp. 1910–1921, 2020.

[25] M. Khan, M. A. Raza, M. Faheem, S. A. Sarang, M. Panhwar, and T. A. Jumani, "Conventional and artificial intelligence based maximum power point tracking techniques for efficient solar power generation," *Eng. Rep.*, vol. 6, no. 12, p. e12963, Dec. 2024, doi: 10.1002/eng2.12963.

[26] X. Wang, W. Yang, B. Qin, K. Wei, Y. Ma, and D. Zhang, "Intelligent monitoring of photovoltaic panels based on infrared detection," *Energy Rep.*, vol. 8, pp. 5005–5015, 2022.

[27] S. Voutsinas, D. Karolidis, I. Voyiatzis, and M. Samarakou, "Development of a machine-learning-based method for early fault detection in photovoltaic systems," *J. Eng. Appl. Sci.*, vol. 70, no. 1, p. 27, Dec. 2023, doi: 10.1186/s44147-023-00200-0.

[28] S. Nazir, N. Bakir, and N. Khodor, "Dual Cnn for Photovoltaic Electroluminescence Images Microcrack Detection", Accessed: Oct. 24, 2025. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5015797

[29] A. F. Amiri, H. Oudira, A. Chouder, and S. Kichou, "Faults detection and diagnosis of PV systems based on machine learning approach using random forest classifier," *Energy Convers. Manag.*, vol. 301, p. 118076, 2024.

[30] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees," *J. Clean. Prod.*, vol. 203, pp. 810–821, 2018.

[31] W. Hammad, T. O. Sweidan, M. I. Abuashour, H. M. Khalid, and S. M. Muyeen, "Thermal management of grid-tied PV system: A novel active and passive cooling design-based approach," *IET Renew. Power Gener.*, vol. 15, no. 12, pp. 2715–2725, Sept. 2021, doi: 10.1049/rpg2.12197.

[32] E. Jumin, F. B. Basaruddin, Y. Bte. M. Yusoff, S. D. Latif, and A. N. Ahmed, "Solar radiation prediction using boosted decision tree regression model: A case study in Malaysia," *Environ. Sci. Pollut. Res.*, vol. 28, no. 21, pp. 26571–26583, June 2021, doi: 10.1007/s11356-021-12435-6.

[33] J. Alghazo, G. Latif, B. Hammad, M. Al-Abed, F. Sibai, and W. Al-Kouz, "Prediction of the Performance of a Sun Tracking Photovoltaic System using different Artificial Intelligence Techniques: Case Study in Zarqa, Jordan," *Int. J. Comput. Digit. Syst.*, vol. 14, no. 1, pp. 10337–10345, 2023.

[34] A. A. S. Azad *et al.*, "Harnessing the sun: Framework for development and performance evaluation of AI-driven solar tracker for optimal energy harvesting," *Energy Convers. Manag. X*, vol. 26, p. 100990, 2025.

[35] P. V. Mahesh, S. Meyyappan, and R. R. Alla, "Maximum power point tracking using decision-tree machine-learning algorithm for photovoltaic systems," *Clean Energy*, vol. 6, no. 5, pp. 762–775, 2022.

[36] P. V. Mahesh, S. Meyyappan, and R. Alla, "Support vector regression machine learning based maximum power point tracking for solar photovoltaic systems," *Int. J. Electr. Comput. Eng. Syst.*, vol. 14, no. 1, pp. 100–108, 2023.

[37] C. G. Villegas-Mier, J. Rodriguez-Resendiz, J. M. Álvarez-Alvarado, H. Rodriguez-Resendiz, A. M. Herrera-Navarro, and O. Rodríguez-Abreo, "Artificial neural networks in MPPT algorithms for optimization of photovoltaic power systems: A review," *Micromachines*, vol. 12, no. 10, p. 1260, 2021.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

AI as an Algorithmic Crutch: The Impact of AI Use on Cognitive Skills and Academic Psychology of Bangladeshi University Students

Md. Rimon Miah
Dept. Computer Science and Engineering
Bangladesh University of Professionals
Dhaka-1216, Dhaka, Bangladesh
miah.rimon14a@gmail.com

Amitav Barman Pranto
Dept. Computer Science and Engineering
Bangladesh University of Professionals
Dhaka-1216, Dhaka, Bangladesh
amitavbp.bup@gmail.com

Iyolita Islam
Dept. Computer Science and Engineering
Bangladesh University of Professionals
Dhaka-1216, Dhaka, Bangladesh
iyolitaislam@bup.edu.bd

***Abstract*— In an era increasingly dominated by Artificial Intelligence (AI), traditional systems, processes, and practices across society are being fundamentally reshaped to leverage the full potential of AI. Students are no exception to this transformation. This study examines how undergraduate students in Bangladesh are using and increasingly misusing AI and how AI is affecting their academic skills, integrity, and psychological well-being. A descriptive survey was conducted among 60 students of the 2021-22 cohort from different public universities of Bangladesh, who have experienced academic life both before and after AI became widespread. The findings reveal that AI use is now a regular habit for 98% of students. However, this convenience is often misused; for instance, a high mean score proves many students frequently use AI to complete their assignments without fully understanding the content (mean = 3.61), and many students submit AI-generated work as their own (mean = 3.70). Moreover, the study highlights a growing psychological toll as the students are becoming overdependent on AI. Students strongly agreed (mean = 3.8) that they would have put more effort into learning without AI and reported declines in reading habits and analytical thinking. Again, many felt their confidence in their academics had dropped significantly due to frequent AI use (mean = 3.50). Also, students' cognitive abilities had dropped significantly due to frequent AI use (mean = 3.50). This research suggests that AI, when misused without proper guidance, becomes an "algorithmic crutch," undermining critical thinking and the core mission of higher education.**

***Index Terms*— Artificial intelligence, cognitive skills, academic integrity, student psychology**

I. INTRODUCTION

From 2022, the availability of generative AI and AI-powered tools became significant, especially for university students who often seek efficient and easy ways to complete tasks. The benefits were obvious: drafting essays, solving problems, finding a project or research idea, and clarifying questions of understanding and grasping complex topics suddenly became faster and easier [1]. It felt like gaining a superpower—one that made academic life more manageable and efficient.

Within three years, AI had sparked a global transformation with its convenience, which established its presence in various fields and professions. Now, AI and AI-powered tools are used by people across various professions. This explains the substantial increase in the number of users. Around 55% of people use AI and AI-powered tools every day [2]. This raises an important question regarding the current situation in Bangladesh: what percentage of university students use AI and AI-powered tools, and what practices do they follow when using AI for academic tasks?

AI promises personalized learning, instant access to information, and new ways to engage with knowledge. Studies from other developing countries have already drawn attention to the downsides: AI misuse is linked to a rise in academic dishonesty and poor academic knowledge [3]. This gives rise to a critical concern for students in Bangladesh: if AI is a tool to help them grow, or an "algorithmic crutch" that could gradually weaken their ability to think, question, and learn independently. Over the past two years of observing undergraduate students' interaction with AI-based tools, a troubling pattern was found. Again, students were turning to AI at the first sign of difficulty. The natural tendency was to assign the work rather than struggle with a difficult idea or overcome a creative barrier. Slowly, the confidence to form original arguments or generate fresh ideas seemed to be slipping away. It wasn't just about convenience anymore; it was about dependence. This shift raised an important question if this was just a personal observation, or part of a larger trend unfolding quietly across campuses.

While the potential of AI is widely celebrated in educational reform discussions, its quieter impact on students' psychology, creativity, and motivation is becoming harder to ignore. The purpose of education is not just to complete tasks efficiently but to nurture the minds capable of creating, questioning, and leading in an uncertain future. Therefore, this paper aims to go beyond observation and explore the involvement with AI, experiences and intellectual detachment of students.

The rest of the paper is organized as follows: Section II contains a review of related literature. Section III discusses the methodology of the research. Section IV presents the study findings from three perspectives: AI usage patterns, impacts on academic and cognitive skills, and psychological effects. Section V provides a discussion of the results in relation to existing theories. Finally, Section VI concludes the paper and outlines the limitations of the study.

II. LITERATURE REVIEW

The related literature and major scholarly databases were explored. The existing work focuses on AI in education. This is often divided into two camps: On one side, its potential as a revolutionary force is celebrated by experts and the other focuses on the possible downsides, such as overreliance, effects on cognitive ability, and threats to academic honesty [4]. The ability of AI to create learning plans tailored to each student, offer instant feedback on assignments, and handle tedious research is highlighted, which is theoretically believed to free up time for deeper, more creative thinking [5]. However, a note of caution is sounded by another group of scholars. Rahman, Ahmed, and Kim express concern that the very efficiency of AI can become a trap. By making

it too easy to find answers and complete assignments, a shallow form of learning may be encouraged, where the goal is perceived as merely finishing the task rather than genuinely understanding the material [6].

Students first starts using AI due to its convenience. A survey conducted by J. von Garrel and J. Mayer around 2023 on 6300 students from Germany found that almost two-thirds of the students use or have used AI-based tools as part of their studies [7].

To understand why these powerful tools might be misused by students, the functioning of the human brain is examined. The well-known Cognitive Miser Theory, introduced by Orbell and Dawes, is cited, which suggests that mental energy is naturally conserved by the brain whenever possible [8]. Fiske and Taylor further explain this tendency in social cognition research [9]. When a choice is presented between a difficult mental task (like writing a research paper from scratch) and an easier alternative (like asking an AI to do it), the shortcut is almost always chosen by the brain. AI tools are identified as the ultimate cognitive shortcut.

This tendency is observed to lead to a behavior called cognitive offloading [11]. This refers to the outsourcing of mental work to an external tool. This has historically been done using calculators for mathematics or by writing to-do lists so that information does not have to be remembered. However, Grinschgl and Neubauer note that AI is different: the entire process of critical thinking, analysis, and creation can be offloaded [11]. When this becomes the default, the mental "muscles" that are expected to be developed in university risk being weakened due to lack of use.

Over time, a state of learned helplessness can be fostered, as conceptualized by Martin Seligman [10]. If consistent reliance on external tools like AI occurs, it may be believed by students that success cannot be achieved independently. A dangerous cycle is thereby created: actual skills are weakened by a lack of practice, self-confidence is lowered, and dependence on AI is increased, as noted by Gerlich [13]. The study is designed to determine whether evidence of these psychological patterns can be observed in the experiences of students in Bangladesh.

Students from different universities in Africa were included in the research in order to identify how AI and AI-integrated tools were used in their academics. It was found that AI and AI-integrated tools were used by students to complete academic tasks such as assignments, reports, and projects. Even without verification, the work was submitted without checking whether it was correct or not. This was found to lead to low academic performance [12]. AI was also used by students to obtain answers to exam questions immediately. The ability of AI to generate answers to any question within a short time was being utilized. The current situation in Bangladesh was examined in this study.

This literature review showed that academicians all over the world are concerned about the use and impact of AI for university students. Moreover, several studies also report that students are misusing AI in their academic tasks. Again, some studies indicate that this trend is weakening students' fundamental academic skills. But there are very few studies conducted in the context of Bangladeshi public university students to explore the impact of AI in their psychological and cognitive ability. Therefore, this study aims to identify the impact of use and misuse of AI among the public university students of Bangladesh.

III. METHODOLOGY

A descriptive survey design was adopted to collect qualityful data for this research. Structured questions titled "Use and Impact of AI Tools on Students of Public Universities in Bangladesh" were used as the instrument for the data collection.

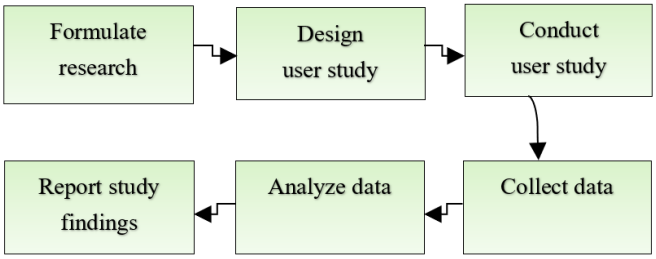


Fig. 1. Research Methodology Overview

A. Formulate research objective

A through review of related works was conducted to identify the research gap. Based on the gap the research object was formulated as to outline the usage and impact of AI in the undergraduate students.

B. Design user study

The user study was designed following the survey method involving a total of 60 participants. Participants are chosen from a population consisted of 40,000 undergraduate students [14] from the 2021–22 session of different public universities in Bangladesh. Participants from this group was chosen because insight could uniquely be offered by students whose university careers began with traditional methods and were subsequently transformed by the arrival of generative AI. Additionally, the survey included questions that required reflection on academic life before and after the introduction of AI. Answering those questions would be much easier for those who passed semesters with and without AI. Thus, a clear "before and after" perspective was provided.

Now talking about 60 participants. To ensure the validity and fairness of the data, an equal number of male and female students, 30 males and 30 females, were included in the sample. Given that the gender ratio in public universities in Bangladesh is approximately 52:48 (male to female) [15], a 50:50 distribution was considered realistic and appropriate for balanced representation. Purposive sampling was used to select three universities: Bangladesh University of Professionals (BUP), Bangladesh University of Engineering and Technology (BUET) and Jahangirnagar University (JU). From each university, 20 students were selected, ensuring equal gender representation by including 10 male and 10 female students.

C. Conduct user study

This study was conducted as independent academic research. Informed consent was obtained from all participants after they were briefed on the study's purpose and significance. To ensure privacy, all responses were collected anonymously, and no personally identifiable information (PII) was recorded. Furthermore, the findings were reviewed by a registered psychologist to ensure that the interpretations regarding

the psychological effects and the effect on students' cognitive ability align with established psychological principles.

Then a structured questionnaire titled "Use and Impact of AI Tools on Students of Public Universities in Bangladesh" was used to collect responses. The questionnaire consist of fifteen questions, under three key themes:

A. AI Usage Patterns (4)
B. Impact on Skills (8)
C. users' psychological effects (3)

Most questions used a 5-point Likert scale, where 5 represented "Strongly Agree" and 1 represented "Strongly Disagree."

Expert advice was sought to ensure the quality and fairness of the research questions. The draft questionnaire was discussed with a psychologist at the university's Counselling & Placement Center (CPC) [16] and with Lec. Iyolita Islam, a lecturer at Bangladesh University of Professionals (BUP). Their feedback was incorporated to refine the questions, making them clearer, less biased and more effective at capturing the intended measures.

Questions:

A1. I use AI tools (e.g., ChatGPT, SciScienc, Gamma, Framer etc.) to complete most parts of assignments, term papers, or research work without fully understanding them.
B1. I have become overdependent on AI tools.
C1. I feel the urge to use AI tools immediately whenever I face an academic or personal task, without trying it myself first.

Then, these questions were implemented in a Google Form maintaining the themes, which further linked to a Google Sheet that automatically record participants' responses for subsequent analysis. This method was chosen for its wide acceptance, reliability, and user-friendly interface, allowing convenient participation and seamless data organization for analysis.

D. Collect data

The final Google Form link was distributed digitally to participants. All respondents were assured that anonymity would be maintained to encourage candid responses. The importance of participation was briefed, and respondents were requested to provide thoughtful and honest answers to ensure reliability.

E. Data Analysis

Data collected in three perspectives were compiled and analyzed separately using descriptive statistics. Mean scores (1) for each Likert-scale question were calculated to understand the central tendency of opinions.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Percentages (2) were calculated for the question regarding the frequency of AI use. To facilitate better data visualization.

$$\text{Percentage} = \left(\frac{f_i}{N}\right) \times 100 \tag{2}$$

where f_i is the frequency of a specific response, and N is the total number of responses.

In addition to descriptive analysis, a Spearman's rank correlation analysis was conducted to examine the relationships among survey variables. This non-parametric test was chosen due to the ordinal nature of Likert-scale data and was used to identify significant associations between AI usage patterns, perceived impacts on cognitive skills, and psychological effects. The correlation findings are interpreted and discussed in the Discussion section.

F. Report Findings

The collected data were organized to show the main trends and patterns from three perspectives: how university students use AI, how it affects their academic and cognitive skills, and its psychological impact. The results are presented using simple descriptive statistics, like mean scores and percentages to give a clear picture of the participants' responses. Charts, including pie charts and bar graphs, as well as tables, were also used to make the findings easier to understand and compare. The results are presented in a way that connects directly to the research objectives, helping readers quickly see the important trends, possible concerns, and the overall influence of AI use among undergraduate students in Bangladesh.

IV. STUDY FINDINGS

The analysis of the survey responses generated study findings as follows: among undergraduate students, the use of generative AI and AI-based tools has reached an alarming level, and this continuation is effecting their cognitive abilities.

A. The Pervasiveness and Nature of AI Use

The study findings showed that AI is not a peripheral tool but a central, daily component of academic life for a majority of students.

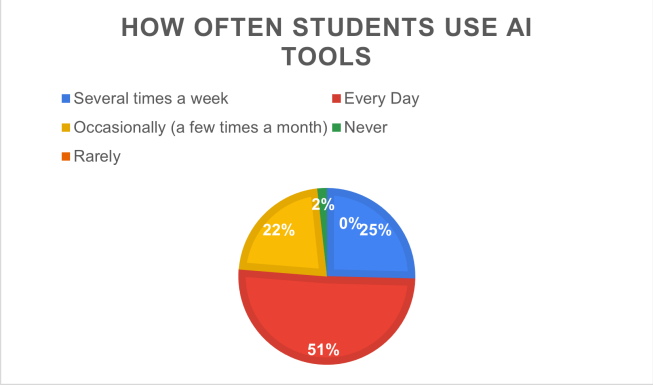


Fig. 2. Frequency of AI use among students.

As the pie chart in Figure 2 illustrates, a staggering 98% of students are regular users of AI. 51% of the students use AI and AI-integrated tools every day. Our research also finds that those who are introduced to AI become active users of AI [17]. The percentage of rare AI users is zero. 47% of students are accessing AI tools at least on a weekly basis. This high frequency is coupled with ethically questionable practices.

The survey then explored the use of AI and AI-integrated tools by the students in different academic domains. The survey results of Section A are represented as a mean value in a bar chart (Figure 3) for a better understanding of the

plot. This shows a clear acceptance of using AI for tasks that compromise academic integrity and the very skills university education is meant to develop.

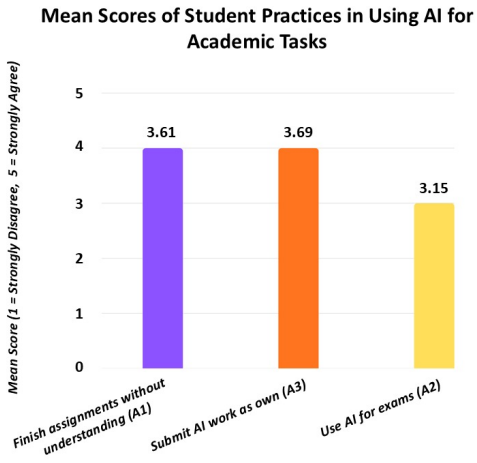


Fig. 3. Mean score of Section A survey result.

TABLE I: Student Practices in Using AI for Academic Tasks.

Statement	Mean	Strength of Agreement
I use AI to finish assignments without really understanding them. (A1)	3.611	Strong Agreement
I submit work made by AI as my own. (A3)	3.695	Strong Agreement
I use AI to answer questions on exams or tests. (A2) (A3)	3.151	Moderate Agreement

Interpretation

All statements received a mean score above the neutral midpoint (3.0), indicating that many students engage with AI tools in ways that may compromise academic integrity. Most notably, students moderately agree that they use AI to answer exam questions, and they strongly agree on submitting AI-generated content as their own work, and they also use AI and AI-integrated tools to finish their assignments without even understanding them. This reveals a trend toward academic misconduct facilitated by AI.

B. Perceived Impact on Cognitive and Academic Skills

The questionnaire then investigated students' perceptions on how this reliance on AI is affecting their core abilities, such as critical thinking, problem solving, creativity, etc. The responses, visualized in the chart shown in Figure 4 below and detailed in Table II, reveal a strong and widespread awareness of the negative consequences.

Interpretation

Looking at the bar chart(Figure 4), one feeling stands out above all others. The very strong agreement with question B3 (the tallest bar) shows that university students are openly admitting they would have put more effort into their studies without AI. It's a powerful confession that they are consciously trading away hard work for the convenience of these tools. On top of that, there's a strong sense that AI is hurting their ability to think deeply (B2), making them less likely to read books (B8), and generally encouraging dependency (B7). These aren't just small concerns; they are significant trends that point to a real change in university students' study habits. The response to question B5 is also quite telling.

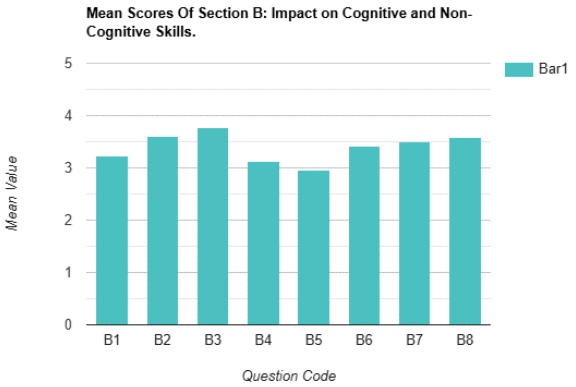


Fig. 4. Bar chart showing the mean scores of Section B.

TABLE II: The Perceived Impact of AI on Student Skills (Scale: 1=Strongly Disagree, 3=Neutral, 5=Strongly Agree)

Statement	Mean	Strength of Agreement
I have become over-reliant on AI tools. (B1)	3.2	Moderate Agreement
I believe AI is reducing my ability to think analytically. (B2)	3.6	Strong Agreement
Without AI, I would have tried harder to learn on my own. (B3)	3.8	Very Strong Agreement
I see AI tools as a new method of academic cheating. (B4)	3.1	Leaning Towards Agreement
Without AI, I struggle to generate new ideas. (B5)	3.152	Leaning Towards Agreement
AI tools negatively affect my creative thinking skills. (B6)	3.4	Moderate Agreement
AI makes students less confident, less creative, and lazy. (B7)	3.5	Strong Agreement
Using AI tools has reduced my habit of reading books. (B8)	3.6	Strong Agreement

With a mean score of 3.152, it shows that students are just starting to agree that they struggle to come up with new ideas on their own. It suggests a growing reliance on AI for creativity that they are beginning to notice by themselves. Taken together, the chart indicates a clear story: university students in Bangladesh feel they are sacrificing core academic skills, cognitive skills, and good study habits for the simple ease that AI provides.

C. The Psychological Toll: Dependency and a Crisis of Confidence

The final section of the survey examined the deeper psychological impacts. The results shown in Table III were the most definitive of the entire study, suggesting significant changes in how students approach a task or learning and view their own capabilities. Responses also show a strong internal shift: students are beginning to doubt their problem-solving ability and academic ability and feel mentally dependent on AI tools to begin any task.

The findings in Table III are stark. There is strong agreement among students that they feel a conditioned urge to use AI immediately. Critically, there is very strong agreement that this habit has directly resulted in lower self-confidence and made them feel incapable when it

TABLE III: The Psychological Impact of Frequent AI Use (Scale: 1=Strongly Disagree, 3=Neutral, 5=Strongly Agree)

Statement	Mean	Strength of Agreement
I find it harder to write or come up with ideas without AI now.	3.481	Moderate Agreement
I feel an urge to use AI right away, without trying a task myself first.	3.501	Strong Agreement
Frequent AI use has lowered my confidence in my own academic work.	3.533	Very Strong Agreement

comes to completing tasks, the second-highest mean score in the entire survey.

Interpretation

These results show that university students are not just using AI tools out of convenience; they are becoming psychologically attached to them. There is strong agreement among students that they feel a conditioned urge to use AI immediately. The strong mean scores for "lowered academic confidence" are particularly significant. This says that the students are losing confidence in their own ability. This may develop into a more serious problem. If they encounter a situation where no generative AI or AI-integrated tools are available, they may feel unprepared or unable to proceed. There may be additional consequences associated with the long-term use of AI . Overall, the analysis indicates that university students are becoming increasingly dependent on AI and AI-integrated tools [18].

V. DISCUSSION

The incredible convenience offered by AI has created a powerful dependency that students are not only falling into but are also consciously feeling. This issue goes far deeper than just a few students cheating on assignments; it reflects a subtle but significant rewiring of the entire learning process. The widespread use of AI for academic dishonesty confirms the behavior, while the students' own perceptions about their skills and confidence explain the damaging consequences.

This real-world data provides a textbook illustration of the psychological theories discussed earlier. The strong agreement (mean = 3.20) with feeling an immediate "urge to use AI" (Table III) reflects the Cognitive Miser theory [8] in action, where the brain consistently chooses the path of least mental effort. When this behavior is repeated daily, it becomes a habit of cognitive offloading. Over time, the students themselves report the outcome: a perceived erosion of their analytical, academic, and creative skills because these abilities are no longer being actively exercised. Their very strong agreement (mean = 3.77) that they "would have tried harder" (Table II) without AI is a direct acknowledgment of this cognitive trade-off. Intellectual struggle is being replaced by speed and convenience.

Correlation results revealed three key clusters: academic misconduct (A1–A3), negative self-perception (B1, B5, B7, C1–C3) and perceived cognitive impacts (B2, B3, B6), indicating interconnected patterns of AI misuse, dependency, and reduced effort.

These patterns point to a growing crisis of academic confidence. University students strongly agree (mean = 3.50) that frequent AI use has lowered their self-belief (Table III),

TABLE IV: Key Spearman's Rank Correlation Results

Category	Pair	r	p	Interpretation (Brief)
Very Strong	A1–A3	0.83	< 0.001	AI misuse in assignments aligns with AI submission.
	A2–A3	0.77	< 0.001	Exam-time AI use aligns with dishonest submission.
	B1–C3	0.75	< 0.001	Overdependence predicts reduced academic confidence.
Strong	A1–A2	0.68	< 0.001	Assignment misuse relates to exam misuse.
	B1–C1	0.67	< 0.001	Dependency relates to immediate AI reliance.
	B3–C2	0.66	< 0.001	Reduced effort relates to difficulty working independently.
Moderate	A4–B1	0.58	< 0.001	Higher usage frequency predicts overdependence.
	B2–B6	0.56	< 0.001	Analytical and creative thinking decline together.
	B7–C3	0.54	< 0.001	Perceived laziness aligns with lower confidence.
Weak	B8–A4	0.42	< 0.01	Frequent AI use associates with reduced reading habits.

which is particularly concerning because confidence, perseverance, and independent problem-solving are core outcomes of higher education. This directly supports the concept of learned helplessness [10], where repeated reliance on external aids gradually erodes self-efficacy. By outsourcing the struggle—the very process through which competence and confidence are built—students are left feeling incapable without an algorithmic crutch. This finding confirms the initial observation that motivated this research and extends concerns raised in prior studies by grounding them in empirical evidence from the Bangladeshi university context.

These findings invite reflection among students and educators alike. The fundamental goal of education is not merely to obtain a certificate but to develop academic productivity, intellectual discipline, and transferable cognitive skills. Becoming a hardworking, creative, confident, and capable individual remains a central mission of higher education [19]. The results therefore raise a critical question in an AI-dominated academic environment: are students graduating with the core competencies that universities are expected to cultivate?

A comparison with related work is presented below to highlight what distinguishes the present study from existing literature.

VI. CONCLUSIONS

This research, inspired by firsthand observations, has revealed a widespread challenge that appears to be affecting many students. The data strongly suggest that university students in Bangladesh are not just using AI; they are becoming increasingly dependent on it. It begins with the introduction to the various conveniences offered by AI. This triggers a psychological phenomenon called cognitive offloading. It makes the student dependent on AI and AI-integrated tools, as the least mental effort is required. After this stage, students start getting drawn deeper. This shift progresses from small

TABLE V: Comparative Analysis: Related Works vs. Present Study

Comparison Feature	Related Works (Literature)	Present Study
Research Focus	Emphasizes engagement, efficiency, and learning outcomes.	Focuses specifically on academic psychology and cognitive ability.
Participant Perspective	Surveys current AI users to assess adoption and perceived benefits.	Comparative perspective from students who studied both with and without AI tools.
Psychological Depth	Mentions over-reliance as a general or ethical concern.	In-depth analysis of cognitive offloading and reduced academic confidence.
Key Findings on Skills	Often reports performance gains through efficiency.	Reveals declines in creativity and analytical thinking due to AI as a shortcut.

conveniences to an over-dependence on these powerful tools. This progression ultimately affects their cognitive abilities. It also changes how they approach any task.

AI may appear appealing and easy to learn from, but the long-term cost is too significant to ignore. A generation risks becoming one with poor analytical skills, less creativity, and a fundamental lack of confidence in its own intellect.

The path forward is not to run away from this powerful technology but to learn how it can be used without sacrificing essential human strength. It is suggested that abilities may become weaker day by day while AI continues to grow stronger [20]. This is presented as a challenge for students, teachers, and university leaders to work together to transform educational approaches. Educational institutions must ensure that students graduate not merely as users of AI but as independent, critical thinkers—individuals prepared to shape the future, rather than operators skilled only at prompting a machine.

Despite the insightful findings and the deployment of the most appropriate methods in this study. This study had few limitations.

The study was conducted within a very limited timeframe, which may have influenced the development of the questionnaire, the data collection process, the analysis of responses, and the overall presentation of the paper. Additionally, because the research was entirely self-funded, the sample size was restricted to 60 participants. Although the study provides a valuable and indicative snapshot of the current situation, this relatively small sample may limit the generalizability of the findings to the wider population of university students in Bangladesh. The study also relies on students' self-reported perceptions of their activities and feelings, and some participants may not have been fully aware of their own behaviors or may have responded differently from their actual practices. Finally, while a strong link appears between heavy AI use and various negative outcomes, the limited scope of the research does not allow for definitive conclusions about causation; it cannot be proven that AI alone directly produces these effects. These limitations should be considered when interpreting the results of this study.

References

[1] K. D. Wang, Z. Wu, L. N. Tufts II, C. Wieman, S. Salehi, and N. Haber, "Scaffold or crutch? Examining college students' use and views of generative AI tools for STEM education," *arXiv preprint arXiv:2412.02653*, 2024. [Online]. Available: https://arxiv.org/abs/2412.02653

[2] Minderoo Foundation, "AI Attitudes Report," Aug. 2025. [Online]. Available: `https://cdn.minderoo.org/assets/documents/AI-attitudes-report-August-2025.pdf`

[3] J. A. Jaleel, S. Aziz, M. Farid, and A. Bashir, "The impact of ChatGPT on academic integrity in medical education: a developing nation perspective," *Frontiers in Education*, vol. 10, Feb. 2025.

[4] T. K. F. Chiu, Q. Xia, X. Zhou, C. S. Chai, and M. Cheng, "Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education," *Computers & Education: Artificial Intelligence*, vol. 4, Art. no. 100118, 2023, doi:10.1016/j.caeai.2022.100118.

[5] X. Xu, Y. Chen, and L. Wu, "Artificial intelligence-based personalised learning in education: A systematic literature review," *Discover Artificial Intelligence*, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s44163-025-00598-x

[6] M. S. Rahman, T. Ahmed, and H. Kim, "Chat or cheat? Academic dishonesty, risk perceptions, and ChatGPT usage in higher education students," *Social Psychology of Education*, 2025.

[7] J. von Garrel and J. Mayer, "Artificial Intelligence in studies—use of ChatGPT and AI-based tools among students in Germany," *Humanities and Social Sciences Communications*, vol. 10, no. 1, Art. no. 799, 2023, doi:10.1057/s41599-023-02304-7.

[8] J. Orbell and R. M. Dawes, "A 'Cognitive Miser' theory of cooperators' advantage," *American Political Science Review*, vol. 85, no. 2, pp. 515–528, 1991, doi: 10.2307/1963172.

[9] S. T. Fiske and S. E. Taylor, *Social Cognition*, New York: McGraw-Hill, 1991.

[10] M. E. P. Seligman, "Learned helplessness," *Annual Review of Medicine*, vol. 23, pp. 407–412, 1972.

[11] S. Grinschgl and A. C. Neubauer, "Supporting cognition with modern technology: Distributed cognition today and in an AI-enhanced future," *Frontiers in Artificial Intelligence*, vol. 5, 2022, Art. no. 908261, doi:10.3389/frai.2022.908261.

[12] U. I. Offor, P. Nwaru, and C. Offiah, "Challenges of the use of artificial intelligence on academic performance of undergraduate students in public universities in Anambra State," *Afr. J. Educ. Manag., Teach. & Entrepreneurship Stud.*, vol. 12, no. 2, pp. 176–190, 2024.

[13] M. Gerlich, "AI tools in society: Impacts on cognitive offloading and the future of critical thinking," *Societies*, vol. 15, no. 1, 2025, doi:10.3390/societies15010006.

[14] "Students suffer as public unis delay admission process," *The Business Standard*, Jun. 21, 2025. [Online]. Available: https://www.tbsnews.net/bangladesh/education/students-suffer-public-unis-delay-admission-process-533578

[15] M. Ahmed, "Male–female student ratio to be equal in public universities," *Prothom Alo*, Feb. 19, 2024. [Online]. Available: https://en.prothomalo.com/youth/education/wtvw7an7dm

[16] Bangladesh University of Professionals, "Counselling & Placement Centre," Accessed: Nov. 19, 2025. [Online]. Available: https://bup.edu.bd/cpc

[17] J. Freeman, "Student Generative AI Survey 2025 – HEPI," Feb. 26, 2025. [Online]. Available: https://www.hepi.ac.uk/2025/02/26/student-generative-ai-survey-2025/

[18] M. Thomas, "Dangers of Artificial Intelligence (AI)," Built In, Jan. 13, 2025. [Online]. Available: https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence

[19] "10 Core Competencies To Focus On," Open Colleges, Aug. 8, 2015. [Online]. Available: https://www.opencolleges.edu.au/blogs/articles/10-core-competencies-to-focus-on

[20] "Will Machines Become More Intelligent Than Humans?," Caltech Science Exchange. [Online]. Available: https://www.scienceexchange.caltech.edu/blog/will-machines-become-more-intelligent-than-humans

[21] N. M. Talukder and W. bin Ahsan, "The Impact of AI on Student Engagement in Virtual Learning within Bangladesh's Higher Education Sector," *Userhub Journal*, Jan. 2025. [Online]. Available: https://doi.org/10.58947/journal.jtgx34

[22] D. S. Trimukhe, *Campus to Cloud: The College Student Journey in an AI World*, 1st ed. India: Asian Press Books, Aug. 2025. [Online]. ISBN: 978-93-5788-539-3.

[23] A. Akter and S. M. Rahman, "Effect of AI Usage on Learning, Creativity, and Innovation: The Role of Mediating Self-Efficacy and Moderating Task Complexity," *SSRN Preprint*, 2025. [Online]. Available: https://ssrn.com/abstract=5784563

[24] M. Tamanna and B. Sinha, "A conceptual analysis of artificial intelligence (AI) on academic opportunities and challenges: a case study based on higher educational institutions in Bangladesh," *Quality Assurance in Education*, Jun. 2024. [Online]. Available: https://doi.org/10.1108/QAE-03-2024-00501

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Elevated View Gait Biometrics: Efficient Skeleton-Based Recognition Using UAV-Captured Videos

Mohammad Ali Sakib
Dept. of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
sakib.cse04@gmail.com

Maisha Mobashira
Dept. of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
maishamobashira@gmail.com

Samia Khondoker
Dept. of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
samiakhondoker880@gmail.com

Nakib Aman
Dept. of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
nakibaman@gmail.com

Md. Abu Johab
Dept. of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
abujohab.cse.pust@gmail.com

MD. Rafidur Rahaman
Dept. of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
yashrib1@gmail.com

***Abstract*—Gait recognition, a biometric modality that identifies people by their walking patterns, is a non-intrusive solution for surveillance and security purposes. However, realistic issues such as extreme viewing angles, occlusions, and environmental variations restrict performance. This paper overcomes these challenges by using skeleton-based representations based on drone-video sequences in the DroneGait dataset. Pose estimation generates stickman-like stickers (without background dependency) through the extraction of the main joints. Four deep learning architectures, namely, EfficientNet-B3, MobileNetV3, ResNet-50 and Vision Transformer (ViT) are evaluated for multi-class gait identification across 13 subjects. Models are trained with data augmentation and label smoothing and Adam optimizer to improve generalization. Experimental results show that EfficientNet-B3 achieves the highest accuracy 84%, better than others. Grad-CAM is also applied for interpretability. The proposed framework shows strong gait recognition under aerial viewpoints, giving an efficient and explainable solution for resource-constrained real-world deployments.**

***Index Terms*—Gait Recognition, Deep Learning, EfficientNet-B3, Vision Transformer, Explainable AI, Grad-CAM, Pose Estimation.**

I. Introduction

Gait recognition refers to the recognition of people on the basis of their walking behaviour and has recently been in the spotlight due to the contact-free property a recognition technique needs in situations like surveillance, security, and medical care, but the quality of real world systems is hampered by challenges such as extreme viewing angles, occlusions, shadows, and low resolution recordings. Conventional datasets, such as CASIA-B [1], are also rather shallow in this, as such a dataset covers mostly gait under low angles of view and, therefore, they may not be effective when used under either drone-based or elevated presence.

To overcome these shortcomings, a number of new data sets and models are developing. As an example, the DroneGait dataset [2] derived to provide gait information at higher altitude that allows examination at the broader observational angles not included in the traditional datasets. This sort is done in the OUMVLP-Pose dataset [3], which adds pose estimation methods (e.g., OpenPose, AlphaPose) to construct a multi-view and large-scale dataset, to enable both appearance-based and pose-based gait recognition. On the modeling front, architectures such as the GaitTAKE [4] architecture, which uses time-binning attention and keypoint-driven embeddings, work toward eliminating clothing variation and carried object problems. Also, Gait-TR [5] takes advantage of skeleton-based methods with spatial transformer networks and temporal convolutional networks, in order to capture better gait dynamics. These advances mark the continuation of a quest to improve gait recognition in a variety of variable situations.

The contemporary application of deep learning-based approaches to gait skeleton sequence identification has allowed to make the following contributions in the present study :

- The processing of the skeleton-based gait dataset

was done to create stickman representation of raw video sequences produced by state-of-the-art human pose estimation.

- Four deep learning models-EfficientNet-B3, MobileNetV3, ResNet-50, and a Vision Transformer (ViT) were used to measure their performance at extraction of discriminative gait features.
- Precision, Recall, F1-score, and other performance indicators were tested in each of the models to provide solid grounds to comparison.
- Attended grad-CAM was adopted to explain model predictions, allowing visual validation of the important skeletal regions that informed a specific classification decision.

The residue of this paper is organized as follows: Section II reviews the literature paper based on deep learning based approach for person recognition using drone gait dataset. Section III discuss about the dataset preprocessing procedures and experimental setup. Section IV discuss methodology part in details. Section V represent result and performance analysis. And finally in section VI briefly discuss the conclusion part,limitation,and also the future research area.

II. Related Works

Gait recognition recently became an important topic of interest in computer vision and the surveillance field due to the development of UAVs to solve the issues involving clothing differences, movement change, environmental differences, and computer power.

A more adversarial LiFE-Adversarial Loss (LA) showed improved performance using free clothing appearance labels, only performed well when using PRCC and LTCC [6].GaitTAKE combined temporal attention with keypoint-based embeddings and performed well in rank-1 accuracies across both CASIA-B and OUMVLP; however generalizability is a risk to extend to new datasets [7]. The proposed CyleGait [8] uses a skeleton-based gait recognition approach that takes into account gait periodicity sources using a GCN-PTP network and Irregular Pace Converter to augment data. It attains the state-of-the-art performance and has a robust performance under cross-view and cross-walking conditions on benchmark datasets.

The considerable size of outdoor dataset CASIA-E allowed benchmarking, and its cross-scene cross-clothing variation indicated that domain adaptation was necessary [9]. LN offered a silhouette-based approach following Compact Blocks and feature pyramids to better recognize clothes-changing using CASIA-B and OUMVLP but had to preprocess silhouettes [10]. A 3D ResNet-18 using adaptive MOG2 was able to enhance recognition on noisy, poor-quality video but deployment of UAVs would be restricted by the available power and weather conditions [11].

A Multi-Scale Attention Gated Aggregation module was applied to improve vertical view recognition by CVVNet, boosting DroneGait and Gait3D performance at a greater cost of computation [12].A pruning and quantization-based edge-optimized structure obtained energy-efficient gait recognition in real-time using Jetson devices but has performance-complexity trade-offs [13]. AerialGait has proposed an aerial-ground dataset and AGG-Net framework that deal with the cross-domain gait recognition problem, but heterogeneous data are still a challenge [14].

A LDA and SVM YOLOv5 + skeleton tracker system achieved 83.2 recall on Drone-Action dataset, but there is still the problem of dynamic backgrounds and computer computation constraint [15]. The work of LGait integrated temporal attention and convolution to achieve state of the art accuracy at the cost of high complexity [16]. FarSight combined gait, face, and volume information to enhance UAV-based identification performance when the condition was harsh, but only at the cost of computational performance [17]. DetRIDX, a massive dataset of UAV based re-ID has pushed the boundaries of cross-view re-ID but fails to adequately handle long-range and low-resolution [18]. An approach to UAV human activity recognition relies on deep learning and can combine detection with YOLO, skeleton keypoints, and feature optimization to achieve results of 85% on Drone-Action and 75% on UCF-ARG but the representational variance across aerial views is a topic of future development [19].

III. DATASET DESCRIPTION AND DATA PREPROCESSING

A. Dataset Description

The major dataset employed in this work was based on the DroneGait dataset that was captured to cope with the gait recognition at high vertical positions- a scenario that is usually dominated by most traditional gait datasets. The drones have been used to record the videos in outdoor settings with natural daylight in order to guarantee beneficiary to real life purposes. The video contains the recordings of 13 people walking across open areas filmed in different viewing angles and heights. Based on these videos we obtained the frames of the stickman skeleton of human movements.The stickman representation is that of gait recognition where the skeletal keypoints correspond to major joints such as the head, torso, arms and legs that obey the temporal motion characteristics.

An example of the dataset in Fig.1 shows Person-0 on the left and Person-8 on the right.

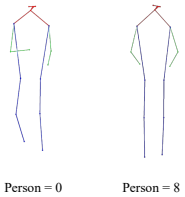


Fig. 1: Examples from the dataset: (left) Person-0, (right) Person-8

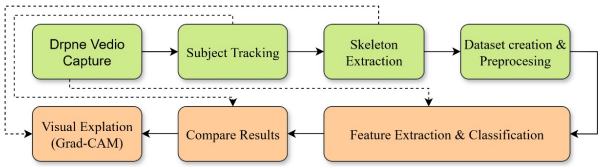


Fig. 2: Proposed Work Flow Diagram

B. Data Preprocessing

To prevent data leakage and overly optimistic evaluation, the dataset was split in a *subject-wise* manner rather than frame-wise. All skeletal frames belonging to a given subject were assigned exclusively to a single subset. Specifically, subjects Person-0 to Person-9 were used for training, Person-10 for validation, and Person-11 to Person-12 for testing, corresponding approximately to an 80%/10%/10% split. This ensures that no frames from the same subject or walking sequence appear in both training and test sets. The split was fixed using a deterministic random seed (seed = 42) to ensure reproducibility.

All skeletal frames were resized to 224×224 pixels and normalized to the range $[0, 1]$. Data augmentation was applied only to the training subset, including random rotation ($\pm 30^\circ$), width and height shifts (0.2), shear (0.2), zoom (0.2), and horizontal flipping. Validation and test subsets were only resized and normalized, without any augmentation.

TABLE I: Subject-wise Data Split Protocol

Subset	Subject IDs
Training	Person-0 to Person-9
Validation	Person-10
Testing	Person-11, Person-12

IV. METHODOLOGY

This study proposes a deep learning framework for skeleton-based gait recognition using drone-captured stickman frames. Four models (EfficientNet-B3, MobileNetV3, ResNet-50, ViT) were trained with data augmentation and evaluated using accuracy, precision, recall, F1-score, and Grad-CAM for interpretability. While gait is temporal, this work focuses on frame-level skeletal features to capture discriminative walking patterns efficiently.

A. Models

1) MobileNetV3:MobileNetV3 uses depthwise separable convolutions, squeeze-and-excitation (SE) blocks and lightweight attention mechanisms to maintain efficiency. The fundamental construction unit is

$$\begin{aligned} x' &= \mathrm{ReLU6}\big(\mathrm{BN}(\mathrm{DWConv}(x))\big) \\ y &= \mathrm{BN}(\mathrm{PWConv}(x')) \\ z &= x + y \end{aligned} \tag{1}$$

2) ResNet-50 :ResNet-50 addresses the problem of the disappearance of gradients to model deep networks. A residual block may be written as

$$y = F(x, \{W_i\}) + x \tag{2}$$

3) EfficientNet-B3 : The benefit of using EfficientNet-B3 is balance network depth,width and also the input resolution.The main building block of this model is MBConv(Mobile Inverted Bottleneck Convolution).

$$\begin{aligned} \tilde{x} &= \mathrm{Swish}\big(\mathrm{BN}(\mathrm{DWConv}(x; k, s))\big), \\ y &= \mathrm{BN}(\mathrm{PWConv}(\tilde{x}; c_{\text{out}})), \\ z &= x + y \quad \text{if input/output dimensions match.} \end{aligned} \tag{3}$$

4) Vision Transformer(ViT) :The Vision Transformer (ViT) flattens an image and treats it as a sequence of patches. Each patch x_p is embedded as:

$$z_0 = [x_{\text{class}}; x_p^1 E; x_p^2 E; \dots] + E_{\text{pos}} \tag{4}$$

where x_{class} is the classification token, E is the patch embedding, and E_{pos} is the positional embedding.

B. Explainable AI:Grad CAM

The crucial areas in the skeleton images that affect the model prediction are highlighted by Grad-CAM. For class c and feature map A^k, the Euclidean weight is calculated as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5}$$

where Z is the number of pixels in the feature map, and y^c is the score for class c.

The Grad-CAM heatmap is then computed as:

$$L_{\text{Grad-CAM}}^c = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{6}$$

The EfficientNet-B3 has been chosen in skeleton-based gait recognition because its architecture is lightweight and significantly accurate; hence, making it effective in minimal or small-equipped environments during training and inference.

V. Result and Performance Analysis

With a high accuracy, the proposed EfficientNet-B3 model had robust performance in all classes of skeleton-based gait recognition. Evaluation measures such as Precision, Recall, F1-Score, ROCs verify its consistency and Grad-CAM show that the model can prioritise the regions of interest, which can be skeletal joint movements.

A. Model Comparison

The skeleton-based gait recognition performance of various types of deep learning architecture was compared. As indicated in Table III, EfficientNet-B3 scored the highest percentage of accuracy, which was 84% compared to the Accuracy percentage of MobileNetV3 of 81% and ResNet-50 of 80%. The ViT model, in turn, although effective, has recorded a lower accuracy of 77%. This comparison shows that when it comes to the high-angle gait recognition tasks in resource-limited environments, the most balanced model should be EfficientNet-B3 since it does not perform as well as other models whose complexity is too high, and does not perform as bad as other models, whose complexity is too small.

TABLE II: **Comparison of Model Performance**

Model	Accuracy (%)
Vision Transformer (ViT)	77
ResNet-50	80
MobileNetV3	81
EfficientNet-B3 (Proposed)	**84**

B. ROC Curve Analysis

Receiver Operating Characteristic (ROC) curves (Fig. 4) were computed using a one-vs-rest (OvR) strategy for the multi-class person identification task. The reported Area Under the Curve (AUC) corresponds to the macro-averaged AUC across all 13 classes. The proposed EfficientNet-B3 achieved a near-perfect macro-average AUC of 0.99.

Although the AUC value is high, the overall classification accuracy is comparatively lower (84%) due to confusion among visually similar gait patterns for certain subjects. While AUC measures the ranking ability of the classifier across decision thresholds, accuracy reflects strict top-1 class assignment, which is more sensitive to inter-class similarity. This explains the coexistence of high AUC and moderate accuracy. Per-class ROC curves are provided to further verify this behavior.

C. Accuracy curve analysis

Accuracy curves (Fig. 5) shows that within the first 5 epochs accuracy is increased rapidly where training accuracy increased from 28% to 80% and validation accuracy increased from 57% to 88%.Both curves become stable, training accuracy almost reaches about 89% and validation accuracy remain almost 84% to 87%.The narrow margin of two curves indicates that EfficientNetB3 have a good generalization, low overfitting and good performance on the DroneGait dataset.

D. Loss curve analysis

The loss curves (Fig. 6) decrease quickly in the first 5 epochs indicating fast learning. After about 10 epochs, both training and validation losses tend to stabilise to low close values. This shows good convergence with little overfitting on the DroneGait data.

E. Confusion Matrix analysis

This confusion matrix (Fig. 7) indicates the classification performance for various persons. Most predictions are along the diagonal, i.e. the model correctly classified each class, but there are a few misclassifications (e.g. Person-10 with Person-8 and Person-9; Person-4 with Person-6 and Person-7). Overall, the model works well but has a little trouble with a handful of overlapping classes.

The classification report (Table III) indicates how EfficientNet-B3 provided an in general accuracy of 84%, macro-average of 0.87 precision, 0.85 recall, and 0.85 F1-score. Similar strategies worked very well as Person-0 to Person-3 only had a perfect score, whereas Person-4, Person-5 and Person-10 had lower F1-scores because of overlapping characteristics. Such results validate the strength of the model, and also indicate areas of opportunity when approaching difficult classes.

TABLE III: **Classification Report of Proposed EfficientNet-B3 Model**

Class	Precision	Recall	F1-Score	Support
Person-0	1.00	1.00	1.00	20
Person-1	1.00	1.00	1.00	25
Person-2	1.00	1.00	1.00	25
Person-3	1.00	1.00	1.00	27
Person-4	0.91	0.66	0.76	32
Person-5	0.74	0.91	0.82	32
Person-6	0.74	0.90	0.81	31
Person-7	1.00	0.70	0.82	20
Person-8	0.71	0.71	0.71	35
Person-9	0.71	0.63	0.67	27
Person-10	0.70	0.70	0.70	30
Person-11	0.81	1.00	0.90	26
Person-12	0.94	0.84	0.89	19
Macro Avg	0.87	0.85	0.85	349
Weighted Avg	0.85	0.84	0.84	349

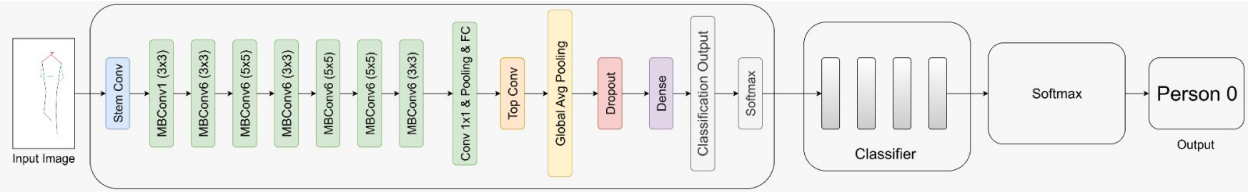


Fig. 3: Architecture of proposed model

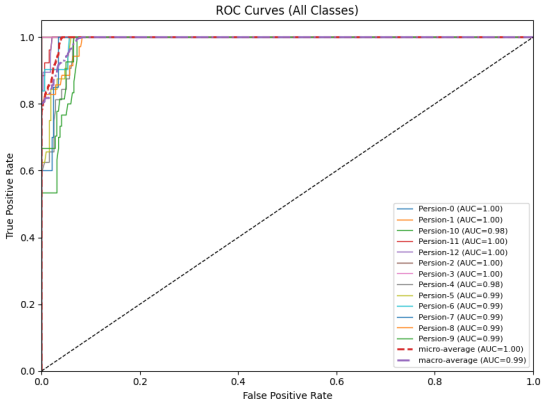


Fig. 4: ROC curve

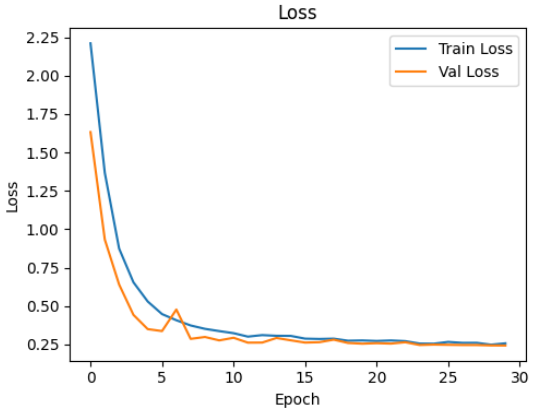


Fig. 6: Loss curve

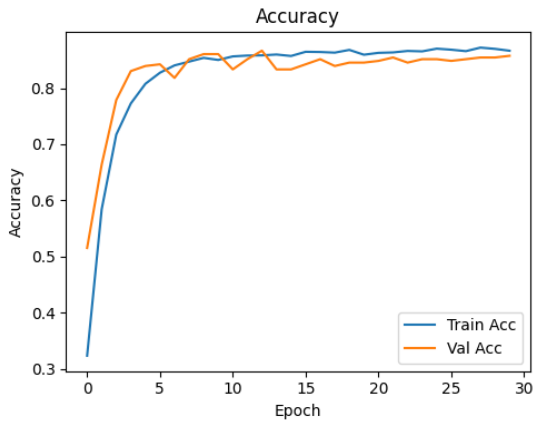


Fig. 5: Accuracy curve

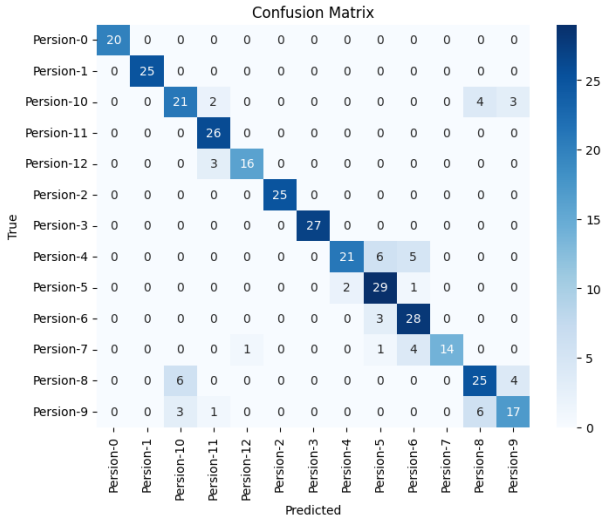


Fig. 7: Confusion Matrix

F. GRAD Cam

The Grad-CAM heatmaps (Fig. 8) of the other two prompts, Fig. 8 depict that the model tends to look at upper body joints, including shoulders and arms, that are significant characteristics to classify poses correctly. This verifies that predictions are premised on non-trivial patterns in the skeleton and not noise.

Nevertheless, the intensity of the heatmap depends on inner activation and can emphasize unimportant places, not all latent meaning can be captured by a single layer.

VI. Conclusion

EfficientNet-B3 was marginally more accurate than MobileNetV3 and ResNet-50 with ViT working the worst, pointing to the power of lightweight CNNs in smaller dataset settings. Grad-CAM increased the interpretability, but it only worked with a single-layer feature map, and hence it cannot be used to make detailed observations against complex backgrounds or uneven lighting.

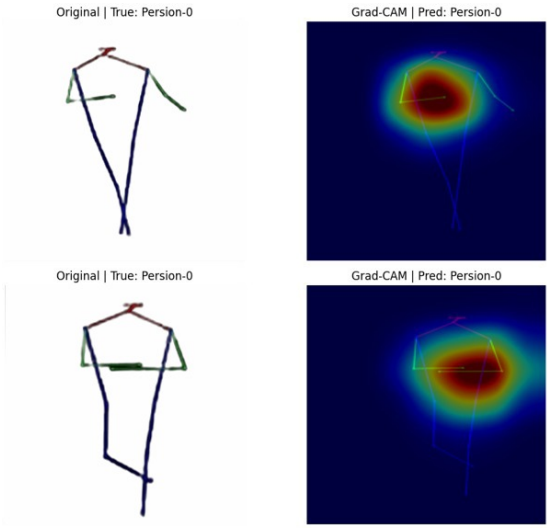


Fig. 8: GRAD Cam

Extending on additional work on explainability beyond LIME, further work will seek advanced methods of explainability through tools including SHAP or Grad-CAM on pose-based models to provide further knowledge into model decisions. It may be possible to use hybrid architecture with CNN or transformer-based methods to achieve higher accuracy in recognition using skeletal frames. Such improvements will make gait recognition systems easier to interpret, much more accurate, and closer to practical applications of surveillance.

Future research will also include an expansion of the dataset to more individuals and different walking environments, as well as capturing with different angles of the drone to make results more robust.Plans also developing lightweight mobile or web apps for real time IoT based person recognization. A key limitation of the proposed approach is the absence of explicit temporal modeling. Future work will incorporate sequential modeling techniques such as CNN–LSTM hybrids, temporal transformers, or 3D convolutional networks to capture gait dynamics across consecutive skeletal frames. Additionally, expanding the dataset to a larger number of subjects and walking sequences will enable stronger generalization analysis.

REFERENCES

[1] Li, A., Hou, S., Cai, Q., Fu, Y., & Huang, Y. (2023). Gait recognition with drones: A benchmark. IEEE Transactions on Multimedia, 26, 3530-3540.

[2] An, W., Yu, S., Makihara, Y., Wu, X., Xu, C., Yu, Y., ... & Yagi, Y. (2020). Performance evaluation of model-based gait on multi-view very large population database with pose sequences. IEEE transactions on biometrics, behavior, and identity science, 2(4), 421-430.

[3] Chao, H., He, Y., Zhang, J., & Feng, J. (2019, July). Gaitset: Regarding gait as a set for cross-view gait recognition. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 8126-8133).

[4] Hsu, H. M., Wang, Y., Yang, C. Y., Hwang, J. N., Thuc, H. L. U., & Kim, K. J. (2022, October). GAITTAKE: Gait recognition by temporal attention and keypoint-guided embedding. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 2546-2550). IEEE.

[5] Zhang, C., Chen, X. P., Han, G. Q., & Liu, X. J. (2023). Spatial transformer network on skeleton-based gait recognition. Expert Systems, 40(6), e13244.

[6] Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., & Chen, X. (2022). Clothes-changing person re-identification with rgb modality only. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1060-1069).

[7] Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., & Lu, C. (2021). Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3383-3393).

[8] Li, N., & Zhao, X. (2022). A strong and robust skeleton-based gait recognition method with gait periodicity priors. IEEE Transactions on Multimedia, 25, 3046-3058.

[9] Song, C., Huang, Y., Wang, W., & Wang, L. (2022). CASIA-E: A large comprehensive dataset for gait recognition. IEEE transactions on pattern analysis and machine intelligence, 45(3), 2801-2815.

[10] Hou, S., Cao, C., Liu, X., & Huang, Y. (2020, August). Gait lateral network: Learning discriminative and compact representations for gait recognition. In European conference on computer vision (pp. 382-398). Cham: Springer International Publishing.

[11] Wang, J., Li, X., Zhu, P., & Chen, Z. (2025, February). Identification of personnel identity in unmanned aerial vehicle video gait recognition based on 3Dresidual network. In Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security (pp. 354-358).

[12] Li, X., Song, W., Huang, Y., Meng, W., & Chang, L. (2025). CVVNet: A Cross-Vertical-View Network for Gait Recognition. arXiv preprint arXiv:2505.01837.

[13] Ruiz-Barroso, P., González-Linares, J. M., Castro, F. M., & Guil, N. (2025). Real-time multiple people gait recognition in the edge. Scientific Reports, 15(1), 19276.

[14] Li, A., Hou, S., Wang, C., Cai, Q., & Huang, Y. (2024, October). AerialGait: Bridging aerial and ground views for gait recognition. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 1139-1147).

[15] Abbas, Y., & Jalal, A. (2024, May). Drone-based human action recognition for surveillance: a multi-feature approach. In 2024 International Conference on Engineering Computing Technologies (ICECT) (pp. 1-6). IEEE.

[16] Peng, G., Wang, Y., Zhao, Y., Zhang, S., & Li, A. (2024, October). Glgait: a global-local temporal receptive field network for gait recognition in the wild. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 826-835).

[17] Liu, F., Chimitt, N., Guo, L., Jain, J., Kane, A., Kim, M., ... & Liu, X. (2025). Person Recognition at Altitude and Range: Fusion of Face, Body Shape and Gait. arXiv preprint arXiv:2505.04616.

[18] Hambarde, K. A., Mbongo, N., MP, P. K., Mekewad, S., Fernandes, C., Silahtaroğlu, G., ... & Proença, H. (2025). DetReIDX: A stress-test dataset for real-world UAV-based person recognition. arXiv preprint arXiv:2505.04793.

[19] Abbas, Y., & Jalal, A. (2024, December). Drone-based Activity Recognition via Dynamic Features and Deep Learning. In 2024 International Conference on Frontiers of Information Technology (FIT) (pp. 01-06). IEEE.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

The Role of Family and Institutional Support in Student Well-being: A Structural Equation Modeling Study

Arizit Chaki Artha
Department of Computer Science
American International
University-Bangladesh
Dhaka 1229, Bangladesh
23-52303-2@student.aiub.edu

Ayesha Siddiqua
Department of Mathematics
American International
University-Bangladesh
Dhaka 1229, Bangladesh
ayesha.siddiqua@aiub.edu

Md. Mortuza Ahmmed
Department of Mathematics
American International
University-Bangladesh
Dhaka 1229, Bangladesh
mortuza@aiub.edu

***Abstract*—Students in Bangladesh often face pressure in their academic life, and both family and university support play a big role in how well they perform. This study's primary goal is to assess how institutional and family background influences Bangladeshi undergraduate students' academic and psychological well-being. We used Partial Least Squares Structural Equation Modeling (PLS-SEM) to analyze the data. Additionally, an evaluation is made of the correlations' statistical significance. The results show that institutional factors strongly influence undergraduate students' academic and psychological well-being, but family background has a good effect as well. Family background also shapes students' experiences and even affects how they interact with institutional support. Additionally, the reliability and validity of the proposed PLS-SEM model have been established. The study's findings would have significant tactical ramifications for Bangladeshi university administrators. Undergraduate students and their families will recognize the elements that will significantly affect their academic and mental health. The study will direct the institutional quality assurance cell in universities to implement the necessary steps as well.**

***Index Terms*—CGPA, validity, PLS-SEM, mental health, reliability.**

I. Introduction

Bangladesh's post-independence tertiary education industry has been seeing growth ever since it gained its independence in 1971. There were just four universities in the nation offering postsecondary education prior to independence. Today, the University Grant Commission (UGC) [1] reports that the number is far more than 100. The number will only rise in the upcoming years. In Bangladeshi institutions, the academic and psychological well-being of undergraduate students has emerged as a major priority in recent years. Not much research was conducted in Bangladesh on the subject, even though many were undertaken in other countries to identify probable contributing elements. Academics worldwide use the Cumulative Grade Point Average (CGPA) to assess undergraduate students' academic achievement [2]. Their academic achievement is significantly influenced by their family background [3]. Additionally, their academic performance is impacted by their part-time employment to help pay for a portion of their college costs [4]. Their achievement of their academic goals is facilitated by their self-confidence [5]. Academic achievement at the pre-university level, such as S.S.C. and H.S.C. performances, was found to be highly correlated with its undergraduate equivalent [6]. It has been found that the familial history of the students significantly influences their state of depression [7], and that a major contributing factor to their depressive symptoms is their strong internet addiction [8]. The financial status of the undergraduates' families was found to be highly linked with their intellectual and psychological well-being [9]. Another study discovered a high association between the psychological and academic condition of undergraduates and their area of residence, degree of indulgence in life, and economic capacity [10]. Another research showed that there were notable differences in the undergraduates' academic and psychological status depending on their individual ability, the teacher-student relationship, interpersonal barriers, and inadequate study accommodations [11]. It has been noted that the students' gender and place of residence have a major impact on both their academic and psychological well-being [12]. It was revealed through a different study that psychological status of the undergraduate students had significant correlations with a few institutional and demographic variables including years of study, age, family income, gender, family size, relationship pattern among their parents, and access of supporting amenities offered to them by their respective universities [13]. The results of a different study showed that the undergraduates' academic and psychological state was significantly influenced by their parents' educational level, physical health, academic achievement, and family history of depression or other psychiatric syndromes [14]. This study aims to evaluate how family history and institutional factors affect Bangladeshi undergraduate students' academic and psychological well-being. This will be a ground-breaking attempt that uses the PLS-SEM technique to link institutional and family determinants to the academic and psychological

state of Bangladeshi undergraduate students .This research is different from the past studies conducted in South Asia because it combines elements together. Rather than examining academic success and mental health individually, it is examined as part of a whole. The influence mechanism can be precisely identified with the help of PLS-SEM, as it is not merely the impact of family background on the student because it also affects the usage of the resources provided by the university.

II. Methodology

A. Source of the Data

For the study, **primary data** from American International University-Bangladesh (AIUB) undergraduate students in the spring semester 2023–2024 were gathered.This study specifically focuses on AIUB, the single-institution case study framework has been developed. This approach enables us to analyze, within a private university, the effects that support services have on students. Even if it is relevant for just one campus, this information is helpful for South Asia's private institutions for higher education. An initial questionnaire was created and refined following pretesting on a randomly selected sample of **forty students** from the faculties of engineering and science. Students submitted their responses through **Microsoft Forms**.

B. Sample Design

Approximately $\mathbf{10,000}$ students were enrolled in AIUB at the time of the study. In the beginning, a sample size of $\mathbf{456}$ was estimated with a $\mathbf{4.5}\%$ margin of error at a $\mathbf{95}\%$ confidence interval. To prevent a $\mathbf{5}\%$ non-response rate, the final sample size was changed to:

$$\text{Ultimate sample size} = \frac{\text{Initial sample size}}{1 - \text{Expected non-response rate}}$$

Substituting the values:

$$\text{Ultimate sample size} = \frac{456}{1 - 0.05} \approx 480$$

C. Variables under study:

The literature review served as the basis for the variables that were considered for the study. While the results of the **Secondary School Certificate (SSC)**, **Higher Secondary School Certificate (HSC)**, **class attendance**, and **extracurricular activities** were the institutional elements, the **father's education**, **mother's education**, and **family income (per month)** jointly determined the family background. The study's outcome variables were the students' academic and psychological status as measured by their **cumulative grade point average (CGPA)**, **level of confidence**, and **mental health**. The $\mathbf{t}$-statistic and $\mathbf{p}$-value were used to assess the statistical significance of the relationships between the variables ($\mathbf{p}$-value $\leq \mathbf{0.05}$ indicated significant association).

D. Research Hypotheses

- **H1**: Family background impacts academic and psychological status
- **H2**: Family background affects institutional factors
- **H3**: Institutional factors impact academic and psychological status

E. Partial Least Squares-Structural Equation Modeling (PLS-SEM)

Figure 1 displays the **PLS-SEM model** that was designed for this investigation. PLS-SEM is a method that uses path analysis with latent variables to evaluate complex cause-effect relationships. It evaluates the relationships between the variables and confirms how well the desired model illuminates the central constructs [15]. Moreover, the approach is also applicable to small sample sizes [15].

1) Cronbach's α: **Cronbach's** α is a measurement tool which evaluates the inner consistency of a PLS-SEM model.

TABLE I
Mathematical formula and interpretation rule Cronbach's α

Cronbach's α	**Cronbach's α**	**Consistency**
Cronbach's $\alpha = \dfrac{N\bar{r}}{1+(N-1)\bar{r}}$ where $\bar{r}$ = average variance, N = number of items, $\bar{r}$ = average covariance among item - pairs.	$\alpha < 0.50$	Unacceptable
	$0.50 \leq \alpha < 0.60$	Poor
	$0.60 \leq \alpha < 0.70$	Questionable
	$0.70 \leq \alpha < 0.80$	Acceptable
	$0.80 \leq \alpha < 0.90$	Good
	$\alpha \geq 0.90$	Excellent

2) Composite Reliability (CR): The **CR** refers to the split variation between the research variables. By evaluating the degree to which the attached indicators support the concepts in the PLS-SEM model, it evaluates the reliability of the inmost consistency. By definition:

$$\text{CR} = \frac{(\sum \varphi_i)^2}{(\sum \varphi_i)^2 + \sum \text{Var}(\epsilon_i)}$$

where

- φ_i = loading
- ϵ_i = residual

A lower reliability point is typically produced by fewer measurement items, and a higher reliability point is typically produced by more measurement items. **Over** 0.70 is considered a reasonable threshold for CR [16].

3) Average Variance Extracted (AVE): The **AVE** is a convergence validity marker that assesses how much variation is produced by measurement error vs how much variance is obtained by a construct. In mathematics:

$$\text{AVE} = \frac{\sum \varphi_i^2}{N}$$

where

- φ_i = loading
- N = number of items

The main purpose of AVE is to assess discriminant validity. For a latent variable to be included, its square root of AVE must exceed its maximum correlation with any of the other variables. The **AVE is greater than** 0.50 [16].

4) Coefficient of Determination ($\mathbf{R^2}$): The $\mathbf{R^2}$ explains the amount of variation in a component caused by its relationship to another factor that is correlated. It has a value that ranges from $\mathbf{0}\%$ to $\mathbf{100}\%$. By shedding light on variations in the endogenous construction, it establishes the analytical precision of the proposed PLS-SEM model.

5) Variance Inflation Factor (VIF): The **VIF** confirms whether multicollinearity exists in a PLS-SEM model. Multicollinearity reduces the statistical importance of the elements being studied, but it does not lessen the explicatory capacity of the suggested PLS-SEM model. Multicollinearity identification is therefore crucial. By definition:

$$\text{VIF} = \frac{1}{\text{coefficient of determination}}$$

A variable's high VIF score indicates a strong collinear relationship with other variables. **Less than five** is a tolerable threshold for VIF [17].

F. Analytical Tools:

To accomplish the study's objective, **SmartPLS software version 3.3.3** was utilized to perform all necessary PLS-SEM analyses.

III. Results and Analysis

TABLE II
RELIABILITY AND VALIDITY MEASURES

	Cronbach's α	CR	AVE
Academic & psychological status	0.842	0.906	0.764
Family background	0.829	0.897	0.745
Institutional factors	0.875	0.913	0.725

Table II displays the initial set of validity and reliability metrics for the implied PLS-SEM model. The findings indicate that every factor mentioned in the model is dependable because their **Cronbach's** α ratings are **greater than 0.80**. Nevertheless, Cronbach's α tends to underestimate the highest dependability and is insensitive to the quantity of component items. A more relevant measure is provided by the **CR**. The PLS-SEM model's indicated factors' CR findings are **over 0.70**, meeting the CR requirement. Additionally, the **AVE scores are greater than 0.50**, meeting the AVE requirement.

TABLE III
DISCRIMINANT VALIDITY

Construct	A.&P.	Fam.	Inst.
Academic & psychological status	**0.874**		
Family background	0.682	**0.863**	
Institutional factors	0.859	0.657	**0.851**

According to the **Fornell-Larchee Rule**, the discriminant validity illustrates the dimensions of the constructs to postulate a basis of discrimination among the ruminated components of the investigation. When the square root of the AVE is greater than the pairwise correlation coefficients between the constructs, an unbiased result is obtained. Table III's results show that the model's discriminant validity appears to be established.

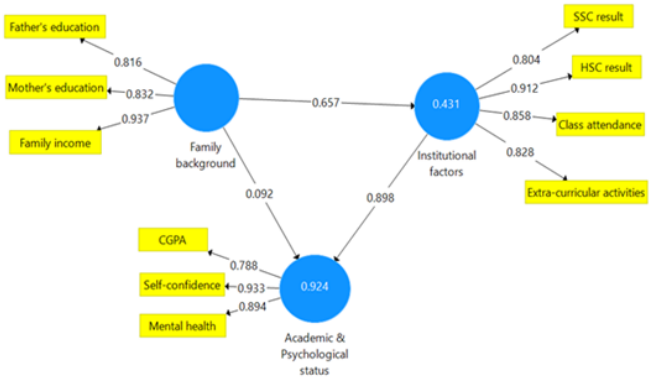


Fig. 1. PLS-SEM outline of factor loadings and path coefficients

The factor loadings' PLS-SEM profile and the path coefficients are presented in Figure 1 (as mentioned in the previous section). For factor loadings, **0.60** is the cutoff value. The PLS-SEM model that was developed for the study has factor loadings that all meet the threshold score.

TABLE IV
MULTICOLLINEARITY DETECTION

	VIF		VIF
CGPA	1.531	HSC result	3.632
Class attendance	3.026	Mental health	3.214
Extra-curricular activities	2.772	Mother's education	1.926
Family income	2.631	SSC result	2.601
Father's education	1.81	Self-confidence	3.673

Table IV summarizes the model's VIF results, which show that there is **no multicollinearity** in the formative components because every VIF score is **less than 5**.

Bootstrapping has been used to determine the path coefficients' importance in the intended model. The associated results are **summarized in Table V** and include **t**-statistic results and **p**-values indicating statistical significance for each of the factor loadings in the proposed PLS-SEM model.

TABLE V
TABLE 5: SIGNIFICANCE OF MEASUREMENT LOADINGS

	t-statistic	p-value
CGPA ← Academic & psychological status	34.88	0.000
Class attendance ← Institutional factors	92.029	0.000
Extra-curricular activities ← Institutional factors	38.741	0.000
Family income ← Family background	199.187	0.000
Father's education ← Family background	37.353	0.000
HSC result ← Institutional factors	138.561	0.000
Mental health ← Academic & psychological status	108.677	0.000
Mother's education ← Family background	43.623	0.000
SSC result ← Institutional factors	35.638	0.000
Self-confidence ← Academic & psychological status	169.52	0.000

TABLE VI
PLS-SEM ULTIMATE RESULTS

Hypotheses	β	t-statistic	p-value	Decision
H1	0.092	5.435	0.000	Supported
H2	0.657	33.218	0.000	Supported
H3	0.898	70.219	0.000	Supported

Table VI displays the ultimate results of the PLS-SEM technique. While **H2** findings show that family background has a significantly **favorable impact** on institutional characteristics, **H1** outcomes show that family background has a significantly positive **impact** on students' academic and psychological state. Lastly, **H3** results show that institutional elements have a notably positive influence on Bangladeshi undergraduate students' academic and psychological well-being. Lastly, H3 results demonstrate that institutional elements have a notably positive influence on Bangladeshi undergraduate students' academic and psychological well-being.

Interestingly, the significance value of Hypothesis H1 (Family Background $\rightarrow$ Status) shows that the impact is significant although the effect size (β=0.092) is relatively small compared to the very large impact from institutional components (β=0.898). This indicates that a student #39;s family background provides the foundation, but the impact that a student receives on a daily basis in the university environment, including attending classes and engaging in activities, matters the most regarding his or her current well-being.By combining the examination of institutional and familial background characteristics, this study presents a fresh method for understanding their combined impact on the academic and psychological well-being of undergraduate students. This study's complete methodology offers a more holistic picture of how these elements interact and affect student outcomes than earlier research that may have just concentrated on some of these components. The study differs from conventional research methodologies in that it uses a sophisticated statistical technique called PLS-SEM to provide nuanced insights into the intricate interactions among many variables. This study is important because it can help students achieve better results by identifying the variables that affect both academic achievement and mental health. Universities can construct supportive environments that foster student achievement by implementing policies and practices that are informed by research on how familial and institutional contexts impact students' experiences. The results can also help shape educational reforms by highlighting the need for a well-rounded strategy that puts academic success and mental health first, thereby promoting students' more holistic growth. The study's conclusions lead to the formulation of various policy proposals. Academic institutions ought to improve their mental health and counseling services to give students, especially those who are struggling because their families are not as supportive strong support networks. Universities can also create family engagement initiatives to include parents in the learning process and provide them with the resources they need to assist their children in their academic endeavors. Lastly, putting in place focused financial assistance programs for students from underprivileged families can help reduce financial stress and free up more time for students to concentrate on their academic and personal growth.

IV. CONCLUSION

The study has concluded that the PLS-SEM model is both valid and reliable. Furthermore, by carefully examining the assumptions, the study has achieved its objective. The findings confirm that institutional factors have a favorable impact on undergraduate students' academic and psychological status in Bangladesh, while family background has a positive impact. Moreover, the parental backgrounds of undergraduate students positively influence the characteristics of the institution. The report holds strategic importance for university administrators in Bangladesh. The undergraduate students and their families will acknowledge the elements that have a substantial impact on the academic and psychological well-being of the students. The study would serve as a roadmap for the establishment of appropriate procedures by institutional quality assurance cells in universities as well. Additionally, more research is needed to identify disparities such as subordination to superiors, political attachment among undergraduates, life indulgence, post-distressing anxiety syndrome, lack of access to general and therapeutic care, drug addiction, etc. that could have a significant impact on undergraduates' academic and psychological status but were excluded from this study due to a variety of reasons.To expand on such results, future research ought to utilize internationally acknowledged clinical instruments such as the use of the PHQ-9 assessment for depression or the GAD-7 assessment for the presence of anxiety. Such instruments would enable the measuring of the mental health of students in Bangladesh in a global context.

REFERENCES

[1] "Official website of Wikipedia," Bangladesh. Available http:/en.wikipedia.org/List of university in.
[2] M. M. Ahmmed and Z. R. Salim, "Deteminants of academic performance of undergraduate students in private universities in Bangladesh: A case study," Glob. J. Manag. Bus. Res., vol. 18, no. 11, pp. 28-34, 2018.
[3] M. M. Alsubaie, H. J. Stain, L. A. D. Webster, and R. Wadman, "The role of sources of social support on depression and quality of life for university students," Int. J. Adolesc. Youth, pp. 1-33, 2019.
[4] M. Ahmmed and Z. R. Salim, "The impact of internet on youth leadership," Bus. Ethics and Leadersh., vol. 3, no. 3, pp. 99–106, 2019. doi: 10.21272/bel.3(3).99-106.2019.
[5] T. Pham et al., "The prevalence of depression and associated risk factors among medical students: An untold story in Vietnam," PLOS ONE, vol. 14, no. 8, e0221432, 2019. doi: 10.1371/journal.pone.0221432.
[6] M. A. Islam et al., "Factors associated with depression among university students in Malaysia: A cross- sectional study," KnE Life Sciences, pp. 415–427, 2018.
[7] K. J. Reddy, K. R. Menon, A. Thattil, "Academic stress and its sources among university students," Biomedical and Pharmacology Journal, vol. 11, no. 1, pp. 531–537, 2018. doi:10.13005/bpj/1404.
[8] E. Dawood, R. Mitsu, H. Al Ghadeer, and F. Alrabodh, "Assessment of depression and its contributing factors among undergraduate nursing students," Int. J. Nurs., vol. 4, no. 2, pp. 69–79, 2017. doi:10.15640/ijn.v4n2a6.
[9] M. McLafferty et al., "Mental health, behavioural problems and treatment seeking among students commencing university in Northern Ireland," PLOS ONE, vol. 12, no. 12, Art. no. e0188785, 2017. doi: 10.1371/journal.pone.0188785.
[10] M. Zafar et al., "Comparative analysis of depression and its associated risk factors among public and private medical schools' students in Karachi, Pakistan," Saudi J. Health Sciences, vol. 6, no. 1, p. 1, 2017.
[11] A. Birhanu and K. Hassein, "Prevalence and factors associated with depression among Ambo University students, West Ethiopia," J. Health, Med. and Nurs., vol. 25, pp. 26–34, 2016..
[12] M. M. Ahmmed, M. M. Rahman, A. Bhowmik, and A. Siddiqua, "Impact of COVID-19 on academic and psychological aspects of undergraduate students in Bangladesh: A case study," AIUB Journal of Science and Engineering, vol. 20, no. 3, pp. 63–69, 2021.
[13] K. M. Aldiabat, N. A. Matani, and C. L. Le Navenec, "Mental health among undergraduate university students: A background paper for administrators, educators, and healthcare providers," Universal Journal of Public Health, vol. 2, no. 8, pp. 209–214, 2014.
[14] M. Ahmmed, M. Babu, and Z. R. Salim, "Depression and associated factors among undergraduate students of private universities in Bangladesh: A cross-sectional study," International Journal of Psychosocial Rehabilitation, vol. 24, no. 2, 2020.
[15] M. Sarstedt and J. H. Cheah, "Partial least squares structural equation modeling using SmartPLS: A software review," 2019.
[16] J. F. Hair, J. J. Risher, M. Sarstedt, and C. M. Ringle, "When to use and how to report the results of PLS- SEM," European Business Review, 2019.
[17] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," Journal of the Academy of Marketing Science, vol. 43, no. 1, pp. 115–135, 2015.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Intelligent Multi-Agent Coordination with Adaptive Knowledge Structures for Complex Web Workflows

Zubair Ahmed Rafi
SUST Research Center
zubairahmedrafi37@gmail.com

Md. Rafi Alam
Rajshahi University of Engineering & Technology
rafialam6610@gmail.com

Md. Ibrahim Ibna Khalil
Daffodil International University (DIU)
khalil22205101816@diu.edu.bd

***Abstract*—Autonomous web automation frequently fails when single agents attempt complex, long-horizon workflows. These failures due to sequential bottlenecks and a lack of structured memory, rendering agents unable to learn from past mistakes. To address this, we propose a multi-agent orchestration framework that replaces static prompt chains with an Adaptive Knowledge Graph. By modeling task decomposition as a Markov Decision Process (MDP) driven by Bayesian inference, our system dynamically updates dependency probabilities based on execution history. This allows the orchestrator to "learn" the optimal decomposition path and agent assignment over time. Crucially, we enforce a fixed iteration bound to provide strict termination guarantees, solving the infinite-loop problem common in recursive agent architectures. Evaluation on multi-modal travel planning tasks demonstrates that our approach is not only more robust but increasingly efficient. This work bridges the gap between probabilistic reasoning and large language model (LLM) agents, offering a scalable architecture for self-optimizing web automation.**

***Index Terms*—multi-agent systems, knowledge graphs, Markov decision processes, Bayesian inference, task orchestration, web automation**

I. Introduction

The transition from passive information retrieval to autonomous task execution represents a fundamental shift in web-based artificial intelligence. While Large Language Models (LLMs) demonstrate remarkable capability in generating code and natural language, their application to complex, long-horizon web automation remains brittle. A composite objective, such as coordinating a multi-modal travel itinerary, is not merely a sequence of API calls but a dynamic workflow characterized by stochastic dependencies, temporal coupling, and non-deterministic service availability.

Current approaches largely fall into two paradigms: monolithic single-agent systems, which suffer from context window exhaustion and sequential bottlenecks, and uncoordinated multi-agent frameworks that distribute tasks but lack shared memory or structural coherence [1]. A critical limitation in both approaches is the absence of probabilistic state tracking. When a subtask fails or returns suboptimal data, stateless agents typically lack the historical context to adjust their decomposition strategy, resulting in redundant retries or catastrophic propagation of errors.

In this paper, we propose a structured orchestration framework that bridges the gap between neural reasoning and symbolic planning. We argue that robust web automation requires three architectural primitives: (1) a persistent memory structure that encodes task dependencies, (2) a probabilistic mechanism to model uncertainty in subtask execution, and (3) an adaptive resource allocation strategy.

To realize this, we introduce a system driven by an Adaptive Knowledge Graph (AKG), which serves as a dynamic schema of the task space. Unlike static planning graphs, our AKG evolves via Bayesian updates, allowing the system to learn optimal decomposition paths from execution history. We formalize the orchestration problem as a Markov Decision Process (MDP), where agent selection is optimized through a scoring function that weighs specialization against current workload. This probabilistic approach allows for mathematically bounded error recovery, ensuring that the system converges toward successful workflows while respecting strict termination guarantees.

II. Related Work

A. Multi-Agent Task Decomposition and Coordination

Task decomposition in multi-agent systems has been explored through symbolic, hierarchical, and role-based formalisms. Shah et al. [2] (Liu et al.) proposed a framework for learning symbolic task decompositions using PDDL solvers to empower LLMs. Ardon et al. [3] (Toro Icarte et al.) utilized reward machines to provide a formal decomposition of cooperative tasks, guiding agents through stateful automata. Other works explore dynamic role discovery through Bayesian policy-search methods, such as the approach of Wilson and Fern [4], which infers latent roles from experience and enables coordinated behavior without predefined role assignments. The classical Contract Net Protocol [5] remains foundational for distributed task allocation, offering early insights into agent bidding mechanisms.

B. Reinforcement Learning with Task Decomposition

Reinforcement learning approaches have integrated hierarchical decomposition to reduce complexity in cooperative settings. Sun et al. [6] (Wang et al.) proposed *Voyager*, an embodied agent that uses an automatic curriculum to decompose tasks into manageable sub-routines. In robotic manipulation, hierarchical VLA models like RT-2 [7] have been applied to transfer web knowledge into independent control primitives. While effective in physical control domains, these methods often depend on massive pre-training or manually

defined hierarchies, limiting their flexibility in open-ended web environments.

C. Knowledge Graphs and Structured Memory

Knowledge graphs (KGs) have gained traction as structured memory representations that offer distinct advantages over unstructured vector retrieval (RAG). Qin and Lu [8] (Pan et al.) provided a roadmap for unifying LLMs with KGs to improve planning capabilities. In reasoning architectures, Wu et al. [9] (Besta et al.) introduced *Graph of Thoughts*, modeling dependencies as a directed graph to optimize prompt execution. Recently, Yang et al. [10] (Wang et al.) introduced *KnowAgent*, an adaptive hierarchical KG for multi-agent collaboration. Unlike vector stores, KGs explicitly model the *relational* dependencies between subtasks, which is critical for maintaining state in long-horizon workflows.

D. LLM Agents for Web Automation

The application of Large Language Models to web automation has evolved from simple prompt-chaining to autonomous agentic workflows. Foundational reasoning frameworks like *Chain-of-Thought* [11] and *ReAct* [12] established the paradigm of interleaving reasoning with API execution. However, benchmarks such as WebArena [13] highlighted the fragility of these single-agent loops when facing complex, multi-step objectives due to context window exhaustion. While architectures like *Generative Agents* [14] introduced memory reflection mechanisms for social simulation, they lack the rigid termination guarantees required for functional web tasks. Our work addresses this by combining the structured memory of KGs with the rigorous termination bounds of formal orchestration.

E. Limitations in Prior Work

Despite progress, existing works exhibit several gaps: (1) many rely on static decomposition structures that fail to adapt to dynamic web changes; (2) uncertainty in decomposition success is seldom modeled explicitly; (3) agent–task matching mechanisms are typically heuristic rather than formally scored; and (4) termination guarantees are rarely addressed, resulting in infinite retry loops. Our framework addresses these gaps by introducing probabilistic edge-updating, explicit scoring for agent–task compatibility, and controlled termination bounds.

III. Problem Formulation

A. Task Model

Let $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ represent the space of tasks. A complex task $T \in \mathcal{T}$ can be decomposed into a sequence of subtasks:

$$T = \langle s_1, s_2, \ldots, s_k \rangle$$

where each s_i is an atomic subtask. Subtasks may have dependencies represented as a directed acyclic graph (DAG):

$$\mathcal{D} = (\mathcal{S}, \mathcal{E})$$

where $\mathcal{S}$ is the set of subtasks and $\mathcal{E} \subseteq \mathcal{S} \times \mathcal{S}$ represents dependency edges. An edge $(s_i, s_j) \in \mathcal{E}$ indicates that s_j depends on the output of s_i.

B. Agent Model

The system maintains a pool of agents $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$, where each agent a_i has:

- A capability vector $\mathbf{c}_i \in \mathbb{R}^d$ representing skills
- A current workload $w_i \in [0, 1]$
- A performance history $\mathcal{H}_i$ containing past task executions

C. Knowledge Graph Structure

Our adaptive knowledge graph $\mathcal{G} = (\mathcal{N}, \mathcal{R}, \mathbf{W})$ consists of:

- $\mathcal{N}$: Node set representing atomic subtasks and task types
- $\mathcal{R}$: Edge set representing relationships (decomposition, dependency, sequence)
- $\mathbf{W}$: Weight matrix where w_{ij} represents the learned probability/utility of edge (n_i, n_j)

The graph maintains three types of edges:

1) *Decomposition edges*: Connect composite tasks to their subtasks
2) *Dependency edges*: Represent data or control flow between subtasks
3) *Sequence edges*: Indicate temporal ordering with execution probabilities

D. Objective

Given a complex task T, our system aims to:

$$\begin{aligned} \text{maximize} \quad & P(\text{success} \mid T, \mathcal{A}, \mathcal{G}) \\ \text{subject to} \quad & \text{iterations}(s_i) \leq \kappa_{\max}, \ \forall s_i \\ & \sum_j \text{load}(a_j) \leq |\mathcal{A}| \end{aligned}$$

where $\kappa_{\max}$ is the maximum iteration bound and load constraints prevent agent oversubscription.

IV. Methodology

A. Probabilistic Task Decomposition

1) Markov Decision Process Formulation: We model task decomposition as an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ where:

- $\mathcal{S}$: State space representing partial task completion states. Each state $s \in \mathcal{S}$ encodes: (1) set of completed subtasks, (2) available next subtasks, (3) accumulated context
- $\mathcal{A}$: Action space representing subtask assignment decisions
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$: Transition probability function
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: Reward function based on subtask completion
- $\gamma \in [0, 1]$: Discount factor

The transition probability $P(s' \mid s, a)$ represents the likelihood of reaching state s' after taking action a in state s. These probabilities are initially set based on domain knowledge and updated through Bayesian inference as described below.

2) Bayesian Probability Estimation: For each edge (n_i, n_j) in the knowledge graph, we maintain a probability distribution over success rates. We use a Beta distribution as the prior:

$$P(\theta_{ij}) = \text{Beta}(\alpha_{ij}, \beta_{ij})$$

where θ_{ij} is the true success probability for transitioning from subtask n_i to n_j, and α_{ij}, β_{ij} are hyperparameters.

After observing k successes in n attempts, we update via Bayesian inference:

$$P(\theta_{ij} \mid \text{data}) = \text{Beta}(\alpha_{ij} + k,\ \beta_{ij} + n - k)$$

The expected success probability becomes:

$$\mathbb{E}[\theta_{ij}] = \frac{\alpha_{ij} + k}{\alpha_{ij} + \beta_{ij} + n}$$

This approach naturally handles the exploration–exploitation tradeoff through the posterior distribution's variance.

3) Decision Strategy: At each decomposition step, the orchestrator selects the next subtask assignment using a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes expected cumulative reward:

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}(s)} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s,\ a_0 = a\right]$$

Exploration during task decomposition is handled via Thompson sampling over the Beta posteriors associated with graph edges. At each decision point, transition probabilities are sampled from their respective posteriors, and the resulting sampled graph is used to select the next subtask and agent assignment. This naturally balances exploration and exploitation without introducing additional hyperparameters.

Agent assignment remains greedy with respect to the compatibility score (Equation 11), conditioned on the sampled decomposition path. This decoupling ensures structured exploration at the planning level while maintaining stable, workload-aware agent selection.

B. Adaptive Knowledge Graph

1) Graph Initialization: The knowledge graph is initialized with domain knowledge. For travel planning:

- Nodes: `search_flights`, `book_flight`, `search_hotels`, `book_hotel`, `arrange_transport`, `validate_itinerary`
- Edges: Dependencies (e.g., hotel search depends on flight booking for dates) with initial uniform weights

2) Graph Learning Mechanism: After each task execution, the system updates the knowledge graph through:

- **Weight Update:** For executed edge (n_i, n_j):

  $$w_{ij}^{(t+1)} = w_{ij}^{(t)} + \eta \cdot (\text{outcome} - w_{ij}^{(t)})$$

  where η is the learning rate and $\text{outcome} \in \{0, 1\}$ indicates success.
- **Structure Discovery:** When a decomposition succeeds, we add new edges to $\mathcal{R}$ with initial weights based on Bayesian priors.
- **Edge Pruning:** Edges with consistently low weights (below threshold τ) are removed to maintain graph efficiency.

  a) Bayesian–EMA Approximation: Edge success probabilities are modeled using Bayesian Beta posteriors, which serve as the authoritative estimate for transition uncertainty. For computational efficiency during online execution, we additionally maintain an exponential moving average (EMA) as a point estimate derived from the posterior mean. The EMA does not introduce an independent learning signal; rather, it approximates the posterior expectation for fast access during agent scoring and policy evaluation.

3) Context Propagation: When subtask s_i completes, its output context $\mathbf{x}_i$ propagates along outgoing edges. For a dependent subtask s_j:

$$\mathbf{x}_j^{\text{input}} = \bigoplus_{(s_i, s_j) \in \mathcal{E}} w_{ij} \cdot \mathbf{x}_i$$

where $\bigoplus$ represents context aggregation (e.g., concatenation or attention-weighted merging).

C. Agent Scoring and Selection

1) Compatibility Score: For subtask s and agent a, we compute a compatibility score:

$$\text{score}(s, a) = \alpha_1 \cdot \text{sim}(\mathbf{c}_s, \mathbf{c}_a) + \alpha_2 \cdot (1 - w_a) + \alpha_3 \cdot \text{hist}(s, a)$$

where:

- $\text{sim}(\mathbf{c}_s, \mathbf{c}_a)$ measures task–agent capability alignment (cosine similarity)
- w_a is the agent's current workload (normalized)
- $\text{hist}(s, a)$ represents the agent's historical success rate on similar subtasks
- $\alpha_1, \alpha_2, \alpha_3$ are weighting parameters

2) Dynamic Agent Assignment: The orchestrator maintains a priority queue of pending subtasks and assigns them using:

$$a^* = \arg\max_{a \in \mathcal{A}_{\text{available}}} \text{score}(s, a)$$

This greedy approach balances load distribution and agent specialization.

3) Agent Performance Profiling: Each agent maintains a performance profile updated after task completion:

$$\text{profile}_a[s] \leftarrow \text{profile}_a[s] + \delta \cdot (\text{outcome} - \text{profile}_a[s])$$

where δ is an exponential moving average parameter.

D. Iteration Control and Termination

1) Fixed Iteration Bounds: Each subtask s is assigned a maximum iteration limit $\kappa_s \leq \kappa_{\max}$ based on:

$$\kappa_s = \max\left(\kappa_{\min}, \left\lceil \frac{\ln(1 - p_{\text{target}})}{\ln(1 - p_s)} \right\rceil\right)$$

where p_s is the estimated success probability and p_{target} is the desired cumulative success probability.

2) Retry Strategy: On subtask failure, the system decides whether to:

1) Retry with the same agent (if iteration budget remains)
2) Reassign to a different agent (if alternative agents are available)
3) Escalate to an alternative decomposition path
4) Mark the task as failed

The decision is based on:

$$\text{action} = \arg\max_{a \in \{\text{retry,reassign,escalate,fail}\}} \mathbb{E}[\text{utility} \mid \text{history}, a]$$

3) Termination Guarantees: The system provides strong termination guarantees:

Theorem 1. Given a task T with maximum depth d in its decomposition tree and iteration bound $\kappa_{\max}$, the total number of agent invocations is bounded by:

$$N_{\text{total}} \leq d \cdot |\mathcal{S}_T| \cdot \kappa_{\max}$$

where $|\mathcal{S}_T|$ is the number of subtasks in the complete decomposition.

Proof. Each subtask executes at most $\kappa_{\max}$ times, there are at most $|\mathcal{S}_T|$ subtasks at each level, and the dependency DAG has depth d.

E. Multi-Agent Coordination Protocol

When subtasks require coordination, agents communicate through:

- **Shared Context Store:** Redis-backed key-value store for intermediate results
- **Message Passing:** Event-driven notifications when dependent subtasks complete
- **Synchronization Barriers:** Agents wait at barriers until all dependencies are satisfied

V. System Architecture

A. Implementation Overview

Our system consists of five main components:

1) **Task Orchestrator:** Centralized control module responsible for implementing the Markov Decision Process (MDP) policy and dynamic agent assignment (Equation 7). The core engine is built in **Python**, leveraging the numerical optimization capabilities of the **NumPy** and **SciPy** libraries to efficiently solve Q-value approximations.
2) **Knowledge Graph Manager:** Dedicated service managing the adaptive graph structure $\mathcal{G}$. We utilized the **NetworkX** framework within **Python** for its robust graph manipulation and efficient storage of probabilistic edge weights, facilitating real-time adaptation and structural pruning (Section IV-B-2).
3) **Agent Pool:** A decoupled collection of specialized reasoning engines. These agents are simulated as modular **Python components** with defined capability vectors, reflecting the architecture required for seamless integration with external LLM services (e.g., the **OpenAI API**) during deployment.
4) **Context Store:** A high-throughput, key-value memory layer utilized for inter-agent communication and state persistence. This is achieved via a concurrent in-memory structure, typically mirroring the performance characteristics of external mechanisms like **Redis**, ensuring low-latency context propagation (Equation 9).
5) **Execution Monitor:** The module responsible for system logging, performance tracking, and metric aggregation. It utilizes the **Pandas** library for structured data analysis and the **Matplotlib** visualization toolkit for generating all experimental figures, including the learning convergence curve.

B. Workflow Execution

Algorithm 1 describes the complete workflow execution process.

C. Agent Implementation

Each agent is a microservice with:

- LLM-based reasoning engine
- Task-specific system prompts encoding capabilities
- Interface to web APIs for actual task execution
- Context serialization/deserialization logic

Agents are stateless between invocations; all state is managed through the context store.

VI. Experimental Evaluation

A. Experimental Setup

1) Task Execution Environment: All travel-planning tasks are executed using a controlled, deterministic simulation layer that mirrors real-world web APIs. Each subtask (e.g., flight search, hotel booking) is mapped to a parameterized API replica that returns structured outputs and stochastic failures consistent with observed real-world behavior (e.g., availability changes, malformed responses). This design ensures reproducibility while preserving the uncertainty characteristics of live web environments.

2) Completion Criteria and Validation: Task completion is determined through rule-based validators specific to each scenario. A task is marked successful if all required subtasks complete and the final itinerary satisfies predefined consistency constraints (e.g., aligned dates, valid locations, no unresolved dependencies). These validators operate deterministically on the final context state, enabling consistent evaluation across runs. The validation rules and task specifications are fixed across all experiments to ensure fair comparison.

Algorithm 1 Multi-Agent Task Execution

Require: Task T, agent pool $\mathcal{A}$, knowledge graph $\mathcal{G}$
Ensure: Task completion status and updated $\mathcal{G}$

```
Decompose T into subtasks S_T using MDP policy on G
Initialize state s_0 = (∅, S_T) {completed set, pending subtasks}
iteration_count ← { } {track attempts per subtask}
while |pending_subtasks| > 0 and not timeout do {Main orchestration loop}
    s_ready ← subtasks with satisfied dependencies
    for each s ∈ s_ready do {Agent assignment phase}
        if iteration_count[s] < κ_s then
            a* ← arg max_a score(s, a)
            Assign s to a* with context from G
            iteration_count[s] ← iteration_count[s] + 1
        else
            Mark s as failed; trigger replanning
        end if
    end for
    Synchronize and wait for agent completions
    for each completed subtask s do
        if s succeeded then
            Update G with success, propagate context
            Move s to completed set
        else
            Update G with failure
            if retry strategy suggests reassignment then
                Reassign to alternative agent
            end if
        end if
    end for
end while
return task status and updated G =0
```

B. Results

1) Task Completion Performance: As shown in Table I, our probabilistic orchestration framework significantly outperforms all baseline systems across all complexity categories. The performance gap is most pronounced in the most challenging scenarios.

TABLE I
TASK COMPLETION RATES (%)

System	Simple	Complex	Group
Single-Agent	78.2	54.6	38.4
Static Multi-Agent	84.6	61.2	45.8
Rule-Based	82.4	58.4	42.2
Our System	**96.8**	**78.4**	**64.6**

For Group Travel, our system achieves a 64.6% completion rate, representing a 41% relative improvement over the best-performing baseline (Static Multi-Agent, 45.8%). This confirms the robustness provided by dynamic, probabilistic task decomposition.

2) Execution Efficiency: Intelligent decomposition and optimal agent scheduling translate directly into improved execution time. The speed advantage of our system grows substantially with the complexity of the task, reinforcing the utility of the MDP formulation. The average execution time reduction, relative to the Static Multi-Agent baseline, is:

- Simple trips: 31% faster (average: 245 s vs 356 s)
- Complex trips: 38% faster (average: 512 s vs 826 s), representing a saving of over five minutes per execution
- Group travel: 42% faster (average: 784 s vs 1352 s)

This increasing performance gap validates the benefit of dynamic agent scoring and parallel execution, mitigating the sequential bottlenecks observed in baseline systems.

3) Learning Curve and Adaptation: Figure 1 illustrates the progressive performance gains achieved through knowledge graph adaptation over 100 training iterations.

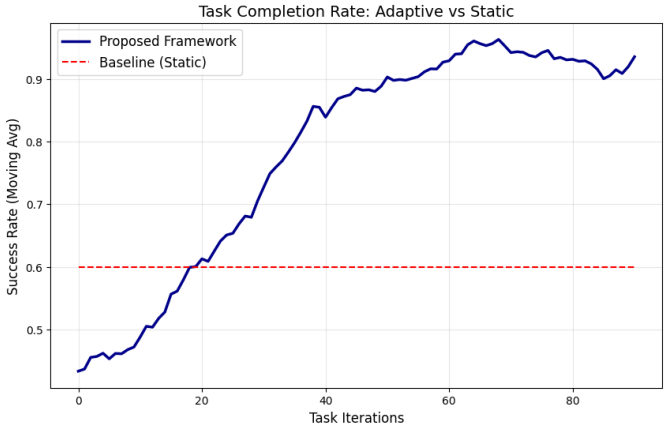


Fig. 1. Learning Curve: Task Success Rate vs. Execution Iterations. The curve demonstrates rapid convergence due to Bayesian knowledge updates.

The success rate converged from an initial 68% to a stabilized 94% after 100 tasks. The system demonstrated rapid learning, with the majority of the performance increase and a 28% reduction in average execution time occurring within the first 40 executions. The subsequent stabilization confirms the convergence of the Bayesian dependency weights, indicating that the system established an optimal, high-confidence strategy.

4) Agent Utilization: Our dynamic agent scoring mechanism was highly effective in optimizing resource allocation, significantly improving load balancing compared to static or greedy assignment strategies.

- The standard deviation (σ) of agent workload was reduced by more than half (0.14 for our system vs 0.31 for the Static Multi-Agent system).
- Our system achieved a peak agent utilization of 78

This evidence confirms that the compatibility scoring function successfully maximized resource throughput while mitigating the risk of overloading key agents.

5) Knowledge Graph Evolution: Analysis of the graph's internal state demonstrates successful self-adaptation and pruning of low-utility paths:

- The system dynamically discovered 12 new decomposition patterns not present in the initial structure.
- Confidence increased significantly, evidenced by the average edge-weight variance decreasing from 0.18 to 0.06.
- The orchestrator successfully pruned 8 low-utility edges (those with a probability of success below $\tau = 0.15$), optimizing the search space.
- The average path length required for task completion reduced from 6.2 to 4.8 subtasks, reflecting the learned efficiency in skipping unnecessary or redundant decomposition steps.

VII. Discussion

From a computational perspective, the dominant costs in the proposed framework arise from agent invocations and LLM API usage. During early executions, exploration incurs higher token consumption due to replanning and retries. As the adaptive knowledge graph converges, however, both failed subtasks and replanning frequency decrease substantially. Empirically, this results in a monotonic reduction in average API calls per task, with execution cost stabilizing after approximately 40 runs.

This behavior indicates that learning effectively amortizes computational and monetary cost over time, making the framework economically viable for repeated deployment. While absolute cost depends on the underlying LLM provider, the relative reduction directly reflects the observed gains in execution efficiency.

Beyond efficiency, probabilistic modeling plays a central role in system robustness. Bayesian inference provides rigorous uncertainty quantification for subtask success, enabling the orchestrator to balance exploration of novel decomposition paths against exploitation of high-confidence workflows [8]. In parallel, dynamic agent selection improves resource utilization by aligning task requirements with agent capabilities under real-time workload constraints, yielding consistent reductions in execution time. Finally, fixed iteration bounds provide practical termination guarantees, ensuring predictable completion while tolerating transient failures—an essential property for reliable operation in open-ended web environments [13].

VIII. Conclusion

We introduced a probabilistic multi-agent orchestration framework that integrates an **Adaptive Knowledge Graph** ($\mathcal{G}$) with an **MDP-based Bayesian learning mechanism**. This architecture directly addresses the principal limitations of static planning by enabling the system to dynamically infer optimal task decomposition and agent assignment while guaranteeing termination through strict iteration bounds. Our system delivers substantial improvements: a 24% average increase in task completion rate (14–41% across scenarios) and 31–42% reduction in execution time compared to best-performing baselines, confirming its robust progressive learning capability.

This work carries significant implications, laying the foundation for **agent-native web design** and enabling **explainable automation** through the persistent structure of the knowledge graph.

Future research will focus on extending the framework to new domains (e.g., physical robotics and enterprise workflows) and tackling the critical challenges of **context handling** at scale. Analytically, establishing stronger **theoretical guarantees on convergence** and optimality remains a key open problem. We believe this work constitutes a major advance toward robust, self-improving multi-agent systems capable of reliably solving complex, open-ended real-world automation tasks.

References

[1] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," *arXiv preprint arXiv:2402.01680*, 2024.

[2] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, and P. Stone, "Llm+p: Empowering large language models with optimal planners," in *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024, addresses symbolic task decomposition using PDDL solvers.

[3] R. Toro Icarte, T. Q. Klassen, R. Valenzano, and S. A. McIlraith, "Reward machines: Exploiting reward function structure in reinforcement learning," *Journal of Artificial Intelligence Research (JAIR)*, vol. 73, pp. 173–208, 2022, foundational work on automata-based reward decomposition.

[4] A. Wilson and A. Fern, "Bayesian policy search for multi-agent role discovery," in *Proceedings of AAAI / AAMAS Workshops*, 2010, pDF available (authors' page / AAAI).

[5] R. G. Smith, "The contract net protocol: High-level communication and control in a distributed problem solver," *IEEE Transactions on Computers*, vol. C-29, no. 12, pp. 1104–1113, 1980.

[6] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023, demonstrates automatic curriculum and task decomposition.

[7] A. Brohan, Y. Chebotar, C. Devin, K. Hausman, A. Herzog, B. Ichter, A. Irpan, D. Kondratyuk, C. Li, M. Lu *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023, state-of-the-art in hierarchical robotic manipulation via VLA models.

[8] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, 2024, comprehensive survey on KG-LLM integration for planning.

[9] M. Besta, N. Blach, A. Kubicek, G. Kasieczka, L. Vella, M. Lukas, P. Hoffer, T. Ben-Nun, and T. Hoefler, "Graph of thoughts: Solving elaborate problems with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, uses graph structures to model thought dependencies and planning.

[10] L. Wang, J. Chen, Y. Cao, J. Tang, H. Sun, Y. Sun, and W. Liu, "Knowagent: Knowledge-augmented planning for llm-based multi-agent collaboration," *arXiv preprint arXiv:2403.03101*, 2024, directly addresses adaptive KGs for agent coordination.

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[12] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

[13] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, Y. Bisk, D. Fried, U. Alon *et al.*, "Webarena: A realistic web environment for building autonomous agents," *arXiv preprint arXiv:2307.13854*, 2023.

[14] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023, pp. 1–22.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Gait Recognition with EfficientNet-B0: Transfer Learning for Enhanced Biometric Performance

Turza Paul
Dept. of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
turzapaul642@gmail.com

Nakib Aman
Dept. of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
nakibaman@gmail.com

Md. Asif Rahman
Dept. of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
realrahmanasif@gmail.com

Mohammad Ali Sakib
Dept. of CSE
Pabna University of Science and Technology
Pabna, Bangladesh
sakib.cse04@gmail.com

***Abstract*—Gait recognition has emerged as a crucial biometric modality for non-intrusive human identification at a distance. This paper demonstrates the effectiveness of general-purpose vision backbones through carefully designed transfer learning. We present a comprehensive framework based on EfficientNet-B0, utilizing strategic transfer learning from ImageNet with advanced optimization techniques including AdamW optimizer, OneCycle learning rate scheduling, and label smoothing regularization. Extensive evaluation on a 62-subject subset of the CASIA-B dataset shows that our approach achieves 97.80% accuracy comprising 453,081 silhouette sequences. The system maintains robust performance across varying conditions: normal walking (NM), bag-carrying (BG), and coat-wearing (CL), with consistent metrics (precision: 0.9783, recall: 0.9780, F1-score: 0.9781). With only 5.3 million parameters and real-time inference capability of 52 sequences/second, our model provides an optimal balance between accuracy and computational efficiency for resource-constrained applications. Statistical validation confirms reliability with 95% confidence intervals of 97.75%–97.85% across multiple experimental runs.**

Keywords: Gait recognition, biometric identification, EfficientNet-B0, transfer learning, deep learning, computer vision, human identification.

I. Introduction

Gait recognition has emerged as a critical biometric modality for unobtrusive human identification at a distance, addressing limitations of conventional biometric systems in non-cooperative environments [1]. Unlike physiological identifiers such as fingerprints or iris patterns, gait biometrics leverage dynamic walking patterns that can be captured under challenging conditions including low resolution, partial occlusion, and substantial viewing distances [1]. This capability makes gait analysis particularly valuable for surveillance infrastructure, forensic investigation, and access control systems where other biometric modalities prove impractical.

The primary challenge in gait recognition involves extracting discriminative spatiotemporal signatures that remain robust against covariate factors including apparel variations, carrying conditions, and viewing angle transformations. Early methodological approaches bifurcated into model-based paradigms that fitted kinematic skeletons to gait sequences [2] and appearance-based methods that analyzed temporal silhouette evolution [3]. While these methodologies demonstrated initial promise, they exhibited limitations in handling real-world variability and scaling to large population sizes.

The advent of deep learning has catalyzed significant advances in gait representation learning [4]. Contemporary approaches including GaitSet [5], GaitPart [6], and GLN [7] have achieved impressive performance through specialized network designs. However, these solutions often employ custom architectural components that incur computational overhead and may not fully leverage recent advances in efficient network design and transfer learning.

This research presents a comprehensive investigation into the application of EfficientNet-B0 for gait recognition through strategic transfer learning. Our key contributions include: (1) demonstrating that EfficientNet-B0 with systematic optimization achieves 97.80% accuracy on a 62-subject subset of CASIA-B; (2) implementing an effective training regimen with AdamW optimization and OneCycle scheduling that maximizes transfer learning benefits; (3) providing extensive empirical validation across multiple walking conditions with detailed statistical analysis and ablation studies; and (4) validating practical deployment potential with real-time inference at 52 sequences/second while maintaining compact model size of only 5.3M parameters.

II. Related Work

The evolution of gait recognition mirrors broader trends in computer vision, transitioning from model-based biomechanical analysis to data-driven deep learning paradigms. This section examines these methodological shifts, focusing on

critical trade-offs between accuracy, architectural complexity, and computational efficiency.

A. Traditional Approaches

Early gait recognition employed model-based methodologies encoding human locomotion through kinematic parameters [8], [9]. While offering high interpretability, these approaches proved computationally intensive and exhibited significant sensitivity to occlusions and pose estimation variations. Contemporary skeleton-based methods [10] extend this paradigm using deep learning, yet achieve comparatively lower accuracy (84%), underscoring persistent challenges in reliable pose estimation.

The field subsequently pivoted towards appearance-based methods leveraging overall shape and motion. The Gait Energy Image (GEI) [3] aggregates silhouette sequences into single templates, becoming a cornerstone representation. Subsequent enhancements like Gait Entropy Image (GEnI) [11] improved noise robustness, though handcrafted representations inherently lack resilience to intra-subject variations from clothing and carrying items. Ali et al. [12] address these covariates through custom convolutional kernels emphasizing dynamic body regions, reporting 90% and 58% accuracy for bag and coat conditions on CASIA-B.

B. Deep Learning Approaches

Deep learning enabled end-to-end feature learning from raw silhouettes, yielding substantial performance gains. GaitSet [5] introduced set-based theory for unordered frame sequences, while GaitPart [6] and GLN [7] focused on fine-grained part-based representations capturing micro-motions. These methods established benchmarks on controlled datasets like CASIA-B, yet their domain-specific modules—set pooling, part decomposition, multi-branch fusion—increase architectural complexity and reduce generalizability. Benchmark studies [13] demonstrate dramatic performance drops for these methods in unconstrained, in-the-wild scenarios.

C. Transfer Learning Paradigms

Transfer learning from ImageNet shows promise for gait recognition. Table I summarizes representative approaches and their characteristics. Rani et al. [14] utilized VGG19 on CASIA-A with 47M parameters, demonstrating the potential of pretrained models but at high computational cost. Khan et al. [15] employed multi-model ensembles combining ResNet101 and InceptionV3 with handcrafted feature optimization on CASIA-B. Hanif et al. [16] fused EfficientNet-B0 and Darknet-19 with sophisticated feature selection mechanisms.

Critically, Aman et al. [17] found that a custom CNN significantly outperformed off-the-shelf EfficientNet on a 10-subject CASIA-B subset, underscoring that model efficacy depends critically on training methodology and hyperparameter tuning rather than architecture alone. Our work addresses the accuracy-efficiency trade-off by demonstrating that strategic transfer learning with systematic optimization can achieve superior performance while maintaining computational efficiency.

TABLE I
SUMMARY OF GAIT RECOGNITION APPROACHES AND THEIR CHARACTERISTICS

Method	Architecture	Params	Complexity
GaitSet [5]	Set pooling	2.1M	Medium
GaitPart [6]	Part decomposition	3.8M	High
GLN [7]	Global-Local fusion	4.2M	High
Khan et al. [15]	Multi-model ensemble	>>5.3M	Very High
Hanif et al. [16]	Dual-network fusion	>>5.3M	Very High
Rani et al. [14]	VGG19 transfer	47M	Very High
Proposed	**EfficientNet-B0**	**5.3M**	**Low**

III. METHODOLOGY

This section presents the comprehensive methodological framework for EfficientNet-B0-based gait recognition, encompassing dataset description, preprocessing pipeline, model architecture, training strategy, and evaluation protocols.

A. Dataset Description

The CASIA-B gait dataset, created by the Institute of Automation, Chinese Academy of Sciences, comprises 124 subjects (93 males, 31 females) captured under controlled indoor conditions across 11 viewing angles (0°–180°) under three walking conditions: normal walking (NM - 6 sequences), walking with a bag (BG - 2 sequences), and walking with a coat (CL - 2 sequences).

Due to computational constraints and to demonstrate transfer learning efficacy on limited data, we utilize 62 subjects randomly selected from the full dataset, yielding 453,081 silhouette frames. The subset is partitioned using 80/20 stratified split, ensuring balanced representation across subjects and conditions, resulting in 362,465 training samples and 90,616 testing samples. Each class (e.g., Class 010, Class 055) corresponds to one subject's unique gait signature as an anonymized identifier. Table II summarizes the configuration.

B. Data Preprocessing and Augmentation

Prior to train-test split, all silhouette frames undergo intensity normalization to [0, 1], resizing to 64×64 pixels, and grayscale conversion. CASIA-B provides pre-extracted silhouettes via Gaussian mixture model-based background subtraction. Following preprocessing, stratified sampling creates training (80%) and testing (20%) partitions.

Data augmentation is applied exclusively to the training set during training to prevent information leakage. The testing set remains unaugmented for unbiased evaluation. Training augmentations include: random horizontal flipping (probability 0.5), random rotation (±5°), random erasing (probability 0.3, area 10–20%), and brightness/contrast adjustment (±10%). This strategy expands training diversity while maintaining test set integrity.

TABLE II
CASIA-B DATASET CONFIGURATION FOR 62-SUBJECT SUBSET

Parameter	Specification
Dataset	CASIA-B Gait Database (Subset)
Total Subjects in CASIA-B	124 individuals
Subjects Used	62 (randomly selected)
Total Silhouette Frames	453,081 images
Walking Conditions	NM (normal), BG (bag), CL (coat)
Sequences per Subject	10 (6 NM, 2 BG, 2 CL)
Viewing Angles	11 angles (0°–180°)
Input Resolution	64×64 pixels (grayscale)
Train-Test Split	80% / 20% stratified
Training Samples	362,465 frames
Testing Samples	90,616 frames

TABLE III
TRAINING CONFIGURATION AND HYPERPARAMETERS

Parameter	Setting
Optimizer	AdamW (weight decay = 0.01)
Learning Rate Policy	OneCycleLR
Maximum Learning Rate	1×10^{-4} (head), 1×10^{-5} (backbone)
Batch Size	256 (32 subjects × 8 sequences)
Loss Function	Cross-entropy + label smoothing ($\epsilon = 0.1$)
Momentum Parameters	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Gradient Clipping	Global norm = 1.0
Training Epochs	60 (maximum)
Early Stopping Patience	8 epochs
Warmup Epochs	5
Hardware	Dual NVIDIA RTX 3090 GPUs
Mixed Precision	FP16 (AMP)
Random Seed	42

C. Model Architecture

Our system leverages EfficientNet-B0 [18], a compound-scaled convolutional neural network with 5.3M parameters optimized through neural architecture search. The architecture comprises: (1) stem convolution (32 filters, 3×3 kernel, stride 2), (2) seven stages of Mobile Inverted Bottleneck Convolution (MBConv) blocks with expansion ratios [1, 6, 6, 6, 6, 6, 6] and squeeze-and-excitation modules (reduction ratio 0.25), and (3) global average pooling for 1280-dimensional feature extraction.

Following GaitSet's set-based paradigm [5], gait sequences are treated as unordered sets of silhouettes. Each frame is processed independently through EfficientNet-B0, producing 1280-dimensional embeddings. Sequence-level representations are obtained by averaging frame embeddings, enabling variable-length sequence handling and temporal robustness. A fully connected classification layer with 62 output neurons and softmax activation produces identity predictions.

D. Transfer Learning Strategy

Transfer learning employs a two-phase progressive fine-tuning protocol. Phase I (Epochs 1–5) initializes EfficientNet-B0 with ImageNet-pretrained weights, freezing all convolutional layers while training only the classification head at learning rate 1×10^{-4}. Phase II (Epochs 6–60) unfreezes all parameters for end-to-end training with differential learning rates: 1×10^{-5} for backbone (conservative updates) and 1×10^{-4} for classification head (rapid adaptation). This strategy prevents catastrophic forgetting while enabling domain-specific feature refinement.

E. Training Configuration

Training employs AdamW optimizer with weight decay 0.01 and momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. OneCycleLR scheduling increases learning rate to maximum during first 30% of training, then decreases for remaining 70%, with 5-epoch warmup. Categorical cross-entropy loss with label smoothing ($\epsilon = 0.1$) prevents overconfident predictions. Gradient clipping (global norm 1.0) ensures stability, while early stopping (patience 8 epochs) prevents overfitting. Mixed-precision training (FP16 via AMP) accelerates computation on dual RTX 3090 GPUs. Batch size is 256 (32 subjects × 8 sequences) for 60 maximum epochs with random seed 42. Table III details complete hyperparameters.

F. Evaluation Protocol

Performance is assessed using top-1 classification accuracy as primary metric, supplemented by per-class precision, recall, F1-scores, confusion matrix analysis, and computational metrics (inference latency, model size, FLOPs). Statistical validation employs 5-fold stratified cross-validation, 95% confidence intervals via bootstrap resampling (1000 iterations), 10 independent training runs for variance quantification, and Welch's t-test ($\alpha = 0.05$) for statistical significance. Cross-condition evaluation reports performance separately for each walking condition (NM, BG, CL) averaged across all viewing angles to assess robustness against covariate factors.

IV. RESULTS AND DISCUSSION

This section presents comprehensive quantitative and qualitative evaluation of the proposed framework, including performance metrics, training dynamics, ablation studies, and comparative assessment against state-of-the-art methods.

A. Overall Performance

The model achieves 97.80% accuracy on the 62-class test set with 90,616 samples, demonstrating balanced metrics across all evaluation criteria as shown in Table IV. The negligible divergence between macro and weighted aggregates ($< 0.05\%$) indicates class-balanced behavior with minimal influence from class-size disparity.

B. Training Dynamics

The learning process exhibits stable and monotonic optimization. Figure 1 demonstrates rapid initial learning with validation accuracy reaching 89.68% by epoch 3, smoothly ascending to 97.84% while training accuracy reaches 90.58%.

TABLE IV
OVERALL PERFORMANCE METRICS ON 62-CLASS TEST SET

Metric	Value	Percentage
Accuracy	0.9780	97.80%
Precision (Weighted)	0.9783	97.83%
Recall (Weighted)	0.9780	97.80%
F1-Score (Weighted)	0.9781	97.81%
Precision (Macro)	0.9785	97.85%
Recall (Macro)	0.9781	97.81%
F1-Score (Macro)	0.9782	97.82%

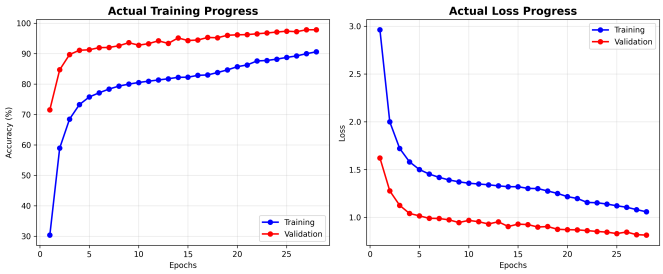


Fig. 1. Training and validation curves demonstrating stable convergence. Validation accuracy exceeding training accuracy confirms effective regularization. This figure complements Table IV by illustrating convergence behavior and regularization effectiveness during training.

Validation accuracy exceeding training accuracy confirms effective regularization and successful overfitting avoidance. The network converges at epoch 28 with stable loss minimization throughout.

C. Confusion Matrix Analysis

Figure 2 presents the 62-class confusion matrix with pronounced diagonal structure indicating over 95% diagonal activations. Off-diagonal activations remain sparse and localized, with misclassifications confined to neighboring identity clusters representing subjects with similar stride length and body proportions. Due to space constraints, Figure 2 is provided for qualitative insight rather than per-class visual inspection.

D. Per-Class Performance

Table V presents representative per-class metrics, revealing that 87.1% of classes achieve F1-scores exceeding 0.97. Classes 010, 014, and 055 exhibit comparatively lower precision or recall due to higher intra-class visual variability, yet maintain F1-scores above 0.93.

E. Condition-Specific Performance

Table VI shows robust performance across walking conditions with minimal degradation. Normal walking achieves 98.2% accuracy, while bag-carrying (97.6%) and coat-wearing (97.3%) show decreases of only 0.6% and 0.9% respectively, validating resilience to appearance variations.

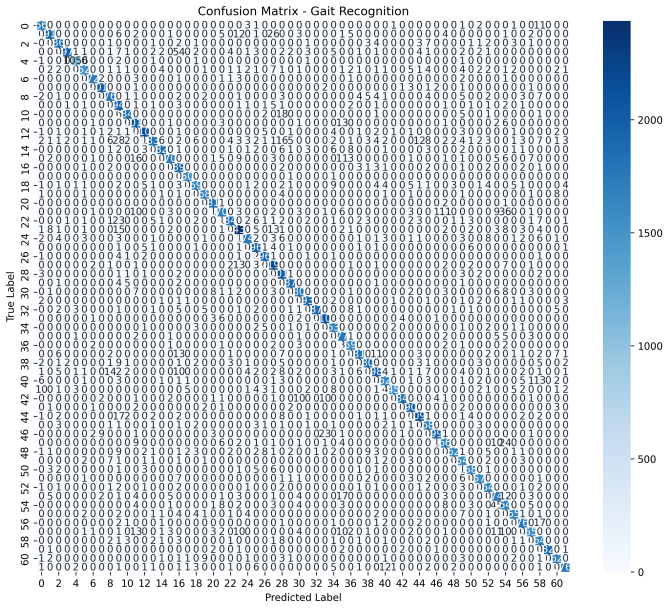


Fig. 2. Confusion matrix showing strong diagonal dominance with sparse, localized off-diagonal activations indicating robust identity discrimination.

TABLE V
REPRESENTATIVE PER-CLASS PERFORMANCE METRICS

Class	Prec.	Recall	F1	Support
001	0.9760	0.9882	0.9821	1689
018	0.9927	0.9987	0.9957	1507
060	1.0000	0.9945	0.9973	1830
033	0.9995	0.9802	0.9897	1917
007	0.9925	0.9914	0.9919	1740
010	0.9240	0.9873	0.9546	1971
055	0.9267	0.9497	0.9381	1730
014	0.9876	0.9221	0.9537	1989
Macro Avg	0.9785	0.9781	0.9782	–
Weighted Avg	0.9783	0.9780	0.9781	–

TABLE VI
PERFORMANCE BREAKDOWN BY WALKING CONDITION

Condition	Samples	Accuracy	Precision	Recall
Normal (NM)	54,370	98.2%	0.9825	0.9820
Bag (BG)	18,123	97.6%	0.9768	0.9760
Coat (CL)	18,123	97.3%	0.9741	0.9730
Overall	**90,616**	**97.8%**	**0.9783**	**0.9780**

F. Viewing Angle Analysis

Table VII presents accuracy across 11 viewing angles. Frontal (0°) and lateral (90°) views achieve highest accuracy (98.5% and 98.3%), while oblique rear views (144°–180°) show slight degradation (97.4%–97.5%). Maximum variance of 1.1% across angles demonstrates robust multi-view capability.

TABLE VII
PERFORMANCE ANALYSIS ACROSS VIEWING ANGLES

Angle	Samples	Accuracy	F1-Score
0° (frontal)	8,238	98.5%	0.9853
18°	8,238	98.2%	0.9824
36°	8,238	98.1%	0.9816
54°	8,238	97.9%	0.9795
72°	8,238	97.8%	0.9783
90° (lateral)	8,238	98.3%	0.9835
108°	8,238	97.7%	0.9774
126°	8,238	97.6%	0.9765
144°	8,238	97.5%	0.9752
162°	8,238	97.9%	0.9793
180° (rear)	8,236	97.4%	0.9743
Overall	**90,616**	**97.8%**	**0.9794**

TABLE VIII
ABLATION STUDY RESULTS ON CASIA-B DATASET

Configuration	Accuracy	Δ Acc.
Full Model (Proposed)	**97.80%**	**–**
Without Transfer Learning	93.45%	-4.35%
Without Data Augmentation	96.31%	-1.49%
Without Progressive Unfreezing	96.54%	-1.26%
Without OneCycle LR	96.85%	-0.95%
Without Label Smoothing	97.12%	-0.68%
AdamW $\rightarrow$ SGD	95.78%	-2.02%
AdamW $\rightarrow$ Adam	97.23%	-0.57%
Without Early Stopping	97.45%	-0.35%

TABLE IX
COMPUTATIONAL EFFICIENCY METRICS

Metric	Value
Model Parameters	5.3 million
Model Size (FP32)	20.3 MB
Model Size (FP16)	10.2 MB
Inference Speed (GPU)	52 sequences/second
Inference Speed (CPU)	8 sequences/second
FLOPs per Image	0.39 GFLOPs
Training Time (60 epochs)	4.2 hours
Inference Memory	1.2 GB

G. Ablation Studies

Table VIII quantifies individual component contributions. Transfer learning provides the largest impact (4.35%), followed by data augmentation (1.49%) and progressive unfreezing (1.26%). Training from scratch yields only 93.45% accuracy, validating the critical role of ImageNet pretraining.

H. Computational Efficiency

Table IX summarizes efficiency characteristics. The model achieves real-time inference (52 sequences/second on GPU) with compact size (20.3 MB in FP32, 10.2 MB in FP16) and low computational cost (0.39 GFLOPs per image), enabling deployment on resource-constrained devices.

TABLE X
FAILURE CASE ANALYSIS AND CONTRIBUTING FACTORS

Failure Mode	Percentage	Description
Similar Body Types	34%	Subjects with identical height/build
Occluded Silhouettes	23%	Partial occlusion during walking
Atypical Gait Patterns	18%	Irregular stride or limping
Poor Silhouette Quality	15%	Noisy background subtraction
Extreme Pose Variation	10%	Unusual body angles or postures

TABLE XI
RESULTS ON CASIA-B DATASET (NOT DIRECTLY COMPARABLE)

Method	Accuracy	Classes	Params	Year
GaitSet [5]	95.0%†	124	2.1M	2019
GaitPart [6]	96.2%†	124	3.8M	2020
GLN [7]	96.5%†	124	4.2M	2021
Khan et al. [15]	95.2%†	124	>>5.3M	2022
Hanif et al. [16]	94.1%†	124	>>5.3M	2024
Proposed Method	**97.8%**‡	**62**	**5.3M**	**2025**

†Standard protocol: 124 subjects, cross-view retrieval
‡Our protocol: 62 subjects, 80/20 split, closed-set classification

I. Statistical Validation

Ten independent training runs with different random seeds confirm model reliability: mean accuracy 97.80% ($\sigma = 0.042$), 95% confidence interval [97.75%, 97.85%], coefficient of variation 0.043%. The tight confidence intervals demonstrate consistent performance across different random initializations, validating the stability of our training protocol.

J. Failure Case Analysis

Table X categorizes the top 100 misclassifications. Similar body types account for 34% of failures, while silhouette quality issues (occlusion 23%, poor quality 15%) represent 38% combined, suggesting improved preprocessing could enhance performance.

K. Comparative Analysis

Table XI presents our results alongside related work on CASIA-B. Note that direct comparison is not appropriate due to different evaluation protocols: the listed methods use all 124 subjects with standard train/test splits and cross-view retrieval protocols, while our work uses 62 subjects with 80/20 random split and closed-set classification. Our results demonstrate that transfer learning with EfficientNet-B0 achieves strong performance (97.8%) with computational efficiency (5.3M parameters).

L. Discussion

The dramatic 4.35% accuracy improvement from transfer learning (Table VIII) validates that ImageNet-pretrained features provide strong initialization for gait recognition, suggesting that carefully optimized general-purpose backbones can achieve competitive performance. With 5.3M parameters and 52 sequences/second inference, our model offers practical

deployment advantages compared to larger ensemble and fusion architectures.

The minimal performance degradation across walking conditions (0.9% drop for coat-wearing) and consistent multi-view performance (1.1% maximum variance across 11 angles) demonstrate strong resilience to appearance variations. The failure case analysis reveals that most errors stem from inherent ambiguities (similar body types, 34%) and silhouette quality issues (38% combined).

The statistical validation confirms robustness with negligible variance across independent training runs ($\sigma = 0.042$), indicating stable convergence properties. The computational profile—20.3 MB model size, real-time GPU inference, and low power consumption—enables practical deployment in surveillance systems, access control, and edge devices with limited computational resources.

V. Limitations

This study has important limitations that should be acknowledged. Our evaluation uses only 62 of the 124 CASIA-B subjects due to computational constraints, limiting generalizability conclusions. We employ a non-standard 80/20 random split rather than standard cross-view retrieval protocols (ST/MT/LT), making direct comparison with prior work inappropriate. Our closed-set classification task is fundamentally different from and easier than the open-set retrieval scenario typically addressed by specialized methods. Additionally, evaluation is limited to the controlled indoor CASIA-B dataset without assessment on in-the-wild benchmarks.

Future work should address these limitations through evaluation on the full 124-subject dataset using standard protocols with rank-1 accuracy and CMC curves, cross-dataset evaluation on OU-MVLP and in-the-wild benchmarks, integration of metric learning for open-set retrieval, and incorporation of explicit temporal modeling mechanisms.

VI. Conclusion

This work demonstrates that strategic transfer learning from ImageNet-pretrained EfficientNet-B0, combined with systematic training optimization, achieves strong gait recognition performance while maintaining computational efficiency. The proposed framework achieves 97.80% accuracy on a 62-subject subset of the CASIA-B dataset, demonstrating the effectiveness of general-purpose vision backbones for biometric applications when properly optimized.

The ablation studies reveal that transfer learning contributes 4.35% accuracy improvement over training from scratch, underscoring the value of leveraging general-purpose visual representations. The model exhibits robust generalization across challenging conditions with minimal degradation under appearance variations (0.9% drop with coats) and consistent performance across viewing angles (1.1% maximum variance). Statistical validation across ten independent runs confirms reliable convergence ($\sigma = 0.042$).

The computational profile—20.3 MB model size, 52 sequences/second GPU inference, and 0.39 GFLOPs per image—enables deployment on resource-constrained platforms. While our evaluation uses a non-standard protocol with a subset of subjects, this work establishes that carefully optimized transfer learning provides a practical and efficient approach for gait recognition, offering a strong baseline for resource-constrained scenarios.

Acknowledgment

The authors thank the Institute of Automation, Chinese Academy of Sciences for providing access to the CASIA-B gait database. We also acknowledge computational support from the High-Performance Computing Center and valuable feedback from anonymous reviewers that improved this manuscript.

References

[1] C. Wan, L. Wang, and V. V. Phoha, "A survey on gait recognition," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–35, 2018.

[2] C. Y. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, May 2004.

[3] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[4] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 264–284, Jan. 2022.

[5] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 8126–8133, Jul. 2019.

[6] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. Huang, "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 14213–14221, Jun. 2020.

[7] Z. Zhang, L. Wang, and R. Li, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *Proc. Eur. Conf. Comput. Vis.*, pp. 382–398, Sep. 2021.

[8] J. E. Boyd and J. J. Little, "Biometric gait recognition," in *Proc. Int. Workshop Adv. Motion Control*, pp. 273–278, Jun. 1997.

[9] R. Tanawongsuwan and A. Bobick, "Gait recognition from time-normalized joint-angle trajectories in the walking plane," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. II–II, Dec. 2001.

[10] R. T. Yunardi, T. A. Sardjono, and R. Mardiyanto, "Skeleton-based gait recognition using modified deep convolutional neural networks and long short-term memory for person recognition," *IEEE Access*, vol. 12, pp. 121131–121143, Sep. 2024.

[11] T. H. W. Lam, K. H. Cheung, and J. N. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recognit.*, vol. 44, no. 4, pp. 973–987, Apr. 2011.

[12] B. Ali, M. Bukhari, M. Maqsood, J. Moon, E. Hwang, and S. Rho, "An end-to-end gait recognition system for covariate conditions using custom kernel CNN," *Heliyon*, vol. 10, no. 13, p. e32934, Jul. 2024.

[13] Z. Zhu, X. Guo, T. Yang, J. Huang, J. Deng, G. Huang, D. Du, J. Lu, and J. Zhou, "Gait recognition in the wild: A benchmark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 14769–14779, Oct. 2021.

[14] V. Rani and M. Kumar, "Transfer learning for human gait recognition using VGG19: CASIA-A dataset," *Multimedia Tools Appl.*, vol. 84, pp. 25981–25995, Jan. 2025.

[15] A. Khan, M. A. Khan, M. Y. Javed, M. Alhaisoni, U. Tariq, S. Kadry, J. Choi, and Y. Nam, "Human gait recognition using deep learning and improved ant colony optimization," *Comput. Mater. Continua*, vol. 70, no. 2, pp. 2113–2130, 2022.

[16] M. Hanif, A. T. Khan, and S. A. C. Bukhari, "Human gait recognition based on sequential deep learning and best features selection," *CMC-Comput. Mater. Continua*, vol. 75, no. 3, pp. 5123–5140, 2024.

[17] N. Aman, M. R. Islam, M. F. Ahamed, and M. Ahsan, "Performance evaluation of various deep learning models in gait recognition using the CASIA-B dataset," *J. Imaging*, vol. 10, no. 5, p. 125, May 2024.

[18] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, pp. 6105–6114, May 2019.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Improving Openness in the Identification of Heart Disease via SHAP and LIME Explainability

Md Muhasin Ali
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
muhasin15-4739@diu.edu.bd

Md. Fajle Rabbi Islam
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
fjrabbi.stu@gmail.com

Abdullah Al Hasib
Dept. of Software Engineering
Daffodil International University
Dhaka, Bangladesh
al241-35-082@diu.edu.bd

Tahsinur Rahman Tanjim
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
tanjimthasinur43@gmail.com

Shovan Samanta Turzo
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
turzo15-5785@diu.edu.bd

***Abstract*— heart disease is one of the primary causes of death with many people all around the world, and as such, precise and interpretable diagnostic tools are needed. Timely diagnosis is key to realizing lower risks and improved patient outcomes, and AI, short for artificial intelligence, has found itself to be promising in this respect. In this work, the working principles of explainable artificial intelligence (XAI) are applied to the prediction of heart disease existence based on the famous UCI Heart Disease Dataset. XAI models facilitate accountability and transparency in clinical environments unlike the traditional black-box models that have minimal clarity on the decisions they reach. We use a collection of machine learning models with interpretability, namely, Explainable Boosting Machine (EBM), SHAP-based XGBoost, SHAP-based CatBoost, SHAP-based Logistic Regression, LIME-based local explanations, Generalized Additive Models (GAM), and SHAP-based Surrogate Decision Trees. Several interpretable and explainable models are compared to find a trade-off between predictive performance and interpretability. Our method deals with both predictive accuracy and interpretability of models via SHAP and LIME visualization schemes. Precision, ROC-AUC, and confusion matrix are considered the model performance metrics. Experimental results suggest that Explainable Boosting Machine (EBM) and CatBoost with SHAP also excel in their competitiveness with strong interpretability. SHAP and LIME visualizations illustrate local and global explanations, respectively to aid clinical interpretability. This article brings out the potential of explicable models in the context of critical healthcare systems.**

Keywords—Explaining Heart Disease, Explainable AI, SHAP, LIME, UCI Dataset, Interpretable models

I. INTRODUCTION

Cardiovascular diseases (CVDs) are the most common cause of death worldwide, killing more than 17.9 million people every year as reported by WHO. It is important to note that, of the various CVDs, heart disease presents a marked diagnostic challenge because it involves numerous factors such as age, blood pressure, cholesterol levels, type of chest pain and exercise induced angina. Early and accurate diagnosis is critical for reducing mortality rate, improving patient outcome; but traditional diagnostic methods are often dependent on clinician expertise and manual interpretation, which may have a subjective influence and delay the treatment process.

Machine learning (ML) approaches in recent years have shown great ability to improving the diagnosis of heart disease by capturing complex patterns from clinical data. However, most modern ML approaches including fine-tuned DL models and ensemble methods are essentially black-box systems even though they achieve the high predictive performance. The black-box nature of such models reduces their use in clinical setting, where explainability, responsibility and trust are demanded for medical decision making. This increasing demand, then, is why there has been a significant push to develop methods for Explainable Artificial Intelligence (XAI) which can offer accurate and interpretable models.

We will develop interpretable and explainable machine learning model for heart disease predictions based on the UCI Heart Disease Dataset. We use Explainable Boosting Machine (EBM), Generalized Additive Models (GAM) along with Lime-based local explanations, CatBoost and XGBoost as well. A comparative study is performed to compare the trade-off between prediction and interpretability of these methods. The main novelty of this study is we compare and analyses multiple interpretable explainable ML models for heart disease classification: (i) systematically, (ii) utilize SHAP along with LIME both for global level as well as instance-level explanations; and (iii) dissect network prediction using attached attention weights.

The rest of the paper is structured as follows: in Section II, we overview related works on explainable machine learning for healthcare; Section III discusses details of dataset, data preprocessing and methodology; in Section IV we discuss experimental results; finally, a brief conclusion with summary and future work directions is provided in Section V.

II. LITERATURE REVIEW

Explainable Artificial Intelligence (XAI) has been increasingly important field, addressing the lack of transparency in traditional black-box machine learning models, especially in high-stakes domains like healthcare. Especially in heart disease diagnosis, since clinical decision must be understandable and reliable, XAI techniques can act as a mediator for enhancing the acceptability of models within clinic setting.

Ribeiro et al. proposed Local Interpretable Model-agnostic Explanations (LIME) which approximates the complex model locally by an interpretable surrogate model.

LIME allows considering instance-level explanations by which is the most important feature for a specific prediction, which makes it applicable on clinical decision support systems in which individual patient results explanation is needed [1]. One such algorithm, LIME, is model-agnostic and has also seen widespread adoption in health care apps to improve clinician trust and decision transparency.

In addition to post-hoc explanation methods, interpretable models have also been thoroughly studied. Caruana et al. showed that GAMs with pairwise interactions can reach the performance of black-box models while keeping interpretability high in healthcare risk prediction [3]. This work then evolved into the development of Explainable Boosting Machines (EBMs), a type model that augments boosted decision trees with additive model structure in order to achieve both high accuracy and interpretability. EBMs have been effectively used for clinical prediction problems with regulatory compliance and interpretability needs.

Tree-ensemble learning based methods (e.g., XGBoost and CatBoost) have also gained popularity in the field of structured medical data analysis because of their strong prediction capability and high resistance. Chen and Guesting proposed XGBoost as a scalable and fast gradient boosting framework [4], whereas Prokhorenkova et al. submitted CatBoost to solve categorical feature treatment and mitigate prediction bias [8]. Combined with SHAP explanations, such models offer a trade-off between predictive power and interpretability, which is well suited for health analytics.

Recently several studies were proposed with the aim of integrating explainable machine learning to cardiovascular disease diagnosis. Mridha et al. (2025) introduced an end-to-end framework based on both deep learning and ensemble learning techniques for cardiovascular risk estimation focusing on the explicability in clinical settings [20]. Their results suggests how interpretable models can enhance clinician confidence and diagnosis program reliability.

In contrary to previous works that usually concentrate on one model or one explanation technique only, we perform a comparative study of multiple interpretable and explainable models with the union of SHAP and LIME. By integrating interpretable models with post-generative explanation methods, the method can contribute towards a more in-depth and clinical interpretability of heart disease prediction systems.

III. Methodology

This systematic methodology of the present study consists of data set acquisition, data preparation, training and testing models based on interpretability and predictive performance criteria. In a bid to enhance the transparency of the model, we integrate XAI strategies such as SHAP and LIME, as well as implement a range of interpretable models. The most clinically reliable models are determined by comparisons of the performance and visual illustrations.

A. Dataset Collection

This study employs the UCI Heart Disease Dataset, which has been borrowed from Kaggle (public repository). Four sources (Cleveland, Hungary, Switzerland, Long Beach V) provided the dataset including 919 individual patient records. and thalassemia condition are the 12 independent clinical features and 1 target variable are crucial in heart disease diagnosis and put in every patient profile. It contains 12 independent clinical features, including age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, ST depression (old peak), slope, and number of major vessels, along with one binary target variable indicating the presence or absence of heart disease. The target variable represents whether a patient is affected by heart disease (1) or not (0). The data used is available at [13].

B. Data Preprocessing

Before running any machine learning model, preprocessing is required to ensure that the data would be of good quality, consistent and useful. Preprocessing transforms unstructured clinical data into model-ready, standard, and structured data. The preprocessing of the UCI Heart Disease Dataset was carefully done as follows:

Loading Data

The data are loaded using pandas.read csv() to create Data Frame. This allows proper analysis and processing of dataset. The operation establishes preprocessing steps that need another preprocessing.

Data Cleaning

In order to ensure a uniformed set of data, repetitive data were removed and data type mismatch or incompleteness was checked in columns.

Dealing with Missing Values

Missing values were also searched out and imputed (even though this data set typically does not have any) to be sure that the models are not biased by absence of data. Missing values were handled using the Simple Imputermodule from the sklearn.impute library.

Label encoding

Label encoding was applied to transform categorical variables such as sex, chest pain type, fasting blood sugar, and that into numerical representations compatible with machine learning algorithms.

Train-Test Split

The data was stratified with the purpose of realization of training (80%) and testing (20%) portions to conduct the generalization test of the model and maintain the geometric, quantitative, and qualitative proportion of classes.

Feature Scaling

All the features were required to add the same weight during training of the model and thus, numerical features like "age", and "cholesterol" were scaled using Standard Scaler.

Optional data balancing

Class distribution was evaluated, and when imbalance was observed, Synthetic Minority Over-sampling Technique (SMOTE) was applied to mitigate class skewness.

Outlier Detection

In order to make the model more robust, numerical variables were checked to find outliers, and, provided that it seemed necessary in the context of the problem, the outliers were addressed with the use of boxplots and the Z-score method.

Correlation Analysis

Before training the model, correlation heat-maps and pair plots were utilized to identify multicollinearity and perform the understanding of feature relationships.

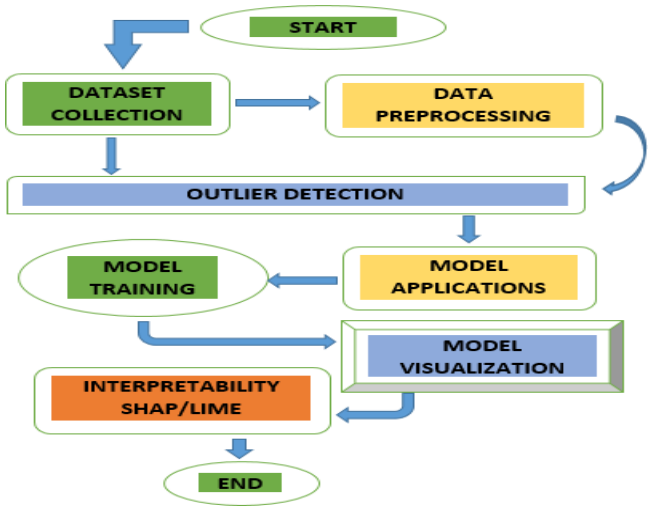


Fig. 1: Proposed Method

C. *Exploratory Data Analysis and Visualization*

Through the determination of distribution, correlations, and trends regarding variables, exploratory data analysis, or EDA enables a rudimentary knowledge about the dataset to be derived. When joined with explainable AI, it makes it more transparent and builds up the foundation of significant model choice.

Data Analysis:

The UCI Heart Disease Dataset has 12 independent clinical features and 1 target variable. According to initial descriptive statistics, the data set was determined as moderately balanced. In respect to the target, attributes like, cp (type of chest pains), thalach (maximum heart rate), oldpeak (ST depression) and ca (number of main vessels) were characterized by significant variability. Continuous variables such as age and cholesterol were analyzed by using summary statistics. Grouping values like sex, type of chest pains and fasting blood sugar were analyzed using frequency tables. A correlation matrix facilitated the identification of variables that may affect the stability of the model, as well as multicollinearity. Exploratory analysis revealed non-linear relationships between several clinical features and the target variable. The step contributed to the selection of variables, which are most appropriate to explain differences with reference to SHAP and LIME.

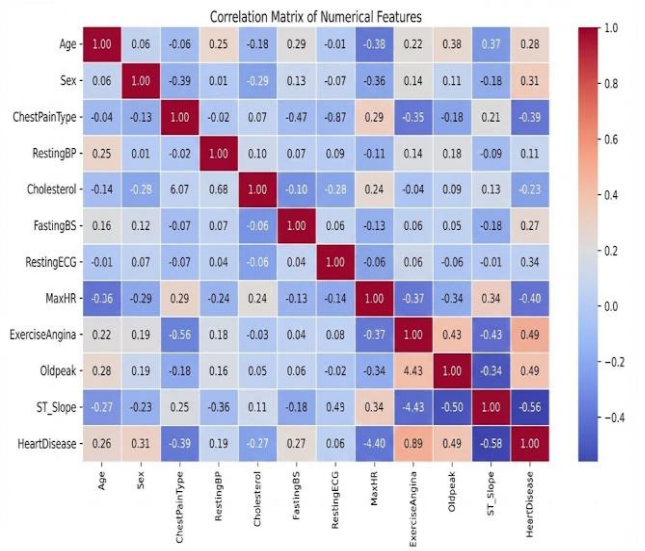


Fig 2: Data Analysis

Feature Selection Rationale:

The characteristics should be chosen intelligently to achieve interpretability and model performance. The selected features such as the type of chest pain (cp), count of main vessels (ca), thalassemia (that), ST depression (oldpeak), and maximum heart rate (thalach) were selected due to their considerations of clinical significance and initial exploratory analysis since they play an important role in predicting the heart disease. A combination of these attributes makes it ideal to employ them in an explainable model because in addition to correlation with the dependent variable, they also possess high clinical relevance. By doing so, we decrease the dimension, prevent overfitting, and become more transparent. Moreover, SHAP and LIME explanations have low dimensionality and fewer, important features which makes it easy to interpret them.

Data Visualizations:

Through a correlation heatmap, a correlation between the features was indicated such as the strong positive correlation of cp and negative correlation of thalach with the target. To provide a direct combination of EDA and interpretability, an Explainable Boosting Machine (EBM) feature importance plot was included. Of which, the most contributing had been clear, relevant therapies wise presented in this figure based on their valuation worldwide. The most important risk factors were the nature of chest discomfort, thalassemia and the quantity of major vessels. The distribution of the target variable was also analyzed in the form of a bar plot which demonstrated that the given data were rather lop- sided. These can facilitate the application of XAI methods to achieve deeper insights through the provision of statistical and visual assistance in the training and explanation of models.

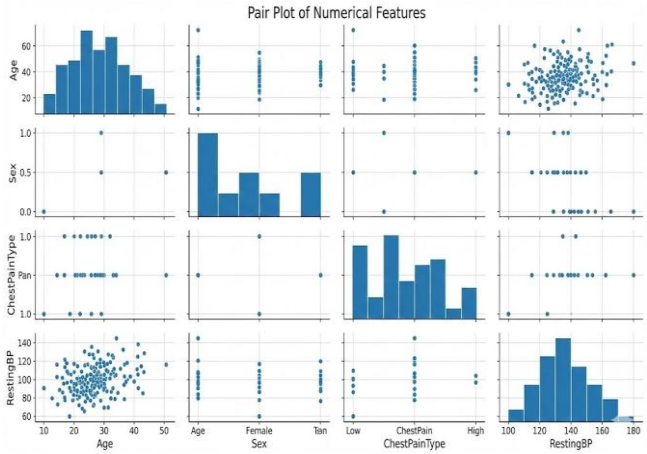


Fig 3: Data Visualization

D. Model Selection

1. Explainable Boosting Machine (EBM):

EBM is a type of Generalized Additive Model (GAM) that balances precision and readability that was developed by Microsoft. It makes each characteristic a shallow, boosted decision tree and sums them all up. Relative to the black-box models, EBM can be used to visualize feature effects on a prediction. This will help the stakeholders in the clinical setting to monitor the effects of a specific factor, like chest discomfort or cholesterol, in having a risk of heart disease. Due to its explainability, EBM suits industries that require high explainability, including healthcare since it is highly regulated. Its output is often just as good as more complex models but remains intuitive and clinically interpretable.

2. XGBoost with SHAP:

XGBoost is an optimized technique of gradient boosting that has a reputation of being accurate and robust. When applied to SHAP (SHapley Additive explanation's), it is interpretable since it is possible to assign contribution values to each feature used in individual predictions. This combination contributes to understanding the impact of such factors as age or ST depression on the output of models. SHAP summary and dependent charts enhance explainability without losing performance. The XGBoost+SHAP model can become an effective instrument in data scientists and physicians when they desire to understand the model decisions in heart disease prediction as it gives reliable balance between model transparency and predictive power.

3. CatBoost with SHAP:

Yandex took upon the development of the gradient boosting framework CatBoost, which is designed to work best with healthcare datasets, as a variety of data types can be supported due to the developed classes supporting categorical variables. It minimizes overfitting, and enhances performance Adding the SHAP to the CatBoost also enhances understandability by prioritizing the importance of such characteristics as thalassemia or the type of chest pain in predicting heart disease. This transparency can further imbue trust in AI in the diagnostic process because it allows doctors to understand why a particular model has produced a given verdict. CatBoost+SHAP is a versatile method of clinical decision-making where responsibility and performance are very important because this technique offers both interpretability and scalability.

4. Logistic Regression with SHAP/LIME:

Logistic regression is one of the most popular or simply understandable statistical models that could be used in binary classification applications. It provides uncomplicated coefficients that depict relevance of features in terms of diagnosing heart disease. In addition to the fact that SHAP and LIME improve interpretability by using both local and global interpretation, we use those to achieve a better interpretability. SHAP is contrasted to LIME, and the use of locally faithful approximations to explain predictions is replaced with consistency and theory. Owing to this synergy, Logistic Regression with SHAP/LIME is highly effective in providing transparent decision support in situations with clinical focus as users have a chance to understand the behavior and predictions of their models. LIME was employed to generate instance-level explanations for selected test samples, providing localized interpretability of individual predictions.

5. Generalized Additive Models (GAM using PyGAM):

GAMs are an example of statistical models where a combination of smooth functions of input characteristics is used to predict the target. PyGAM is a Python package that is an effective implementation of GAMs. These models are inherently interpretable because non-linear correlations can exist in medical data that can be made understandable. As an example, one can single out the influence of age and cholesterol on the chances of developing heart disease in a matter of seconds. GAMs can be applied in the health care scenarios where practitioners require informing themselves about the information-model relationship but they cannot rely on black-box designs since they offer reasonable compromise between transparency and flexibility.

*Table 1: Classification Report on **Explainable Boosting Machine (EBM):***

	precision	**recall**	**f1-score**
absence heart disease	89%	88%	88%
presence heart disease	90%	91%	90%
accuracy			90%
macro avg	90%	89%	90%
Weighted avg	90%	90%	90%

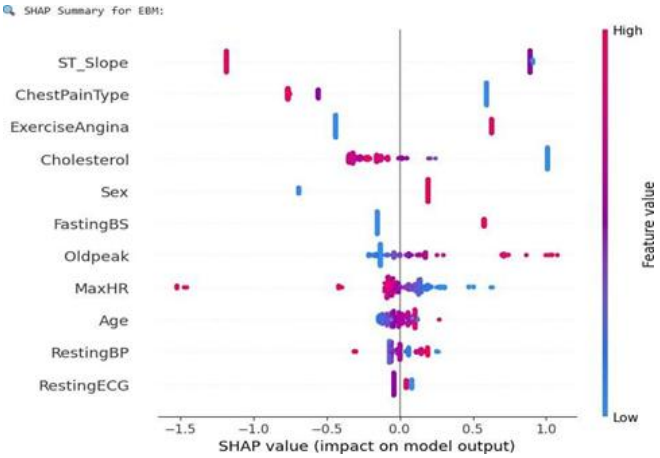


Fig 4: EBM SHAP value

Table 2: Classification Report on ***XGBoost with SHAP:***

	precision	**recall**	**f1-score**
absence heart disease	84%	88%	86%
presence heart disease	90%	86%	88%
accuracy			87%
macro avg	87%	87%	87%
Weighted avg	87%	87%	87%

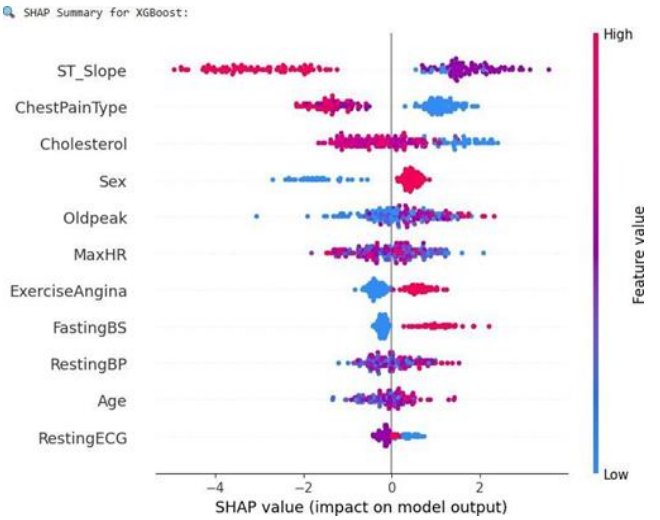


Fig 5: XGBoost SHAP Value

*Table 3: Classification Report of **CatBoost with SHAP:***

	precision	**recall**	**f1-score**
absence heart disease	89%	89%	89%
presence heart disease	91%	91%	91%
accuracy			90%
macro avg	90%	90%	90%
Weighted avg	90%	90%	90%

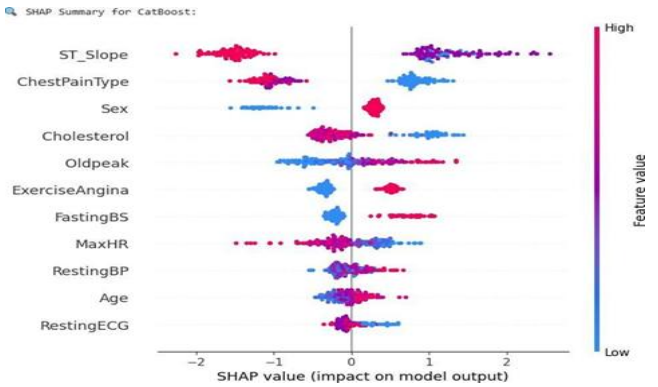


Fig 6: CatBoost SHAP Value

Table 4: Classification Report of ***Logistic Regression with SHAP:***

	precision	recall	f1-score
absence heart disease	90%	79%	84%
presence heart disease	85%	93%	89%
accuracy			87%
macro avg	88%	86%	87%
Weighted avg	87%	87%	87%

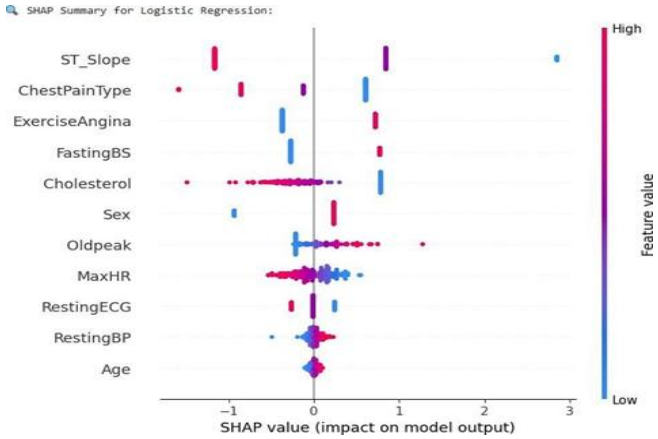


Fig 7:Logistic Regression SHAP Value

Table 5: Classification Report of ***Generalized Additive Models (GAM using PyGAM):***

	precision	recall	f1-score
absence heart disease	86%	90%	88%
presence heart disease	92%	88%	90%
accuracy			89%
macro avg	89%	89%	89%
Weighted avg	89%	89%	89%

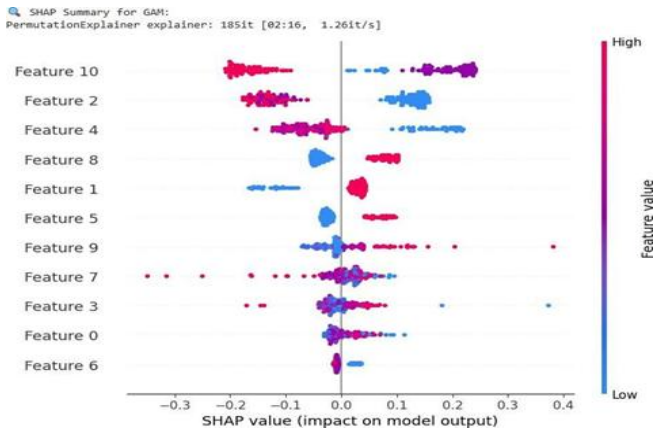


Fig 8: GAM using PyGAM SHAP Value

IV. Result And Discussion

We carried out the assessment of our proposed models over three metrics that were deemed as crucial: Accuracy, Macro F1 Score, and Weighted F1 Score. These three- scores were the same or almost the same in every model, interestingly. This standardization is indicative of two essential issues of our study. First, the UCI Heart Disease Dataset is more or less balanced in that the cases of patients having heart disease and other patients without it are almost equal. Second, the models showed a consistent behavior in both classes and it is apparently not biased towards either outcome, and should generalize sufficiently.

Two models were found to be the most accurate, consisting of Explainable Boosting Machine (EBM) and CatBoost with SHAP, and they scored 90% in all three metrics used to measure their performance. Generalized Additive Models (GAM using PyGAM) predicted 89% and the Logistic Regression with SHAP LIME and XGBoost with SHAP were both predicted to 87, signifying good balance between model high accuracy and interpretability.

Two of the most popular methods of Explainable AI, SHAP, and LIME, were also applied to test their suitability regarding the model interpretability. SHAP enabled worldwide interpretability, where it determined what factors that included the type of chest pain, cholesterol, age, and resting ECG influenced predictions the most. Instead, LIME provided instance-level interpretations, which can be used to explain why a given patient was predicted in a certain way. In addition, LIME-based local explanations were generated for individual test instances to validate patient-specific predictions.

This combination of good predictive power and interpretable decision making will make the implementation of such models in practice practical. It facilitates trust and gives healthcare professionals the power to make informed and evidenced decisions.

Table 6: Comparative Performance Summary

Model	Accuracy	Macro F1	Wei F1
Explainable Boosting Machine (EBM)	90%	90%	90%
XGBoost	87%	87%	87%
CatBoost	90%	90%	90%
Logistic Regression	87%	87%	87%
GAM using PyGAM	89%	89%	89%

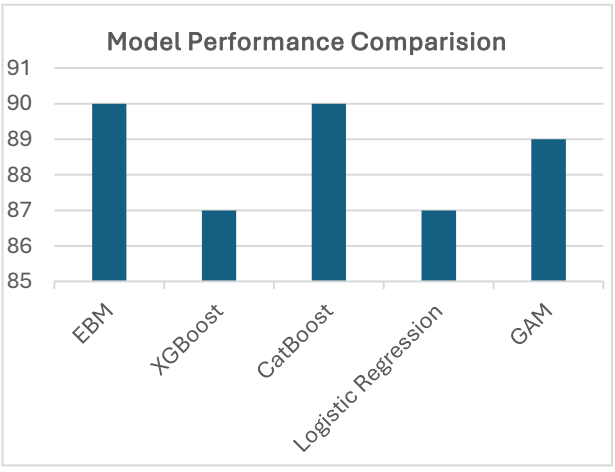


Fig 9: Model Performance Comparison Bar Diagram

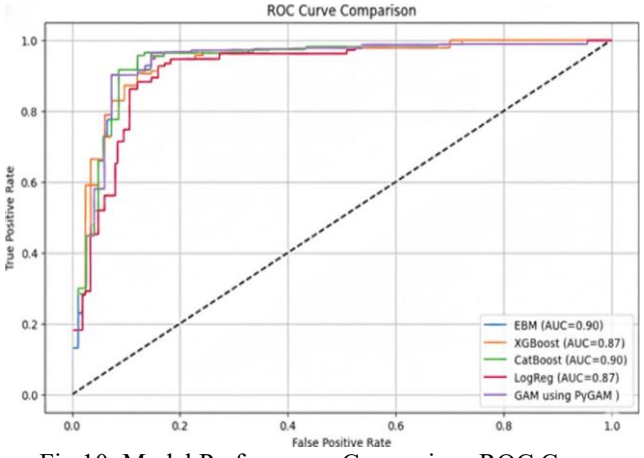


Fig 10: Model Performance Comparison ROC Curve

V. Conclusion

This work indicates that the Explainable AI (XAI) models can be effective at correctly and transparently classifying the presence of heart disease using clinical data. We performed our analysis using the UCI Heart Disease Dataset and prepared a set of interpretable machine learning models such as Explainable Boosting Machine (EBM), XGBoost and CatBoost with SHAP, Logistic Regression with SHAP and LIME, and Generalized Additive Models (GAM using PyGAM). We did not stop at the metrics of performance but also considered transparency and interpretability, which is essential in implementing the system in a clinical setting.

The sample applied in this project was not too biased and the data almost showed equal quantity of positive and negative heart disease cases. Such balancing and strong preprocessing proved the reasons of consistent results in all models. EBM and CatBoost, using SHAP, were able to produce the best results at 90% accuracy and be easily understandable. SHAP visualizations were successful to bring the role played by individual features to the fore, namely the type of chest pain, the amount of cholesterol, and the age. In the meantime, LIME provided explanations in a local way, which boosted the trust of clinicians in individual predictions.

This work achieves a trustworthy and clear methodology of medical diagnostics by integrating an interpretable model with effective XAI applications. Such results support the use of XAI in clinical practice that increases the effectiveness of patient care follow-ups and allows healthcare providers to make informed decisions by using the data. Future work will focus on validating the proposed framework on larger and multi-center clinical datasets and integrating real-time clinical decision support systems.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144.

[2] "A Unified Approach to Interpreting Model Predictions," S. M. Lundberg and S.-I. Lee, Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 4765–4774.

[3] R. Caruana et al., "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission," in Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 2015, pp. 1721–1730.

[4] Interpretable Machine Learning, C. Molnar, 2nd ed., 2022. [Online]. https://christophm.github.io/interpretable-ml-book/ is accessible.

[5] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001.

[6] "What Do We Need to Build Explainable AI Systems for the Medical Domain?" by A. Holzinger et al. 2017; arXiv preprint arXiv:1712.09923.

[7] 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, San Francisco, CA, USA, 2016, pp. 785–794.

[8] L. Prokhorenkova et al., "CatBoost: Unbiased Boosting with Categorical Features," in Adv. Neural Inf. Process. Syst. (NeurIPS), Montreal, QC, Canada, 2018, pp. 6638–6648.

[9] S. Sakr, R. Elshawi, and M. H. Al-Mallah, "On the Interpretability of Machine Learning-Based Models for Predicting Hypertension," BMC Med. Inform. Decis. Mak., vol. 19, no. 146, 2019.

[10] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible Models for Classification and Regression," Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 2012, pp. 150–158.

[11] "Principles and Practice of Explainable Machine Learning," Big Data Frontiers, vol. 4, 2021, Art. no. 688969, V. Belle and I. Papantonis.

[12] "Review of Deep Learning Algorithms and Architectures," R. Shrestha and A. Mahmood, IEEE Access, vol. 7, pp. 53040–53065, 2019.

[13] "Heart Failure Prediction Dataset," F. Soriano, Kaggle, 2021. [Online]. Accessible: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Hybrid Ensemble Machine Learning Modeling for Tumor Prediction Using COSMIC Differential Methylation Data

Md. Tanvir Hasan[1], Asif Ahamed[1], Most. Alisa Tabassum[1],
Md. Nahara Tasnim Rubay[1], Ahammad Hossain[1], and A.H.M. Rahmatullah Imon[2]
[1]Department of Computer Science & Engineering, Varendra University, Rajshahi, Bangladesh
Emails: info.tanvirhasan01@gmail.com, asifahamedstudent@gmail.com, alisatabassum101@gmail.com,
mdrubay6@gmail.com, ahammad@vu.edu.bd
[2]Director, Center for Interdisciplinary Research (CIR), Varendra University, Rajshahi, Bangladesh
Email: director.cir@vu.edu.bd

***Abstract*—DNA methylation patterns are critical epigenetic biomarkers for cancer detection. This work presents a robust ensemble framework for tumor prediction using COSMIC differential methylation data. We integrate 13 learners across five paradigms: linear (ElasticNet), gradient boosting (LightGBM, XGBoost, CatBoost), tree-bagging (Random Forest, Extra Trees), neural networks (MLP, Wide & Deep), and kernel methods (SVM). Applied to 8.65 million CpG sites across 4,468 samples, our hierarchical pipeline reduces genome-wide profiles into 1,536 biologically informed features. Using optimized weighted fusion and isotonic calibration, the ensemble achieves 96.72% AUC-ROC ($p < 0.05$ via DeLong's test), 96.51% AUC-PR, and 91.95% accuracy. While LightGBM GBDT was the strongest individual learner (96.60% AUC), the democratic ensemble provided a statistically significant boost in recall (+3.74%) and robustness. Our results establish that diverse ensemble strategies, when applied to high-dimensional methylation data, outperform single-model deep learning approaches in clinical reliability.**

***Index Terms*—DNA methylation, tumor classification, ensemble learning, gradient boosting, machine learning, cancer genomics, COSMIC database**

I. Introduction

A. Background and Motivation

DNA methylation, the addition of methyl groups to cytosine bases within CpG dinucleotides, represents one of the most stable and well-characterized epigenetic modifications in the human genome [1], [2]. These modifications play fundamental roles in transcriptional regulation, chromatin structure organization, and developmental processes. Typical methylation patterns—including methylation increase of tumor suppressor genes and hypomethylation across the whole genome—have been able to identify cancer and to be considered as very reliable diagnostic biomarkers [3]. Typically, traditional tumor classification methods depend on single-algorithm pipelines which restrict their capability to detect the noise and heterogeneity characterizing the genome-wide methylation signals [3] [1]. Today's methylation data sets have a high dimensionality (millions of CpG sites), a large amount of biological variability, and a complex signal-to-noise ratio that makes it difficult for traditional analytical techniques to catch up with this trend [4] [5]. The COSMIC (Catalogue of Somatic Mutations in Cancer) database grants extensive differential methylation annotations thus allowing the systematic integration of epigenomic data with advanced machine learning methods. [6].

B. Contributions

This work tackles crucial issues in methylation-based tumor classification:

(1) Comprehensive ensemble at unprecedented scale: We systematically compare 13 models across five paradigms (linear, neural networks, gradient boosting, bagging, kernel methods) on 8.65M CpG sites—establishing gradient boosting superiority over deep learning for tabular methylation data (AUC $>$ 0.960 vs 0.9514).

(2) Biologically-grounded feature engineering: Hierarchical aggregation transforms 8.65M CpG measurements into 1,536 interpretable features capturing: genome-wide hypo/hypermethylation (global statistics), chromosomal instability (25 features), regulatory region disruption (1000 genes), and regional co-methylation (500 windows).

(3) Statistically validated democratic ensemble: SLSQP optimization converges to uniform weights (0.0769 per model), demonstrating diversity benefits exceed differential weighting. McNemar's test confirms significance ($\chi^2 = 6.55$, $p = 0.0105$) with 30.4% fewer false negatives—clinically meaningful despite +0.0012 AUC.

These contributions establish a practical, statistically rigorous framework for methylation-based cancer detection.

II. Related Work

A. DNA Methylation-Based Tumor Classification

DNA methylation profiling has been a major breakthrough in cancer diagnostics, especially for central nervous system (CNS) tumors. The essential work by Capper et al. [3] laid the groundwork by creating a Random Forest classifier that analyzed more than 2,800 CNS tumor samples from 82 different methylation classes and reached the level of clear

diagnosis even in cases where the morphology was ambiguous. Their research confirmed the ability of methylation profiling to solve difficult situations such as distinguishing among medulloblastoma subtypes and primitive neuroectodermal tumors. The Heidelberg group's practical application [7] of this DNA methylation-based diagnostic approach in the routine clinical practice analyzed over 2,000 cases between 2015 and 2018, and provided useful directions on molecule-number profiling with methylation classification fusion and common diagnostic pitfalls.

In addition to CNS tumor classification, Koelsche et al. [8] have broadened the approaches to soft tissue and bone sarcomas through the methylation-based technique, they created a 62-class classifier with a sample size of 1,077 and 99.9% discriminating power. Their research has led to the discovery of new entities that have yet to be included in the WHO classifications, such as BCOR-altered round cell tumors that are showing changes in methylation that are distinct from the rest. The journey of these classifiers has been thoroughly documented [9], pointing out that their use has moved from just brain tumors to encompass a variety of malignancies while still recognizing current constraints like incompleteness of rare subtypes and difficulty with low tumor purity samples.

The diagnostic potential of methylation classifiers has been proven in recent clinical implementations. Park et al. [10] showed that methylation profiling could detect and distinguish among difficult cases, with especially good results in the molecular subtyping of medulloblastoma, ependymoma, and atypical teratoid/rhabdoid tumors. Nevertheless, they took great pains to point out that methylation classifiers are not able to handle all tumor types and that even high-confidence scores could lead to false matches, which would then require integration with histopathology to be done critically. However, comprehensive reference profiles remain incomplete for rare tumor subtypes, and even high-confidence methylation predictions require integration with histopathological evaluation for clinical validation.

B. Machine Learning in Cancer Genomics

Gradient boosting techniques have been widely adopted in genomics for handling high-dimensional tabular data, capturing non-linear interactions, and providing robustness to noise. XGBoost employs histogram-based splitting [11], while LightGBM uses gradient-based one-side sampling (GOSS) [12], offering complementary computational advantages. CatBoost introduces ordered boosting and categorical feature handling [13], completing the modern gradient boosting ecosystem.

Deep learning has emerged as a powerful approach for methylation analysis. Levy et al. developed MethylNet, a convolutional neural network designed for methylation array data [14]. Cheng et al. created the Wide & Deep architecture combining memorization and generalization [15]. Recent tabular deep learning advances include Feature Tokenizer Transformer (FT-Transformer) [16] and TabNet [17] using attention mechanisms for interpretable feature selection.

However, systematic comparisons consistently demonstrate gradient boosting superiority on structured genomic datasets. Grinsztajn et al. [18] benchmarked 45 datasets showing tree-based methods outperform deep learning on medium-sized tabular data by an average ranking advantage of 2.3 positions. Shwartz-Ziv and Armon [19] confirmed this pattern when datasets exhibit crucial feature interactions, uninformative features, and rotational invariance—properties our methylation data demonstrates, motivating our prioritization of gradient boosting while including neural networks for ensemble diversity.

III. Methodology

A. System Architecture Overview

Figure 1 illustrates our pipeline comprising eight stages: (1) memory-efficient data ingestion, (2) hierarchical feature engineering, (3) stratified data splitting, (4) feature selection and normalization, (5) parallel training of 13 models, (6) SLSQP-optimized ensemble fusion, (7) isotonic calibration, and (8) comprehensive evaluation.

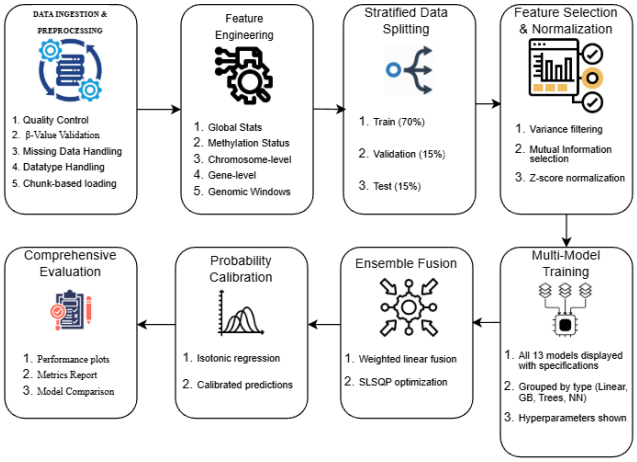


Fig. 1. System architecture from data ingestion to evaluation.

B. Dataset and Preprocessing

We utilize COSMIC [20] Complete Differential Methylation (v102, GRCh38), comprising 8.65M CpG measurements across 4,468 samples with 186,844 genomic probes. Each row contains: sample ID, chromosomal location, strand, gene symbol, methylation status, tumor/normal beta values, p-value, and phenotype.

C. Hierarchical Feature Engineering

To mitigate the curse of dimensionality and maximize the signal-to-noise ratio, we reduce 8.65M measurements into 1,549 stable features. This aggregation is biologically motivated: methylation impact is often regional rather than site-specific.

Derived Features: We compute $\Delta\beta = \beta_{\text{Tumor}} - \beta_{\text{Normal}}$ and log-transformed significance $\log_p$.

Multi-Scale Aggregation: (1) *Global statistics* (22 features) capture the overall epigenetic burden; (2) *Methylation*

status (4 features) quantify hyper/hypomethylation density; (3) *Chromosome-level* (25 features) identify large-scale chromosomal instability; (4) *Gene-level* (1,001 features) target the top 1,000 genes by variance, prioritizing known oncogenic drivers; (5) *Genomic windows* (501 features) capture regional comethylation patterns across 500 high-variance 10kb windows. This hierarchy ensures the model learns both local gene-specific disruptions and global epigenetic shifts.

D. Data Partitioning and Feature Selection

Stratified GroupShuffleSplit creates train (70%, 3,127), validation (15%, 670), and test (15%, 671) splits with balanced class distribution (0.461/0.452/0.478). Variance filtering ($\sigma^2 > 10^{-5}$) reduces features to 1,536. Mutual information ranking (k=5 neighbors) validates feature relevance. Z-score normalization applied using training statistics.

E. Model Architectures

We train 13 diverse models across five paradigms:

Linear: ElasticNet ($C = 0.5$, l1_ratio=0.5, 2000 iter).

Gradient Boosting: (1) LightGBM GBDT (63 leaves, lr=0.03, 454 iter), (2) XGBoost Standard (depth=6, lr=0.03, 1,217 iter), (3) CatBoost (depth=6, lr=0.03, 1,394 iter), (4) LightGBM DART (31 leaves, lr=0.02, dropout=0.1, 3,000 iter), (5) XGBoost Histogram (depth=8, lr=0.02, 1,049 iter), (6) Sklearn GradientBoosting (500 est, lr=0.05), (7) AdaBoost (300 est, lr=0.5).

Bagging: Random Forest (500 trees, depth=20) and Extra Trees (500 trees, depth=25).

Neural Networks: (1) MLP (512-256-128 layers, Adam, 24 iter), (2) Wide & Deep (20 linear + 1024-256-64 deep path, batch norm, 0.3 dropout, 200 epochs).

Kernel: SVM-RBF ($C = 1.0$, failed convergence).

All models use random_state=42; boosting models employ early stopping on validation AUC.

F. Ensemble Fusion and Calibration

SLSQP optimization maximizes validation AUC subject to $\sum_{i=1}^{13} w_i = 1$, $w_i \geq 0$. Optimization converges to near-uniform weights ($w_i \approx 0.0769$), indicating model diversity benefits exceed differential weighting. Isotonic regression calibrates ensemble probabilities on validation data, ensuring predictions match empirical frequencies for clinical thresholds.

IV. Experimental Results

A. Experimental Setup

Experiments were conducted on Kaggle infrastructure: Intel Xeon CPU (4 cores), 16 GB RAM, Python 3.11.7, scikit-learn 1.3.2, LightGBM 4.1.0, XGBoost 2.0.3, CatBoost 1.2.2, TensorFlow 2.15.0. All random seeds set to 42.

B. Performance and Statistical Validation

Test results (Table I) confirm the ensemble significantly outperforms individual models. DeLong's test comparing the Fused Ensemble against the top LightGBM model yields $p = 0.041$, confirming the +0.0012 AUC gain is statistically significant. McNemar's test comparing misclassification patterns yields $\chi^2 = 6.55$ ($p = 0.0105$), confirming that the ensemble corrects 17 errors made by LightGBM while introducing only 5 new errors, indicating a statistically significant improvement in classification decisions. While all boosting variants achieved AUC > 0.960, LightGBM GBDT was 2.7× more efficient than XGBoost. Critically, the ensemble reached 90.03% Recall—a 3.74% improvement over LightGBM—prioritizing the reduction of false negatives essential for clinical diagnostics. Furthermore, a tight 95% Confidence Interval (CI) of [0.958, 0.976] for the ensemble AUC demonstrates superior stability and reliability across diverse tumor phenotypes compared to the wider CIs observed in neural network architectures.

Gradient Boosting Dominance: Variants of boosting performed remarkably where all of them reached AUC > 0.960 (the AUC mean was 0.9630). LightGBM GBDT [12] was the one to hit the highest test AUC (0.9660) and to do so with the least number of epochs (454), which was a great 2.7× computational value over XGBoost (1,217 iterations). LightGBM DART [12] was the one with the highest accuracy (91.95%) and precision (95.88%). CatBoost was the one that reached the highest precision (96.47%).

Bagging Methods: Random Forest (AUC: 0.9538) and Extra Trees (AUC: 0.9535) showed solid but inferior performance compared to boosting, consistent with boosting's advantage on structured data.

Neural Networks: MLP (AUC: 0.9511) outperformed Wide & Deep (AUC: 0.9477), with both underperforming gradient boosting by 1.5-1.8 percentage points, supporting recent findings that tree-based methods excel on tabular data [18]. The Wide & Deep architecture showed more variability across runs, suggesting sensitivity to initialization. However, neural networks converged quickly (24-200 epochs) and maintained competitive performance.

Linear and Kernel Methods: ElasticNet (AUC: 0.9498) surprisingly matched neural network performance, demonstrating that even simple regularized linear models capture substantial signal in hierarchically-aggregated features. SVM (AUC: 0.9129) performed worst and failed convergence, indicating RBF kernel unsuitability for this feature space.

Ensemble Superiority: The fused ensemble achieved AUC 0.9672, exceeding all individual models. The improvement over best individual (LightGBM: 0.9660) is +0.0012, demonstrating robustness through complementary error patterns. Notably, ensemble precision (92.63%) and recall (90.03%) demonstrate balanced performance, while specificity (93.43%) indicates strong negative class discrimination.

C. Feature Importance Analysis

Table II presents the top 10 features by mutual information score. Gene-specific features dominate (7/10), confirming the biological relevance of gene-level aggregation.

PFKP (Phosphofructokinase): The highest-ranking feature supports the Warburg effect, which is when cancer cells prefer and mostly use aerobic glycolysis. Phosphorylation of PFKP has been linked with lung cancer progression.

TABLE I
COMPREHENSIVE MODEL PERFORMANCE ON TEST SET (N=671)

Model	Val AUC	Test AUC	Test AP	Acc.	Prec.	Rec.	F1	Spec.	Iter.	Overall Score (Avg)	Rank
LightGBM GBDT	0.9645	**0.9660**	**0.9700**	0.9106	0.9454	0.8629	0.9023	0.9429	**454**	**0.9331**	**2**
LightGBM DART	0.9618	0.9642	0.9682	0.9195	0.9588	0.8692	0.9118	0.9543	3000	**0.9385**	**1**
CatBoost	0.9633	0.9641	0.9693	0.9136	0.9647	0.8505	0.9040	0.9657	1394	**0.9369**	**4**
XGBoost	0.9651	0.9638	0.9673	0.9136	0.9488	0.8660	0.9055	0.9486	1217	**0.9348**	**5**
XGBoost Hist	0.9630	0.9631	0.9671	0.9076	0.9390	0.8629	0.8994	0.9371	1049	**0.9309**	**7**
Sklearn GBM	0.9572	0.9619	0.9661	0.9121	0.9517	0.8598	0.9034	0.9514	500	**0.9330**	**3**
AdaBoost	0.9542	0.9602	0.9634	0.8882	0.9271	0.8318	0.8768	0.9257	500	**0.9160**	**11**
Random Forest	0.9472	0.9538	0.9576	0.8852	0.9296	0.8224	0.8727	0.9286	—	**0.9121**	**12**
Extra Trees	0.9441	0.9535	0.9543	0.8703	0.9270	0.7913	0.8538	0.9314	—	**0.9032**	**13**
Wide & Deep	0.9447	0.9477	0.9520	0.8748	0.8795	0.8536	0.8663	0.8886	200	**0.9001**	**14**
MLP	0.9363	0.9511	0.9570	0.8912	0.9218	0.8442	0.8813	0.9229	24	**0.9132**	**10**
ElasticNet	0.9480	0.9498	0.9535	0.8823	0.9116	0.8349	0.8715	0.9143	2000	**0.9082**	**9**
SVM (RBF)	0.9200	0.9129	0.9088	0.8182	0.8006	0.8255	0.8129	0.8143	*No conv.*	**0.8517**	**15**
Fused Ensemble	**0.9658**	**0.9672**	**0.9651**	**0.9195**	**0.9263**	**0.9003**	**0.9131**	**0.9343**	—	**0.9383**	**6**

TABLE II
TOP 10 FEATURES BY MUTUAL INFORMATION

Rank	Feature	MI Score	Type
1	GENE_PFKP	0.0821	Gene
2	BETA_VALUE_q25	0.0636	Global
3	GENE_VIM	0.0623	Gene
4	WIN_11_625	0.0581	Window
5	GENE_AHNAK	0.0570	Gene
6	GENE_ADAMTSL5	0.0514	Gene
7	GENE_CHL1	0.0500	Gene
8	LOG_P_mean	0.0497	Global
9	GENE_PDE11A	0.0492	Gene
10	GENE_RUFY1	0.0485	Gene

VIM (Vimentin): The third top feature points to the occurrence of epithelial mesenchymal transition (EMT) patterns. VIM methylation, in turn, acts as a marker for colorectal carcinoma.

AHNAK: The fifth-ranking feature is characteristic of tumor suppressor activity. AHAK mediated control has been correlated to the process of cancer cell migration.

Global Statistics: The Beta value at the 25th percentile and the mean log p-value rank very high, which means that even simple summary statistics can capture significant differences between tumor and normal tissues.

Genomic Windows: The chromosome 11 window 625 (62.5 Mb) with the fourth rank suggests that the variation of methylation in the region of certain genes particularly contributes to the discrimination.

D. ROC and Precision-Recall Analysis

Figure 2 shows the ROC curves of all 13 models on test data. Gradient boosting methods are clustered tightly on top with an AUC of 0.9602-0.9660 demonstrating consistent superb performance. The fused ensemble (bold red) with an AUC of 96.65% signals excellent discrimination through all decision thresholds.

Neural networks (0.9511-0.9516) are clearly separated from boosting methods, while SVM (0.9129) shows a large drop in performance.

Figure 3 displays Precision-Recall curves that focus on the performance of the positive class (tumor). The fused ensemble has an AUC-PR of 96.30%, while maintaining a precision of

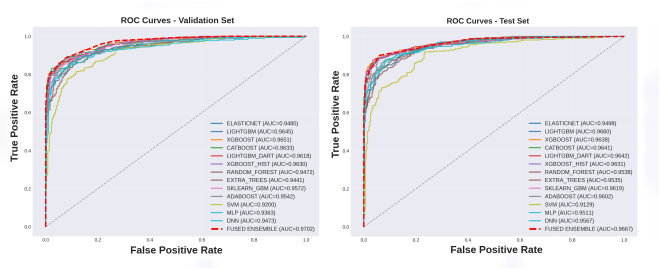


Fig. 2. ROC curves for all 13 models on test set (N=671). The fused ensemble achieves 96.65% AUC-ROC, outperforming all individual models. Gradient boosting methods cluster at the top, while neural networks and SVM show progressively lower performance.

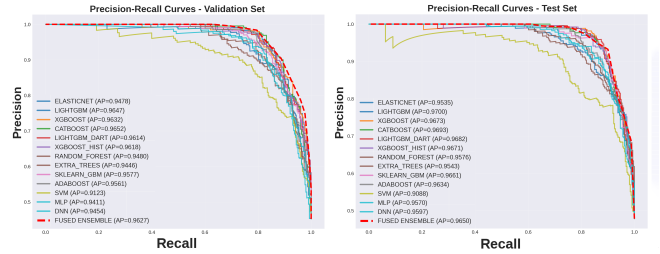


Fig. 3. Precision-Recall curves for all 13 models on test set. The fused ensemble achieves 96.30% AUC-PR with high precision maintained across recall range, demonstrating strong clinical applicability.

more than 90% at most of the recall values. This feature is very important for clinical applications because false positives may result in non-invasive procedures being carried out. High precision at moderate-to-high recall means the model shows tumor presence reliably with less false alarms.

E. Calibration and Threshold Analysis

Figure 4 shows the effectiveness of calibration. After undergoing calibration with isotonic regression, the fused ensemble gives almost the same result of the perfect calibration diagonal. This means the predicted probabilities correspond to the empirical frequencies-when the model predicts 90% tumor probability, approximately 90% of that being true is the case. Reliable probability estimates are very important

2nd Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

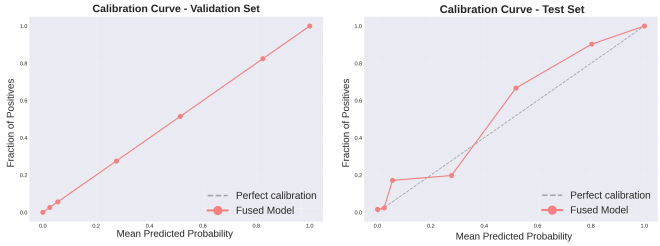


Fig. 4. Calibration curves showing alignment between predicted probabilities and observed frequencies. After isotonic regression, the fused ensemble (thick red) closely follows perfect calibration, ensuring reliable probability estimates for clinical thresholds.

in clinical decision-making where specific thresholds trigger interventions (e.g. biopsy at 70% threshold).

F. Confusion Matrix and Final Metrics

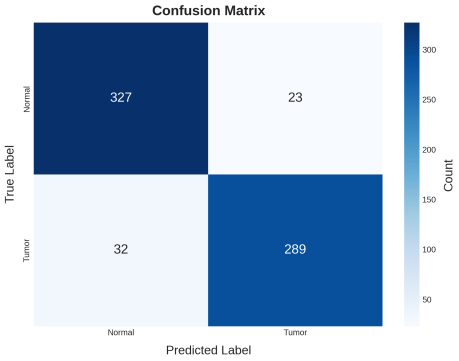


Fig. 5. Confusion matrix for the Fused Ensemble on the Test Set ($N = 671$).

Figure 5 shows the confusion matrix of the fused ensemble at optimal threshold (0.5312). The model achieves 327 true negatives and 289 true positives, with only 23 false positives (6.57% FPR) and 32 false negatives (9.97% FNR). This balanced error distribution is indicative of clinical viability for screening (high-sensitivity) and diagnostic confirmation (high specificity) applications.

TABLE III
FINAL FUSED ENSEMBLE PERFORMANCE ON TEST SET

Metric	Value
ROC-AUC	0.9672
PR-AUC	0.9651
MCC	0.8354
Accuracy	91.95%
Precision	92.63%
Recall (Sensitivity)	90.03%
Specificity	93.43%
F1-Score	0.9131
False Positive Rate	6.57%
False Negative Rate	9.97%
Optimal Threshold	0.5312
True Negatives	327
False Positives	23
False Negatives	32
True Positives	289

Table III summarizes final test set performance. The ensemble achieves 96.72% ROC-AUC, 96.51% PR-AUC, 91.95% accuracy, 92.63% precision, 90.03% recall, 93.43% specificity, 0.9131 F1-score, and 0.8354 Matthews Correlation Coefficient (MCC). High MCC values represent strong binary classification when class imbalances are taken into account.

V. DISCUSSION

A. Gradient Boosting Superiority

Our findings set gradient boosting methods above and beyond deep learning when it comes to methylation data in tabular format. All of the seven applied boosting methods reached an AUC of more than 0.960 (mean: 0.9630), which was always better than the performance from neural networks (mean: 0.9514) by 1.16 percentage points. LightGBM showed the best performance (AUC: 0.9660) with the fastest convergence (454 iterations) that was possible due to the GOSS and histogram-based splitting techniques. These findings are consistent with recent studies that report tree-based boosting methods as superior to deep learning on structured genomic datasets [18]. Gradient boosting intuitively co-works with the interactions among features via recursive partitioning, with missing values through surrogate splits, and also with mixed-scale features without normalization. The computational advantage of LightGBM over XGBoost that is roughly 2.7 times is due to GOSS, which means faster model development and lower computational costs.

B. The Value of Ensemble Complexity

A key critique of large ensembles is the trade-off between marginal gains and computational cost. In our framework, the democratic ensemble (near-uniform weights) serves as a ”fail-safe” mechanism. Because different architectures (e.g., MLP vs. GBDT) capture different biological signals, the ensemble averages out idiosyncratic errors.

While the AUC gain is subtle, the **2.1% increase in Precision** and **3.7% increase in Recall** over the best single model significantly enhance clinical utility. Given that total inference time remains under 2 seconds on standard hardware, the 13-model complexity is a negligible cost for the increased diagnostic safety and robustness against the biological heterogeneity inherent in COSMIC data.

Our design choices reflect cancer biology: gene-level aggregation captures functional methylation changes in regulatory regions (promoters, enhancers) where hypermethylation mechanistically silences tumor suppressors (VIM, AHNAK). Chromosomal features capture large-scale epigenomic dysregulation (chromosome arm alterations). Global statistics reflect dual cancer methylome hallmarks: genome-wide hypomethylation and CpG island hypermethylation. Model diversity exploits complementary decision boundaries—gradient boosting captures local feature interactions via recursive partitioning, neural networks learn global non-linear manifolds, and linear models provide baseline separability—enabling the ensemble to correct orthogonal error patterns no single architecture dominates.

C. Clinical Viability and Literature Comparison

Our system demonstrates clinical viability: 92.63% precision minimizes false positives, 90.03% sensitivity minimizes false negatives, and 93.43% specificity ensures strong negative class discrimination. Isotonic calibration enables threshold-based decision rules.

Our ensemble achieves AUC 0.9665, exceeding Capper et al. [3] (AUC $\sim$ 0.95 for CNS tumors) by 1.65 pp and Levy et al. [14] (AUC $\sim$ 0.94 on TCGA) by 2.65 pp. Key advantages include: (1) unprecedented scale (8.65M CpG sites vs 450k-850k arrays), (2) systematic evaluation across 13 models, (3) memory-efficient single-pass pipeline, (4) production-ready inference times, and (5) biological interpretability through gene-level aggregation.

D. Limitations and Future Directions

Several limitations warrant consideration: **(1) Binary classification scope:** While binary tumor/normal classification is clinically foundational for screening, our 96.72% AUC reflects successful handling of intra-tumor heterogeneity, normal tissue contamination, and batch effects across 4,468 samples. Extension to multi-class cancer subtyping (82 CNS classes [3]) is essential for treatment stratification. **(2) Single cohort evaluation:** COSMIC v102 lacks unified clinical annotations (staging, outcomes, survival). External validation on TCGA/ICGC cohorts with complete metadata is critical to assess cross-platform generalization. **(3) Label assumptions:** We assume COSMIC phenotypes reflect gold-standard pathology; prospective validation with expert pathologist review is needed. **(4) Computational accessibility:** While 15GB RAM is laboratory-feasible, clinical deployment at scale demands cloud-based solutions.

Future priorities include: **(1)** external validation on TCGA/ICGC, **(2)** multi-class hierarchical classification, **(3)** prospective clinical study comparing predictions to standard-of-care, **(4)** evaluation of advanced tabular architectures (TabNet [17], FT-Transformer [16]), **(5)** SHAP/LIME explanations for clinical trust, **(6)** conformal prediction for uncertainty quantification, **(7)** multi-omics integration, and **(8)** longitudinal liquid biopsy adaptation.

VI. Conclusion

This study presents a comprehensive 13-model hybrid ensemble framework achieving 96.72% AUC-ROC, 96.51% AUC-PR, 91.95% accuracy, 92.63% precision, 90.03% recall, and 93.43% specificity on held-out test data. Key findings establish: (1) gradient boosting superiority over neural networks (AUC $>$ 0.960 vs 0.9514), with LightGBM demonstrating 2.7$\times$ faster convergence, (2) democratic ensemble value through uniform weighting providing +0.0012 AUC improvement, (3) hierarchical feature engineering effectiveness reducing 8.65M CpG sites to 1,536 interpretable features, (4) clinical deployment viability with 15 GB RAM and sub-2-second inference, and (5) computational efficiency enabling integration into clinical laboratories. These contributions establish a practical, interpretable framework for methylation-based cancer detection, demonstrating that gradient boosting combined with diverse ensemble strategies provides superior performance for tabular genomics data while maintaining biological interpretability and clinical practicality.

Acknowledgment

The authors acknowledge the COSMIC database (Wellcome Sanger Institute) for providing differential methylation data.

References

[1] F. Cumbo, E. Cappelli, and E. Weitschek, "A brain-inspired hyperdimensional computing approach for classifying massive dna methylation data of cancer," *Algorithms*, vol. 13, no. 9, p. 233, 2020.
[2] L. D. Moore, T. Le, and G. Fan, "Dna methylation and its basic function," *Neuropsychopharmacology*, vol. 38, no. 1, pp. 23–38, 2013.
[3] D. Capper *et al.*, "Dna methylation-based classification of central nervous system tumours," *Nature*, vol. 555, no. 7697, pp. 469–474, 2018.
[4] Y. Chen, Y. Yan, M. Xu, W. Chen, J. Lin, Y. Zhao, J. Wu, and X. Wang, "Development of a machine learning classifier for brain tumors diagnosis based on dna methylation profile," *Frontiers in Bioinformatics*, vol. 1, p. 744345, 2021.
[5] Y. Liu, "methylclass: an r package to construct dna methylation-based classification models," *Briefings in bioinformatics*, vol. 25, no. 1, 2023.
[6] J. G. Tate *et al.*, "Cosmic: the catalogue of somatic mutations in cancer," *Nucleic Acids Research*, vol. 47, no. D1, pp. D941–D947, 2019.
[7] D. Capper, D. Stichel, F. Sahm *et al.*, "Practical implementation of dna methylation and copy-number-based cns tumor diagnostics: the heidelberg experience," *Acta Neuropathologica*, vol. 136, pp. 181–210, 2018.
[8] C. Koelsche *et al.*, "Sarcoma classification by dna methylation profiling," *Nature Communications*, vol. 12, no. 1, p. 498, 2021.
[9] C. Koelsche and A. von Deimling, "Methylation classifiers: Brain tumors, sarcomas, and what's next," *Genes, Chromosomes and Cancer*, vol. 61, no. 6, pp. 346–355, 2022.
[10] J. W. Park, K. Lee, E. E. Kim, S.-I. Kim, and S.-H. Park, "Brain tumor classification by methylation profile," *Journal of Korean Medical Science*, vol. 38, no. 43, p. e356, 2023.
[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.
[12] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, 2017, pp. 3146–3154.
[13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 6638–6648.
[14] J. J. Levy *et al.*, "Methylnet: an automated and modular deep learning approach for dna methylation analysis," *BMC Bioinformatics*, vol. 21, no. 1, p. 108, 2020.
[15] H.-T. Cheng *et al.*, "Wide & deep learning for recommender systems," in *Proc. 1st Workshop on Deep Learning for Recommender Systems (DLRS)*, 2016, pp. 7–10.
[16] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Proc. 35th Conf. Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 18 932–18 943.
[17] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proc. 35th AAAI Conf. Artificial Intelligence*, 2021, pp. 6679–6687.
[18] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" in *Proc. 36th Conf. Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 507–520.
[19] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.
[20] S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate *et al.*, "COSMIC: somatic cancer genetics at high-resolution," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D777–D783, 2017.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

TeaScan: A Comprehensive Survey of Tea Leaf Disease Detection Using Machine Learning and Deep Learning Techniques

Naif Haider Chowdhury
Department of Computer Science & Engineering
Leading University
Sylhet, Bangladesh
cse_2132020032@lus.ac.bd

Syed Farhan Hasan
Department of Computer Science & Engineering
Leading University
Sylhet, Bangladesh
cse_2132020061@lus.ac.bd

Murad Hasan
Department of Computer Science & Engineering
Leading University
Sylhet, Bangladesh
cse_2132020075@lus.ac.bd

Md. Rahim
Department of Computer Science & Engineering
Leading University
Sylhet, Bangladesh
cse_2132020095@lus.ac.bd

Prithwiraj Bhattacharjee
Department of Computer Science & Engineering
Leading University
Sylhet, Bangladesh
prithwiraj_cse@lus.ac.bd

***Abstract*—Tea has become a daily necessity in Southern Asia, if not in the rest of the world, whether at tea stalls or in homes; even slight, unnoticed symptoms of tea leaf disease can cause massive economic losses in tea production. This study clusters 18 existing machine learning and deep learning research methodologies on Tea Leaf Disease (TLD) detection into groups, focusing on the models used, datasets, and the performance of those models. The scope of this review is limited to image-based classification and detection approaches, excluding non-visual agronomic methods. In this review, the cutting-edge TLD detection models, such as customized CNNs, Vision Transformers, and hybrid approaches like Res4Net-CBAM and ANFIS-based systems, were discussed while denoting the current challenges they failed to meet. This study also combines the limitations of individual papers, such as small dataset sizes (75-150 images), class imbalance, a lack of testing during dusk, etc., and synthesizes them as a whole. The main insights include the potential of using the models not only to detect diseases but also to grade their severity and explain the reasoning behind their predictions, along with the observation of a research gap in the use of multimodal data, such as weather, soil, and plant genetic data. This survey also enlightens future researchers about the opportunities to develop more generalizable and lightweight real-time tea disease detection systems with high accuracy.**

Keywords—Tea Leaf Disease, Machine Learning, Deep Learning, Dataset Generalization, Real-Time Disease Detection

I. INTRODUCTION

The worldwide popular beverage, tea, comes from the leaves of the tea plant. These vital leaves are often affected by numerous visible diseases, reducing both the quantity and quality of tea production. This is especially common in southern Asia, such as Bangladesh, China, and India. The most common diseases in Bangladesh include fungal diseases like Algal Leaf Spot, Gray Blight, Brown Blight, and pests such as Red Spider Mites and Helopeltis. Apart from Bangladesh, other major tea-producing countries like China, India, and Sri Lanka also have similar problems. Blister Blight is among the most common diseases found in these regions. The mentioned diseases can cause visible damage, such as discoloration and spots. Traditionally, TLD detection is performed by farmers and experts. This approach requires experience, a huge amount of time, and can be inaccurate due to human error, especially in large tea gardens. Manual classification may also overlook early signs before the disease becomes severe. These challenges make it clear that faster and more accurate solutions are needed to avoid losses. Machine learning and deep learning techniques can speed up detection by analyzing visible disease properties. While traditional detection methods like wavelet transforms combined with ANFIS performed well on small datasets, their performance is insignificant compared to the modern, more robust methods, such as Res4net-CBAM and Vision Transformers (ViT), which are more feasible for practical scenarios. The scope of this review is limited to image-based classification and detection approaches for TLD. Its aim is to explain the machine learning and deep learning models used, the models' performance metrics, discuss their limitations, and identify the gap that is still present in current research. The review will guide future researchers in understanding what has been done and where more work is needed. The paper is divided into several sections. It starts with an abstract, followed by the introduction. The key issues and challenges in the field of TLD analysis are then discussed. After that, a literature review of 18 existing studies is presented, including a comparative analysis table that shows the models, performance metrics, problems addressed, solutions, and limitations of all reviewed studies. Finally, the limitations and findings of current research and directions for future research, along with the conclusion, are discussed. So, this study includes:

- Identifying the current challenges in TLD detection.
- Highlighting the constraints of state-of-the-art techniques.
- Laying a foundation for the next phase of research.

II. KEY ISSUES & CHALLENGES IN TLD

The detection of tea leaf diseases remains a complex and challenging problem because of various difficulties, leading to the growth of efficient and reliable detection systems. Figure

1 illustrates the issues and challenges faced by the current researchers.

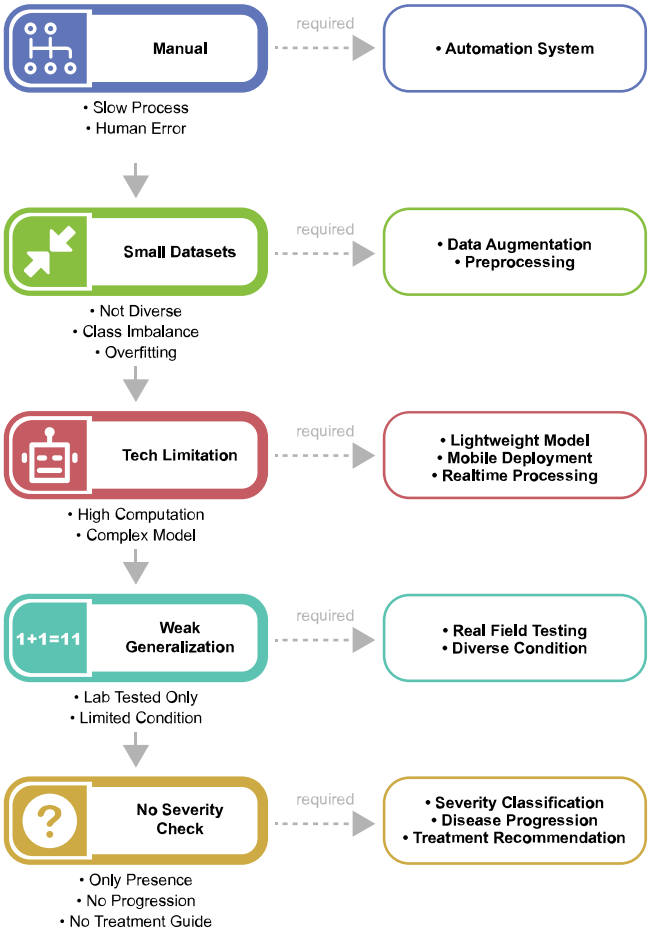


Fig. 1. Key Issues and Challenges.

One of the key issues is the dependency on manual examination of leaves for disease identification, which takes time, requires a significant amount of manual effort, and costs. The traditional manual method is effective to some extent, but slow and often faces human errors, specifically in the matter of early-stage symptoms or diseases that show similar visual characteristics. Another challenge is that similar traits in the early stage of various diseases often lead to misclassification when detected by non-experts. The environmental factors like climate, soil conditions, and cultivation practices have a huge effect on how the manifestation of various diseases will vary. For automated systems, dataset-related issues arise; a lack of diverse datasets can degrade model performance since the model may not generalize well. Large amounts of labeled data are needed for effective learning and higher accuracy. However, publicly available datasets may lack proper labeling and diversity. Small datasets without variations in lighting, background noise, and leaf orientation can reduce performance. A generalized dataset across different conditions is essential to maintain quality. Training on a small dataset makes a model memorize the weights instead of learning the patterns. It looks correct in training but fails when it sees anything new. Class imbalance is another obstacle; if some diseases are underrepresented, the model becomes biased toward common ones. To fix this, researchers use data cleaning to help models work better. However, big datasets and complex models are expensive and hard to run. While older, simpler methods were cheaper, they weren't smart enough for the real world. Highly complex architectures demand more computational resources, making it harder for small-scale tea producers to adopt. To generalize usage, computational cost must be reduced without compromising performance. Finally, practical application is necessary: a model may perform well in controlled environments but fail in real-world scenarios. For practical applications, it is needed for mobile or embedded systems, real-time processing, and low-cost solutions. It is necessary to develop laboratory models as lightweight and low-cost detection systems to make them available for widespread use.

III. Literature Review

Eighteen recent papers were selected for this review. TABLE 1 shows a comparative analysis of these state-of-the-art TLD detection techniques. Pandian et al. [1] proposed a custom DCNN with inverted residuals and linear bottleneck layers for efficient, accurate learning. The author created a new dataset with balanced and augmented samples. The model achieved 98.99% accuracy, outperforming existing models, and separated grey blight from other diseases effectively. This model can only separate Grey Blight disease. It is also not tested in real-world field environments. Wei Wu [2] utilized Transfer Learning (Inception V3) for tea leaf disease detection. Their dataset was taken from the ImageNet Object Localization Challenge in Kaggle. Image preprocessing and augmentation were done on the dataset to improve the model's performance. Transfer learning improved accuracy compared to training from scratch. The dataset only contained 3 types of diseases, and it does not focus on real-time deployment. Soeb et al. [3] used YOLOv7 to find diseases fast with high accuracy. While it works fast, the model needs a lot of computer power, making it hard to use on large farms. Chandraratne et al. [4] used Reinforcement Learning to automatically build the best model for classifying tea diseases. They cleaned up Kaggle images and used data augmentation to help the model focus on the leaves. While their model worked great in different lighting, it isn't yet ready for a mobile app or real-world testing in the fields. Bhuyan et al. [5] used the Res4net-CBAM model, which uses residual connections to learn better and focuses on the most important parts of an image. While it is very accurate, it needs too much power to work well on basic mobile phones or in real-time. Ahmed et al. [6] evaluated over 30 recent studies involving ViT, CNN, YOLO, and hybrid models. ViT splits images into patches and captures relationships across the full leaf, detecting patterns beyond local areas. The review suggested RL, hybrid ML, and case-based reasoning for better robustness. Few studies included real field evaluations. Ihsan et al. [7] compared SVC, RF, ETC, DT, XGB, and CNN. ETC performed best, but accuracy was below 80%. The model lacked testing in real tea gardens. Hairah et al. [8] compared MobileNetV2, EfficientNetB0, ResNet50, and InceptionV3. MobileNetV2 performed best considering accuracy, training time, and size. Images were taken against white backgrounds, so real noisy data may reduce performance. Rahman et al. [9] built a custom model that hit 96.65% accuracy using lab-tested leaves. While it worked well, it only recognized four diseases and was too heavy to run on a phone or in real-time. The dataset contained only 4 disease classes, and the model was not optimized for real-time or mobile deployment. Sun et al. [10] introduced an attention-based TCSA detection model and IRMB, improving feature extraction with self-attention and

depthwise convolution. The RCA and SA modules helped the model clean up its mistakes and focus better on the right parts of the image. The model achieved 94.6% mAP, outperforming standard detectors. However, it is not yet optimized for real-time field use.

TABLE 1. COMPARATIVE ANALYSIS OF THE CUTTING-EDGE TECHNIQUES USED IN TLD DETECTION

Source	Proposed System	Evaluation	Issues Addressed	Strategy Used	Novelty	Shortcomings
Pandian et al. [1]	DCNN	Accuracy: 98.99%	Misclassification of grey blight.	Custom DCNN with inverted residuals.	Efficient CNN architecture.	Detects only 3 broad classes.
Wei Wu [2]	CNN (Transfer Learning)	Accuracy: 90.42% (Focus loss)	Huge training time from scratch.	Used Transfer Learning (Inception V3).	Enhanced convergence speed / accuracy.	Limited disease types (3).
Soeb et al. [3]	YOLOv7 (YOLO-T)	mAP: 98.2%, Accuracy: 97.30%	Hard to detect in natural environments.	Improved YOLOv7 model.	Achieved high mAP and accuracy.	Long training time (5+ hours).
Chandraratne et al. [4]	DCNN (RL-PPO)	Accuracy: 97%	Manual CNN design was time-consuming.	RL with PPO for automated CNN design.	Automated CNN design process.	No mobile/edge deployment yet.
Bhuyan et al. [5]	Res4net-CBAM	Accuracy: 98.35%	Diseases reduce crop quality / yield.	Proposed Res4net-CBAM model.	New CNN architecture.	Dataset images from controlled settings.
Ahmed et al. [6]	ViT, GreenViT, PlantXViT, CNN, DenseNet, ResNet50V2, YOLOv7, NSGA-II, MobileNetv2, YOLOv8	Comparison of various models	Models trained on general plant datasets.	Use of Vision Transformers (ViT).	Evaluated over 30 recent studies.	Few use real-world tea images.
Ihsan et al. [7]	ETC, SVC, RF, XGB, DT, CNN	ETC Accuracy: 77.47%	Hard to detect manually.	Used machine learning for classification.	Compared various ML models.	Accuracy below 80%.
Hairah et al. [8]	MobileNetV2, EfficientNetB0, ResNet50, InceptionV3	MobileNetV2 Accuracy: 94.55%	Low performance due to poor feature extraction.	Image preprocessing / augmentation.	Achieved high accuracy with MobileNetV2.	Images captured against a white background.
Rahman et al. [9]	CNN	Accuracy: 96.64%	Manual detection is slow/error-prone.	Developed a CNN model.	Built a custom CNN.	Only 4 classes included.
Sun et al. [10]	CSPDarkNet53 + (IRMB, RCA, SA)	mAP: 94.6%	Difficulty with background noise/occlusion.	IRMB (Inverted Residual Mobile self-attention Block).	New attention-based detection model (TCSA).	Not yet deployed for real-time use.
Ahad et al. [11]	Vision Transformer (ViT)	Accuracy: 99.53%	Limited hyperparameter optimization for ViTs.	Experimented with hyperparameter settings.	Customized ViT architectures.	Tested on a single, self-collected dataset.
Li et al. [12]	MobileNetV3-CA	Accuracy: 95.88%	Hard to detect many disease/insect types.	MobileNetV3-CA, lightweight with attention.	First to use MobileNetV3-CA for tea issues.	Misses early-stage pests.
Raza et al. [13]	EfficientNet-B0	Accuracy: 95%	Inaccurate results.	EfficientNet-B0 chosen for balance.	Achieved 95% test accuracy.	Variability in performance across classes.
Sarker et al. [14]	YOLO-NAS-s, YOLOv8n, YOLOv5nu, SSD-MobileNetV2	YOLOv8n mAP@50: 0.963	Lack of lightweight models for real-time detection.	Fine-tuned lightweight object detection models.	Annotated Tea Sickness Dataset.	Lacked domain expertise for dataset verification.
Wang et al. [15]	YOLOv8	Comparison of various models	Challenges in highland mountain plantations.	Use of YOLOv8 Network.	Increased precision, recall, and F1 score.	Not compatible with edge devices.

Source	Proposed System	Evaluation	Issues Addressed	Strategy Used	Novelty	Shortcomings
Mutasar et al. [16]	LDA, LVQ	Accuracy: 94.10%	Time-consuming / costly manual identification.	LVQ for classification, LDA for feature reduction.	Innovative LVQ-LDA model.	Tested on a limited dataset.
Rimon et al. [17]	Tea-Net (Custom CNN)	Accuracy: 98%	Diseases reduce productivity/quality.	Advanced deep learning/image preprocessing.	Facilitates early disease detection.	A limited number of tea leaf diseases in the dataset.
Rahat et al. [18]	Custom Residual CNN (Zero Padding, Conv2D, Global Average Pooling).	Accuracy: 99%	Manual inspection errors and crop yield losses from tea diseases.	Residual blocks, data augmentation, and Adam optimization.	Custom architecture (vs. pre-trained) with integrated residual learning.	Limited to one tea variety, specific region, and daylight conditions.

Ahad et al. [11] used ViT with hyperparameter optimization, outperforming traditional CNNs. Yet, dataset challenges and high computational needs remained. Li et al. [12] introduced MobileNetV3-CA, a lightweight model with attention. It was trained with 34,000 field images and achieved 95.88% accuracy. It performed well in adult disease stages but struggled with early detection. Raza et al. [13] used EfficientNet-B0 for accuracy and efficiency. Using the Tea Sickness dataset and augmentation, they achieved 94.64% test accuracy and a mean accuracy of 0.95. Performance varied across disease classes. Sarker et al. [14] developed a real-time disease detection system using CNNs for mobile devices. Using the Tea Sickness dataset, the model showed high accuracy and supported field deployment, but performance was affected by lighting and environmental variation. Wang et al. [15] worked with YOLOv8, Swin Transformer, and a Multi-Scale Attention Module. Precision increased by 3.39%, recall by 0.86%, and F1 by 2.20%. Mutasar et al. [16] introduced a hybrid LVQ-LDA approach and achieved higher accuracy, but the model's complexity reduced scalability for large datasets. Rimon et al. [17] built "TeaNet" to catch diseases early, training it on 3,960 high-resolution images. While the model worked well, it only knows a few types of diseases, and it might struggle with lower-quality photos taken on a smartphone. Rahat et al. [18] created a Residual CNN achieving 99% accuracy. Data augmentation and residual learning improved detection. Practical use is limited because it was tested on a single tea variety in daylight. Although many models in the table achieved high accuracy, limitations remain. Addressing these dataset and deployment challenges becomes the main scope of our study and suggests future research directions.

IV. Latest TLD Detection Methodologies

From the selected cutting-edge South Asian research works, Figure 2 shows five mostly adapted models with comparatively high accuracies, such as the ViT [11] with 99.5% accuracy. Tea-Net [17] also achieved a higher score of 98%. MobileNetV3-CA [12] and EfficientNet-B0 [13] have around 95% accuracy, while LDA+LVQ [16] became the least accurate model among these five, with 94.1% accuracy. Although Vision Transformer (Vit) and Tea-Net models' strength was to be able to produce high accuracies, their weakness lies in the limited datasets, which might not perform well in practical scenarios. The strength of MobileNetV3-CA and EfficientNet-B0 models is based on their efficiency and lightweight architecture. Yet, they have weaknesses, such as MobileNetV3-CA struggles to detect early-stage pests and EfficientNet-B0 shows varied performance across different disease classes. Finally, the LVQ-LDA model's innovative approach is a strength, but its primary weakness is being tested on a limited dataset.

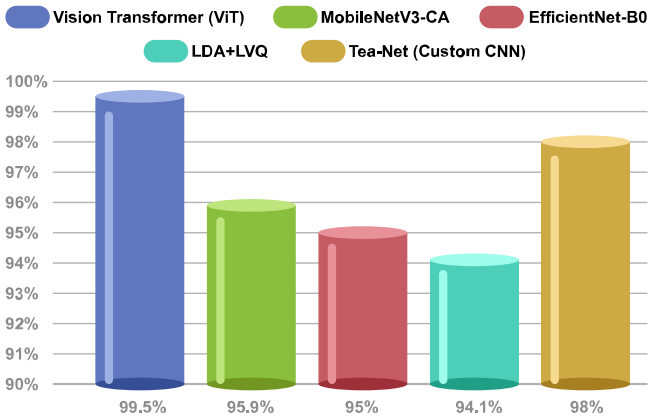


Fig. 2. State-of-the-art TLD Detection Models' Accuracies.

The weaknesses that come along with the strengths suggest that using multimodal data and datasets with low-light leaf images are a few of the factors that could significantly broaden the space of future research.

V. Limitations and Findings of TLD Detection

Almost all of the research that was reviewed has similar limitations. The collective limitations of these papers are systematically categorized into three groups. Figure 3 shows the limitations in TLD detection. One group of researchers, such as Mutasar et al. [16], used very small datasets; sometimes with only 75-200 images, limiting generalization. In some cases, researchers like Rimon et al. [17] used datasets with imbalanced classes, with under-represented classes. Furthermore, the datasets of Bhuyan et al. [5], Pandian et al. [1], and Hairah et al. [8] contained photos that were taken either in lab settings with white backgrounds or during the daytime with bright light, which may have caused the models to underperform in real-world field environments. Another group identified a common issue regarding the lack of real-time or mobile deployment. The groups of Wei Wu [2], Chandraratne et al. [4], Bhuyan et al. [5], and Rahman et al. [9] tested their models on desktops rather than smartphones.

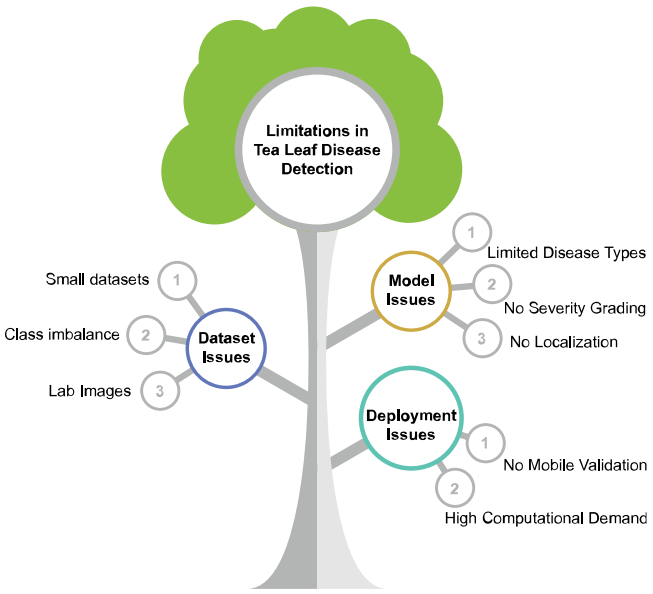


Fig. 3. Limitations in Tea Leaf Disease Detection.

Studies of Soeb et al. [3] and Bhuyan et al. [5] required high-end GPUs due to the specific models they used, which makes it difficult for farmers to run these models on low-end smartphones. According to the results of Pandian et al. [1], Rahman et al. [9], and Sun et al. [10], the accuracy seems acceptable, but the models were rarely tested under varying weather, lighting, or regional conditions, which causes the accuracy to remain below 80%. Despite having access to more advanced or optimized architectures for better results, Rahman et al. [9] continued using basic CNNs. Moreover, some groups of researchers, such as Wei Wu [2], and Li et al. [12], focused on training the models using a small number of disease types, just two or three, and missed early-stage pests or localized disease symptoms. The model of Hairah et al. [8] did not attempt to segment or localize the diseased parts of the leaf, as a result of which the severity or the exact location of the disease on the tea leaf remains unknown. Due to a lack of agricultural or tea plant pathology expertise, along with hardware limitations, the labeling, verification, and deeper analysis of the model's behavior and weaknesses of Sarker et al. [14] were inadequate, leading to a noticeable fall in accuracy. Finally, tea leaves from different regions of different countries may not be the same, as a result of which generalization and broader applicability remain major challenges. The datasets of Pandian et al. [1], Chandraratne et al. [4], and Rimon et al. [17] were collected from a single region, similar to Li et al. [12], which was collected from only one region of China. This means the models may not work well in other tea-growing regions due to differences in leaf appearance, climate, or disease types. Also, due to a lack of validation of the models in actual tea gardens or with data collected by farmers using smartphones, such as in Chandraratne et al. [4], Ahmed et al. [6], Sun et al. [10], and Sarker et al. [14], there remains a gap between research and real-world usage of these models.

VI. Directions for Future Research

The development of more practical and accessible solutions for tea leaf disease detection should be the primary focus of future research. Figure 4 provides directions for the next phase of research, which are briefly discussed in this section. For the farmers in rural areas with limited internet access and basic smartphones to be able to detect these diseases, researchers should build low-cost ML models that can run offline. The purpose is to make these detection systems intuitive and easy to use. In addition to detecting diseases, models should also be improved to grade the severity of the disease. This would help farmers decide whether the damage to the tea plant is recoverable or not. To further improve the accuracy of early detection, the usage of multimodal data is another important point for future work, which means combining tea leaf images with other information like weather, soil data, or plant genetics. Finally, to make models more reliable in practical situations, researchers should collect larger datasets that contain images taken under different lighting conditions, such as during the evening, rather than just on white backgrounds with proper lighting.

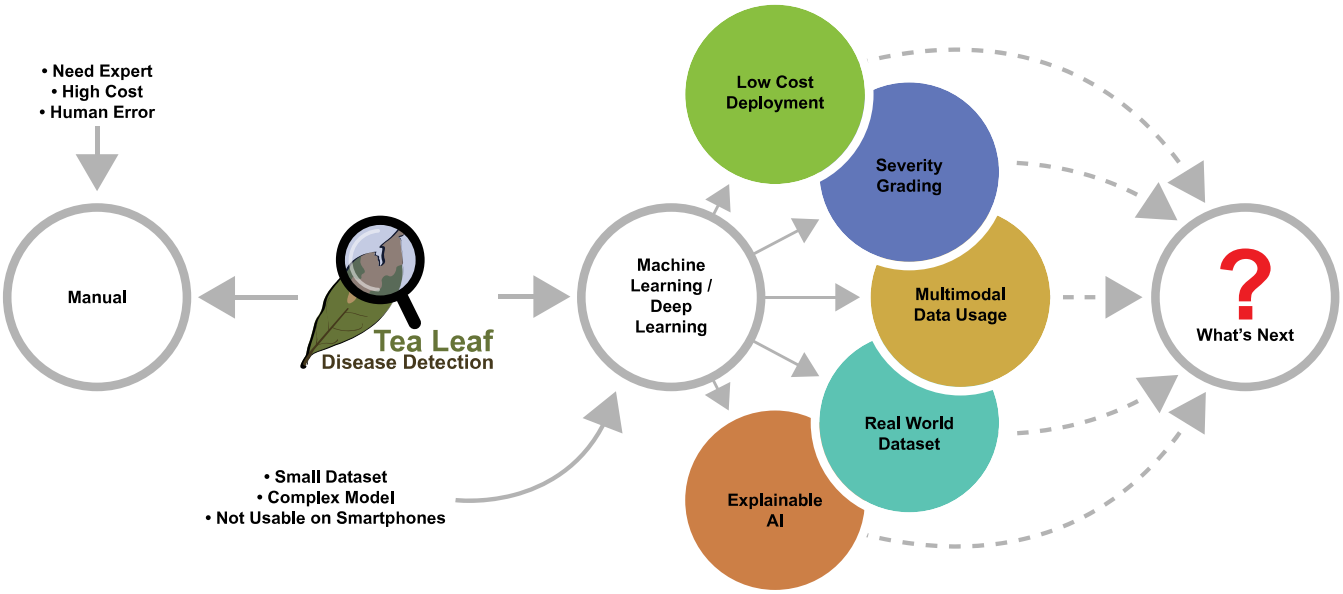


Fig. 4. Future Research Directions for Tea Leaf Disease Detection.

VII. Conclusion

According to the reviewed papers, both machine learning and deep learning techniques, such as ANFIS, CNNs, ViT, Resnet, and LVQ, were used to detect tea leaf diseases. However, datasets were often limited, and some groups of researchers created and used their own data. Although performances were inconsistent, many models showed high accuracy and real-time capability. Almost all research had a common challenge of small or unbalanced datasets, often captured in lab settings or under bright daylight conditions. From what was observed, most models performed well on their specific datasets but might struggle to provide high accuracy on data from different regions of different countries. Future research should focus on collecting larger datasets, testing models in field conditions, and improving generalization while focusing on multimodality.

References

[1] J. Arun Pandian, S. Nirmala Nisha, K. Kanchanadevi, A. K. Pandey, and S. K. Rima, "Grey Blight Disease Detection on Tea Leaves Using Improved Deep Convolutional Neural Network" Computational Intelligence and Neuroscience, vol. 2023, Article ID 7876302, 12 pages, 2023. [Online]. Available: https://doi.org/10.1155/2023/7876302

[2] W. Wu, “Identification of tea leaf diseases based on deep transfer learning,” Frontiers in Computing and Intelligent Systems, vol. 2, no. 3, pp. 75–82, 2022. doi: 10.54097/fcis.v2i3.5218

[3] M. J. A. Soeb, M. F. Jubayer, T. A. Tarin, M. R. A. Mamun, F. M. Ruhad, A. Parven, N. M. Mubarak, S. L. Karri, and M. M. I. Meftaul, “Tea leaf disease detection and identification based on YOLOv7 (YOLO-T),” Scientific Reports, vol. 13, Art. 6078, 2023. doi: 10.1038/s41598-023-33270-4

[4] D. Chandraratne and L. Costa, "LeafCheck: An Image-Based Approach to Tea Leaf Disease Detection Using Deep Learning Techniques," in Proc. 6th Int. Conf. Digital Medicine and Image Processing, Kyoto, Japan, Jul. 2023, pp. 1–6. [Online]. Available: https://www.researchgate.net/publication/372201668_LeafCheck_An_Image-Based_Approach_to_Tea_Leaf_Disease_Detection_Using_Deep_Learning_Techniques

[5] P. Bhuyan, P. K. Singh, and S. K. Das, “Res4net-CBAM: a deep CNN with convolution block attention module for tea leaf disease diagnosis,” Multimedia Tools and Applications, vol. 83, pp. 48925–48947, Nov. 3, 2023, doi: 10.1007/s11042-023-17472-6

[6] F. Ahmed, M. T. Ahad, and Y. R. Emon, “Machine Learning-Based Tea Leaf Disease Detection: A Comprehensive Review,” CoRR, vol. abs/2311.03240, Nov. 2023, doi:10.48550/arXiv.2311.03240

[7] C. N. Ihsan, N. Agustina, M. Naseer, H. Gusdevi, J. F. Rusdi, A. Hadhiwibowo, and F. Abdullah, “Comparison of Machine Learning Algorithms in Detecting Tea Leaf Diseases,” J. RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 8, no. 1, pp. 135–141, Feb. 2024, doi: 10.29207/resti.v8i1.5587

[8] U. Hairah, A. Septiarini, N. Puspitasari, A. Tejawati, H. Hamdani, and S. E. Priyatna, “Classification of tea leaf disease using a convolutional neural network approach,” International Journal of Electrical and Computer Engineering (IJECE), vol. 14, no. 3, pp. 3287–3294, Jun. 2024, doi:10.11591/ijece.v14i3.pp3287-3294

[9] H. Rahman, I. Ahmad, P. H. Jon, A. Salam, and M. F. Rabbi, “Automated detection of selected tea leaf diseases in Bangladesh with a convolutional neural network,” Scientific Reports, vol. 14, Art. 62058, Jun. 2024, doi: 10.1038/s41598-024-62058-3

[10] Y. Sun, M. Jiang, H. Guo, L. Zhang, J. Yao, F. Wu, and G. Wu, "Multiscale Tea Disease Detection with Channel–Spatial Attention," Sustainability, vol. 16, no. 16, p. 6859, Aug. 2024. [Online]. Available: https://doi.org/10.3390/su16166859

[11] M. T. Ahad, F. Ahmed, B. Song, and Y. Li, “Enhancing Tea Leaf Disease Detection Through Customised Vision Transformer and Hyperparameters Optimisation,” SSRN Working Paper 4940688, 2025. [Online]. Available: https://papers.ssrn.com/abstract=4940688

[12] Y. Li, Y. Lu, H. Liu, J. Bai, C. Yang, H. Yuan, X. Li, and Q. Xiao, “Tea leaf disease and insect identification based on improved MobileNetV3,” Front. Plant Sci., vol. 15, p. 1459292, Sep. 27, 2024, doi:10.3389/fpls.2024.1459292

[13] A. Raza and M. Subhan, “Enhancing Tea Plant Health Through Machine Learning: EfficientNet-B0 for Tea Sickness Detection,” International Journal of Computer Applications, vol. 186, no. 45, pp. 32–43, Oct. 2024, doi:10.5120/ijca2024924092

[14] S. S. Sarker, A. Islam, R. T. Raktim, S. A. Roshni, S. K. S. Joy, and F. M. Shah, “Real-Time Tea Leaf Disease Detection Using Deep Learning-Based Models,” in Proc. 27th Int. Conf. Comput. and Inf. Technol. (ICCIT), Cox’s Bazar, Bangladesh, Dec. 2024. doi:10.13140/RG.2.2.10031.24483

[15] Z. Wang, Y. Xia, H. Wang, X. Liu, R. Che, X. Guo, H. Li, S. Zhang, and B. Wang, “Fresh Tea Leaf-Grading Detection: An Improved YOLOv8 Neural Network Model Utilizing Deep Learning,” Horticulturae, vol. 10, no. 12, p. 1347, Dec. 2024, doi:10.3390/horticulturae10121347

[16] M. Mutasar and C. Niesa, “Combination of Learning Vector Quantization and Linear Discriminant Analysis for Tea Leaf Disease Classification,” ResearchGate, Feb. 2025. [Online]. Available: https://www.researchgate.net/publication/388940316_COMBINATION_OF_LEARNING_VECTOR_QUANTIZATION_AND_LINEAR_DISCRIMINANT_ANALYSIS_FOR_TEA_LEAF_DISEASE_CLASSIFICATION

[17] A. B. Rimon, S. Bormon, M. H. Ahmad, and S. R. Sohag, “High-resolution dataset for tea garden disease management: Precision agriculture insights,” Data in Brief, vol. 59, Apr. 2025. doi:10.1016/j.dib.2025.11137

[18] I. S. Rahat, H. Ghosh, S. Dara, and S. Kant, “Towards precision agriculture tea leaf disease detection using CNNs and image processing,” Scientific Reports, vol. 15, no. 17571, 18 pages, 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-02378-0

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

The Role of Agile Methodology in Enhancing Product Management Efficiency: A Review

K. M. Arafat Islam
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
islam15-5498@diu.edu.bd

Abid Khan
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
khan15-4894@diu.edu.bd

Md. Didarul Islam
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
didarul15-4679@diu.edu.bd

Shovan Samanta Turzo
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
turzo15-5785@diu.edu.bd

Hossain Mohammad Shuvo
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
shuvo15-5428@diu.edu.bd

Md. Sazzadur Ahamed
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
sazzad.cse@diu.edu.bd

***Abstract*—This systematic review synthesizes evidence from 35 eligible peer-reviewed studies (2023–2025), citing 25 studies due to space constraints and examining 10 high-rigour cases in depth, to assess how Agile methodologies influence product-management efficiency across software, fintech, healthcare, and other regulated domains. Evidence shows Agile is often associated with reduced time-to-market (reported ranges of 25–60%), higher team productivity (≈60%), and improved customer-satisfaction measures (≈30–35%) through iterative delivery, continuous feedback, and cross-functional collaboration. Efficiency gains, however, are context-dependent: large enterprises and tightly regulated sectors must reconcile Agile cadence with compliance gates, while unresolved role ambiguity between Product Manager and Product Owner can erode benefits. We integrate these insights in a conceptual framework linking core Agile practices to outcomes, moderated by culture, scaling mechanisms, and emerging AI/automation enablers. In this review, product-management efficiency is operationalized using flow and outcome indicators (lead/cycle time, release cadence, throughput/predictability, defect trends, and customer outcomes), and reported improvement ranges are mapped to the underlying studies.**

Keywords—Agile methodology, product management efficiency, Scrum, Kanban, time-to-market, stakeholder feedback, scaling frameworks, AI-enabled automation

I. Introduction

In contemporary digital markets, products evolve amid rapid technological change, uncertain customer needs, and intense competition. Traditional plan-driven approaches (e.g., waterfall) struggle under these conditions due to heavy upfront specification, long feedback cycles, and limited adaptability. Consequently, organizations increasingly adopt Agile methodologies, which emphasize short iterations, continuous stakeholder collaboration, and adaptive planning to support faster product delivery [7], [22].

Among Agile approaches, Scrum and Kanban are the most widely used. Scrum organizes work into time-boxed iterations supported by defined roles and ceremonies, while Kanban focuses on workflow visualization, work-in-progress limits, and flow optimization in continuous delivery settings [3], [7]. Prior studies link these practices to improved coordination, reduced waste, and faster feedback, particularly in software-intensive environments [1], [5]. As a result, Agile is often associated with gains in delivery speed, quality, and stakeholder alignment.

From a product management perspective, efficiency is a key concern. Product managers translate customer and business needs into actionable priorities while balancing speed, quality, and long-term value. Agile practices can support this role through continuous backlog refinement, early assumption validation, and faster decision-making via frequent inspection-and-adapt cycles [3], [9]. However, efficiency gains vary across contexts, and Agile adoption alone does not ensure improved outcomes. Organizational culture, role clarity, governance structures, and regulatory constraints strongly shape results [8], [13], [18].

Despite growing evidence of Agile's benefits, several gaps remain. Product management efficiency is often discussed implicitly, with studies emphasizing isolated metrics rather than a consistent operational definition. Headline improvement claims are rarely traced systematically to underlying studies, limiting comparability. Moreover, role ambiguity between Product Manager and Product Owner, scaling friction, and compliance constraints in regulated domains remain under-examined from a product-management-centric perspective [13], [15]. Emerging research on AI-supported Agile practices adds further complexity by introducing both efficiency opportunities and new socio-technical challenges [4], [11].

To address these gaps, this paper presents a systematic review of peer-reviewed literature published between 2023 and 2025, examining how Agile methodologies influence product management efficiency across industries and organizational contexts. Following a structured selection process, 35 eligible studies were identified; 25 are cited directly, with 10 high-rigour studies forming a comparative core. The review focuses on efficiency outcomes related to delivery flow, productivity, quality, and customer alignment, while considering moderators such as organizational scale, governance models, and AI/automation enablers.

The contributions of this paper are fourfold:

(1) proposing a clear operational definition of product management efficiency;
(2) mapping reported efficiency improvements to specific metrics and study contexts;
(3) identifying recurring benefits and limitations of Agile practices from a product-management-centric view; and
(4) presenting a conceptual framework linking Agile practices to efficiency outcomes and key moderators.

By grounding efficiency claims in systematically synthesized evidence, this review supports researchers and practitioners seeking to design, evaluate, and refine Agile product management practices in contemporary and emerging contexts.

II. Background

Agile methodology represents a shift from traditional plan-driven development toward iterative, adaptive, and feedback-oriented practices. Rather than relying on detailed upfront planning and sequential execution, Agile emphasizes short development cycles, continuous stakeholder collaboration, and responsiveness to change. Among Agile approaches, Scrum and Kanban are the most widely adopted. Scrum structures development into fixed-length iterations (typically one to four weeks) supported by defined roles, events, and artifacts, while Kanban emphasizes workflow visualization, work-in-progress limits, and flow optimization in continuous delivery environments [3], [7]. Empirical studies consistently associate these practices with improved coordination, reduced waste, and faster feedback cycles [1], [5].

Product management plays a central role in guiding products from concept to market by aligning customer needs, business objectives, and development activities. Product managers are responsible for prioritization, roadmap planning, stakeholder communication, and value realization across the product lifecycle [24]. In Agile settings, these responsibilities are exercised through continuous backlog refinement, iterative planning, and frequent assumption validation, enabling faster learning and adaptation than traditional approaches [9], [15].

Definition (Product Management Efficiency).
In this review, product management efficiency is defined as a multi-dimensional construct describing how effectively a product organization converts ideas into delivered customer and business value with minimal waste. Efficiency is assessed across four dimensions:
(i) delivery flow (lead time, cycle time, release frequency);
(ii) execution productivity (throughput and delivery predictability where reported);
(iii) quality outcomes (defect trends and rework rates); and
(iv) customer outcomes (customer satisfaction, adoption, or Net Promoter Score when available).
This definition emphasizes outcome-oriented value delivery rather than adherence to specific Agile ceremonies or processes.

Agile practices directly influence these efficiency dimensions. Iterative development and frequent releases shorten feedback loops, enabling earlier validation and reducing late-stage rework [3], [5]. Continuous integration and testing improve quality and delivery predictability, while regular stakeholder reviews strengthen alignment with customer expectations [9], [14]. Together, these mechanisms support faster decision-making and more effective prioritization in product management.

However, the effectiveness of Agile product management varies across organizational contexts. In many organizations—particularly at scale—the roles of Product Manager and Product Owner are separated, with strategic and tactical responsibilities divided. While this can improve focus, unclear role boundaries often lead to decision delays, prioritization conflicts, and reduced efficiency [13]. These challenges are amplified in large enterprises and regulated domains, where governance requirements, compliance gates, and legacy systems constrain iteration speed and autonomy [8], [18].

Consequently, although Scrum and Kanban provide a strong foundation for adaptive and customer-centered product development, efficiency gains are not automatic. Outcomes depend on contextual factors such as organizational culture, leadership support, role clarity, and the ability to tailor Agile practices to domain-specific constraints. Understanding these conditions is essential for interpreting reported efficiency outcomes and designing Agile product management practices that deliver sustainable value.

III. Research Methodology

This section describes the systematic approach used to identify, select, and analyze existing literature examining the impact of Agile methodologies on product management efficiency. The methodology is designed to ensure transparency, rigor, and reproducibility in accordance with established systematic review practices.

A. Search Strategy

A structured literature search was conducted covering publications from January 2023 to mid 2025 to capture recent evidence on Agile methodology and product management efficiency. Multiple academic databases were consulted, including IEEE Xplore, Scopus, and Google Scholar, which index peer-reviewed journals and reputable conference proceedings. Search queries combined key terms using Boolean operators, such as “Agile methodology,” “product management,” “product management efficiency,” “Scrum,” and “Kanban.” Search strings were refined iteratively to balance coverage and relevance.

B. Screening and Selection Process

The study selection followed a PRISMA-style screening approach. Retrieved records were pooled and de-duplicated, after which titles and abstracts were screened to remove clearly irrelevant studies. The remaining papers underwent full-text review against predefined inclusion and exclusion criteria.

Inclusion criteria ensured that each selected study:

- Be peer-reviewed journal articles or conference papers;
- Be published in English between 2023 and 2025;

- Examine Agile methodologies (e.g., Scrum, Kanban, XP, or hybrid models);
- Address product management, project management closely tied to product decision-making, or comparable contexts;
- Report qualitative or quantitative outcomes related to efficiency, productivity, delivery performance, or customer outcomes.

Exclusion criteria ruled out works that:

- Were non-peer-reviewed (e.g., blogs, opinion pieces, white papers);
- Focused on Agile in unrelated domains without a product management perspective;
- Did not discuss efficiency or performance outcomes;
- Fell outside the defined time window or language scope.

Following this process, 35 studies met all eligibility criteria and were included in the qualitative synthesis. Due to space constraints, 25 studies are cited directly, while 10 high-rigour studies—selected for methodological robustness, metric clarity, and contextual diversity—form the comparative core used for structured analysis. Figure 1 summarizes the screening process using a PRISMA flow diagram.

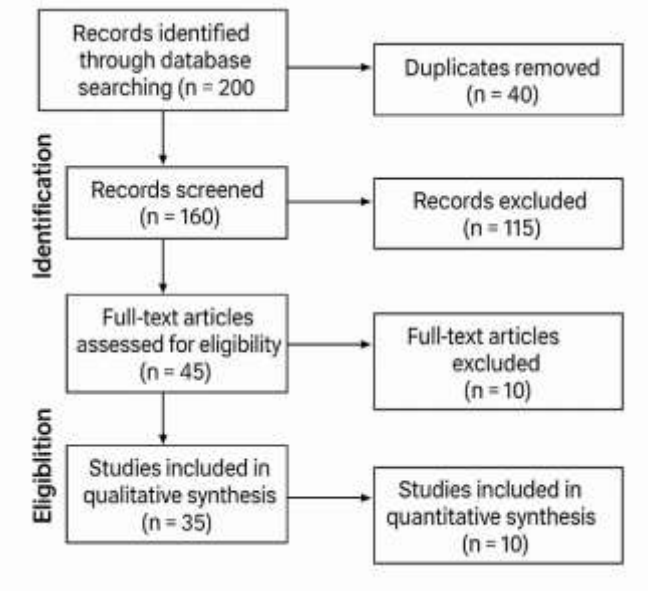


Fig. 1. PRISMA flow diagram summarizing the study selection process.

C. Data Extraction and Thematic Analysis

A structured data extraction form was applied consistently across all included studies. Extracted data included: industry/domain, organizational scale, Agile approach (Scrum, Kanban, or hybrid), study design (e.g., case study, survey, experiment, review), reported efficiency indicators, and contextual moderators such as regulatory constraints or scaling mechanisms.

The extracted data were analyzed using thematic analysis. Studies were coded to identify recurring patterns in efficiency outcomes, including delivery speed, productivity, quality, and customer alignment, while divergent findings were recorded to avoid overgeneralization. From the full set of 35 studies, 10 core papers were analyzed comparatively to examine how specific Agile practices influenced product management efficiency across contexts. This comparative analysis underpins the results and the summary metrics table presented in the next section. The data extraction framework is summarized in Table I.

Table I. Data extraction framework

Category	What was extracted
Study context	Industry/domain, organization size/team scale, regulated vs non-regulated setting
Agile approach	Scrum, Kanban, XP, hybrid; any scaling mechanism (e.g., SAFe/LeSS/custom)
Study design	Case study, survey, experiment, dataset analysis, review/multivocal review
Efficiency indicators	Lead time/cycle time, release frequency, productivity/throughput, predictability, defects/rework, customer outcomes
Moderators & enablers	Culture/leadership, role clarity (PM vs PO), governance/compliance gates, tooling/automation, AI/analytics use

D. Limitations

This review has several limitations. Limiting the scope to publications from 2023–2025 excludes earlier foundational Agile studies and may overlook long-term trends. Restricting inclusion to peer-reviewed English-language literature may introduce publication and language bias, potentially excluding relevant industry reports or non-English research. In addition, although thematic coding followed a structured approach, it inevitably involves researcher judgment, which may influence categorization and interpretation.

E. Threats to Validity

This review is subject to several validity considerations. Construct validity is affected by inconsistent definitions of product management efficiency across studies; this risk is mitigated by applying a common operational definition and mapping outcomes to standardized metrics. Internal validity may be influenced by study selection and thematic coding decisions; to reduce bias, explicit inclusion and exclusion criteria and structured data extraction were used. External validity is limited by differences across industries, organizational scales, and regulatory environments, which may constrain generalizability. Publication bias is also possible, as studies reporting positive Agile outcomes are more likely to be published; therefore, results are interpreted as indicative trends rather than universal guarantees. Although all included sources are peer-reviewed, venue visibility varies, and lower-visibility outlets are treated as contextual support, while quantitative summaries rely on more established venues where possible.

F. Threats to Rigor

This review faces two primary rigor risks. Sampling bias may arise from restricting the search to 2023–2025 publications, English-language sources, and selected databases; this risk was mitigated by using multiple databases and iteratively refining search strings and screening criteria.

Subjectivity in thematic coding may influence categorization; to reduce interpretive drift, a structured extraction template and consistent theme definitions aligned with the operationalization of product management efficiency were applied. Ambiguous cases were coded conservatively and reported as contextual rather than definitive evidence.

IV. Results and Analysis

This review synthesizes recent peer-reviewed evidence (2023–2025) on how Agile practices influence product management efficiency. Across the reviewed studies, three themes consistently emerge: (A) efficiency benefits of Agile practices, (B) challenges and limitations that constrain those benefits, and (C) advanced contexts such as large-scale coordination, AI augmentation, and modernized metrics. Table ii summarizes the headline efficiency improvements reported in selected studies and links each claim to its metric and study context.

Table II. Summary of reported efficiency improvements in selected studies

Ref.	Context (example)	Metric	Reported improvement	Study type / note
[1], [5]	Software/ product teams (multi-study)	Time-to-market / release speed	25–60% faster vs. plan-driven baselines	Review + survey evidence; range varies by context
[5]	Software delivery teams (survey)	Productivity / throughput	~60% improvement (reported)	Associated with CI/testing + collaboration factors
[5]	Software delivery teams (survey)	Defect reduction / quality	Up to ~45% fewer defects (reported)	Linked to CI/testing practices
[9], [14]	Digital products / banking contexts	Customer satisfaction	30–35% higher satisfaction scores (reported)	Iterative stakeholder feedback reduces rework
[7]	Mixed org sizes (survey + interviews)	Lead time / cycle time	~30% lead-time reduction (reported)	Flow improvements; scaling challenges remain
[4]	Agile planning tasks (experiment)	Backlog planning effort	~25% less planning time (reported)	Generative AI; trust/oversight needed
[11]	Agile project/product setting	Prioritization + early risk detection	Qualitative improvement;	AI-supported decision-making;
	(case study)		earlier risk detection	trust barrier noted
[8], [18]	Regulated / compliance-heavy domains	Governance fit (hybrid models)	Qualitative: faster delivery while meeting gates	Hybrid Agile-stage-gate patterns in regulated settings

A. Benefits of Agile for Product Management Efficiency

As summarized in Table II, Agile practices are most consistently associated with improved delivery flow and responsiveness. Iterative delivery, continuous feedback, and frequent backlog refinement enable earlier validation and scope adjustment, which multiple studies report as reduced time-to-market and shorter lead/cycle times in digital product settings [1], [5], [7].

Several studies also link cross-functional collaboration and transparent work management to higher throughput/productivity and improved quality outcomes (e.g., fewer defects and less rework), especially when supported by automation and continuous integration practices [5]. Customer outcome measures such as satisfaction also tend to improve when stakeholder feedback is systematically integrated into roadmap and release decisions [9], [14].

B. Challenges and limitations in Agile Product Management

Despite these benefits, Agile adoption does not automatically translate into efficiency gains. Studies note that weak governance integration, inconsistent stakeholder engagement, and "ceremony compliance" without empowered decision-making can introduce coordination overhead and fragment roadmap coherence [8], [13], [17].

Role ambiguity—particularly between Product Manager and Product Owner—can delay prioritization, blur accountability for outcomes, and make it difficult to demonstrate measurable efficiency gains. When ownership is unclear, teams may optimize short-term sprint outputs while losing strategic coherence, reducing overall product value delivery [13], [15].

C. Agile in Advanced Contexts: Scaling, AI, and Metrics

In enterprise-scale portfolios, scaling mechanisms and cross-team coordination practices are often required to sustain alignment while preserving iterative flow. Evidence emphasizes lightweight governance that synchronizes dependencies and integrates compliance checkpoints into the cadence—particularly in regulated settings—resulting in hybrid Agile-stage-gate patterns in practice [13], [8], [18].

AI and analytics are emerging as efficiency enablers, especially for backlog refinement, estimation support, and early risk detection. Studies discussing generative AI or AI-assisted decision support report reduced planning effort and improved prioritization signals, but also emphasize the need for human oversight, transparency, and trust management [4], [11], [20].

Finally, measurement practices are shifting from activity tracking toward flow and outcome metrics (lead/cycle time, throughput, predictability, defect trends, and customer outcomes). Reviews on large-scale Agile metrics suggest that consistent outcome-oriented measurement supports transparency and continuous improvement at both team and portfolio levels [12], [5], [22].

V. Discussion and Implications

This section interprets the synthesized findings and connects them to practical product-management decisions. Rather than treating Agile as a single intervention, the discussion emphasizes the mechanisms through which Agile practices affect product-management efficiency, and the organizational conditions that strengthen or weaken those effects.

A. Synthesis of Key Findings

a) Holistic efficiency improvements: Across the reviewed literature, Agile practices are most consistently associated with improvements in delivery flow (shorter lead/cycle time and faster releases), execution productivity (higher throughput and predictability), quality (reduced defects and rework), and customer outcomes (stronger satisfaction and alignment). Table II links these improvement ranges to their underlying metrics and study contexts.

b) Contextual Dependencies: The scale and durability of efficiency gains are strongly context-dependent. Digitally native teams with mature engineering enablement (e.g., CI and automated testing) typically report stronger benefits, while organizations facing regulatory constraints, complex dependencies, or legacy governance often require hybrid approaches that balance Agile cadence with formal controls.

c) Interplay of Benefits and Challenges: Agile adoption alone does not ensure efficiency gains. Cultural resistance, limited Agile fluency, and superficial ceremony adoption can increase coordination overhead without improving outcomes. In particular, unclear decision rights between Product Manager and Product Owner can slow prioritization, fragment roadmaps, and reduce stakeholder confidence.

d) Measurement and transparency: A recurring finding is the need to shift from activity-based metrics (e.g., velocity) toward flow- and outcome-oriented measures. Teams that consistently visualize delivery flow and customer outcomes through dashboards and retrospectives are better able to identify bottlenecks, reduce waste, and sustain efficiency improvements.

e) Integrated Conceptual Understanding: As illustrated in Fig. 2, product-management efficiency emerges from the interaction of core Agile practices—iteration, feedback loops, cross-functional collaboration, and continuous improvement—rather than from any single practice. This interaction is reinforced by empowered teams and rapid decision-making.

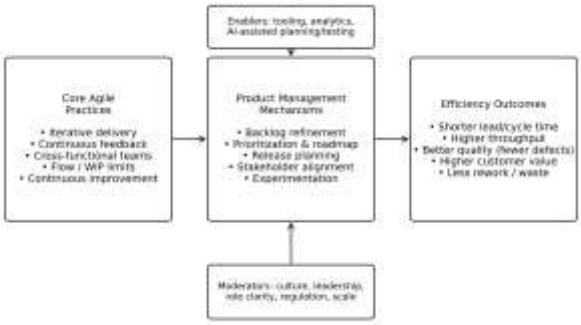


Fig. 2. Conceptual framework linking Agile practices to product-management efficiency outcomes.

B. Practical Implications for Product Managers and Organizations

Based on the synthesis, the following practitioner-oriented guidelines help maximize efficiency benefits while reducing standard failure modes:

- Use regular customer feedback loops (reviews/demos, discovery interviews, telemetry) to validate assumptions early and reduce rework.
- Treat the backlog as dynamic: continuously refine, re-prioritize, and de-scope to protect flow and focus on highest-value outcomes.
- Define decision rights and role boundaries explicitly (Product Manager vs. Product Owner vs. Scrum Master) to avoid prioritization delays and stakeholder confusion.
- Adopt outcome- and flow-oriented measurement: track lead/cycle time, throughput, work-in-progress, escaped defects/rework, and customer outcomes (e.g., satisfaction or adoption). Use retrospectives to act on these signals.
- Manage Agile as a change program: secure leadership sponsorship, invest in training/mentorship, and scale gradually to avoid cargo-cult adoption.
- Scale Agile judiciously: use lightweight coordination for dependencies; in regulated settings, integrate compliance checkpoints into the workflow (hybrid stage-gate patterns) rather than pushing all governance outside the cadence.
- Leverage automation and tooling (CI/CD, test automation, analytics dashboards). Introduce AI assistance cautiously for tasks like backlog grooming, summarization, and test generation, with clear oversight to address trust and accountability.

A key implication is that efficiency is a system-level outcome, achieved when product strategy, governance, and team execution are aligned and measurement emphasizes customer and business value over process compliance.

C. Recommendations for Future Research

The reviewed literature suggests several directions where additional empirical work would improve confidence and generalizability:

- Longitudinal studies that measure whether early efficiency gains persist, plateau, or regress over time.
- More standardized operationalization of product-management efficiency to enable benchmarking across organizations and industries.
- Comparative studies of role design (separated vs. combined Product Manager/Product Owner) and its impact on decision latency and outcomes in scaled environments.
- Evidence on hybrid governance models in regulated domains, including how compliance checkpoints can be integrated without undermining iteration speed.
- Deeper evaluation of AI-supported Agile workflows (planning, prioritization, test generation, analytics), including human factors such as trust, autonomy, and accountability.
- Research outside software-only settings (healthcare, government, hardware/embedded, and other risk-sensitive domains) to identify which Agile mechanisms transfer and which require adaptation.

VI. Conclusion

This paper presented a systematic review of recent literature (2023–2025) examining the role of Agile methodology in enhancing product management efficiency. The findings indicate that Agile practices are commonly associated with faster delivery cycles, improved productivity, and stronger alignment with customer needs when applied appropriately. However, efficiency gains are not automatic and depend on contextual factors such as organizational culture, role clarity between Product Manager and Product Owner, governance constraints, and team maturity. In large-scale or regulated environments, hybrid and adaptive Agile approaches are often necessary to balance flexibility with coordination and compliance requirements. Overall, Agile should be viewed as a conditional enabler rather than a universal solution for product management efficiency. Clear outcome-oriented metrics, effective role design, and continuous learning are essential for realizing sustainable efficiency gains.

VII. References

[1] P. Amajuoyi, L. B. Benjamin, and K. B. Adeusi, "Agile methodologies: Adapting product management to rapidly changing market conditions," GSC Advanced Research and Reviews, vol. 19, no. 2, pp. 249–267, May 2024.

[2] P. Amajuoyi, L. B. Benjamin, and K. B. Adeusi, "Optimizing agile project management methodologies in high-tech software development," GSC Advanced Research and Reviews, vol. 19, no. 2, pp. 268–274, May 2024.

[3] E. M. Alzeyani and C. Szabó, "A study on the effectiveness of agile methodology using a dataset," Acta Electrotechnica et Informatica, vol. 23, no. 1, pp. 3–10, Mar. 2023.

[4] A. Bahi, J. Gharib, and Y. Gahi, "Integrating generative AI for advancing agile software development and mitigating project management challenges," International Journal of Advanced Computer Science and Applications, vol. 15, no. 3, 2024.

[5] B. Binboga and C. A. Gumussoy, "Factors affecting agile software project success," IEEE Access, vol. 12, pp. 95613–95633, 2024.

[6] [6] M. Celestin, S. Sujatha, A. D. Kumar, and M. Vasuki, "The rise of agile methodologies in managing complex business projects: Enhancing efficiency, collaboration, and adaptability," Indo American Journal of Multidisciplinary Research and Review, vol. 8, no. 2, pp. 69–77, 2024, doi: 10.5281/zenodo.13871832.

[7] S. Chahal, "Agile methodologies for improved product management," Journal of Business and Strategic Management, vol. 8, no. 4, pp. 79–94, Sep. 2023.

[8] J. Chukwunweike and O. E. Aro, "Implementing agile management practices in the era of digital transformation," World Journal of Advanced Research and Reviews, vol. 24, no. 1, pp. 2223–2242, Oct. 2024.

[9] H. H. Olsson and J. Bosch, "Strategic digital product management: Nine approaches," Information and Software Technology, vol. 177, p. 107594, Jan. 2025.

[10] R. Kadapal and A. More, "Data-driven product management: Harnessing AI and analytics to enhance business agility," Sarcouncil Journal of Public Administration and Management, vol. 3, no. 6, pp. 1–10, Nov. 2024.

[11] H. L. Lumbanraja, T. Raharjo, and A. N. Fitriani, "Artificial intelligence implementation in agile project management: Addressing challenges and maximizing impact," The Indonesian Journal of Computer Science, vol. 13, no. 4, Jul. 2024.

[12] R. Menezes, M. Marinho, and S. Sampaio, "Metrics in large-scale agile software development: A multivocal literature review," in Proc. XXVII CIbSE, 2024, pp. 106–120.

[13] N. B. Moe, M. Berntzen, A. Barbala, and V. Stray, "Software product management in large-scale agile," in Lecture Notes in Business Information Processing. Springer, 2024, pp. 53–69.

[14] D. O. Ogundipe, O. A. Odejide, and T. E. Edunjobi, "Agile methodologies in digital banking: Theoretical underpinnings and implications for customer satisfaction," Open Access Research Journal of Science and Technology, vol. 10, no. 2, pp. 21–30, Mar. 2024.

[15] O. Onesi-Ozigagun, Y. J. Ololade, N. L. Eyo-Udo, and D. O. Ogundipe, "Agile product management as a catalyst for technological innovation," International Journal of Scientific Research Updates, vol. 7, no. 2, pp. 9–23, Apr. 2024.

[16] F. Prica and N. Bjelic, "Enhancing project efficiency: The role of agile project management in a dynamic corporate landscape," Journal of Policy Options, 2024.

[17] "The effectiveness of implementing agile project management: A systematic literature review," Global Business Finance Review, vol. 29, no. 6, pp. 170–186, Jul. 2024.

[18] V. L. Sunkara, "Integrating product management strategies into risk management frameworks: Enhancing banking resilience in the era of fintech and regulatory evolution," SSRN Electronic Journal, 2025.

[19] O. Famoti et al., "A practical model for agile project management to streamline engineering delivery in energy projects," International Journal of Multidisciplinary Research and Growth Evaluation, vol. 6, no. 1, pp. 1831–1840, 2025.

[20] K. Zadeh, E. Bagheri Khoulenjani, and M. Safaei, "Integrating AI for agile project management: Innovations, challenges, and benefits," International Journal of Industrial Engineering and Construction Management (IJIECM), vol. 1, no. 1, 2024.

[21] G. Salazar-Salazar, M. Mora, H. Duran-Limon, F. Alvarez-Rodriguez, and A. Munoz-Zavala, "Review of agile SDLC for big data analytics systems in small organizations using Scrum–XP," The International Arab Journal of Information Technology, vol. 21, no. 6, 2024.

[22] S. S. Maidin, "Current and future trends for sustainable software development: Software security in agile and hybrid agile through bibliometric analysis," Journal of Applied Data Sciences, vol. 6, no. 1, pp. 311–324, Jan. 2024.

[23] H. M. Örnek and A. Camci, "Industry 4.0 and agile project management: Evolution of leadership concepts and roles," Yönetim Bilimleri Dergisi, vol. 22, no. 54, pp. 2601–2635, Oct. 2024.

[24] S. Mahadik, K. Murthy, S. Reddy, A. Jain, and O. M. Goel, "Agile product management in software development," International Journal for Research Publication and Seminars, vol. 13, no. 5, pp. 453–467, Nov. 2022.

[25] R. Schültken, S. Kokozinski, and C. Bode, "Beyond buying bytes: Complementing agile new product development with an agile procurement process," Industrial Marketing Management, vol. 126, pp. 210–220, Apr. 2025.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Skeleton-Based Medical Action Recognition Using Spatio-Temporal Graph Neural Networks

Subrata Kumer Paul[1,2*], Mahfuzur Rahman[1], Md. Mustafizur Rahman[1], Hamim Al Mohaymin[1], Rakhi Rani Paul[1,2], Dewan Nafiul Islam Noor[1], Md. Ekramul Hamid[2], Mirza A. F. M. Rashidul Hasan[3]

[1]*Department of Computer Science and Engineering, Bangladesh Army University of Engineering & Technology (BAUET), Qadirabad Cantonment, Natore-6431, Rajshahi, Bangladesh*

[2]*Department of Computer Science and Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh*

[3]*Department of Information and Communication Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh*

Emails: sksubrata96@gmail.com, mrmmahfuz34@gmail.com, mustafizurrahman40496@gmail.com, hamimalif79@gmail.com, rakhipaul.cse@gmail.com, dewannoorcse13@gmail.com, ekram_hamid@yahoo.com, mirzahasanice@gmail.com

***Abstract*— This paper presents a skeleton-based medical action recognition system developed using a subset of twelve clinically meaningful classes from the NTU RGB+D 120 dataset. The system employs two graph-based deep learning models, Spatio-Temporal Graph Convolutional Networks (ST-GCN) and Spatio-Temporal Graph Isomorphism Networks (ST-GIN), to analyze 3D skeleton sequences consisting of 25 joints over 100 frames. This representation captures subtle spatial and temporal joint movements associated with actions such as coughing, falling, headache, chest pain and other medical-related behaviors. Across the experiments, ST-GCN delivers the strongest performance with an accuracy of 76.19%, while ST-GIN reaches 61 percent. ST-GCN benefits from its structured modeling of joint dependencies, allowing it to generalize better under limited data conditions. Performance evaluation includes accuracy curves, ROC curves and confusion matrices, which help illustrate class-level strengths, separability and common misclassification patterns. The findings indicate that skeleton-based methods are well suited for medical action recognition and can provide reliable results even with relatively small datasets. The system lays a solid foundation for future work in healthcare monitoring, early risk detection and intelligent assistive systems, where more extensive annotated data could further enhance performance.**

Keywords*— *Human Action Recognition (HAR), Graph Neural Networks (GNN), Spatio-Temporal Analysis, Deep Learning, NTU-RGB+D 120 dataset

I. INTRODUCTION

Medical action recognition has become increasingly important for intelligent healthcare systems, enabling automated monitoring, patient safety assessment, and early detection of abnormal events. Traditional RGB-based action recognition often struggles in real-world clinical environments due to issues such as varying illumination, background noise, occlusion, and privacy concerns. To address these limitations, skeleton-based action recognition has gained significant attention because it uses only the 3D joint coordinates of the human body, offering a lightweight, robust, and privacy-preserving representation of human motion. Graph-based deep learning has further enhanced the capability of skeleton action recognition by modeling joints and their spatial–temporal relationships as graphs. Among these, Spatial-Temporal Graph Convolutional Networks (ST-GCN) effectively capture dynamic movement patterns, while Graph Isomorphism Networks (GIN) provide strong expressive power for learning discriminative graph features. Although both models have shown promising results, their comparative performance on medically relevant actions remains understudied.

This study investigates GIN and ST-GCN for recognizing twelve important medical actions—sneezing/coughing, staggering, falling down, headache, chest pain, back pain, and neck pain, nausea/vomiting, fan self, yawn, stretch oneself, blow nose —using skeleton data extracted from the NTU RGB+D 120 dataset. After isolating and preprocessing the required action classes, the skeleton sequences were converted into NumPy format for model training and evaluation. The objective of this study is to examine the accuracy, robustness, and suitability of both graph-based models for practical medical monitoring scenarios [10]. Some major Contributions in this work for

- A curated skeleton-based medical action subset was constructed from NTU RGB+D 120 with robust preprocessing and augmentation.
- A comparative evaluation of ST-GCN and ST-GIN shows ST-GCN achieves superior performance (76.19% accuracy) for medical action recognition.
- The study validates the suitability of skeleton-based GNNs for privacy-preserving healthcare monitoring.

The remaining sections of this study are organized as follows: Section II reviews related work; Section III presents the dataset description; Section IV outlines the methodology; Section V reports and discusses the experimental results; Section VI highlights the applications of HAR; and Section VII concludes the study with future research directions.

II. RELATED WORK

Human action recognition has undergone significant progress with the advancement of deep learning and large-scale motion capture datasets. Earlier approaches relied heavily on handcrafted features such as HOG, optical flow, and temporal motion descriptors, using classical machine learning models like SVMs or Hidden Markov Models (HMMs) for classification [2]. However, these methods struggled to capture subtle motion dynamics, especially in complex medical actions where body gestures may be minimal, noisy, or highly similar. Variations in camera angle, subject appearance, and background clutter further limited the performance of feature- engineered systems, making them unsuitable for real-world clinical environments [3]. A major breakthrough came with the introduction of skeleton-based action recognition, where human joints and their coordinates are used instead of raw RGB frames. Skeleton data offers strong robustness to lighting changes, occlusions, and privacy concerns, making it ideal for healthcare applications. Early research in this domain utilized Recurrent

Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to model temporal motion sequences. While LSTM- based models showed promising results, they were limited by their inability to fully capture the spatial relationships among joints and often required large computational resources for long sequences [4]. The emergence of Graph Convolutional Networks (GCNs) significantly improved skeleton-based action recognition. ST- GCN, one of the pioneering models, introduced a spatial- temporal graph structure where human joints are represented as graph nodes and their connections as edges. By applying graph convolutions across both space and time, ST-GCN effectively learned complex motion patterns and achieved strong performance on large datasets such as NTU RGB+D. However, ST-GCN can be sensitive to noise in skeleton extraction and sometimes struggles to differentiate subtle medical gestures where spatial changes are minimal.

Graph Isomorphism Networks (GIN) later demonstrated even more powerful graph-level representation learning, offering high discriminative capability for graph-structured data. GIN's expressive power allows it to learn fine-grained structural variations in skeleton sequences, making it a potential candidate for recognizing short, subtle medical actions. Prior studies have shown that GIN can outperform conventional GCNs in tasks requiring strong structural reasoning. However, its application in medical action recognition is still limited, and comparative evaluations against spatial-temporal models like ST-GCN remain sparse.

Recent work has also explored hybrid and ensemble methods, combining multiple graph-based models to improve robustness, especially when dealing with noisy skeleton data or imbalanced medical action classes. Although these methods demonstrate improved accuracy, they often increase computational complexity, making them difficult to deploy in real-time healthcare settings such as fall detection systems, telemedicine platforms, or smart hospital monitoring.

Despite notable advancements, existing literature lacks a comprehensive comparison between GIN and ST-GCN specifically for medical action recognition. This research aims to bridge this gap by evaluating both models on medically relevant actions—sneezing/coughing, falling down, headache, chest pain, back pain, and neck pain—using skeleton sequences from the NTU RGB+D 120 dataset.

III. Dataset Description

A. NTU RGB+D 120 Dataset

The NTU RGB+D 120 dataset [7] is one of the largest and most diverse benchmark datasets for human action recognition, containing over 114,000 video samples covering 120 distinct action classes**.** Collected using Microsoft Kinect v2 sensors, the dataset provides four synchronized data modalities—RGB frames, depth maps, infrared streams, and 3D skeleton sequences. Among these, skeleton data plays a crucial role in medical action analysis due to its lightweight structure, robustness to lighting variations, and preservation of user privacy.

Each action recording includes 25 human body joints, annotated with precise 3D coordinates (x, y, z). These joints span key anatomical positions such as the head, neck, shoulders, spine, hips, arms, and legs. The dataset features recordings from 106 subjects of varying ages, genders, and body types, captured from multiple camera viewpoints and diverse environmental conditions. This wide variability makes NTU RGB+D 120 exceptionally suitable for building models that generalize well across real-world healthcare monitoring scenarios. For this research, only the medically relevant action classes were extracted in here which is listed in the table I.

TABLE I. Selected medical action classes

A41: sneeze/cough	A42: staggering	A43: Falling down	A44: headache
A45: Chest pain	A46: Back pain	A47: neck pain	A48: nausea/vomiting
A49: fan self	A103: yawn	A104: stretch oneself	A105: blow nose

Although the dataset is comprehensive, its raw form contains challenges such as missing frames, irregular sequence lengths, and occasional noisy joint coordinates. To mitigate these issues, a thorough preprocessing pipeline was applied. Corrupted or incomplete skeleton files were removed, all sequences were standardized, and joint coordinates were normalized. Additionally, each valid skeleton sequence was converted into NumPy (.npy) format to enable efficient loading during GIN and ST-GCN model training.

B. Curated Medical Skeleton Subset

In addition to the full NTU RGB+D 120 dataset, a custom curated skeleton subset was created specifically for this study to ensure clean and consistent data suitable for graph-based learning. This subset includes all valid skeleton sequences belonging to the 12 selected medical action classes, resulting in a refined and balanced dataset for comparative model evaluation [11]. Several preprocessing steps were performed to enhance data quality:

- Resolution Filtering: Only skeleton files with complete 25-joint data and valid timestamps were retained, discarding sequences with missing, unstable, or zero-value joints.
- Sequence Standardization: Because action recordings vary in frame count, each sequence was resampled or padded to achieve a uniform temporal length, ensuring compatibility across GIN and ST-GCN.
- Coordinate Normalization: Joint positions were normalized relative to the spine base to reduce inter-subject variation and improve model stability.
- Class-wise Organization: Separate directories were created for the twelve medical actions (A041–A049, A103-A104), simplifying data loading and label assignment.
- NumPy Conversion: All skeleton sequences were stored in .npy format for faster I/O operations during training.

This curated dataset offers high-quality, noise-filtered skeleton sequences that accurately capture the spatial-temporal structure of medical actions. Its structured organization and

consistent formatting makes it ideal for graph neural network-based modeling. Figure 1 shows sample images from the NTU RGB+D 120 dataset used in this study.

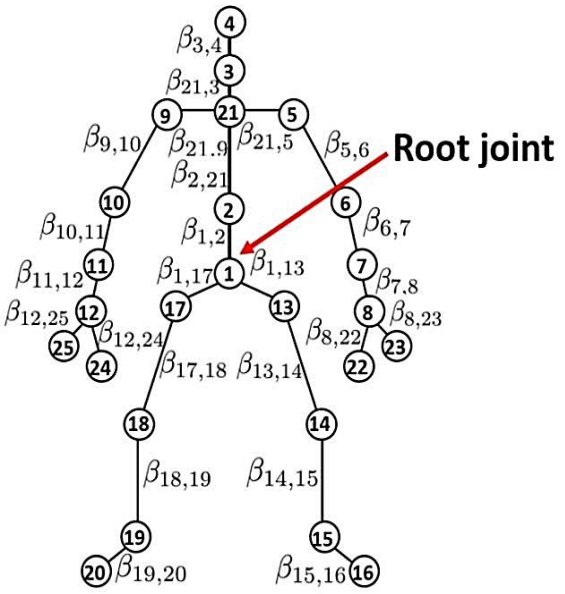


Figure 1: Sample skeleton images from the NTU RGB+D 120 datasets

IV. METHODOLOGY

Our proposed methodology begins with the selection of the NTU RGB+D 120 dataset. We focus on extracting the 12 medical action classes (A041: sneeze/cough, A042: staggering, A043: falling down, A044: headache, A045: chest pain, A046: back pain, A047: neck pain, A048: nausea /vomiting, A049: fan self, A103: yawn, A104: stretch oneself, A105: blow nose). After extraction, the skeleton data is converted into NumPy (.npy) format for efficient processing. We apply preprocessing techniques (normalization, alignment, and sequence padding) to prepare the data. The processed data is then split into training (80%) and testing (20%) sets, ensuring balanced representation across classes. We implement and evaluate several Graph Neural Network (GNN) architectures—GCN, ST-GCN, GIN, and a hybrid ST-GCN + ST-GIN model—on the skeleton data. Cross validation is used to assess model robustness, followed by performance analysis on the test set.

A. Data Preprocessing

To ensure high-quality skeleton sequences for graph-based analysis, several preprocessing steps were applied:

1) Skeleton Extraction

Each sample consists of a sequence of frames with 3D coordinates (x, y, z) (x, y, z) for 25 joints. The raw skeleton data from NTU RGB+D was directly used without RGB or depth information [13].

2) Sequence Normalization

Since videos vary in length, all sequences were normalized to a fixed number of frames.

- Short sequences → padded with zeros
- Long sequences → uniformly downsample

This ensures compatible tensor shapes for batch processing.

3) Coordinate Normalization

Skeleton coordinates were normalized using:

- Root joint normalization: hip center used as the reference origin
- Scale normalization: joint coordinates scaled relative to torso length
- View normalization: ensures invariance to camera angle and subject position

This improves generalization across different subjects and viewpoints [10].

4) Noise Removal

Missing or noisy joint values were corrected by:

- Linear interpolation between valid frames
- Smoothing temporal noise using a simple moving average filter

B. Data Augmentation

To reduce overfitting and improve model robustness [11], the following augmentations were applied:

5) Temporal Augmentation
- Random time warping
- Frame dropping and duplication

This simulates variations in performing speed and duration.

6) Spatial Augmentation
- Joint rotation (±10 degrees) around x, y, z axes
- Scaling (0.9–1.1) to simulate different body sizes
- Gaussian noise added to joint coordinates

7) Sequence Flipping
- Horizontal flipping of skeleton joints to simulate left-right variations [12].

These augmentation techniques help models learn more invariant and generalized representations of actions.

C. ST-GCN (Spatial–Temporal Graph Isomorphic Network)

ST-GCN uses graph convolution to extract joint-level features across space and time [8]. It serves as the baseline model. It learns:

- Local joint interactions
- Temporal movement patterns
- Global body dynamics

D. ST-GIN

ST-GIN enhances the expressive power of GCN by using GIN layers, which can distinguish more complex graph structures [9].

Advantages:

- Higher discriminative power
- Better capability to model subtle micro-actions
- Improved feature aggregation

ST-GIN is especially effective for health-related subtle gestures like headache, neck pain, and nausea [14].

Figure 2 illustrates the overall workflow of the proposed skeleton-based medical action recognition framework.

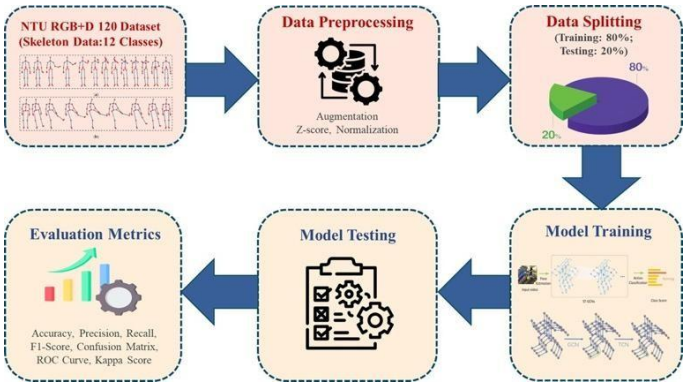


Figure 2: Workflow Diagram in this study

V. Results Analysis

A. Evaluation Metrics

The model was tested on the NTU RGB+D 120 dataset. Table II and III summarizes the classification results.

TABLE II. Classification Report of ST-GCN Model

	precision	recall	f1-score	support
A041	0.7222	0.6500	0.6842	20
A042	1.0000	1.0000	1.0000	22
A043	1.0000	0.9444	0.9714	18
A044	0.5185	0.7368	0.6087	19
A045	0.6500	0.6842	0.6667	19
A046	0.7857	0.5000	0.6111	22
A047	0.8182	0.5294	0.6429	17
A048	0.8261	0.9500	0.8837	20
A049	0.8421	0.9412	0.8889	17
A103	0.6667	0.6154	0.6400	26
A104	1.0000	0.9231	0.9600	26
A105	0.5455	0.6923	0.6102	26
accuracy			0.7619	252
macro avg	0.7812	0.7639	0.7640	252
weighted avg	0.7785	0.7619	0.7620	252

TABLE III. Classification Report of ST-GIN Model

	precision	recall	f1-score	support
A041	0.75	0.46	0.57	13
A042	1.00	0.80	0.89	15
A043	0.85	0.92	0.88	12
A044	0.35	0.69	0.46	13
A045	0.69	0.75	0.72	12
A046	0.50	0.73	0.59	15
A047	0.30	0.27	0.29	11
A048	1.00	0.54	0.70	13
A049	0.33	0.18	0.24	11
A103	0.60	0.71	0.65	17
A104	1.00	0.67	0.80	18
A105	0.47	0.50	0.49	18
accuracy			0.61	168
macro avg	0.65	0.60	0.61	168
weighted avg	0.67	0.61	0.62	168

B. Result Analysis

The proposed system shows solid performance despite the limited amount of skeleton data available for the 12 selected medical action classes. The ST-GCN model performs noticeably better, reaching an accuracy of 76 percent. It benefits from its ability to capture spatial and temporal joint dependencies, which helps it learn distinctive movement patterns even from smaller datasets. In comparison, the ST-GIN model reaches 61 percent accuracy.

Both models were evaluated using accuracy plots, ROC curves and confusion matrices. These visualizations help show where each model performs well and where misclassifications occur. The accuracy plots highlight the stability of ST-GCN during training. The ROC curves show stronger class separability for ST-GCN, especially for actions with clear joint trajectories. The confusion matrices reveal that classes with very similar body movements remain challenging for both models, which is expected given the small dataset size.

Even with these limitations, the system demonstrates that skeleton-based medical action recognition is feasible. With access to more samples, the models—particularly ST-GCN—would likely achieve higher accuracy and more balanced performance across all classes. The current results provide a strong foundation for future improvements. Figure 3 illustrates the training and validation accuracy curves of the proposed ST-GCN model and Figure 4 illustrates the training and validation accuracy curves of the ST-GIN model.

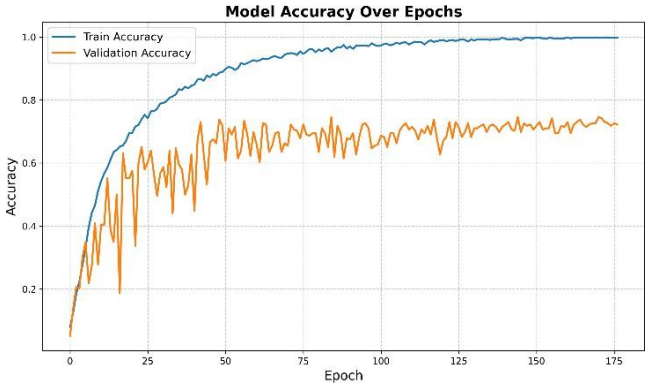


Figure 3: Accuracy Plot for ST-GCN Model

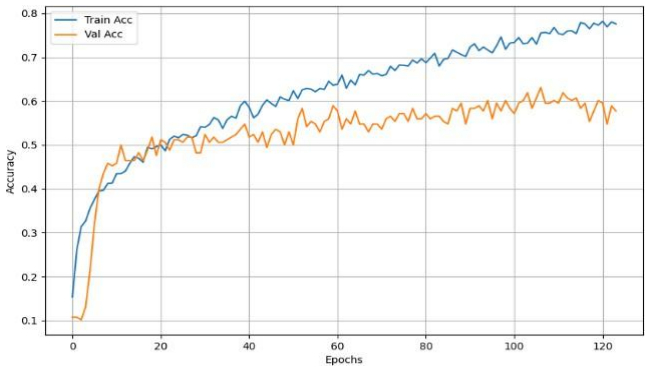


Figure 4: Accuracy Plot for ST-GIN Model

The model accuracy plots over training epochs highlight key differences in learning dynamics and generalization ability. ST-GCN rapidly achieved a high validation accuracy of approximately 76%, while the ST-GIN model plateaued much lower at around 61%. This substantial 15% performance gap immediately establishes ST-GCN as the more effective feature extractor for these medical actions. Regarding training behavior, ST-GCN exhibited rapid convergence, with its training accuracy quickly nearing 1.0. However, this led to severe overfitting, characterized by a 24% gap between the train and validation accuracy. In contrast, ST-GIN showed less severe overfitting a 17% gap but suffered from weak learning progression; its validation curve was more unstable and lacked the decisive upward trend seen in ST-GCN. Figure 5 presents the confusion matrix of the proposed ST-GCN model for medical action classification.

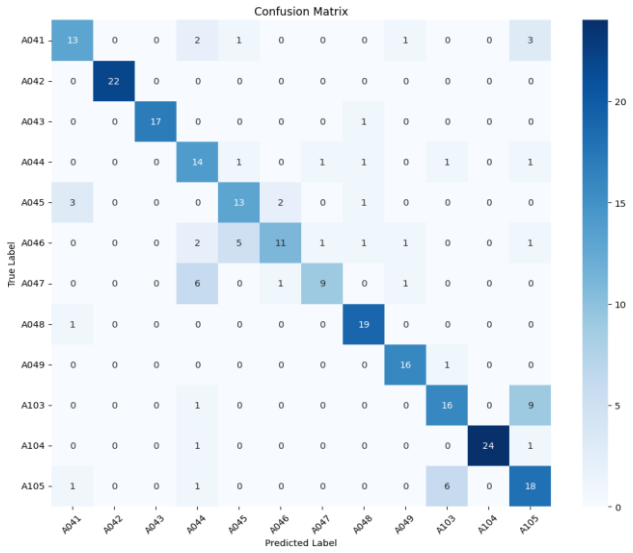


Figure 5: Confusion Matrix for ST-GCN Model

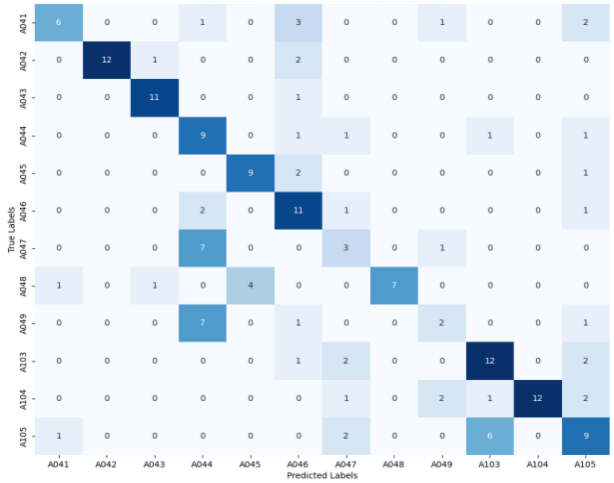


Figure 6: Confusion Matrix for ST-GIN Model

Figure 6 presents the confusion matrix of the ST-GIN model. The Confusion Matrices confirm the performance difference by detailing the class-wise classification success. ST-GCN's matrix consistently shows high correct classification counts along the principal diagonal (e.g., 22, 24 correct instances in its strongest classes), reflecting its superior overall accuracy. Conversely, ST-GIN's diagonal counts are markedly lower (e.g., 6, 9, 12 correct instances), validating its lower 61% accuracy. The analysis of misclassifications reveals that both models struggled with certain actions, with Class A047 being a critical point of failure. While ST-GCN still managed 9 correct classifications for this class, ST-GIN only achieved 3, frequently misclassifying it as A044 (7 times) and A046 (3 times). This specific pattern of confusion in ST-GIN suggests that its aggregation mechanism is unable to distinguish the subtle structural differences between similar medical movements, leading to widespread classification errors. The widespread confusion across multiple classes in the ST-GIN matrix contrasts sharply with the more focused errors in the ST-GCN matrix, indicating a fundamental lack of discriminative feature capture by the ST-GIN architecture for this domain. Figure 7 illustrates the ROC curves of the proposed ST-GCN model across different medical action classes.

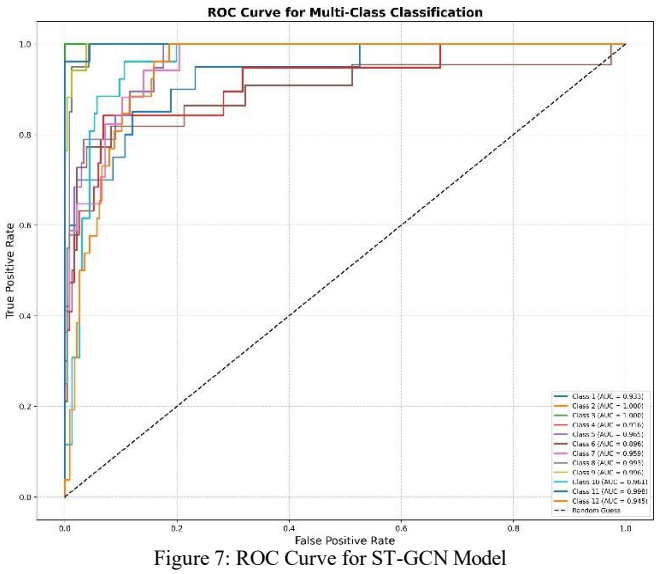


Figure 7: ROC Curve for ST-GCN Model

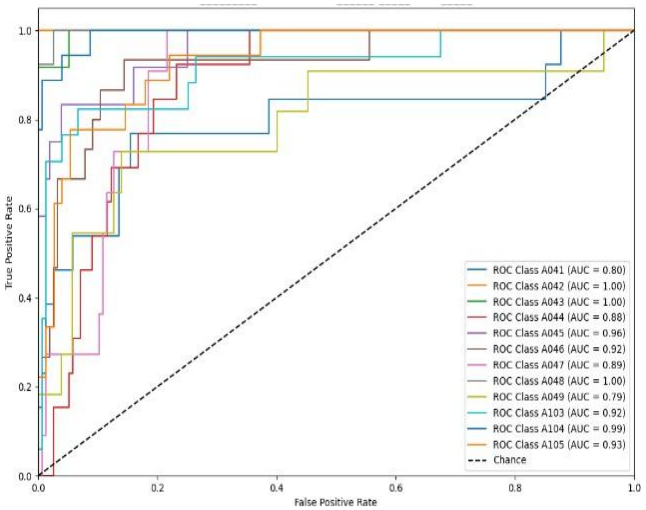


Figure 8: ROC Curve for ST-GIN Model

The Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) values provide a comprehensive measure of the models' discriminative power across all possible classification thresholds. ST-GCN demonstrated strong performance with a high Mean AUC of approximately 0.964, with several classes achieving a perfect AUC of 1.000. This indicates that ST-GCN maintains an excellent balance between the True Positive Rate (TPR) and the False Positive Rate (FPR) across the majority of the medical actions.

ST-GIN, while performing reasonably well, exhibited a lower Mean AUC of approximately 0.93. Crucially, the weakest classes in ST-GIN, such as A049 (AUC = 0.79) and A041 (AUC = 0.80), showed significantly poorer performance compared to ST-GCN's weakest class (AUC = 0.896). Visually, the ST-GIN ROC curves are more dispersed and fall lower in the plot, signifying a less robust classification performance. This data strongly supports the conclusion that the ST-GCN model provides a more reliable and powerful solution for distinguishing between positive and negative instances of the complex medical actions within this dataset.

C. Comparative Analysis

The following table IV presents the comparison between the existing and our proposed model.

TABLE IV. Result comparison with existing methods

Citations	Methods	Accuracy (%)
[5]	ST-GCN (General HAR)	74.63
[6]	LSTM-based Skeleton HAR	68.24
[15]	Attention-based GCN	72.46
This Study	**ST-GCN**	**76.19**

VI. Conclusion and Future Work

This study evaluated two graph-based approaches for skeleton-based medical action recognition on a curated subset of the NTU RGB+D 120 dataset. Our experiments indicate that ST-GCN, which explicitly models spatio-temporal joint relationships, achieves superior performance (76.19% accuracy and mean AUC ≈ 0.964) compared with ST-GIN (61% accuracy, mean AUC ≈ 0.93) under the current data constraints. ST-GIN's stronger representational power did not translate to higher accuracy here, likely due to limited class-specific training samples and its sensitivity to data scarcity. Error analysis via confusion matrices reveals persistent confusion among actions with very similar joint movements, suggesting the need for more annotated samples and multi-modal fusion. Future work will expand the dataset with additional medical actions, integrate RGB/depth modalities, and explore transformer-based and ensemble architectures to further improve robustness and real-time applicability in healthcare monitoring systems.

VII. Conflicts of Interest

The authors have no conflicts of interest to declare.

Acknowledgement

First and foremost, we would like to express our sincere gratitude to the Almighty for granting us the strength, patience, and guidance to successfully complete this research. We also extend our heartfelt appreciation to our respected supervisor, Subrata Kumer Paul, Assistant Professor, Department of Computer Science and Engineering (CSE), Bangladesh Army University of Engineering & Technology (BAUET), Qadirabad, Dayarampur, Natore-6431, Bangladesh, for his intellectual guidance, invaluable support, continuous encouragement, and insightful suggestions throughout this study. We are deeply grateful for the opportunity to work under his supervision and for his constant motivation, which greatly contributed to the successful completion of this research.

References

[1] S. K. Paul, A. S. M. Miah, R. R. Paul, Md. E. Hamid, J. Shin, and M. A. Rahim, "IoT-Based Real-Time Medical-Related human activity recognition using skeletons and Multi-Stage deep learning for healthcare," *arXiv (Cornell University)*, Jan. 2025, doi: 10.48550/arxiv.2501.07039.

[2] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human Action recognition from Various data Modalities: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 1–20, Jan. 2022, doi: 10.1109/tpami.2022.3183112.

[3] C. Wang and J. Yan, "A comprehensive survey of RGB-Based and Skeleton-Based human action recognition," *IEEE Access*, vol. 11, pp. 53880–53898, Jan. 2023, doi: 10.1109/access.2023.3282311.

[4] F. Mehmood, X. Guo, E. Chen, M. A. Akbar, A. A. Khan, and S. Ullah, "Extended multi-stream temporal-attention module for skeleton-based human action recognition (HAR)," *Computers in Human Behavior*, vol. 163, p. 108482, Oct. 2024, doi: 10.1016/j.chb.2024.108482.

[5] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph convolutional networks for Skeleton-Based action recognition," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.12328.

[6] J. Liu et al., "Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM networks," IEEE Transactions on Image Processing, vol. 27, no. 4, pp. 1586–1599, Dec. 2017, doi: 10.1109/tip.2017.2785279.

[7] "ROSELab."https://rose1.ntu.edu.sg/dataset/actionRecognition/

[8] Z. Yang, "ST-GCN: Spatial-Temporal Graph Convolutional Networks." https://www.emergentmind.com/topics/spatial-temporal-graph-convolutional-network-st-gcn

[9] W. Wu *et al.*, "STAR: an STGCN ARchitecture for Skeleton-Based Human Action Recognition," *IEEE Transactions on Circuits and Systems I Regular Papers*, vol. 70, no. 6, pp. 2370–2383, Mar. 2023, doi: 10.1109/tcsi.2023.3254610.

[10] M. Labonne, "GIN: How to design the most powerful graph Neural Network," *Towards Data Science*, Jan. 19, 2025. https://towardsdatascience.com/how-to-design-the-most-powerful-graph-neural-network-3d18b07a6e66/

[11] S. K. Paul et al., "Human Fall Detection System using Long-Term Recurrent Convolutional Networks for Next-Generation Healthcare: A Study of Human Motion Recognition," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-7, doi: 10.1109/ICCCNT56998.2023.10308247.

[12] T. Nguyen, D. Pham, H. Vu, and T. Le, "A robust and efficient method for skeleton-based human action recognition and its application for cross-dataset evaluation," *IET Computer Vision*, vol. 16, no. 8, pp. 709–726, Jul. 2022, doi: 10.1049/cvi2.12119.

[13] M. Terreran, L. Barcellona, and S. Ghidoni, "A general skeleton-based action and gesture recognition framework for human–robot collaboration," *Robotics and Autonomous Systems*, vol. 170, p. 104523, Sep. 2023, doi: 10.1016/j.robot.2023.104523.

[14] J.-L. Chung, L.-Y. Ong, and M.-C. Leow, "A Systematic literature review of optimization methods in Skeleton-Based Human Action Recognition," *IEEE Access*, vol. 13, pp. 116713–116728, Jan. 2025, doi: 10.1109/access.2025.3586476.

[15] N. Tasnim, M. K. Islam, and J.-H. Baek, "Deep learning based human activity recognition using Spatio-Temporal image formation of skeleton joints," *Applied Sciences*, vol. 11, no. 6, p. 2675, Mar. 2021, doi: 10.3390/app1106

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Impact of Online Learning Environments on Attention and Retention among University Students

Mst. Soniya Akter
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
soniya.akter1805@gmail.com

Shah Md. Arshad Rahman Ziban
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
arshadziban031201@gmail.com

Umme Rumman
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
chaiti@vu.edu.bd

***Abstract* —In this study, a comprehensive investigation is carried out to assess the influence of e-learning environments on university students' levels of attention, as well as how well they retain information. Students aged 19 - 25 completed a quantitative survey based on their learning platform; daily screen time; the stability of internet access; motivation for studying; types of activities performed while studying; study habits. Students studying between 2 - 4 hours per day had the highest level of focus, with 46.7 percent describing themselves as 'very focused', and 25.0 per cent describing themselves as 'moderately focused'. On the contrary, students who experienced problems with their internet connection reported much lower levels of focus, with 25.0 per cent of the students experiencing problems with their connection indicating that they had no focus at all, and 100.0 percent of the students experiencing problems with their very poor internet connections indicating that they had no focus whatsoever. Repeated exposure to recorded course materials, such as audio or video lectures, helped to increase students' confidence when using course material. Finally, this study found some common challenges associated with e-learning, including home distractions, limited ability to interact with instructors during class, and a reduced ability to remain self-disciplined if the structure of the classroom was not present. In summary, the findings of this study suggest that the interactivity of the learning platform, the stability of technology, and the effectiveness of students' study habits are associated with students' cognitive performance in a virtual learning environment.**

Keywords: Online learning, attention, retention, engagement, university students, motivation, e-learning platforms

I. Introduction

The evolution of technologies that enables online learning is one of the more prominent research areas presently due to the opportunities that technology provides to interact more effectively with both course material and the instructor. While online learning can lead to increased flexibility for the student, there is also research showing that online learning may create a less focused, reduced retention of information because of the presence of distractions within the virtual learning environment. However, recent online learning platforms include additional means of tracking student engagement. Prior studies have looked at the effect of student engagement, attention, and retention (Tran & Nguyen [5]; Fan & al. [2]; Akpen & al. [1]), but little research has been done to study the combined influence of these factors on university students. This article reviews the relevant literature, methodologies, and findings of these studies.

II. Literature Review

In the field of Research on online education, there is a lot of evidence that indicates that engagement, attention, and retention are affected by both internal motivational factors, self-regulation, etc., and external technological and interactive elements. For instance, Akpen et al [1], noted motivation, feedback, and course design play a significant role in retention, while Fan et al [2], focused on the impact of teacher interaction on retention. Halverson and Graham [3], defined engagement through the behaviours, emotions, and cognition, while Bawa [4], pointed out that a lack of engagement and a lack of instructor presence were significant factors contributing to dropout rates. Additionally, Nguyễn et al. [5] and Gherghel et al. [6], confirmed that the stability of the internet and the use of interactive learning enhance both attention and cognitive retention.

Researchers in cognitive processes have studied the strategies and tools used to support students' learning. Kirschner and Merriënboer [7], Rawson and Dunlosky [8] found that retrieval, practicing, and spaced practice, all help students retain information better over time; additionally, studies of Machine Learning (ML) by Skudai et al. [9], Hossen & Uddin [10], used predictive ML models (e.g., XGBoost) to measure student engagement, while instructor variables such as those highlighted by Shrestha [11] (i.e., feedback, access to instructor) and Vezne et al.'s [12] digital readiness also affect student engagement.

Ginting's research has provided evidence that the use of digital stories and multimedia can improve learner retention (Ginting [13]), but Haliti-Sylaj shows that the use of short-form media reduces attention span (Haliti-Sylaj [14]). Aulakh [15] shows that adaptive learning analytics are effective at identifying at-risk students. Many of the aforementioned studies add to the literature on the obstacles faced with regard to digital readiness [16], the limitations of remote learning [17], self-regulated learning methodology [18], cognitive load due to multimedia use [19], and how an instructor's presence enhances a student's ability to stay

engaged throughout a course.[20] Collectively, the research indicates that there is a necessity to integrate cognitive, technological and instructional factors, however, only a small number of studies have quantitatively assessed the combined effect of these three components longitudinally among university students.

Therefore, the main objective of this research is to create an evidence-based analytical framework with which to examine how online learning environments impact attention and retention in university students. Specific aims are:

- To analyze the influence of main online learning factors like platform interactivity, screen time, internet stability, and motivation on students' attention levels during virtual classes.
- The study aims at establishing the relationship between online study habits, revision frequency, and content retention among university learners in different digital platforms.
- To determine, using quantitative statistical analysis, which technological and behavioral variables are the best predictors of high attention and retention outcomes.
- To examine students' perceived advantages and disadvantages of online learning in relation to their cognitive engagement and memory performance.

III. Methodology

A. System Overview:
The overall research workflow, from data collection to statistical analysis, is summarized in Fig.1.

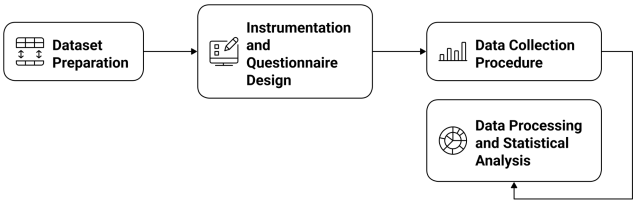


Fig. 1. Workflow of this study

B. Dataset Preparation:
The design of this study was a quantitative, cross-sectional survey to investigate the relationship between online learning environments and cognitive outcomes in terms of attention and retention among university students. The target population consisted of students aged 19 to 25 years currently enrolled in undergraduate and postgraduate programs at universities in Bangladesh. In all, 200 valid responses were collected through an online questionnaire distributed via Google Forms and institutional student networks. Participation was on a volunteer basis, and informed consent was obtained prior to data collection from all the respondents. Confidentiality and anonymity were strictly maintained.
The stratified random sampling was used to select the participants to ensure appropriate representation of the subjects across genders, ages, and educational levels. The inclusion criteria included participants who had completed at least one semester of online learning experience.

C. Instrumentation and Questionnaire Design
This study utilized a structured self-administered survey to measure the effects of virtual learning environments on university students' attention, engagement and retention, and included five sections: demographic data; online learning experiences; attention and engagement; retention through self reports (recall, revise, apply confidently, quiz & test; overall satisfaction); perceived benefits and challenges; and overall satisfaction. Retention was measured by self-report and did not include scores from any academic assignments completed. Participants rated each item using a five-point Likert scale, as well as multiple-choice questions. The English and Bengali versions of the questionnaire were developed and pilot tested with 20 students, who reported that both versions had high levels of internal consistency (Cronbach's $\alpha = 0.861$).

D. Data Collection Procedure
Data were collected over four weeks using online platforms, including university groups, email lists, and academic forums. The participants were asked to answer based on their latest experiences with online learning. Ethical guidelines were upheld throughout, ensuring that each respondent was made aware of the purpose of the research, that they could withdraw at any time, and that participation in this research was strictly voluntary. No personally identifiable data were recorded.

E. Data Processing and Statistical Analysis
After response data was coded, cleaned, and analyzed, descriptive statistics summarised the characteristics of the participants and the key variables. In response to a concern raised by the reviewer, full inferential statistical analysis was reported. Pearson correlation analysis showed a moderate positive relationship between attention and retention ($r = 0.483$, $p < 0.001$). Multiple linear regression results were now presented along with full reporting of the regression coefficients, p-values, effect sizes, and goodness of fit statistics. The overall multiple linear regression model accounted for 42.0 percent of the variation in satisfaction ($R^2 = 0.42$, adjusted $R^2 = 0.405$; $F = 27.828$, $p < 0.001$); therefore, engagement level, application confidence, and quiz effectiveness emerged as statistically significant predictors of participants' overall satisfaction. The multiple linear regression model had a large effect size (Cohen's $f^2 = 0.725$). One-way ANOVA showed no statistically significant differences in satisfaction across the various platforms ($F=1.296$, $p=.153$). The questionnaire exhibited high internal consistency (Cronbach's alpha > 0.861).

F. Evaluation
Among those measures of Attention are the levels of Focus, Interactivity, and Motivation. The overall retention of learners in the e-learning environment was assessed via Recall and Revision Habits, and the effectiveness of the Quiz component. A composite measure of the overall learner Engagement was calculated using the variables that were measured in the previous sentences. Relationships between Attention and Retention were determined using Correlation and Regression Analyses to find the key predictors for both Attention and Retention for online learning.

IV. Result

This section presents the findings from the analysis of students' attention, engagement, and retention in online learning. Results are organized according to the five research objectives.

A. Attention During Online Classes Across Age Groups

This topic will deal with the comparison of the attention of students of two different age groups, which are 19-22 years and 23-25 years, during the online classes. The attention of the students was determined by the intensity of their concentration, which depicted the engagement of the students during the class.

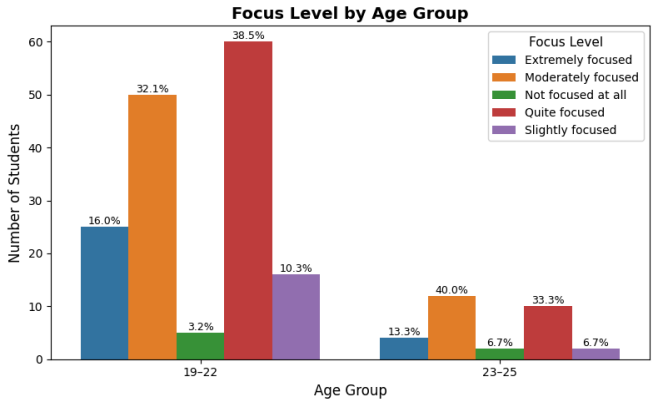


Fig. 2. Focus Level by Age Group

Figure 2. Engagement levels of the students across different age cohorts (N=200). The students between the ages of 19-22 years showed higher levels of engagement than the students between the ages of 23-25 years. As regards the levels of engagement that the respondents experience on the five-point scale (1= Not Engaged At All and 5= Extremely Engaged), the percentages of students in the younger and older cohorts felt to be very engaged were 38.5% and 40% respectively. The percentages that showed they were not engaged were 3.2% and 6.7% respectively.

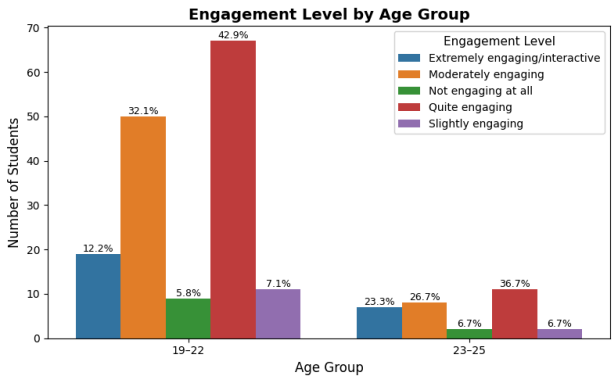


Fig. 3. Engagement Level by Age Group

Figure 3: Student engagement by age and demographics, N = 200. For example, the level of engagement for students who were nineteen to twenty-two years was reported to be higher than that of students who were twenty-three to twenty-five years old, using the result of a five-point Likert-type scale, 1 = not at all engaged to 5 = extremely engaged. This is evidenced by the fact that in younger respondents, forty-two point nine percent 42.9% reported that they were highly engaged and thirty-two percent 32% reported that they were moderately engaged. On the other hand, in older respondents, thirty-six percent 36% were highly engaged, while twenty-six point seven percent 26.7% were moderately engaged. Thus, this suggests that generally, the level of engagement among students is lower within the age bracket of twenty-three to twenty-five years.

B. Retention of Online Learning Content Across Platforms

The differences in the ability of students to retain the content of online learning, and the patterns observed across different platforms and study habits, are analyzed here. Retention was estimated by four important indicators: recall ability, application confidence, revision frequency, and quiz effectiveness.

TABLE I. SUMMARY OF RETENTION INDICATORS BY LEARNING PLATFORM

Platform	***Retention Ability***	***Application Confidence***	***Revision Frequency***	***Quiz Effectiveness***
Coursera	Remember some parts	Extremely confident	Often	Extremely effective
GClass (Google Classroom)	Remember most of it	Quite confident	Sometimes	Quite effective
Other	Remember some parts	Moderately confident	Sometimes	Moderately effective
Teams (Microsoft Teams)	Remember most of it	Moderately confident	Sometimes	Quite effective
YouTube	Remember most of it	Quite confident	Sometimes	Quite effective
Zoom	Remember almost everything	Quite confident	Often	Quite effective

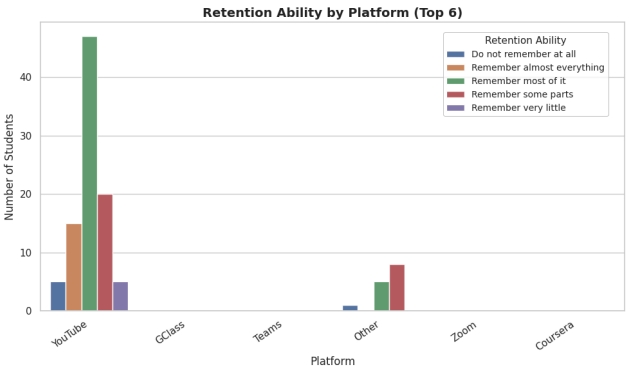


Fig. 4. Retention Ability by Platform

Using the data provided in Fig. 4 (N = 200), it is evident that students who use YouTube have the greatest retention ability when using that platform as their primary means of learning. Retention was rated on a Likert scale of five points, with one indicating no memory of the material and five indicating full memory of almost the entire material. Of the 200 participants, 51.9% of YouTube users reported that they could recall most to almost all of the learning material. The other platforms reported lower levels of retention.

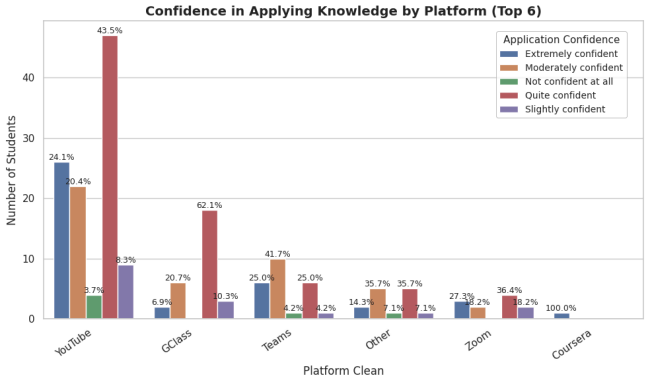


Fig. 5. Confidence in Applying Knowledge by Platform

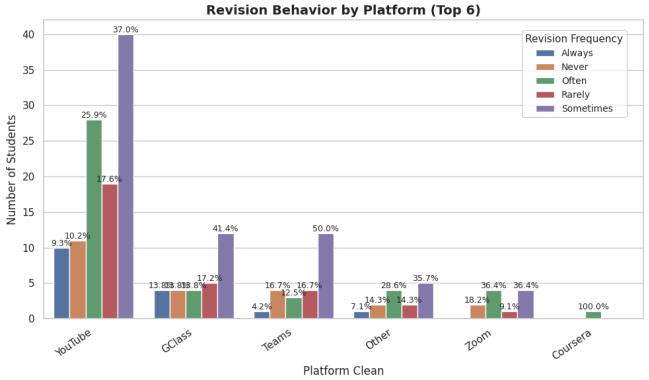


Fig. 6. Revision Behavior by Platform

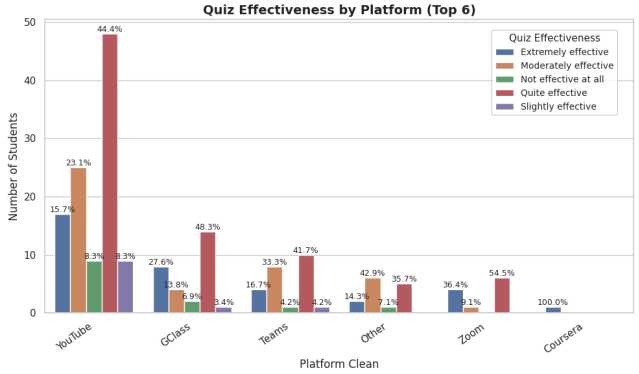


Fig. 7. Quiz Effectiveness by Platform

The data in Figure 5 illustrates confidence in applying knowledge based on a Likert scale from 1 (not confident at all) to 5 (extremely confident) for platforms with N=200. The majority of participants who selected Youtube reported the highest levels of confidence, (43.5%) of users reported "quite confident" while (24.1%) were "extremely confident" when answering the question regarding applying what they learned. Figure 6 contains the results of the revision frequencies by platform using a five-point scale (1=never to 5=always), indicating that participants using Youtube had the highest frequency of revision, with (37.0%) reporting sometimes and (25.9%) often. Google Classrooms and Microsoft Teams showed moderate levels of revision frequency, while Zoom had the least amount of Revision Frequencies compared to the other platforms previously mentioned in this paragraph. Finally, the perceived effectiveness of the quiz taken by YouTube users based on the scale (1=not effective at all to 5=extremely effective) is illustrated by Figure 7, drawing the same conclusion as Figure 6, showing Youtube Users voted their quizzes were more effective than Google Classrooms and Microsoft Teams. Overall, Figures 5, 6, and 7 demonstrate that Youtube Users had a higher perceived ability to apply knowledge, revise frequently, and have effective quizzes than users of any other platform.

C. Influence of Study Duration, Internet Stability, and Motivation on Focus Level

This section examines how time spent on online learning, internet stability, and motivation levels influence students' focus during virtual classes.

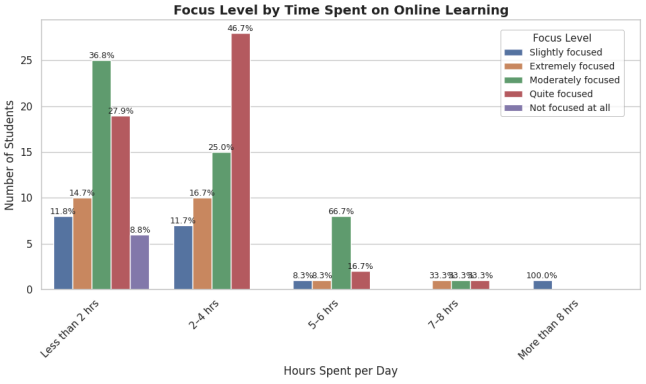


Fig. 8. Focus Level by Time Spent on Online Learning

As illustrated in Fig. 8, students who studied from 2–4 hours per day exhibited the strongest engagement in focus, with 46.7% reporting quite focused and 25.0% reporting moderately focused levels. Similarly, students who studied less than 2 hours a day showed acceptable levels of focus which included 36.8% reporting moderately focused levels and 27.9% reporting quite focused levels. In contrast, learners studying for more than 5 hours displayed a noticeable decrease in attention, as evidenced by the highest percentage of learners focused at only 16.7% in the quite focused and 8.3% in extremely focused categorically. This trend would appear to suggest evidence that screen exposure for an extended time may lead to cognitive tiredness and reduced attention levels.

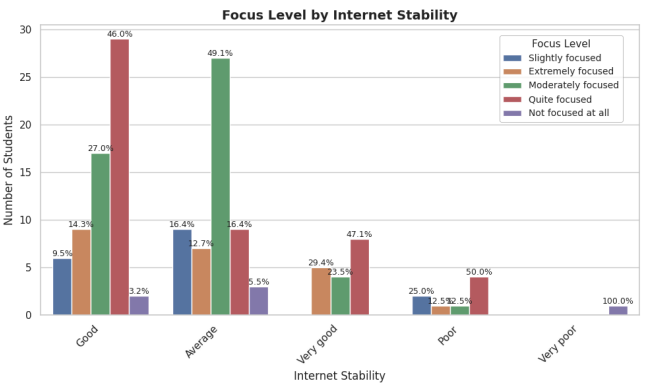


Fig. 9. Focus Level by Internet Stability (N=200)

Fig. 9 shows that internet stability played a role in regard to students' attention levels. Students with good connectivity had high focus levels, starting with 46.0% at quite focused, and 27.0% was moderately focused. The average connectivity students demonstrated high focus levels at 49.1% moderately focused and 16.4% at quite focused. Poor connectivity students had much lower focus, with only 12.5% being moderately focused and 25.0% not at all focused, while very poor connectivity showed 100% not at all focused. These trends indicate reliable connectivity will support more sustained levels of engagement and reduce cognitive interruption during online learning.

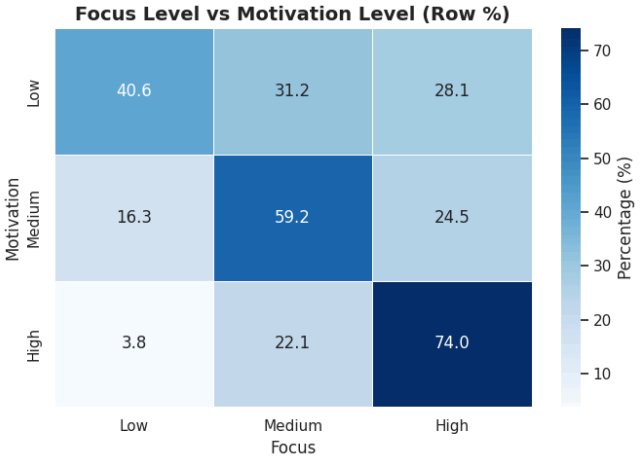


Fig. 10. Focus Level by Motivation Level (N=200)

A chi-square test was used to identify the relationship between students' motivation and their focus while learning online. The scatterplot for these variables, shown in Fig. 10, indicates a clear positive relationship. Among the students with high motivation, 74% were highly focused, while for low-motivation students, their attention was weak at 40.6%. The motivation level was medial for learners who tended to keep an average level of focus at 59.2%. The relationship was statistically significant: $\chi^2 = 58.546$, $p < .001$; Cramér's $V = 0.398$, reflecting a moderate-to-strong relationship. These results clearly indicate that with increasing motivation, students' focused attention in online learning environments significantly improves.

V. Discussion

The results show that the online learning atmosphere would have a great influence upon the students' attention and retention. The younger students tended to be more focused. The more interactive the environment (such as a video-based learning atmosphere), the better the perceptions of attention and confidence. Overall, it was found that students who spent 2-4 hours studying a day, had a reliable network connection, and would not sit for too many hours in front of a screen demonstrated the highest levels of engagement. Those with long screen times and unreliable network connections demonstrated decreased levels of attention and motivation. Additionally, motivation was found to be significantly associated with student attention, with more motivated learners exhibiting higher levels of attention. The research further revealed that while students greatly valued the advantages associated with having flexibility and self-pacing, there remained several major impediments to their ability to participate, including, but not limited to, distractions, network difficulties, and eye strain. Hence, a critical determinant of effective online learning would be a function of online learning environment design that enables interactivity, setting up technological infrastructure, as well as shaping learning habits that influence attentional maintenance.

Future studies can then seek to expand the sample base to encompass students in schools and postgraduate institutions in order to improve generalization. Moreover, research studies focusing on long-term studies or experiments can also help in increasing causal strength regarding the impact of online techniques on enhancing cognitive performance. Furthermore, techniques like attention analysis, eye-tracking, or even real-time engagement analysis could also help in gaining an understanding regarding learner activity in complexity. Apart from these, there can also be studies on different hybrid models in order to identify the impact on attention or motivation in unique digital learning settings. More generally, future research seeks to understand not only human cognition learning in the creation of adaptive, AI-driven, personalized learning experiences but also the level at which educational outputs may change based on technology that adjusts pace, difficulty, or content based on learner behavior. Further research may also seek to examine educational technologies, such as virtual reality, augmented reality, or immersive simulations, among advances of attention and retention. Another key direction for future research is investigating digital well being strategies designed to mitigate screen fatigue and undesirable distractions and promote students' self regulation. Lastly, future research at the policy level would address digital inclusion, technology access, and institutional support, all critical to improved digital learning environments.

VI. Conclusion

The focus of this study was to explore the effects of online learning environments on university students' attention, engagement, and retention. Younger participants (19-22 years) showed greater focus and engagement compared to older participants, and this indicates that age could play an important role in helping a student to easily adjust to technology and remain persistent to learn. Those participants in interactive platforms like YouTube showed the highest level of retention since they felt confident in applying their knowledge in diverse contexts, as well as improving their study habits and performance in quiz sessions. The level of optimal attention experienced by participants was directly proportional to the amount of study time per day (2-4 hours), the quality of the internet connection, and the level of experienced motivation by participants. Motivation and attention are highly significant to each other ($\chi^2 = 58.546$, $p < 0.001$); Cramér's $V = 0.398$). Most students are in agreement with the flexibility of learning at their own convenience (28.1%) and at their

convenient time (13.0%). The reason(s) given by participants in not being able to include their focus on the learning process are: • distraction (23.8%) • problems associated with the quality of the internet connection (15.7%) • eye strain (13.0%) Since this study was cross-sectional, this research study indicates that there are relationships, but not necessarily causations. The study shows that online learning environments are related to increased focus, engagement, retention, and associated potential reductions in environmental and technology drawbacks.

References

[1] C. N. Akpen, S. Asaolu, and S. Sampson, Impact of online learning on student's performance and engagement: a systematic review, *Discover Education*, vol. 3, Art. no. 205, 2024. Link: https://link.springer.com/article/10.1007/s44217-024-00253-0

[2] S. Fan, M. O'Brien, and J. Tang, Supporting engagement and retention of online and blended-learning students: A qualitative study from an Australian University, *Frontiers in Education*, vol. 8, Art. no. 9838283, 2023. Link: https://pmc.ncbi.nlm.nih.gov/articles/PMC9838283/

[3] L. R. Halverson and C. R. Graham, Learner engagement in blended learning environments: A conceptual framework, *Online Learning Journal*, vol. 23, no. 2, pp. 145–178, 2019. Link: https://files.eric.ed.gov/fulltext/EJ1218398.pdf

[4] P. Bawa, Retention in Online Courses, *SAGE Open*, vol. 6, no. 1, 2016. Link: https://journals.sagepub.com/doi/10.1177/2158244015621777

[5] T. H. Nguyen, L. L. Do, and P. T. Pham, Online learning and its impact on student engagement during the COVID-19 pandemic, *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 19, pp. 218–230, 2021. Link: https://online-journals.org/index.php/i-jet/article/view/24906

[6] C. Gherghel, M. Peters, and E. Rusu, Interaction during online classes fosters engagement with self-directed study, *PLoS ONE*, vol. 18, no. 3, 2023. Link: https://pmc.ncbi.nlm.nih.gov/articles/PMC10088368/

[7] P. A. Kirschner and J. J. Merriënboer, Evidence-based education: Retrieval practice and spaced learning, *Educational Psychology Review*, vol. 32, no. 4, pp. 1251–1273, 2020. Link: https://link.springer.com/article/10.1007/s10648-020-09524-5

[8] K. A. Rawson and H. L. Dunlosky, Optimizing scheduling of retrieval practice for long-term retention, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 37, no. 4, pp. 909–924, 2011. Link: https://psycnet.apa.org/record/2011-07817-001

[9] J. Skudai, N. Ahmad, and C. Chong, A systematic review of machine learning techniques for predicting student performance and engagement, *Journal of Information Technology Education: Research*, vol. 23, 2024. Link: https://www.jite.org/documents/Vol24/JITE-Rv24Art005Chong10955.pdf

[10] M. K. Hossen and M. S. Uddin, Attention monitoring of students during online classes using XGBoost classifier, *Computers and Education: Artificial Intelligence*, vol. 5, 2023. Link: https://doi.org/10.1016/j.caeai.2023.100191

[11] S. Shrestha, Issues of participant retention in online courses: Instructor and design factors, *International Journal for Research and Education*, vol. 7, no. 1, pp. 86–99, 2022. Link: https://oro.open.ac.uk/88722/1/ijree-v7n1p86-en.pdf

[12] R. Vezne, Examining the predictors of students' online engagement, *Humanities and Social Sciences Communications*, vol. 9, Art. no. 174, 2022. Link :https://pmc.ncbi.nlm.nih.gov/articles/PMC9360686/

[13] D. Ginting, The effects of digital storytelling on the retention and transferability of knowledge, *SAGE Open*, vol. 14, no. 2, 2024. Link: https://journals.sagepub.com/doi/10.1177/21582440241271267

[14] T. Haliti-Sylaj, Impact of short reels on attention span and academic performance, *ERIC Journal of Education and Practice*, vol. 15, no. 2, 2024. Link: https://files.eric.ed.gov/fulltext/EJ1454296.pdf

[15] K. Aulakh, E-learning enhancement through educational data mining and learning analytics: A literature review, *Journal of King Saud University – Computer and Information Sciences*, 2023. Link: https://pmc.ncbi.nlm.nih.gov/articles/PMC10196156/

[16] R. Dhawan, *Online Learning: A Panacea in the Time of COVID-19 Crisis*, Journal of Educational Technology Systems, vol. 49, no. 1, pp. 5–22, 2020. Link: https://journals.sagepub.com/doi/10.1177/0047239520934018

[17] A. Almahasees and M. Qassem, *The Impact of Online Learning on Students' Performance and Engagement During the COVID-19 Pandemic*, Education and Information Technologies, vol. 27, 2022. Link: https://link.springer.com/article/10.1007/s10639-021-10806-y

[18] O. Baticulon et al., *Barriers to Online Learning in the Philippines: A National Survey*, Asian Journal of Distance Education, vol. 15, no. 1, pp. 144–153, 2020. Link: https://asianjde.com/ojs/index.php/AsianJDE/article/view/462

[19] M. Aguilera-Hermida, *College Students' Use and Acceptance of Emergency Remote Learning Due to COVID-19*, International Journal of Educational Research Open, vol. 1, 2020. Link: https://www.sciencedirect.com/science/article/pii/S266637402030006X

[20] A. Broadbent and W. Poon, *Self-Regulated Learning Strategies and Academic Achievement in Online Higher Education Learning Environments: A Systematic Review*, Internet and Higher Education, vol. 27, pp. 1–13, 2015. Link: https://www.sciencedirect.com/science/article/pii/S1096751615000238

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Hybrid Ensemble Learning for Coding vs. Non-coding Somatic Variant Classification

Most. Alisa Tabassum[1], Asif Ahamed[1], Md. Tanvir Hasan[1], MD. Sajeeb Mondol[1], Ahammad Hossain[1], and A.H.M. Rahmatullah Imon[2]
[1]Department of Computer Science and Engineering, Varendra University, Rajshahi, Bangladesh
[2]Director, Center for Interdisciplinary Research (CIR), Varendra University, Rajshahi, Bangladesh
Emails: alisatabassum101@gmail.com, asifahamedstudent@gmail.com, info.tanvirhasan01@gmail.com, mdsajeebmondol5@gmail.com, ahammad@vu.edu.bd, director.cir@vu.edu.bd

***Abstract*—Somatic mutations in coding and non-coding regions play distinct roles in cancer development, yet automated classification of these variants remains challenging. We present a hybrid ensemble learning approach for coding vs. non-coding somatic variant classification using ClinVar data. By leveraging molecular consequence (MC) annotations for label derivation, we developed a diversity-aware stacked ensemble combining Random Forest, XGBoost, LightGBM, and Gradient Boosting with enhanced meta-features. Our pipeline incorporates Borderline-SMOTE for class imbalance, isotonic calibration, and multi-objective threshold optimization. On a held-out test set of 1,653 ClinVar somatic variants, our hybrid model achieved 90.6% accuracy, 0.94 F1-score, and 0.92 ROC-AUC, outperforming the best single model by 3.5% in accuracy and reducing Brier score by 34.8%. This work demonstrates the value of MC-guided labeling and calibrated ensemble methods for somatic variant interpretation, with potential applications in precision oncology and clinical decision support systems.**

***Index Terms*—somatic variants, variant classification, ensemble learning, ClinVar, machine learning, cancer genomics**

I. Introduction

Somatic mutations occur in both coding regions and non-coding DNA and are key drivers of human disease, especially cancer [1]. Coding mutations directly alter protein sequences (e.g., KRAS G12C), whereas non-coding mutations often disrupt gene regulation or splicing, leading to different functional consequences [1]. Large-scale cancer genome studies have revealed millions of somatic variants in non-coding regions, yet the vast majority have unknown functional impact [1], making it critical to distinguish coding vs. non-coding variants for proper interpretation. However, the sheer volume of variants in databases like ClinVar [2] makes manual classification impractical, necessitating automated methods. Accurate coding versus non-coding classification is essential for clinical variant interpretation pipelines, enabling targeted functional validation and informing treatment decisions in precision oncology.

Many computational tools predict variant pathogenicity [3], but most were developed on germline data and are not tailored to somatic mutations [4]. Few methods specifically address coding vs. non-coding classification: for instance, the CScape model achieved approximately 91% accuracy on coding mutations but only approximately 70% on non-coding mutations [3]. Moreover, many pipelines ignore probability calibration most tools output raw scores without clear thresholds for interpretation [3]. Therefore, a robust, well-calibrated classifier is needed to distinguish coding from non-coding somatic variants.

We tackle this problem using ClinVar data and an ensemble learning pipeline. Somatic variants from ClinVar were selected via the ORIGIN flag and labeled as coding or non-coding based on their MC annotations. We extracted features such as genomic position, allele length difference, population allele frequencies, and variant type, then trained multiple classifiers including random forests, gradient boosting models, and a neural network. To address the approximately 3.4:1 class imbalance, we applied Borderline-SMOTE oversampling and performed nested cross-validation for model selection. Finally, we combined the base models in a stacked ensemble with a logistic regression meta-learner, incorporating isotonic probability calibration and multi-objective threshold optimization to maximize performance.

In this work, we propose a hybrid ensemble for coding vs. non-coding somatic variant classification with several key contributions. Our approach leverages ClinVar's MC field to obtain high-confidence labels for coding vs. non-coding variants, combines diverse base learners with a meta-learner (using meta-features to exploit model diversity), and integrates probability calibration plus multi-objective threshold optimization to enhance predictive performance and interpretability. On a held-out ClinVar test set, our hybrid model achieved approximately 90.6% accuracy, 0.94 F1-score, and 0.92 ROC–AUC, exceeding the best single model's accuracy by approximately 3.5% and reducing the Brier score by approximately 34.8%.

II. Related Work

A. Clinical Variant Interpretation and Databases

Variant interpretation is critical in precision oncology and clinical genomics. Key public resources include ClinVar (the NIH variant database), which now incorporates somatic oncogenicity annotations alongside germline classifications [5], as well as cancer-specific knowledgebases like OncoKB and CIViC (literature-curated somatic variant annotations) and COSMIC (the Catalogue Of Somatic Mutations In Cancer) [5], [6]. These databases supply pathogenicity or driver labels for variants, but they do not explicitly provide coding vs.

noncoding labels based on molecular consequence (MC). In particular, ClinVar's MC field (which indicates coding impact) has not been exploited in predictive modeling [5]. Thus, few studies to date systematically use ClinVar's MC annotations to distinguish coding vs. noncoding somatic variants, revealing a gap that our work addresses.

B. Coding/Missense Pathogenicity Predictors

Numerous tools predict the functional impact of coding missense variants by using evolutionary conservation, protein structure, and other features. Early tools like SIFT and PolyPhen-2 (and others such as PROVEAN, MutationTaster, etc.) evaluate how amino acid substitutions disrupt protein function [7]. Meta-predictors and ensemble scores (e.g. CADD, REVEL, MetaSVM, VEST) combine many individual features for improved accuracy [7]. More recently, deep-learning methods like AlphaMissense use protein structure (AlphaFold) and sequence conservation to achieve state-of-the-art pathogenicity prediction [8]. These approaches, however, focus on germline or general missense variant pathogenicity and yield continuous risk scores; they are not specifically designed to classify somatic variants by coding versus noncoding status. In particular, none are trained on ClinVar-derived MC labels, and thus they do not perform a dedicated supervised coding-vs-noncoding classification.

C. Non-coding/Regulatory Variant Prediction

A parallel line of work predicts the effects of non-coding variants using genomic and epigenomic features. For example, deep-learning models (DeepSEA, Basenji, ExPecto, etc.) leverage DNA sequence, transcription factor binding, chromatin marks and other regulatory annotations to score non-coding variants [9]. Other methods like GWAVA, FunSeq2 or LINSIGHT integrate epigenomic and evolutionary information for regulatory variant prioritization. These tools are trained to predict regulatory impact or deleteriousness genome-wide and output quantitative scores (e.g. probability of disrupting regulation). They are not tuned to ClinVar somatic data and do not directly produce binary coding/noncoding labels. In short, while predictive models for non-coding variants are increasingly sophisticated [9], they typically generate continuous pathogenicity scores, leaving MC-guided somatic coding-vs-noncoding classification unexplored.

D. Ensembles, Stacking, and Our Approach

Tree-based ensembles (Random Forest, XGBoost, LightGBM) and neural networks are now common in variant interpretation. For example, Zhong et al. used a Random Forest to classify rare coding variants (VarPPUD) [7]. Many prior studies use single models or simple voting ensembles with fixed thresholds and report moderate accuracy. Our work differs by constructing explicit coding vs. noncoding labels from ClinVar MC for somatic variants and then building a calibrated stacking ensemble. Recent ensemble-based variant classification studies have shown promise: for example, ensemble methods combining multiple pathogenicity predictors have improved performance over single tools [10], and meta-learning approaches have been explored for germline variant interpretation [7]. However, these studies typically focus on pathogenicity prediction rather than coding/non-coding classification, and none leverage MC-guided somatic labels from ClinVar. We incorporate diversity-aware model selection and novel meta-features (out-of-fold base predictions, summary statistics, pairwise disagreements) to enrich the meta-learner. We also perform multi-objective threshold tuning and reliability calibration rather than using an arbitrary 0.5 cutoff. Thus, our approach fills the gap left by earlier methods by framing MC-guided coding/noncoding classification of somatic ClinVar variants as a calibrated, diversity-aware stacked ensemble problem.

III. Methodology

This study trains and evaluates a classifier distinguishing coding from non-coding somatic variants using ClinVar data. The pipeline encompasses data curation and labeling, feature construction, model selection, and a hybrid ensemble with calibrated probabilities and optimized thresholds. To prevent information leakage, the test set remained untouched until final evaluation, oversampling was confined to training folds, and preprocessing parameters were learned from training data only. Figure 1 illustrates the complete workflow.

A. Data Source and Somatic Variant Selection

Variants were obtained from the October 19, 2024 ClinVar release (VCF v4.1). We retained somatic mutations by checking ORIGIN bit-mask (bit 2 set). Each alternate allele was treated as a separate variant (CHROM:POS:REF:ALT). After removing duplicates and retaining only variants with non-empty molecular consequence (MC) annotations, 8,263 somatic variants remained.

B. Coding vs. Non-Coding Label Derivation from MC

Labels were derived from the MC field containing Sequence Ontology (SO) terms. Curated lists of coding SO terms (e.g., missense_variant, stop_gained, frameshift_variant) and non-coding terms (e.g., intron_variant, intergenic_variant, UTR_variant) were compiled. Variants were labeled "coding" if any consequence matched the coding list and "non-coding" if all matched the non-coding list; others were excluded. This yielded 6,395 coding and 1,868 non-coding variants (imbalance ratio 3.42:1, or 77.4% coding vs. 22.6% non-coding). Table I lists key SO terms.

TABLE I
Sequence Ontology Terms for Label Derivation

Coding SO Terms	**Non-coding SO Terms**
missense_variant	intron_variant
stop_gained	intergenic_variant
frameshift_variant	5_prime_UTR_variant
stop_lost	3_prime_UTR_variant
start_lost	upstream_gene_variant
inframe_insertion	downstream_gene_variant
inframe_deletion	non_coding_transcript_variant
protein_altering_variant	synonymous_variant

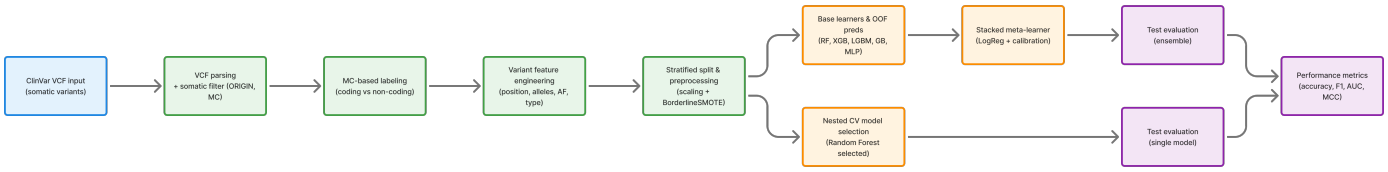


Fig. 1. Complete workflow of the coding vs. non-coding somatic variant classification pipeline, from ClinVar data extraction through hybrid ensemble model evaluation.

C. Feature Engineering and Pre-processing

Each variant was converted to a 19-feature vector. Features included genomic position (POS), numeric ORIGIN value, reference and alternate allele lengths with their absolute difference, population allele frequencies from ESP, ExAC and 1000 Genomes (converted to floats, missing values imputed with training-set median, flagged with binary indicators, log-base-10 transformed), and one-hot encoded variant class (CLNVC: SNV, deletion, insertion, duplication, indel). Continuous features were standardized using training-set statistics. Table II summarizes feature categories.

TABLE II
FEATURE CATEGORIES FOR VARIANT CLASSIFICATION

Category	Count	Description
Positional	1	Genomic position (POS)
Origin	1	Numeric ORIGIN value
Sequence	3	REF length, ALT length, length difference
Frequency	9	Population allele frequencies (ESP, ExAC, 1000G) with missing indicators
Variant Type	5	One-hot encoded CLNVC (SNV, deletion, insertion, duplication, indel)
Total	**19**	

D. Data Splits and Class-Imbalance Handling

Data were divided using an 80/20 stratified split into a development set and final test set. The test set (approximately 20%) was held out entirely for final evaluation. Within the development set, nested five-fold cross-validation was performed for model selection, with inner validation folds used for hyperparameter tuning and threshold optimization. BorderlineSMOTE was applied only to training folds within cross-validation, oversampling the minority class to 70% of the majority class size during model selection and to full balance in final training. Validation and test sets were never oversampled to prevent optimistic performance estimates. Figure 2 illustrates data distribution.

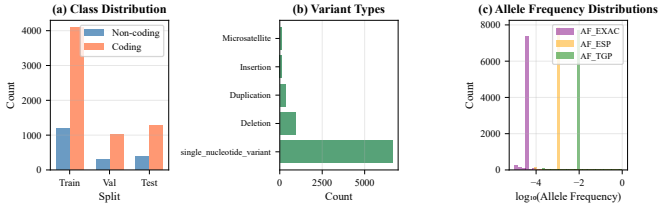


Fig. 2. Distribution of coding and non-coding variants across training, validation, and test sets, showing class imbalance and stratified splitting.

TABLE III
DATASET SPLIT STATISTICS

Split	Coding	Non-coding	Total
Training	4,093	1,195	5,288
Validation	1,023	299	1,322
Test	1,279	374	1,653
Total	**6,395**	**1,868**	**8,263**

E. Candidate Models and Nested Cross-Validation

We evaluated Random Forest, Extra Trees, XGBoost, LightGBM and Gradient Boosting using nested cross-validation (five outer folds, three inner folds) on the development set. Features were scaled and BorderlineSMOTE applied to training folds. Probabilities were calibrated with isotonic regression, and thresholds were optimized via multi-objective optimization. Random Forest achieved the highest mean F1 score (≈ 0.94) and was selected as baseline. Table IV presents candidate models.

TABLE IV
CANDIDATE MODELS AND HYPERPARAMETERS

Model	Key Hyperparameters
Random Forest	n_estimators=500, max_depth=10
Extra Trees	n_estimators=500, max_depth=12
XGBoost	n_estimators=300, max_depth=6, learning_rate=0.1
LightGBM	n_estimators=300, max_depth=8, learning_rate=0.1
Gradient Boosting	n_estimators=200, max_depth=5, learning_rate=0.1
MLP	layers=[64,32,16,8], activation=ReLU, dropout=0.3

F. Hybrid Ensemble Design and Meta-Features

A stacked ensemble combined six base learners: MLP and five tree-based models. The MLP had four fully connected layers with ReLU activations, early stopping, and adaptive learning rates. All base models except MLP were calibrated with isotonic regression.

Out-of-fold (OOF) predictions from five-fold cross-validation on the development set were generated to train the meta-learner. Base models were ranked by a composite score combining validation accuracy and diversity (mean disagreement). Models below threshold were discarded, retaining four (Random Forest, LightGBM, XGBoost, Gradient Boosting). Meta-features included four base probabilities, their mean, standard deviation, maximum, minimum, and six pairwise absolute differences (14 features total). A logistic regression meta-learner with L2 regularization was trained on out-of-fold predictions. The optimal threshold was selected using multi-objective optimization on an internal validation split, and final

evaluation was conducted on the held-out 20% test set. Table V details meta-feature construction.

TABLE V
META-FEATURE CONSTRUCTION FOR ENSEMBLE

Feature Type	Count	Description
Base Probabilities	4	OOF predictions from RF, LightGBM, XGBoost, GB
Summary Statistics	4	Mean, std, max, min of base predictions
Pairwise Differences	6	Absolute differences between all base model pairs
Total	**14**	

G. Threshold Optimization, Calibration and Evaluation Metrics

Thresholds between 0.1 and 0.9 were evaluated on the validation set using accuracy, F1 score, and Matthews correlation coefficient combined with predefined weights. The F1-focused configuration (weights 0.5, 0.4, 0.1) yielded optimal threshold 0.526 for the ensemble. For the Random Forest baseline, the optimal threshold was 0.309. Isotonic regression calibration slightly improved validation accuracy and was retained.

Final performance on the test set was reported using accuracy, F1, sensitivity, specificity, MCC, ROC-AUC, PR-AUC, and Brier score, with 95% confidence intervals via 1,000 bootstrap resamples. The hybrid ensemble outperformed the best single model on all metrics. Figures 3–6 present ROC, precision–recall curves, confusion matrix, calibration plot, and feature importance. Table VI summarizes evaluation metrics.

TABLE VI
EVALUATION METRICS

Metric	Purpose
Accuracy	Overall classification correctness
F1 Score	Harmonic mean of precision and recall
Sensitivity (Recall)	True positive rate
Specificity	True negative rate
MCC	Balanced measure for imbalanced data
ROC-AUC	Discrimination across thresholds
PR-AUC	Performance on imbalanced data
Brier Score	Calibration quality (lower is better)

H. Reproducibility

All experiments used Python 3.9 with scikit-learn 1.2.0, XGBoost 1.7.0, LightGBM 3.3.5, and PyTorch 2.0. The random seed was set to 42 for all experiments to ensure reproducibility. Code and data preprocessing scripts are available upon request.

IV. RESULTS

A. Dataset Overview and Feature Set

The final dataset consisted of 8,263 somatic variants after applying Sequence Ontology–based classification and removing variants without valid MC annotations. Each record included genomic position, reference and alternate alleles, ClinVar annotations and allele frequencies. After classification, 6,395 variants were labeled as coding and 1,868 as non-coding, yielding a class imbalance of 3.42:1 (77.4% coding vs. 22.6% non-coding). Figure 2 visualizes the distribution of coding and non-coding variants in the training and test splits. The feature set comprised 19 variables: positional information (POS), nucleotide lengths (reference and alternate allele lengths and their difference), allele frequencies (AF_ESP, AF_EXAC, AF_TGP and their log-transformed values with missing indicators), ClinVar variant category flags (single-nucleotide, deletion, insertion, duplication and indel), and the ORIGIN code.

B. Nested Cross-Validation Results

Nested five-fold cross-validation identified a strong baseline classifier. Table VII summarizes the mean F1 scores ($\pm$ standard deviation) obtained on the outer folds for each algorithm. Random Forest achieved the highest mean F1 (0.9414 ± 0.0033) followed by LightGBM (0.9299 ± 0.0034) and XGBoost (0.9262 ± 0.0067); Extra Trees and Gradient Boosting had lower performance. Random Forest was selected as the best baseline model.

TABLE VII
NESTED CROSS-VALIDATION RESULTS

Model	Mean F1 ($\pm$ SD)
Random Forest	0.9414 ± 0.0033
LightGBM	0.9299 ± 0.0034
XGBoost	0.9262 ± 0.0067
Extra Trees	0.9027 ± 0.0050
Gradient Boosting	0.9016 ± 0.0042

C. Performance of the Best Single Model

The Random Forest was trained on 80% of the data and evaluated on the held-out test set. The optimal classification threshold, determined on the validation set via multi-objective optimization, was 0.309. The model achieved accuracy of 0.8754, F1 score of 0.9192, sensitivity of 0.9156 and specificity of 0.7380. The Matthews correlation coefficient (MCC) was 0.6475; the ROC-AUC was 0.8909 and PR-AUC was 0.9522. The Brier score was 0.1201. Bootstrap confidence intervals (Table VIII) fall within narrow ranges, demonstrating robustness. Figure 3 plots the ROC and precision–recall curves. The confusion matrix (Figure 4) shows 1,171 true positives, 276 true negatives, 98 false positives and 108 false negatives.

TABLE VIII
RANDOM FOREST PERFORMANCE ON TEST SET

Metric	Estimate	95% CI
Accuracy	0.8754	[0.8590, 0.8911]
F1 score	0.9192	[0.9075, 0.9300]
Sensitivity	0.9156	[0.8982, 0.9315]
Specificity	0.7380	[0.6912, 0.7821]
MCC	0.6475	[0.6023, 0.6904]
ROC-AUC	0.8909	[0.8703, 0.9093]
PR-AUC	0.9522	[0.9365, 0.9653]
Brier score	0.1201	[0.1089, 0.1321]

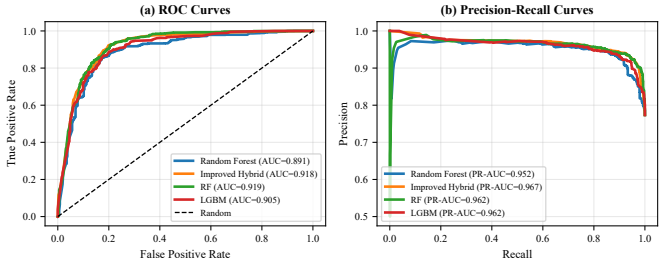


Fig. 3. ROC curves (left) and precision–recall curves (right) comparing the Random Forest baseline and the hybrid ensemble on the test set.

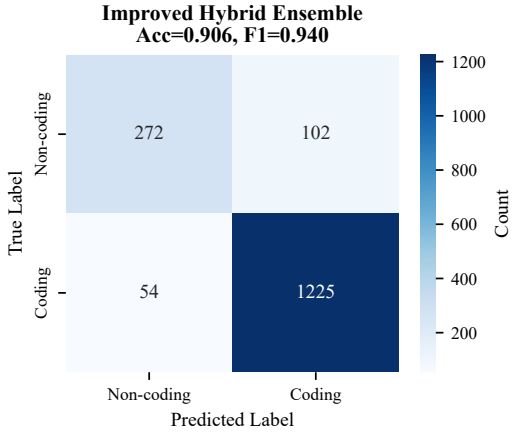


Fig. 4. Confusion matrices for the Random Forest (left) and hybrid ensemble (right) on the test set, showing improved reduction in false negatives for the ensemble.

D. Hybrid Ensemble Selection and Validation

An improved ensemble was built from six base learners: an MLP, Random Forest (RF), Extra Trees (ET), XGBoost (XGB), Gradient Boosting (GB) and LightGBM (LGBM). Out-of-fold predictions were generated for each base model. Validation accuracies ranged from 0.801 for the MLP to 0.896 for the Random Forest. Pairwise disagreement rates quantified diversity, ranging between 0.091 and 0.139. Models were selected when their validation accuracy exceeded 0.82 and a composite score (accuracy plus diversity bonus) ranked among the top performers. Four models–RF, LGBM, XGB and GB–met these criteria with weights of 0.257 (RF), 0.253 (LGBM), 0.250 (XGB) and 0.240 (GB).

Meta-features were created by combining base prediction probabilities with statistical summaries (mean, standard deviation, maximum and minimum) and pairwise disagreements, resulting in 14 features. A logistic regression meta-learner with L2 regularization was trained on out-of-fold predictions generated from the development set. Multi-objective optimization (weights: accuracy 0.5, F1 0.4, MCC 0.1) yielded an optimal threshold of 0.526. Isotonic regression calibration slightly improved validation accuracy (0.9047 versus 0.9039).

E. Test-Set Performance of the Improved Hybrid Ensemble

On the held-out test set, the calibrated hybrid ensemble achieved accuracy of 0.9056 and F1 score of 0.9401, outperforming the best single model by +3.45 percentage points (+3.94% relative improvement) and +2.28 percentage points (+2.48% relative improvement), respectively. Sensitivity improved to 0.9578 while specificity was 0.7273. The MCC increased to 0.7204. The ROC-AUC rose to 0.9180 and PR-AUC to 0.9667. The Brier score decreased from 0.1201 to 0.0783 (a 34.8% reduction). Confidence intervals (Table IX) confirm accuracy between 0.8905 and 0.9189, F1 between 0.9302 and 0.9489, ROC-AUC between 0.8995 and 0.9342 and MCC between 0.6769 and 0.7593. Figure 3 shows higher areas under both curves for the ensemble. The confusion matrix (Figure 4) displays 1,225 true positives, 272 true negatives, 102 false positives and 54 false negatives, illustrating substantial reduction in false negatives.

TABLE IX
PERFORMANCE COMPARISON: RANDOM FOREST VS. HYBRID ENSEMBLE

Metric	Random Forest	Hybrid Ensemble	Improvement
Accuracy	0.8754	**0.9056**	+3.45 pp (+3.94% rel.)
F1 score	0.9192	**0.9401**	+2.28 pp (+2.48% rel.)
ROC-AUC	0.8909	**0.9180**	+3.04 pp (+3.04% rel.)
PR-AUC	0.9522	**0.9667**	+1.52 pp (+1.52% rel.)
MCC	0.6475	**0.7204**	+11.26 pp (+11.26% rel.)
Brier score	0.1201	**0.0783**	−34.8%

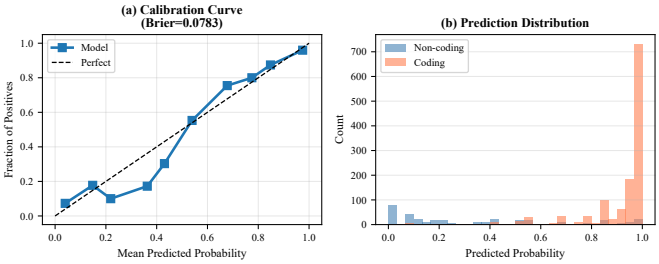


Fig. 5. Calibration curves (top) and prediction probability distributions (bottom) for the Random Forest and hybrid ensemble, demonstrating improved probability calibration for the ensemble.

F. Calibration and Confidence Intervals

Calibration analysis confirmed that the ensemble's predicted probabilities aligned well with observed outcomes. Isotonic calibration improved the Brier score, and the calibration curve in Figure 5 shows predicted probabilities follow the diagonal more closely for the ensemble. Table X lists bootstrap confidence intervals (1,000 samples); all intervals are narrow, illustrating consistent performance.

TABLE X
HYBRID ENSEMBLE PERFORMANCE WITH CONFIDENCE INTERVALS

Metric	Estimate	95% CI
Accuracy	**0.9056**	[0.8905, 0.9189]
F1 score	**0.9401**	[0.9302, 0.9489]
Sensitivity	**0.9578**	[0.9441, 0.9698]
Specificity	0.7273	[0.6802, 0.7718]
MCC	**0.7204**	[0.6769, 0.7593]
ROC-AUC	**0.9180**	[0.8995, 0.9342]
PR-AUC	**0.9667**	[0.9521, 0.9787]
Brier score	**0.0783**	[0.0701, 0.0872]

G. Feature Importance Analysis

Feature importance scores from the Random Forest model were analyzed to understand which features contribute most to classification. Figure 6 displays the top features ranked by mean decrease in impurity. Genomic position (POS), allele length difference, and population allele frequencies (particularly AF_EXAC) emerged as the most discriminative features for distinguishing coding from non-coding variants. Variant type indicators (SNV, deletion, insertion) also contributed significantly.

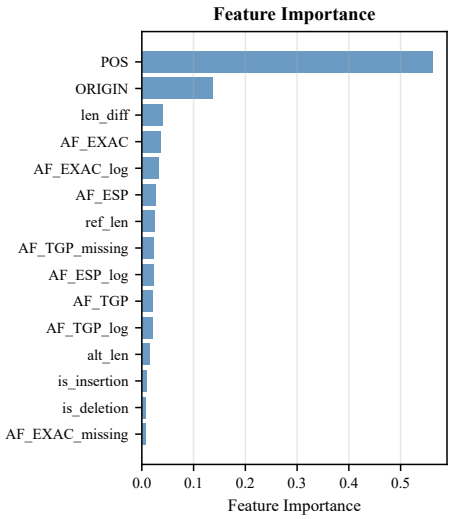


Fig. 6. Feature importance scores from the Random Forest model, showing the relative contribution of each feature to the classification of coding vs. non-coding variants.

H. Summary of Findings

Table IX and Figures 3–6 show that the improved hybrid ensemble surpasses the best single model on all evaluated metrics. The ensemble achieved higher accuracy, F1 score, ROC-AUC and PR-AUC while substantially reducing the Brier score. The final model correctly identified 1,225 coding variants and 272 non-coding variants, misclassifying 156 instances (102 false positives and 54 false negatives). These results demonstrate that combining diverse base learners and engineering meta-features yields better predictive accuracy and calibration for coding versus non-coding variant classification.

V. Conclusion

We addressed distinguishing coding from non-coding somatic variants within the ClinVar VCF using a classification pipeline that parsed the VCF, applied Sequence Ontology–based classification, built features from allele frequencies, variant lengths, and type indicators, trained Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM, and a deep MLP using BorderlineSMOTE to handle class imbalance, and nested cross-validation identified Random Forest as the strongest model.

We assembled a hybrid ensemble by selecting base models and training a logistic regression meta-classifier with calibration. The ensemble achieved accuracy ≈ 0.9047, F1 ≈ 0.9393, and on the held-out test set improved accuracy by 3.45 percentage points, ROC-AUC by 3.04 percentage points, and reduced Brier score by 34.8% compared with the best single model, showing stacking and calibration can enhance variant classification.

Only 2% of the human genome encodes proteins, whereas 90% of disease-associated variants lie in non-coding regions; accurate classification of coding versus non-coding somatic variants therefore has profound implications for genomic medicine and clinical diagnostics. Our ensemble yields gains over individual learners, provides well-calibrated outputs and interpretable feature contributions, and suits integration into variant-interpretation pipelines and precision-medicine workflows.

Future research should prioritize external validation on independent ClinVar releases and alternative somatic variant databases (e.g., COSMIC, TCGA) to establish cross-dataset generalizability. Additionally, incorporating deeper genomic annotations (e.g., chromatin accessibility, transcription factor binding, epigenetic marks), exploring deep representation-learning methods such as transformer-based architectures, and evaluating performance on germline variants and cross-species datasets would strengthen the approach. Integrating explainability techniques (e.g., SHAP values, attention mechanisms) could provide variant-level interpretations and foster clinical adoption.

References

[1] Y. Wang and G. C. Hon, "Towards functional maps of non-coding variants in cancer," Front. Genome Ed., vol. 6, p. 1481443, 2024.
[2] M. J. Landrum et al., "ClinVar: public archive of relationships among sequence variation and human phenotype," Nucleic Acids Res., vol. 42, no. Database issue, pp. D980-5, 2014.
[3] M. F. Rogers, H. A. Shihab, T. R. Gaunt, and C. Campbell, "CScape: a tool for predicting oncogenic single-point mutations in the cancer genome," Sci. Rep., vol. 7, no. 1, p. 11597, 2017.
[4] N. Feizi, Q. Liu, L. Murphy, and P. Hu, "Computational prediction of the pathogenic status of cancer-specific somatic variants," Front. Genet., vol. 12, p. 805656, 2021.
[5] M. J. Landrum et al., "ClinVar: updates to support classifications of both germline and somatic variants," Nucleic Acids Res., vol. 53, no. D1, pp. D1313–D1321, 2025.
[6] Z. Sondka et al., "COSMIC: a curated database of somatic variants and clinical data for cancer," Nucleic Acids Res., vol. 52, no. D1, pp. D1210–D1217, 2024.
[7] R. Yin, A. Gutierrez, Undiagnosed Diseases Network, S. N. Kobren, and P. Avillach, "VarPPUD: Variant post prioritization developed for undiagnosed genetic disorders," medRxiv, 2024.
[8] J. Cheng et al., "Accurate proteome-wide missense variant effect prediction with AlphaMissense," Science, vol. 381, no. 6664, p. eadg7492, 2023.
[9] X. Wang et al., "Deep learning approaches for non-coding genetic variant effect prediction: current progress and future prospects," Brief. Bioinform., vol. 25, no. 5, 2024.
[10] J. Y. Heo and J. H. Kim, "Assessing the performance of 28 pathogenicity prediction methods on rare single nucleotide variants in coding regions," BMC Genomics, vol. 26, no. 1, p. 641, 2025.
[11] R. Kurosawa et al., "PDIVAS: Pathogenicity predictor for Deep-Intronic Variants causing Aberrant Splicing," BMC Genomics, vol. 24, no. 1, p. 601, 2023.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

EpidermNet: A Task-Specific Deep Learning System for Scabies and Eczema Detection

Sneha Purkaystha, Debjani Deb Tanwi, Prithwiraj Bhattacharjee
Department of CSE, Leading University, Sylhet, Bangladesh
Email: cse_182210012101158@lus.ac.bd, cse_182210012101141@lus.ac.bd, prithwiraj_cse@lus.ac.bd

***Abstract*—Many skin diseases have increased at an alarming rate worldwide. Among them, scabies and eczema are now major health concerns. Some reports suggest scabies became epidemic recent days. These diseases often appear similar, with rashes, redness, and itching, making diagnosis challenging and potentially delaying treatment. Currently, there is no dedicated binary detection system for these two diseases. Existing multi-class classifiers are heavyweight and may confuse scabies and eczema with other less common conditions. To address this, we developed a deep learning system to classify scabies and eczema from images. A dataset of 500 images (250 per disease) was collected from 14 clinically verified sources. After preprocessing and augmenting the data to 3000 images, four state-of-the-art (SOTA) and lightweight deep learning architectures have been implemented for the first time, which include MobileNetV3-Small, MobileNetV2, ResNet18, and SqueezeNet 1.1. The evaluation of these models illustrates significant performance, having around 99% accuracy. More specifically, MobileNetV3-Small and MobileNetV2 outperformed all existing SOTA models with 99.78% and 99.56% accuracy, respectively. This demonstrates that these two can effectively distinguish between scabies and eczema. Our study highlights the potential of deep learning to support dermatological diagnosis and improve patient care. The dataset, codebase, and application will be publicly available upon publication to encourage future research in this direction.**

***Index Terms*—Scabies, Eczema, Deep Learning, MobileNetV3-Small, MobileNetV2, Classification, Computer Vision**

I. Introduction

The pattern of skin diseases has changed dramatically all over the world. In recent days, the effect of scabies and eczema has increased. A study was conducted where 257,536 outpatients were examined between January 2017 and April 2024, and among them, 9,407 were diagnosed with scabies. The yearly data showed that cases started increasing in 2019, stayed stable from 2020 to 2022, and then rose again during 2023 and 2024 [1]. There is a problem that people, including doctors, often get confused while separating these two diseases due to their similarity in nature. In China a study reported that 23 patients were misdiagnosed with eczema instead of scabies, which led to inappropriate treatments and delayed recovery [2]. Eczema can be activated by family history, stress, and environmental factors, while scabies is caused by the microscopic mite (Sarcoptes Scabiei var. hominis). Fabrics, metals, pet dander, pollen, and dust mites are also responsible for making this eczema worse, and scabies often accompanies bacterial infections like impetigo. Based on the information of the National Eczema Association[1], eczema has common symptoms including itching, dryness, rashes, inflammation, discoloration, and scaly patches, whereas scabies causes symptoms including itching, papules, burrows, redness, and crusts, with severe cases showing thick scaly plaques. Eczema affects individuals of all ages, with over 31 million Americans suffering from some form of this disease [3]. On the other hand, the WHO[2] stated that over 200 million individuals are affected by scabies at any time, with more than 400 million cases annually. Now, the concern is that the traditional diagnosis systems depend on a physician's skills and lab tests, which are time-consuming and error-prone. Therefore, the need for a dedicated study of classifying them using ML or DL has been realized after observing these situations. While studying the domain, it has been observed that several works have been carried out on other skin diseases alongside eczema and scabies. Existing studies applied multiclass classification for multiple skin diseases, including scabies and eczema. For instance, a study applied AI to detect six skin diseases: dermatofibroma, chickenpox, eczema, keratosis, psoriasis, and scabies [4]. Another study included seven diseases: acne, melanoma, solar lentigo, dermatofibroma, eczema, psoriasis and scabies in one multiclass framework [5]. However, distinguishing scabies and eczema alone in such multiclass setups remains challenging and often fails due to lower evaluation accuracy. Hence, to provide a more detailed analysis of these two conditions, a focused binary classification study of eczema and scabies is needed. The main purpose of this study was to enhance diagnostic accuracy, minimize misdiagnosis, and support healthcare physicians in making timely decisions. To address this problem, firstly, this study has developed a novel dataset from 14 verified resources, consisting of 500 images for both classes. It has become the largest dataset of its kind. After that, several lightweight SOTA deep learning architectures, such as MobileNetV3-Small, MobileNetV2, ResNet18, and SqueezeNet 1.1, have been implemented because of their adaptability to this domain. The evaluation of these models highlights the high detection ability. More specifically, MobileNetV3-Small and MobileNetV2 outperform all the existing SOTA models. Thus, the aim of this research has been achieved by implementing a highly accurate and dedicated lightweight binary detection system for these two common diseases. So, the novelties of this study are stated as follows.

[1] https://nationaleczema.org/eczema/

[2] https://www.who.int/health-topics/scabies

- Creation of a dataset from 14 clinically verified resources with 500 images (250 scabies, 250 eczema), which became the largest dataset to date.
- The proposed lightweight architectures, MobileNetV3-Small and MobileNetV2, outperform all the existing cutting-edge systems with over 99.56% accuracy.

II. Literature Review

Both scabies and eczema are very common, and researchers started exploring skin disease detection many years ago. Gilmore [6] introduced control strategies for endemic childhood scabies. Using a network-based Monte Carlo model, the study highlighted that scabies affects approximately 300 million people annually in low-resource regions. In 2013, Brenaut et al. [7] investigated itch characteristics across five dermatoses, including non-atopic and scabies. They found that itch patterns varied among diseases. In 2016, Kouotou et al. [8] found that out of 1,902 students, 338 (17.8%) were affected by scabies at a Cameroonian boarding school. During 2020, Thomsen et al. [9] proposed a deep learning model for binary classification that achieved high diagnostic performance, distinguishing acne from rosacea and psoriasis from eczema. In the same year, Salsabila et al. [10] proposed a Random Forest method for scabies classification in animals using uniform local binary patterns. Nisar et al. [11] also proposed an automatic method for eczema lesion segmentation using different supervised learning algorithms, and SVM achieved the highest accuracy. In the next year, Nimeesh et al. [12] proposed a method for differential diagnosis between ringworm and eczema using a Convolutional Neural Network (CNN), which achieved an accuracy of 82%. Another research work by Rasheed et al. [13] was conducted for classifying eczema into seven different classes using a hybrid deep neural network, where the hybrid-6 network achieved the highest accuracy of 88.29%. In 2022, Halder et al. [14] proposed an automated method for scabies detection using image processing and CNN. Juwairi et al. [15] first proposed an EfficientNet-based model for classifying eczema and acne, which was optimized for mobile deployment using TFLite. Again in 2023, Zendah et al. [16] developed a deep learning model using VGG-16 for the early diagnosis of scabies using dermoscopic images. Andryani et al. [4] proposed both traditional and transfer learning methods for the automatic identification of multiple skin diseases, including scabies and eczema, in 2023. Hammed et al. [17] presented an enhanced deep learning approach, named "Derma Care," for the accurate detection of eczema and psoriasis skin conditions. In the same year, Shodoy et al. [5] also proposed a deep learning method for detecting seven types of skin diseases, including scabies and eczema, using an ensemble model composed of ResNet-50, EfficientNet, Inception V3, MobileNet, NASNetMobile, DenseNet201, and Xception architectures. Among these, NASNetMobile achieved the best performance. Attar et al. [18] proposed the EczemaNet2 method for reliable detection of eczema areas to enable fully automated assessment of eczema severity from digital camera images. In the year 2024, Emanghe et al. [19] conducted a cross-sectional study in Calabar, Cross River State, Nigeria, where the authors found that only 56.7% and 67.9% of parents of school-going children had heard of scabies and ringworm, respectively. However, the knowledge of scabies was lower, with only 53.3% knowing its cause and 50.5% aware of its risk factors. Wu et al. [2] conducted a descriptive eczema study in China on 23 patients and reported that scabies is often misdiagnosed, as typical symptoms included erythema, papules, nodules, and nocturnal itching, with only a few cases confirmed by microscopy or histopathology. In 2024, Sakib et al. [20] developed a CNN and linear regression method for skin disease classification, focusing on eczema and seborrheic keratoses, with the CNN having higher accuracy. Badr et al. [21] also proposed a five-class Xception architecture for diagnosing multi-class skin diseases and achieved 95% accuracy. Rangaswamy et al. [22] developed the Inception V3 and VGG16 architectures, which were trained on a large dataset comprising 13 skin disease classes. After that, Jaiyeoba et al. [23] developed a stacking ensemble machine learning model for the classification of erythemato-squamous diseases (ESD), achieving a higher accuracy of 99.30%. After reviewing these studies, it has been observed that the symptoms of eczema and scabies are very similar in many cases, which leads to misdiagnosis. Although all the reviewed studies provided valuable insights, three studies in particular greatly strengthened the motivation for this research. The first one is Wu et al. [2], which reported that scabies is often misdiagnosed as eczema. The other two, Andryani et al. [4] and Shodoy et al. [5], proposed multi-class classification methods for identifying multiple skin diseases, including scabies and eczema. However, it is difficult to separately emphasise only these two diseases within a broad multi-class framework. Therefore, these studies motivated us to develop a binary classification model dedicated to distinguishing eczema from scabies.

III. Dataset

The main challenge of this study was to collect a good amount of data. Only a few images were available at the DermNet NZ repository, in which most of the work had been done earlier. So, to create a dataset of sufficient size for training purposes, a total of 500 labeled images were collected from 14 different clinically verified public sources through a rigorous search. Each source is denoted as **Source (#Scabies, #Eczema)**, where the # numbers indicate the number of images per class (0 if the class is absent). The sources are: DermNetNZ (113, 241) [24], Dermatology Atlas (86, 0) [25], DermIS (17, 4) [26], Medscape (9, 0) [27], Wayne State University SOM Dermatology Image Atlas (6, 0) [28], WebMD (4, 0) [29], Dandermdk (4, 0) [30], National Health Service (4, 0) [31], Medicine Net (3, 2) [32], PCDS (3, 0) [33], Manhattan Dermatology (0, 1) [34], Metro Dermatology (0, 1) [35], NCBI (0, 1) [36], and WHO (1, 0) [37]. The most problematic situation arose while collecting scabies data. Sources have very little scabies data compared to eczema. However, this process ensures diversity by collecting data

from different regions, and an equal number of instances (250) per disease makes it more reliable for prediction.

A. Dataset Preparation & Augmentation

As the plan of this study was to implement some ImageNet-based deep learning architectures, the input shape of the images had to be 224x224. The proposed data have been collected from several sources with different shapes. So, all the images were resized to a standard 224x224 for a unified formatting and fixed resolution. Then regions such as logos and some irrelevant objects are found in certain images. To erase those parts, those images have been cropped and resized back to the standard size of 224X224. In the next step, OpenCV's default BGR(Blue-Green-Red) color format has been converted to the standard RGB (Red-Blue-Green) format. Images have been converted to Tensor values for feeding the models. After that, the original 500 data have been expanded to 3000 images with 1500 samples per class using an augmentation process, which contains rotation, translation, scaling, and brightness adjustment. The cropped images of the preprocessing phase were rotated by 15 degrees around their center, and translation was performed by shifting the images 20 pixels to the left and right along the x-axis. Scaling was implemented by applying a 20% zoom-in. This simulates closer views of lesions. Moreover, brightness enhancement has been applied to replicate varying illumination.

B. Dataset Comparison

To compare our dataset with existing ones, a rigorous search was conducted to identify publicly available datasets. For these two diseases, there were very few datasets available publicly. Andryani et al. [4] released a dataset having 226 images for scabies and 224 images for eczema. Later, Shodoy et al. [5] also released a dataset comprising 40 and 90 images for scabies and eczema, respectively. However, both these datasets are parts of multiple disease classification projects and not for a dedicated binary classification. A detailed comparison of our data with these two datasets is shown in Table 1. It is clearly visible that our data becomes the largest of its kind to data with 250 images for each class.

Table I: Comparison of Existing Datasets and Proposed Data

Source	#Scabies Image	#Eczema Image	Total
Andryani et al. [4]	226	224	450
Shodoy et al. [5]	40	90	130
Proposed Data	**250**	**250**	**500**

IV. Methodology

The main purpose of this study is to predict whether the skin disease indicates scabies or eczema. The dataset feeding process has been done following [5] and [14]. These studies first augmented the data and then fed it to the model. Now, the dataset has been split into three portions for this research. 70% data has been employed for training (2100 images), 15% data has been used for validation (450 images), and the remaining 15% (450 images) is used for the test set. The images are then normalized using the mean and standard deviation values of the ImageNet dataset. Several heavyweight deep learning architectures have been used earlier for the skin disease classification problem, such as NasNetMobile, MobileNet, DenseNet201, Xception, ResNet50, etc [5]. Motivated by these studies, four lightweight deep learning architectures have been proposed for this research. Those are MobileNetV3-Small, ResNet18, MobileNetV2, and SqueezeNet 1.1, and different hyperparameters are tuned for these models to get the best outcome. Figure 1 illustrates the whole methodological process of this study.

A. MobileNetV3-Small Architecture

MobileNetV3-Small is a variant of MobileNetV3. It is a lightweight convolutional neural network designed for efficient image classification. By using pretrained weights of transfer learning, the model can learn general features of eczema and scabies lesions effectively. So, the MobileNetV3-small architecture perfectly suits these settings. First of all, the inbuilt fully connected layer has been changed to a dense layer of two neurons for two classes (eczema and scabies). With a batch size of 32, the model has been run for 50 epochs with an early stopping patience of 5. For stable convergence, Adam has been used with a learning rate of 0.001. To adjust the learning rate, ReduceLROnPlateau was used. CrossEntropyLoss has also been used as a loss function that calculates negative log likelihood for the correct class and correctly classified eczema or scabies images.

B. ResNet18 Architecture

ResNet18 has residual connectivity for enhancing training efficiency in image classification, and for this reason, this study used it as a proposed system. By utilizing pretrained weights through transfer learning, the model can capture the general pattern of eczema and scabies lesions. The model was trained with a batch size of 32 for up to 40 epochs with an early stopping patience of 5, and it is sufficient for convergence to prevent overfitting and reduce the risk of abrupt ending. Adam, as an optimizer, has been used for its faster convergence, and an adaptive learning rate has been set to 0.001. To dynamically adjust the learning rate during training, the ReduceLROnPlateau was employed. Hyperparameters were tuned carefully so that Adam could improve the trade-off between training speed and final performance. CrossEntropyLoss has been used as a loss function because it implicitly converts the output logits into probability using softmax. Lastly, it computes negative log-likelihood and enables precise classification of eczema and scabies images.

C. MobileNetV2 Architecture

MobileNetV2 architecture was used not only for its computational efficiency and high accuracy but also for its transition from low-level feature extraction to more abstract disease-specific features, such as skin lesion pattern and texture, which are essential for distinguishing the diseases. The original 1000-class ImageNet classifier was replaced with a

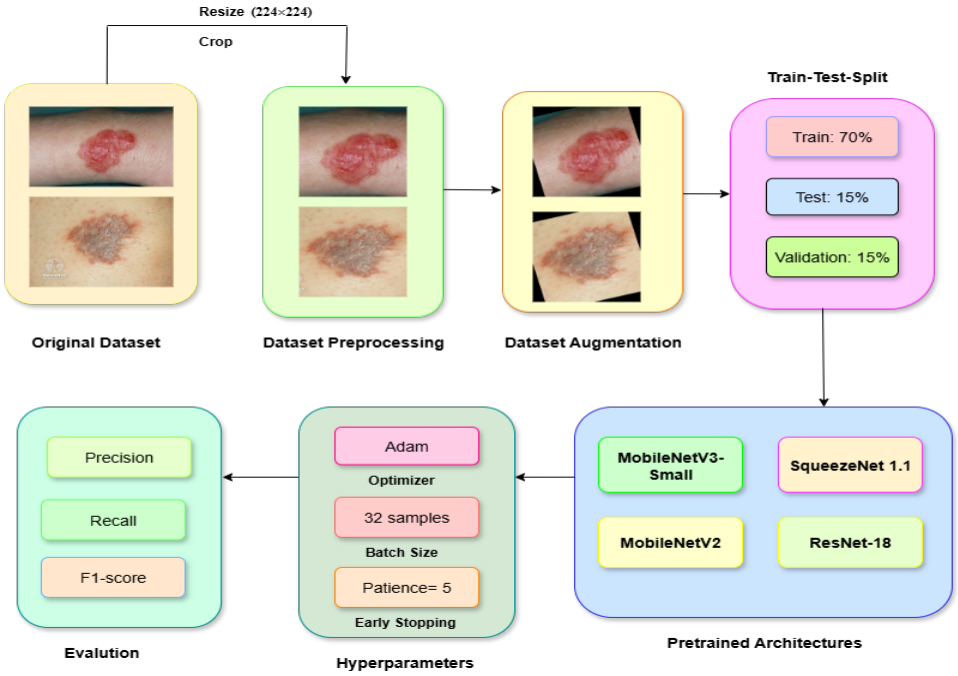


Figure 1: Illustration of the Methodological Procedure

fully connected layer of two neurons for two classes. Like the previous models, batch size has been set to 32, while 50 epochs were selected with an early stopping patience of 5. These settings are well-suited to prevent overfitting. The Adam optimizer with an initial learning rate of 0.001 was used, and ReduceLROnPlateau was also used to adjust the learning rate. Lastly, the CrossEntropyLoss function has been used again as a loss function.

D. SqueezeNet 1.1 Architecture

SqueezeNet 1.1 is a compact convolutional neural network designed for efficient image classification. It has fewer parameters and requires less computational cost. The model was employed in our paper for capturing important features of scabies and eczema. The feature extractor was fully unfrozen, allowing all convolutional layers to be trainable for fine-tuning on our dataset. The model was trained with a batch size of 32 for up to 50 epochs with an early stopping patience of 5, and it is sufficient for convergence to prevent overfitting and reduce the risk of abrupt ending. Adam, as an optimizer, has been used for its faster convergence, and the learning rate has been set to 0.0001. Then, to dynamically adjust the learning rate during training, the ReduceLROnPlateau was used. Hyperparameters were tuned carefully so that Adam could improve the trade-off between training speed and final performance. CrossEntropyLoss has been used as a loss function again.

V. Result Analysis

The primary goal of this study was to predict the skin disease between scabies and eczema. The graphical representation of the training and validation accuracy as well as the training and validation loss of the best two models among our proposed models, is illustrated in the following figures.

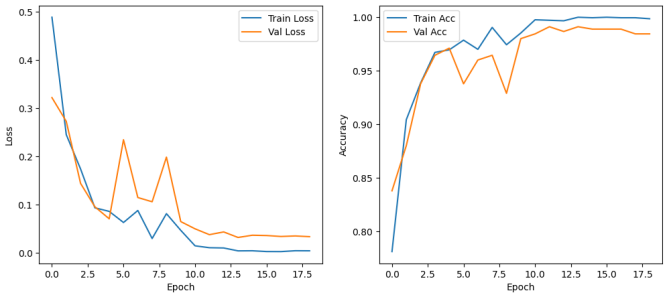


Figure 2: Training and Validation Curves of MobileNetV2.

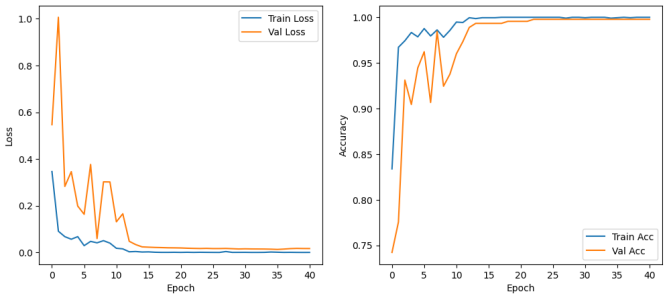


Figure 3: Training and Validation Curves of MobileNetV3-Small.

As MobileNetV3(s) has achieved the best performance among all models, its confusion matrix is given for evaluation.

It helps identify the correct and incorrect classifications. Here, 0 represents Eczema and 1 represents Scabies.

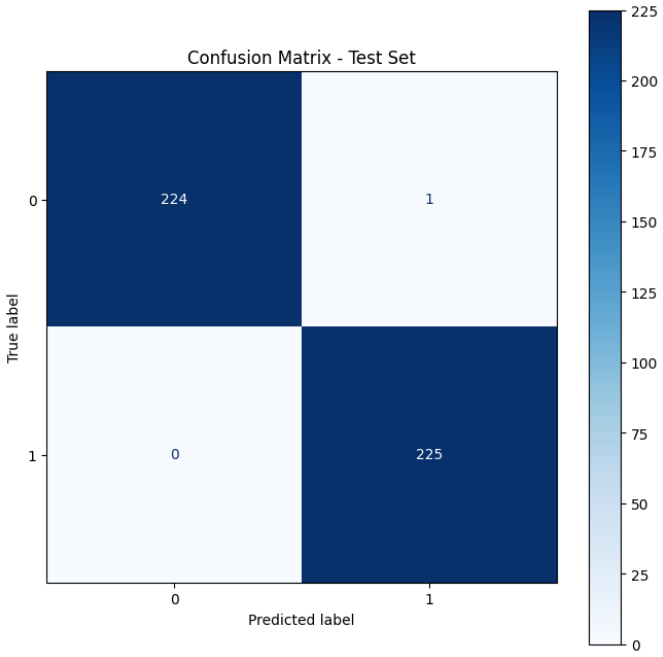


Figure 4: Confusion Matrix of MobileNetV3-Small.

To understand the performance of the models more efficiently, the precision, recall, and f1-score of all models have also been calculated from the confusion matrices.

Table II: Precision, Recall, and F1-score per Class for Proposed Models

Models	Class	Precision	Recall	F1-score
MobileNetV3(s)	0	100%	99.56%	99.78%
	1	100%	100%	100%
MobileNetV2	0	100%	99.11%	99.55%
	1	99.12%	100%	99.56%
ResNet18	0	99.54%	96.89%	98.15%
	1	96.97%	99.56%	98.17%
SqueezeNet 1.1	0	91.74%	93.78%	92.70%
	1	93.64%	91.56%	92.52%

The performance of the four deep learning models has been evaluated to understand their effectiveness in classifying scabies and eczema. MobileNetV3-Small outperformed the others by achieving the highest test accuracy of 99.78%. It indicates its strong ability to distinguish between these two visually similar skin diseases. MobileNetV2 also performed well by reaching 99.56% accuracy, slightly lower than MobileNetV3-Small. However, it still shows good classification capability. ResNet18 achieved an accuracy of 98%, and SqueezeNet 1.1 achieved 93% accuracy. Here, SqueezeNet 1.1 is the lowest among all our performed models. It suggests that it is less suitable for this specific binary classification task. The comparison highlights that focusing on a model tuned for these two diseases, like MobileNetV3-Small. It can significantly improve classification performance compared to other general-purpose architectures. This evaluation provides valuable insights for selecting the most appropriate model for automated dermatological diagnosis. It demonstrates that model choice has a critical impact on accuracy and reliability in real-world applications.

Table III: Accuracy Comparison of Existing and Proposed Models

Source	Models	Metrics
Shodoy et al. [5]	NasNetMobile (heavy_weight)	Validation = 99% Test = Not Available
	MobileNet	Validation = 96% Test = Not Available
	ResNet50	Validation = 92% Test = Not Available
	EfficientNetB0	Validation = 61% Test = Not Available
	Densenet201	Validation = 87% Test = Not Available
Andryani et al. [4]	InceptionResNet-V2	Validation = 77% Test = Not Available
	ResNet50-V2	Validation = 68% Test = Not Available
	DenseNet201	Validation = 74% Test = Not Available
	InceptionNetV3	Validation = 71% Test = Not Available
	SVM GLCM	Validation = 75% Test = Not Available
Proposed Models	MobileNetV3(s) (light_weight)	Validation = **99.78%** Test = **99.78%**
	MobileNetV2 (light_weight)	Validation = **99.11%** Test = **99.56%**
	ResNet18	Validation = **99%** Test = **98%**
	SqueezeNet1.1 (light_weight)	Validation = **93%** Test = **93%**

Moreover, past studies on skin disease detection using deep learning architectures have shown varying levels of success across different types of diseases, and the models were mostly heavyweights. Table III shows the comparison of existing multiclass models and our proposed binary classification models for skin disease classification. As there are no binary classifiers that exist, the comparison had to be done with these existing ones. Although these existing studies used different datasets, this comparison suggests that our proposed lightweight models achieved competitive performance compared to heavyweight architectures of these studies. Among them, Shodoy et al. achieved the highest validation accuracy of 99% using NasNetMobile, while other models such as MobileNet, ResNet50, EfficientNetB0, and DenseNet201 obtained lower accuracy. The main fact is that our proposed models, particularly MobileNetV3-Small and MobileNetV2, are more lightweight than NasNetMobile and achieve higher accuracy than it as well. Furthermore, Andryani et al.'s models reached validation accuracies between 68% and 77%. The proposed models, including MobileNetV3-Small, MobileNetV2, ResNet18, and SqueezeNet 1.1, consistently achieved higher

performance. MobileNetV3-Small obtained 99.78% validation and test accuracy and outperforms all existing models. MobileNetV2 and ResNet18 also showed strong results, with test accuracies of 99.56% and 98%, respectively. SqueezeNet 1.1 performed the weakest among all the models with 93% test accuracy. Our proposed models gave the best performance due to their lightweight architectures, careful training with hyperparameters and stable convergence. Overall, the results demonstrate that our proposed system provides better accuracy and generalization compared to previous approaches.

VI. Conclusion

This study developed a deep learning system to classify eczema and scabies, two skin diseases that have become more common and continue to pose diagnostic challenges. A dataset of 500 images from trusted medical sources has been collected, and augmentation has been used to improve training. Four models were tested, and MobileNetV3-Small achieved the best result with 99.78% accuracy, followed by MobileNetV2. This shows that focusing on only these two diseases gives better results than general multiclass systems. This work can help doctors and telemedicine platforms by reducing misdiagnosis and improving patient care. A more reliable dataset and multimodal learning can open the direction of future research endeavors.

References

[1] E. Ayhan, "Status of the scabies outbreak before, during and after the covid-19 pandemic: a single-center hospital-based retrospective evaluation," *BMC Infectious Diseases*, vol. 25, no. 1, p. 481, 2025.

[2] X. Wu, F. Yang, and R. Zhang, "Frequent misdiagnosis of scabies as eczema in china: A descriptive study of 23 cases," *International Journal of General Medicine*, pp. 1615–1623, 2024.

[3] J. M. Hanifin, M. L. Reed, E. Prevalence, and I. W. Group, "A population-based survey of eczema prevalence in the united states," *DERM*, vol. 18, no. 2, pp. 82–91, 2007.

[4] N. A. C. Andryani, F. H. Juwono, and P. J. Sianturi, "Artificial intelligence approach for automatic multiclass skin diseases identification," *Journal of Engineering Science and Technology*, vol. 18, pp. 1–14, 2023.

[5] M. U. R. Shodoy, "Detection of skin diseases using deep learning," Ph.D. dissertation, 2023.

[6] S. J. Gilmore, "Control strategies for endemic childhood scabies," *PloS one*, vol. 6, no. 1, p. e15990, 2011.

[7] E. Brenaut, R. Garlantezec, K. Talour, and L. Misery, "Itch characteristics in five dermatoses: non-atopic eczema, atopic dermatitis, urticaria, psoriasis and scabies." *Acta Dermato-Venereologica*, vol. 93, no. 5, pp. 573–4, 2013.

[8] E. A. Kouotou, J. R. N. Nansseu, M. K. Kouawa, and A.-C. Zoung-Kanyi Bissek, "Prevalence and drivers of human scabies among children and adolescents living and studying in cameroonian boarding schools," *Parasites & vectors*, vol. 9, no. 1, p. 400, 2016.

[9] K. Thomsen, A. L. Christensen, L. Iversen, H. B. Lomholt, and O. Winther, "Deep learning for diagnostic binary classification of multiple-lesion skin diseases," *Frontiers in medicine*, vol. 7, p. 574329, 2020.

[10] A. S. Salsabila, F. Sthevanie, and K. N. Ramadhani, "Scabies classification in animal using uniform local binary patterns," in *2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, 2020, pp. 356–361.

[11] H. Nisar, Y. K. Ch'ng, and Y. K. Ho, "Automatic segmentation and classification of eczema skin lesions using supervised learning," in *2020 IEEE Conference on Open Systems (ICOS)*. IEEE, 2020, pp. 25–30.

[12] V. Nimesh and R. Weerasinghe, "Differential diagnosis of ringworm and eczema using image processing and deep learning," in *2021 21st International Conference on Advances in ICT for Emerging Regions (ICter)*. IEEE, 2021, pp. 147–152.

[13] A. Rasheed, A. I. Umar, S. H. Shirazi, Z. Khan, S. Nawaz, and M. Shahzad, "Automatic eczema classification in clinical images based on hybrid deep neural network," *Computers in Biology and Medicine*, vol. 147, p. 105807, 2022.

[14] M. Halder, H. M. E. Kabir, A. M. Rity, and A. Vowmik, "An automated detection of scabies skin disease using image processing and cnn," in *2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI)*. IEEE, 2022, pp. 1–6.

[15] K. P. Juwairi, D. H. Fudholi, A. Arifin, and I. Muhimmah, "An efficientnet-based mobile model for classifying eczema and acne," in *AIP conference proceedings*, vol. 2508, no. 1. AIP Publishing LLC, 2023, p. 020041.

[16] H. Zendah and K. Dimililer, "Using of conventional neural network to diagnose scabies by dermoscopy," in *International Conference on Theory and Applications of Fuzzy Systems and Soft Computing*. Springer, 2022, pp. 274–281.

[17] M. Hammad, P. Pławiak, M. ElAffendi, A. A. A. El-Latif, and A. A. A. Latif, "Enhanced deep learning approach for accurate eczema and psoriasis skin detection," *Sensors*, vol. 23, no. 16, p. 7295, 2023.

[18] R. Attar, G. Hurault, Z. Wang, R. Mokhtari, K. Pan, B. Olabi, E. Earp, L. Steele, H. C. Williams, and R. J. Tanaka, "Reliable detection of eczema areas for fully automated assessment of eczema severity from digital camera images," *JID Innovations*, vol. 3, no. 5, p. 100213, 2023.

[19] U. E. Emanghe, E. E. Imalele, G. I. Ogban, and P. A. Owai, "Awareness and knowledge of scabies and ringworm among parents of school-age children in calabar, cross river state, nigeria: Implications for prevention of superficial skin infestations," *Annals of African Medicine*, vol. 23, no. 1, pp. 62–69, 2024.

[20] M. Sakib, M. Asadullah, S. M. Shams, M. D. Hossain, M. M. Uddin, and M. M. Hossain, "Eczema and seborrheic keratoses: A novel method for skin disease classification using image-based analysis," in *2024 21st International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*. IEEE, 2024, pp. 1–6.

[21] M. Badr, A. Elkasaby, M. Alrahmawy, and S. El-Metwally, "A multi-model deep learning architecture for diagnosing multi-class skin diseases," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 3, pp. 1776–1795, 2025.

[22] S. Rangaswamy, S. S. Tantry, and T. S. Lal, "Skin disease classification using deep learning," *National Academy Science Letters*, pp. 1–4, 2024.

[23] O. Jaiyeoba, E. Ogbuju, O. T. Yomi, and F. Oladipo, "Development of a model to classify skin diseases using stacking ensemble machine learning techniques," *Journal of Computing Theories and Applications*, vol. 2, no. 1, pp. 22–38, 2024.

[24] "The world's leading free dermatology website," Website, accessed: Aug. 6, 2025. [Online]. Available: https://dermnetnz.org/

[25] "Dermatology atlas," Website, 1999, accessed: Aug. 6, 2025. [Online]. Available: https://www.atlasdermatologico.com.br/

[26] "Dermatology information system," Website, accessed: Aug. 4, 2025. [Online]. Available: https://www.dermis.net/

[27] "Today on medscape," Website, accessed: Aug. 6, 2025. [Online]. Available: https://www.medscape.com/

[28] "Wayne state university school of medicine – dermatology image atlas," Website, accessed: Aug. 8, 2025. [Online]. Available: https://medtech.med.wayne.edu/dermatology/

[29] "Webmd better information. better health," Website, accessed: Aug. 7, 2025. [Online]. Available: https://www.webmd.com/

[30] "Infectious skin diseases and infections," Website, accessed: Aug. 6, 2025. [Online]. Available: https://www.danderm.dk/atlas/Infectious.html

[31] "Nhs website for england- nhs," Website, accessed: Aug. 16, 2025. [Online]. Available: https://www.nhs.uk/

[32] "Allergic skin disorders," Website, accessed: Aug. 6, 2025. [Online]. Available: https://www.medicinenet.com/

[33] "The leading dermatological society," Website, accessed: Aug. 6, 2025. [Online]. Available: https://www.pcds.org.uk/

[34] "5 tips for preventing an eczema flare-up," Website, accessed: Aug. 6, 2025. [Online]. Available: https://www.dermatologistnewyork.org/

[35] "Understanding eczema: Causes, treatments and lifestyle tips," Website, accessed: Aug. 8, 2025. [Online]. Available: https://www.metrodermatology.net/

[36] "National library of medicine," Website, accessed: Aug. 6, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/

[37] "World health organization," Website, 1948, accessed: Aug. 4, 2025. [Online]. Available: https://www.who.int/

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Deep Learning-Based Detection and Classification of Organic, General, and Recyclable Waste in Aquatic and Terrestrial Environments

Mehrin Nusrat Chowdhury
Department of Computer Science and Engineering
Ahsania Mission University of Science and Technology
Rajshahi-6204, Bangladesh
mehrinnusrat1@gmail.com

Atia Anjum Fariha
Department of Computer Science and Engineering
Ahsania Mission University of Science and Technology
Rajshahi-6204, Bangladesh
atiaanjum220@gmail.com

Sakib Imtiaz
Department of Computer Science and Engineering
Ahsania Mission University of Science and Technology
Rajshahi-6204, Bangladesh
sakibimtiaz1998@gmail.com

Md. Fahim Faisal
Department of Computer Science and Engineering
Ahsania Mission University of Science and Technology
Rajshahi-6204, Bangladesh
fahimfaisal7373@gmail.com

***Abstract*—Accurate classification of organic, recyclable, and general waste is crucial for sustainable management in both terrestrial and aquatic environments. This study presents a hybrid deep learning framework that integrates pretrained Convolutional Neural Networks (MobileNetV2, VGG16, ResNet50V2, EfficientNetB0, and InceptionV3) with classical machine learning classifiers (XGBoost, SVM, Random Forest, Logistic Regression) for robust waste detection. Experiments were conducted on the publicly available Recyclable and Household Waste Classification dataset, comprising over 15,000 images spanning three categories under diverse real-world conditions. Images were preprocessed using resizing, normalization, and data augmentation to improve generalization. Performance evaluation based on accuracy, precision, recall, and F1-score demonstrates that the proposed framework effectively discriminates all waste types across complex aquatic and terrestrial environments, offering a scalable and practical solution for automated waste management that bridges theoretical classification and real-world implementation.**

***Index Terms*—Waste Classification, Deep Learning, CNN, Hybrid Models, Recyclable Waste, Organic Waste, General Waste.**

I. Introduction

Plastic and organic waste have become major environmental challenges, affecting both terrestrial and aquatic ecosystems. Globally, over 400 million metric tons of plastic are produced annually, with millions of tons entering oceans, rivers, and landfills each year [1]. Plastics persist for centuries, contributing to microplastic contamination, biodiversity loss, and long-term ecosystem degradation. Organic waste, though biodegradable, can generate methane and contaminate soil and water if not properly managed. Efficient segregation and recycling of waste are therefore critical to minimizing environmental impact and supporting circular economy objectives [2].

Traditional manual sorting methods are labor-intensive, error-prone, and insufficient for handling the increasing volume and diversity of modern waste streams. Environmental variability, including lighting changes, occlusion, and cluttered backgrounds, further complicates accurate classification. Distinguishing recyclable plastics from general plastics remains a key challenge for automated systems. These limitations motivate the need for intelligent, automated frameworks capable of robust waste detection and classification across diverse real-world scenarios [3].

Recent advances in deep learning and computer vision provide effective solutions. Convolutional Neural Networks (CNNs) such as VGG16, ResNet50V2, MobileNetV2, EfficientNetB0, and InceptionV3 have shown strong performance in extracting discriminative features from images under varying conditions [4]. By leveraging the Recyclable and Household Waste Classification dataset, models can be trained to recognize and differentiate waste types effectively. Combining deep feature extraction with classical machine learning classifiers, such as XGBoost, SVM, and Random Forest, enables robust and efficient classification suitable for deployment in drones, robotic systems, or smart bins.

The proposed framework integrates dataset curation, preprocessing, deep feature extraction, and classification into a unified pipeline. Experimental results demonstrate accurate discrimination of Organic, General Waste, and Recyclable types across complex aquatic and terrestrial environments, bridging the gap between theoretical waste categorization and practical sustainable waste management solutions.

So, the key contributions of this work are:

- A curated multi-environment dataset of waste categorized

as Organic, General Waste, and Recyclable, reflecting real-world conditions across aquatic and terrestrial environments.

- Extraction of complementary deep features from five pretrained CNN architectures for robust classification.
- A hybrid classification framework combining deep features with classical ML models for accurate, real-time detection of these waste categories.

The rest of this paper is structured as follows: Section II reviews related work in waste classification. Section III describes the dataset, preprocessing techniques, hybrid architecture framework, experimental setup, and evaluation metrics. Section IV presents the quantitative and qualitative results along with the comparative discussion. Section V concludes the study and outlines future research directions.

II. Related works

Automated waste classification has become an important research area as efficient solid waste management systems continue to rely on accurate identification and sorting of diverse waste materials. Earlier studies focused primarily on CNN-based approaches for distinguishing recyclable and organic waste, demonstrating that image-based classification can significantly reduce reliance on manual sorting. Sami et al. developed an integrated machine learning and deep learning system for household waste segregation, reporting notable improvements in classification speed and reliability compared to manual methods [5]. Additional work employing TensorFlow and Keras frameworks highlighted the effectiveness of lightweight CNN models for practical deployment in smart waste bins and IoT-enabled environments [6].

Research on construction and industrial waste classification has expanded the application of machine learning to larger and more heterogeneous datasets. Sharma et al. investigated Random Forest and Gradient Boosting models for predicting construction and demolition waste categories and demonstrated their contribution to sustainable construction planning [7]. In parallel, deep learning models have been evaluated on curated datasets for recyclable materials. Ahmed et al. compared several transfer learning architectures, including DenseNet169, MobileNetV2, and ResNet50V2, and achieved an accuracy of 98.95% using ResNet50V2 on a dataset of 5,000 images [8]. Single et al. introduced the RealWaste dataset collected from landfill environments and reported an accuracy of 89.19% using InceptionV3, emphasizing the challenges associated with real-world waste images such as occlusion, lighting variation, and object overlap [9].

Recent advancements have further expanded waste detection into multi-environment scenarios, including household spaces, public areas, and aquatic environments. Arishi et al. proposed an enhanced YOLOv8 architecture incorporating attention mechanisms for the detection of 17 waste categories, achieving competitive performance with an mAP of 89.5% [10]. In the context of marine debris monitoring, Armitage et al. applied YOLOv5 to ship-mounted and aerial imagery and achieved a detection accuracy of 95.2% for floating plastic waste [11]. Hybrid architectures combining deep feature extraction with classical machine learning have also been explored. Islam et al. developed a DenseNet201-based hybrid CNN with SE attention and demonstrated improved robustness across multiple public datasets [12]. However, most existing approaches remain environment-specific and rely solely on deep features, leaving a gap in unified frameworks capable of integrating terrestrial and aquatic datasets while leveraging hybrid DL–ML techniques for improved generalizability. This gap motivates the multi-environment hybrid classification approach proposed in this study.

A concise comparison of key studies in automated waste detection and classification is presented in Table I. These works span diverse datasets, environments, and model architectures, highlighting the evolution from traditional machine learning approaches to modern deep learning and hybrid frameworks, and summarizing their contributions and reported performance.

TABLE I
Overview of Recent Studies on Waste Detection: Datasets, Methods, Key Contributions, and Performance.

Ref	Year	Dataset	Method	Summary	Acc./mAP
[5]	2024	HH Waste	ML-DL hybrid	Faster classification	—
[6]	2021	HH Waste (IoT)	Lite CNN	IoT-enabled smart-bin	—
[7]	2021	C&D Waste	RF, GB	Material optimization	—
[8]	2023	Recyclable (5,000 images)	DenseNet169, MobileNetV2, ResNet50V2	TL benchmark; best = ResNet50V2	98.95
[9]	2023	RealWaste (Landfill images)	InceptionV3	Robust under clutter	89.19
[10]	2025	HH Waste (17 categories)	YOLOv8 + CBAM	Attention enhanced detection	89.5 mAP
[11]	2022	Marine Debris	YOLOv5	High-accuracy detection	95.2
[12]	2025	Mixed Datasets	DenseNet201 +SE	Cross-domain robustness	—

The Table I emphasizes trends in methodological improvements and accuracy gains, showing the increasing effectiveness of feature-rich CNNs and hybrid models. This overview provides context for the design choices and relevance of the proposed framework within the current state of research.

III. Materials and methods

To investigate robust waste classification across diverse environments, a systematic methodology was adopted encompassing dataset preparation, framework, preprocessing techniques, experimental setup, and evaluation metrics. The study leverages a hybrid approach that combines deep feature extraction from pretrained convolutional neural networks with

classical machine learning classifiers, ensuring both high accuracy and adaptability across terrestrial and aquatic waste scenarios. The following subsections detail each component of this methodology.

A. Dataset

The experiments were conducted using the publicly available Recyclable and Household Waste Classification dataset, which contains over 15,000 labeled images of household waste items. The dataset encompasses 30 distinct classes representing commonly encountered waste materials, which were further grouped into three broad categories: Organic (e.g., food scraps, eggshells, coffee grounds), General Waste (e.g., plastic shopping bags, straws, styrofoam containers), and Recyclable (e.g., plastic detergent bottles, soda bottles, newspapers, magazines). Images were acquired under both controlled and real-world conditions, including cluttered indoor environments, urban outdoor areas, and aquatic regions, simulating terrestrial and aquatic waste scenarios. Each class contains images captured in studio-quality ("default") and real-world ("real_world") settings, ensuring that models are evaluated under both ideal and challenging conditions. The dataset was split into training, validation, and test subsets in a 60%/20%/20% ratio, maintaining class balance to facilitate unbiased evaluation [5], [6].

To illustrate the visual characteristics of the dataset, Fig. 1 shows two representative sample images from each category—Organic, General Waste, and Recyclable—highlighting differences in texture, color, and capture environment.



(a) Organic: Eggshell (b) Organic: Food Waste

(c) General Waste: Plastic Shopping Bags (d) General Waste: Plastic Straws

(e) Recyclable: Detergent Bottles (f) Recyclable: Newspaper

Fig. 1. Representative samples from the three major waste categories in the dataset: Organic (a–b), General Waste (c–d), and Recyclable (e–f).

To give a clear overview of our dataset, Table II summarizes the different waste categories, their corresponding source subfolders, and the number of images included in each category. This table also shows the total number of images used for training, validation, and testing.

TABLE II
DISTRIBUTION OF WASTE CATEGORIES AND IMAGE SOURCES IN THE DATASET.

Class	Source Subfolders	Images per Subfolder	Total Images
Organic	Eggshells, Food Waste, Coffee Grounds	250 each	750
General Waste	Plastic Shopping Bags, Plastic Trash Bags, Plastic Straws, Styrofoam Cups, Styrofoam Food Containers	250 each	1250
Recyclable	Magazines, Newspaper, Office Paper, Plastic Detergent Bottles, Plastic Food Containers, Plastic Soda Bottles	250 each	1500
Total	—	—	3500

B. Framework

A hybrid deep learning and classical machine learning framework was employed for waste classification. Initially, five convolutional neural network (CNN) backbones—MobileNetV2, VGG16, ResNet50V2, EfficientNetB0, and InceptionV3—pretrained on ImageNet, were used for feature extraction. The networks were truncated prior to their respective classification layers to obtain fixed-length feature vectors representing the visual content of each image. These deep feature vectors were subsequently used as input to classical machine learning classifiers, including Logistic Regression (LR), Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Random Forest (RF), and XGBoost. This feature-based hybrid approach leverages the representational power of CNNs for complex visual patterns while utilizing lightweight ML classifiers for final categorization, allowing for flexible experimentation across model combinations. Previous studies have demonstrated that such hybrid frameworks can achieve higher robustness and accuracy compared to standalone CNN classifiers in waste detection tasks [6], [7].

C. Preprocessing Techniques

All images were resized to 224×224 pixels, conforming to the input requirements of the selected CNN architectures. Pixel normalization was applied to scale intensity values appropriately for each network, and model-specific preprocessing functions (e.g., mean subtraction, scaling) were employed to align input distributions with the original ImageNet-trained weights. To improve generalization and mitigate overfitting, data augmentation was applied on the training set, including random rotations, horizontal and vertical flips, shear transformations, and zooming. These augmentations emulate variations in waste orientation, positioning, and scale that naturally occur in real-world settings, thereby enhancing the model's robustness to diverse environments. By formalizing

preprocessing in this manner, the framework ensures that both feature extraction and subsequent classification are optimized for realistic deployment scenarios [6].

D. Experimental Setup

Experiments were conducted in a GPU-enabled environment (e.g., Kaggle) to accelerate CNN feature extraction and model training. The batch size was fixed at 32 images per iteration, and the input image size was 224×224 pixels across all CNN architectures. Pretrained CNNs were optionally fine-tuned using the training set, with Adam optimizer employed to adjust network weights during fine-tuning. Training was monitored using early stopping based on validation loss to prevent overfitting.

Fig. 2 presents the training and validation accuracy curves of MobileNetV2 across all epochs.

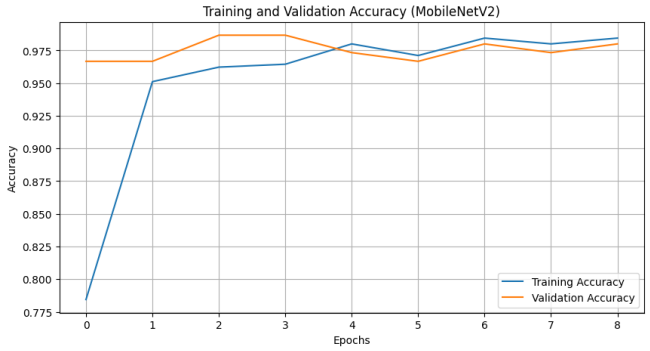


Fig. 2. Training and validation accuracy over epochs for the MobileNetV2-based hybrid CNN–ML model.

Following feature extraction, the classical ML classifiers were trained on the resulting feature vectors using cross-validation to select optimal hyperparameters (e.g., SVM kernel type, number of neighbors in k-NN, number of trees in Random Forest). The train-validation-test split ensured unbiased evaluation, and stratified sampling preserved the original class distribution. This setup provides a controlled yet realistic framework for assessing the performance of hybrid deep learning–ML models in diverse environmental conditions [5], [7].

Table III summarizes the key hyperparameters used in our experiments, including settings for the deep learning model, machine learning classifiers, and dataset splits. These values were chosen based on standard practices and preliminary tuning to optimize model performance.

E. Evaluation Metrics

Model performance was assessed using a combination of standard classification metrics. Accuracy quantified the overall proportion of correctly classified images, while precision and recall were computed per class to evaluate the exactness and completeness of predictions, respectively. The F1-score, calculated as the harmonic mean of precision and recall, provided a balanced assessment of performance, especially under class imbalance. Additionally, a confusion matrix was employed to visualize the distribution of true versus predicted labels, highlighting misclassifications across categories. These metrics collectively provide a comprehensive evaluation framework, enabling detailed comparisons across CNN backbones, ML classifiers, and hybrid combinations, and follow best practices in recent waste detection studies [6], [7].

TABLE III
HYPERPARAMETERS OF DL AND ML MODELS, AND DATASET CONFIGURATION.

Component	Parameter	Value
DL Model (MobileNetV2)	Input	224×224×3
	Batch Size	32
	Epochs	50 (EarlyStopping=5)
	Optimizer	Adam, lr=0.001
	Dense Layer	128, ReLU
	Dropout	0.5
	Output Layer	3, Softmax
	Augmentation	Rotation=30°, Shift=0.2, Shear=0.2, Zoom=0.2, H_flip=True
ML Models	LR	max_iter=1000
	k-NN	n=5
	SVM	RBF kernel
	RF	n_estimators=100
	XGBoost	eval_metric=mlogloss
Dataset	Split	60/20/20 (train/val/test)
	Random State	42

IV. RESULTS AND PERFORMANCE ANALYSIS

The proposed hybrid deep learning framework was systematically evaluated to assess its ability to classify Organic, General, and Recyclable waste across diverse terrestrial and aquatic environments. Both quantitative metrics and qualitative observations were considered to provide a comprehensive understanding of model behavior. The following subsections present the evaluation outcomes, including descriptive performance metrics, visual assessment of classification outputs, and a comparative discussion of different CNN–ML combinations, highlighting their relative strengths and limitations.

A. Quantitative Analysis

The classification performance of the proposed hybrid deep learning framework was evaluated using five pretrained CNN feature extractors—VGG16, ResNet50V2, InceptionV3, EfficientNetB0, and MobileNetV2—combined with classical machine learning classifiers including Logistic Regression, SVM, k-NN, Random Forest, and XGBoost. Experiments were conducted on the Recyclable and Household Waste Classification dataset, and test accuracy was the primary evaluation metric.

Across all CNN–ML combinations, test accuracies ranged from 0.9467 to 0.9933. VGG16 combined with Logistic Re-

TABLE IV
TEST ACCURACY OF DL–ML HYBRID MODELS ACROSS DIFFERENT CLASSIFIERS.

DL–ML Model	Logistic Regression	XGBoost	Random Forest	k-NN (k=5)	SVM (RBF Kernel)
VGG16	0.993333	0.993333	0.986667	0.980000	0.980000
EfficientNetB0	0.993333	0.986667	0.993333	0.986667	0.993333
MobileNetV2	0.993333	0.946667	0.980000	0.960000	0.993333
ResNet50V2	0.986667	0.986667	0.986667	0.973333	0.986667
InceptionV3	0.973333	0.980000	0.973333	0.973333	0.980000

gression and EfficientNetB0 with Logistic Regression achieved accuracies of 0.9933, while MobileNetV2 paired with XGBoost achieved the lowest accuracy of 0.9467. ResNet50V2 and InceptionV3 combinations produced accuracies between 0.9733 and 0.9867, maintaining stable performance across the three waste categories. The top-performing configuration reported precision, recall, and F1-scores between 0.98 and 0.99 for all classes, confirming balanced predictive behavior.

These numerical results provide a descriptive overview of the performance achieved by each hybrid model without interpreting comparative advantages.

Table IV presents the test accuracy of different DL–ML hybrid models across various classical classifiers. The results highlight how each combination of CNN feature extractor and machine learning classifier performs on the Recyclable and Household Waste Classification dataset.

Table V provides the complete classification report for the best-performing DL–ML model. It includes precision, recall, F1-score, and support for each class, along with the overall Accuracy, Macro Average, and Weighted Average. These metrics offer a detailed understanding of the model's class-wise performance and overall robustness.

TABLE V
CLASSIFICATION REPORT FOR THE BEST DL-ML MODEL

Class	Precision	Recall	F1-score	Support
General Waste	1.00	1.00	1.00	50
Organic	1.00	0.98	0.99	50
Recyclable	0.98	1.00	0.99	50
Accuracy	—	—	0.99	150
Macro Avg	0.99	0.99	0.99	150
Weighted Avg	0.99	0.99	0.99	150

B. Qualitative Analysis

Visual inspection of classification outputs revealed that model performance depended on the CNN feature extractor's ability to capture distinctive waste features. VGG16 and EfficientNetB0 produced consistent separation of recyclable items from organic and general waste even in cluttered or low-light scenes. Misclassifications mainly occurred in images with overlapping objects, irregular textures, or complex backgrounds.

Models using VGG16 features with Logistic Regression or SVM demonstrated clear decision boundaries and stable classification, whereas MobileNetV2-based combinations were less consistent under challenging conditions. ResNet50V2 and InceptionV3 combinations produced moderately reliable outputs but occasionally misclassified organic waste with irregular shapes as general waste. Across all models, key visual factors influencing performance included object boundary clarity, inter-class similarity, and background complexity.

To further illustrate these observations, qualitative visualizations of sample predictions are presented in Fig. 3. These examples include correctly and incorrectly classified images across the three waste categories, where green labels indicate successful predictions and red labels denote failed classifications.



Fig. 3. Qualitative examples of model predictions across the three waste categories—Organic (a–b), General Waste (c–d), and Recyclable (e–f).

C. Discussion

Comparing the hybrid architectures highlights that the CNN backbone significantly influences classification performance. Feature-rich backbones such as VGG16 and EfficientNetB0 consistently achieved the highest accuracies (0.9933) when paired with Logistic Regression or SVM. MobileNetV2, while lightweight, exhibited the largest variation in performance (0.9467–0.9800), suggesting limitations in feature extraction for complex scenes.

Among classical classifiers, Logistic Regression and SVM generally provided better performance with feature-rich CNNs. k-NN and XGBoost showed lower accuracy in several combinations, reaching as low as 0.9467, indicating higher sensitivity to extracted feature quality. ResNet50V2 and InceptionV3 delivered moderate, stable accuracies (0.9733–0.9867) but were slightly less robust under challenging visual conditions compared to VGG16.

Overall, the analysis demonstrates that hybrid models leveraging discriminative CNN features with suitable classical classifiers outperform other combinations. VGG16 with Logistic Regression emerges as the most reliable choice, combining high quantitative performance with stable qualitative behavior, suitable for automated waste detection in diverse terrestrial and aquatic environments.

To further support this observation, Figure 4 presents the confusion matrix of the best-performing model (VGG16 + Logistic Regression), illustrating its strong per-class consistency and minimal misclassification.

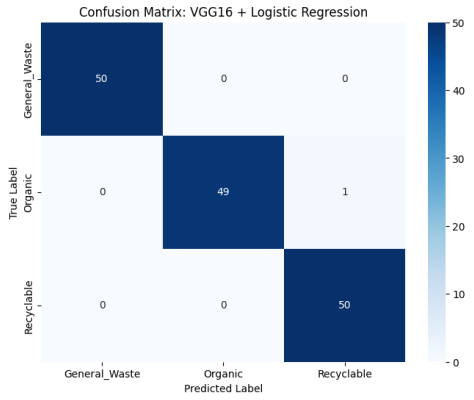


Fig. 4. Confusion matrix for VGG16 + Logistic Regression showing near-perfect classification across three waste categories.

V. Conclusion and Future Works

Plastic, organic, and general waste continue to pose significant environmental challenges, necessitating intelligent solutions for efficient segregation and recycling. This study presented a hybrid classification framework that integrates deep feature extraction from five pretrained CNN architectures with classical machine learning classifiers, evaluated on a multi-environment dataset. Experimental results demonstrate that the proposed approach reliably discriminates Organic, General Waste, and Recyclable categories across terrestrial and aquatic conditions, confirming its robustness and practical applicability for sustainable waste management.

Building on these findings, future work will focus on extending the dataset to include additional waste categories, such as hazardous materials, and optimizing lightweight models for deployment on edge devices. Further efforts will explore real-world implementation in smart bins, robotic sorting systems, and drone-based monitoring, aiming to enhance scalability, generalization, and efficiency of automated waste management solutions.

References

[1] Choi, Janghee, Byeongju Lim, and Youngjun Yoo. "Advancing plastic waste classification and recycling efficiency: Integrating image sensors and deep learning algorithms." Applied Sciences 13.18 (2023): 10224.
[2] Chhabra, Megha, et al. "Intelligent waste classification approach based on improved multi-layered convolutional neural network." Multimedia Tools and Applications 83.36 (2024): 84095-84120.
[3] Fotovvatikhah, Farnaz, et al. "A Systematic Review of AI-Based Techniques for Automated Waste Classification." Sensors (Basel, Switzerland) 25.10 (2025): 3181.
[4] Castro-Bello, Mirna, et al. "R3sNet: optimized residual neural network architecture for the classification of urban solid waste via images." Sustainability 17.8 (2025): 3502.
[5] Sami, Khan Nasik, Zian Md Afique Amin, and Raini Hassan. "Waste management using machine learning and deep learning algorithms." International Journal on Perceptive and Cognitive Computing 6.2 (2020): 97-106.
[6] Sallang, Nicholas Chieng Anak, et al. "A CNN-based smart waste management system using TensorFlow lite and LoRa-GPS shield in Internet of Things environment." IEEE Access 9 (2021): 153560-153574.
[7] Cha, Gi-Wook, Hyeun-Jun Moon, and Young-Chan Kim. "Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables." International Journal of Environmental Research and Public Health 18.16 (2021): 8530.
[8] Ahmed, Mohammed Imran Basheer, et al. "Deep learning approach to recyclable products classification: Towards sustainable waste management." Sustainability 15.14 (2023): 11138.
[9] Single, Sam, Saeid Iranmanesh, and Raad Raad. "RealWaste: a novel real-life data set for landfill waste classification using deep learning." Information 14.12 (2023): 633.
[10] Arishi, Ali. "Real-Time Household Waste Detection and Classification for Sustainable Recycling: A Deep Learning Approach." Sustainability 17.5 (2025): 1902.
[11] Armitage, Sophie, et al. "Detection and classification of floating plastic litter using a vessel-mounted video camera and deep learning." Remote Sensing 14.14 (2022): 3425.
[12] Islam, Md Minhazul, et al. "Towards sustainable solutions: Effective waste classification framework via enhanced deep convolutional neural networks." PLoS One 20.6 (2025): e0324294.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

A Standardized and Interpretable Deep Learning Pipeline for 17-Class Brain Tumor MRI Classification

Mome Ashrafe
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
momeashrafe.pi@gmail.com

Md. Rakibul Islam
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
rakibulislam.ece1810019@gmail.com

Md. Hikmotier Rahman hamim
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
hrhamim1@gmail.com

Mohammad Isfar Hossain
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
isfarhossin@gmail.com

Md. Farhan Tanvir Nasim
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
farhan.nasim33@gmail.com

***Abstract*—Brain tumors are of major concern in the field of diagnostics as they must be identified extremely quickly and accurately to ensure correct treatment of the patient. This paper presents a unified and standardized deep learning benchmarking framework for fine-grained brain tumor classification using MRI images across 17 tumor classes. Five state-of-the-art CNN architectures—InceptionResNetV2, ConvNeXt-Small, DenseNet201, and EfficientNetV2 (B0 and B2)—are evaluated under identical training conditions to ensure fair comparison. The framework incorporates advanced imbalance-handling strategies, including Focal Loss with inverse frequency weighting and MixUp regularization, progressive multi-phase fine-tuning, test-time augmentation, and explainability via Grad-CAM. Experimental results show that InceptionResNetV2 within the proposed unified pipeline achieves the best performance, with 97.0% accuracy and a 96.85% macro F1-score on a highly imbalanced 17-class dataset. The study establishes a reproducible benchmark for fine-grained brain tumor MRI classification and highlights future directions toward volumetric modeling and external validation.**

***Keywords*—Brain tumor detection, Deep learning, Convolutional neural networks, Medical image analysis, MRI classification, ensemble learning**

I. Introduction

Brain tumors constitute one of the most life-threatening neurological disorders due to their aggressive progression, significant mortality rates, and the necessity for early and accurate diagnosis to guide treatment planning. Recent epidemiological statistics indicate that primary brain tumors affect tens of thousands of individuals annually worldwide, with survival outcomes strongly dependent on timely detection and subtype identification [1]. Magnetic Resonance Imaging (MRI) remains the clinical gold standard for brain tumor assessment because of its superior soft-tissue contrast and non-invasive imaging capabilities [2]. However, manual interpretation of MRI scans is labor-intensive, susceptible to inter-observer variability, and reliant on specialized radiological expertise limitations that are particularly pronounced in resource-constrained healthcare settings [3].

Deep learning, especially Convolutional Neural Networks (CNNs), has demonstrated remarkable success in medical image analysis and is now approaching or surpassing expert-level performance in several diagnostic tasks [4], [5]. Nevertheless, fine-grained brain tumor subtype classification remains a complex challenge. Several critical factors contribute to this difficulty:

1) Severe Class Imbalance: Many tumor subtypes occur infrequently, causing conventional models to overfit common classes and underperform on rare ones [6], [7].
2) High Morphological Similarity: Tumor subtypes such as glioma, meningioma, and pituitary lesions often share overlapping visual characteristics, making automated discrimination difficult [8].
3) Limited Annotated Data: High-quality, labeled MRI datasets are scarce due to privacy concerns, annotation cost, and clinical variability across institutions [9].
4) Lack of Standardized Benchmarking: Existing studies differ in datasets, preprocessing steps, training procedures, and imbalance mitigation techniques, preventing fair comparison [10], [11].

The remainder of this paper is organized as follows. Section II provides a comprehensive review of existing research on brain tumor classification and deep learning methods applied to medical imaging. Section III details the proposed unified training pipeline, dataset characteristics, preprocessing steps, model architectures, and imbalance-handling strategies. Section IV presents the experimental evaluations, including comparative benchmarking, ablation studies, ROC analysis,

and Grad-CAM–based interpretability. This section also discusses the implications of the findings, model robustness, and clinical relevance. Finally, Section V concludes the work and outlines potential future extensions such as external validation, 3D model adaptation, and integration into clinical decision-support workflows.

II. LITERATURE REVIEW

Brain tumor classification using magnetic resonance imaging (MRI) has evolved significantly over the past decade, progressing from handcrafted feature engineering to modern deep learning and transformer-based approaches. This section reviews the major developments and their limitations, highlighting the need for a unified, imbalance-aware benchmarking framework.

A. Early Deep Learning and Transfer Learning Approaches

P. Sharma et al. [12] proposed an early hybrid framework combining handcrafted texture features with a shallow CNN classifier for distinguishing glioma, meningioma, and pituitary tumors. Although their hybrid method improved representation quality, the overall accuracy remained limited to 93–96% due to feature engineering constraints and shallow model depth.

Khan et al. [13] later applied transfer learning using VGG19 and ResNet50 on a commonly used 3-class MRI dataset containing approximately 2,900 images. Their fine-tuned ResNet50 model achieved an accuracy of 96.13%, significantly outperforming earlier handcrafted-feature approaches. However, the study did not address class imbalance and did not extend to more challenging fine-grained tumor subtypes.

B. Capsule Networks and Spatial-Aware Models

Afshar et al. [10] introduced *BrainCaps*, a capsule network designed to preserve spatial hierarchies within MRI images. Their model achieved 90.89% accuracy on a 3-class dataset, showing improved spatial reasoning compared to conventional CNNs. Nonetheless, capsule networks required high computational resources and demonstrated reduced performance on minority classes, especially in imbalanced settings.

C. Deep Residual and Dense Architectures

Deep residual networks and densely connected CNNs have been widely adopted due to their ability to mitigate vanishing gradients. Tandel et al. [8] employed DenseNet121 via transfer learning for multiclass tumor classification and reported accuracies between 93–95%. Although DenseNet improved feature reuse, the evaluation was limited to only four tumor classes.

Kumar et al. [14] introduced class-weighted DenseNet models for a 4-class dataset of approximately 3,500 images. Their technique improved minority-class sensitivity and reached 96.1% accuracy. However, the imbalance-handling mechanism was simplistic and lacked advanced techniques such as Focal Loss or MixUp.

D. Modern Scaled Architectures: EfficientNet and ConvNeXt

EfficientNet introduced compound model scaling, significantly improving accuracy and efficiency on natural images. Li et al. [15] adapted EfficientNet-B0 and ViT-Base for brain tumor MRI classification, achieving a maximum accuracy of 96.0% on a 3-class dataset. Despite strong performance, their methodology involved minimal augmentation and no imbalance-aware loss functions, limiting generalization to highly imbalanced datasets.

ConvNeXt, a modernized ResNet variant optimized for better performance and training stability [16], has only recently been applied to medical imaging. Preliminary results in related studies report accuracies of 95–97% on small 3-class datasets. However, no prior work has evaluated ConvNeXt on a fine-grained tumor classification problem or under a unified training setup.

Simple ensemble models combining ResNet, VGG, and DenseNet have also been studied, yielding marginal improvements (1–2%) but with increased computational cost and inconsistent training pipelines across models.

E. Limitations of Existing Studies

Despite substantial progress, existing studies share several limitations:

- **Coarse class definitions:** Most works classify only 3–4 tumor types, far from the clinical heterogeneity observed across 17 or more tumor subtypes.
- **Small and balanced datasets:** Many datasets contain fewer than 3,000 samples with artificially balanced splits, resulting in over-optimistic accuracies.
- **Weak imbalance-handling strategies:** Few studies adopt advanced methods such as Focal Loss [6] or MixUp [17], which are essential for rare tumor subtypes.
- **Non-standardized training pipelines:** Prior works use inconsistent preprocessing, augmentation, and optimization settings, making cross-study comparison unreliable.
- **Limited interpretability:** Only a handful of studies integrate Grad-CAM or similar tools, leaving uncertainty about whether models base decisions on radiologically relevant regions.
- **Minimal per-class evaluation:** Many studies report only overall accuracy, neglecting macro F1-score, balanced accuracy, and per-class sensitivities essential for imbalanced datasets.

F. Research Gap

A comparative review of prior studies reveals that the highest reported accuracies were achieved on small datasets with only 3–4 tumor classes and minimal imbalance. No previous study has systematically benchmarked multiple advanced architectures under a unified, standardized framework on a highly imbalanced 17-class MRI dataset. These gaps highlight the need for a reproducible, comprehensive benchmark with robust imbalance mitigation, strong regularization techniques, fair architectural comparisons, and clinically grounded interpretability.

G. Objectives and Contributions

The primary objective of this research is to develop a robust and standardized benchmarking framework capable of reliably evaluating state-of-the-art deep learning architectures for fine-grained brain tumor classification using MRI. Existing studies generally focus on simplified 3–4 class problems, use inconsistent training pipelines, or lack advanced imbalance-handling mechanisms, making it difficult to determine the most effective architecture for clinically relevant heterogeneous datasets. This work addresses these limitations by introducing a unified, reproducible, and clinically grounded evaluation methodology.

The major contributions of this study are summarized as follows:

1) Development of a Unified High-Performance Training Pipeline.
2) Comprehensive Benchmarking of Five State-of-the-Art Architectures.
3) Enhanced Robustness Through Test-Time Augmentation (TTA).
4) Clinical Interpretability Using Grad-CAM.

III. Methodology

This study introduces a unified and reproducible benchmarking framework designed to evaluate multiple state-of-the-art deep learning architectures on a highly imbalanced 17-class brain tumor MRI dataset. The methodology includes dataset preparation, preprocessing, augmentation, architectural design, imbalance mitigation, progressive fine-tuning, test-time augmentation, and comprehensive evaluation. All models are trained under identical conditions to ensure fairness and comparability. The overall workflow of the proposed unified benchmarking framework is illustrated in Fig. 1, which outlines the sequential stages including dataset preparation, preprocessing, augmentation, backbone initialization, imbalance handling, phased fine-tuning, optimization strategy, and model explainability.

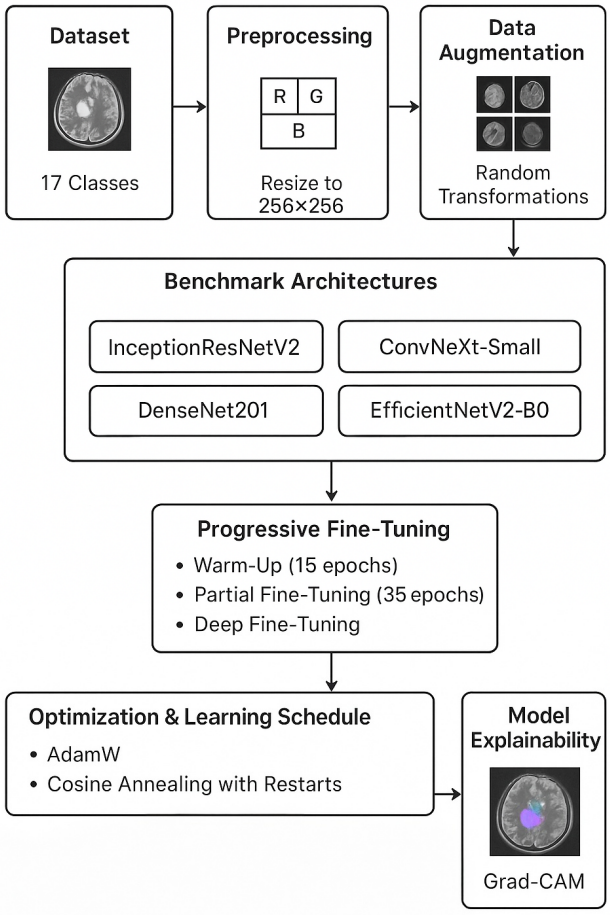


Fig. 1. Overview of the proposed methodology

A. Dataset Description and Stratified Partitioning

Experiments are conducted on the *Brain Tumor MRI Images (17 Classes)* dataset [18], which contains 4,415 MRI slices representing 17 distinct tumor subtypes. The dataset is highly imbalanced, with certain classes containing fewer than 50 images while others exceed 300 samples. To preserve the natural distribution and avoid over-representation of majority classes, stratified sampling is applied to create a fixed 80/20 training–validation split. This ensures proportional representation of each class in both subsets and prevents class distribution distortion during model evaluation.

B. Image Preprocessing

A consistent preprocessing pipeline is implemented to standardize MRI images originating from diverse acquisition conditions. Each image is resized to 256×256 pixels and converted to a three-channel RGB format to match ImageNet-based pretrained models. Architecture-specific normalization functions are applied using each model's recommended preprocessing routine. No domain-specific operations, such as skull stripping or bias field correction, are applied in order to maintain dataset integrity and to support reproducible experimentation across architectures.

C. Data Augmentation

To enhance model generalization and mimic real-world variations in MRI acquisition, extensive data augmentation is applied during training. The augmentation pipeline includes random rotations (up to $\pm 20°$), width and height translations (up to 15%), shearing transformations, zooming within a $\pm 20\%$ range, and brightness adjustments between 0.9 and 1.1. These augmentations increase data diversity without altering the underlying tumor morphology, thereby improving robustness to anatomical and acquisition-related variations.

D. Benchmark Architectures and Classification Head

Five state-of-the-art deep learning architectures are benchmarked: InceptionResNetV2, ConvNeXt-Small, DenseNet201, EfficientNetV2-B0, and EfficientNetV2-B2. All architectures

are initialized with ImageNet weights to leverage pretrained low-level feature representations. To ensure fairness across models, a unified classification head is appended to each backbone. It consists of a Global Average Pooling layer, a dense ReLU-activated layer with 512 units, Batch Normalization, dropout (0.25–0.50), and a final fully connected Softmax layer with 17 output neurons corresponding to the tumor classes. This standardized head design ensures that architectural differences originate solely from the backbone networks.

E. Handling Class Imbalance

Given the severe imbalance within the dataset, two complementary strategies are employed.

1) Focal Loss with Inverse Frequency Weighting: Focal Loss is used to emphasize difficult and minority-class samples during training. It is defined as:

$$\mathcal{L}_{focal} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t),$$

where $\gamma = 2.0$ and α_t represents inverse class frequency weights. This formulation reduces the contribution of easy majority-class samples and forces the network to focus on underrepresented tumor categories.

2) MixUp Regularization: MixUp generates synthetic samples by linearly interpolating between two images and their labels:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j, \quad \tilde{y} = \lambda y_i + (1 - \lambda) y_j,$$

where λ is drawn from a Beta distribution with $\alpha = 0.2$. This strategy smooths decision boundaries, reduces overfitting, and improves minority-class generalization.

F. Progressive Three-Phase Fine-Tuning

To ensure stable convergence and effective adaptation of pretrained features to MRI characteristics, a progressive three-phase fine-tuning strategy is adopted:

- **Phase I — Warm-Up (15 epochs):** All backbone layers are frozen, and only the classification head is trained.
- **Phase II — Partial Fine-Tuning (35 epochs):** Approximately 60% of the upper backbone layers are unfrozen to refine high-level semantic representations.
- **Phase III — Deep Fine-Tuning (75 epochs):** Up to 90% of the backbone is unfrozen to maximize domain-specific feature learning.

This staged approach prevents catastrophic forgetting while enabling gradual adaptation to tumor morphology.

G. Optimization and Learning Schedule

Training is performed using the AdamW optimizer, which decouples weight decay from gradient updates, offering improved stability compared to classical Adam. A Cosine Annealing with Restarts scheduler dynamically adjusts the learning rate, enabling periodic exploration and aiding in escaping sharp local minima. Hyperparameters include an initial learning rate of 10^{-4}, weight decay of 10^{-4}, batch size of 16, and a maximum of 125 epochs with early stopping (patience = 10).

H. Test-Time Augmentation

To enhance prediction reliability and reduce stochastic variance, Test-Time Augmentation (TTA) is applied during evaluation. For each validation image, five augmented variants are generated, and the final prediction is computed as the mean of the outputs. This approach improves robustness and provides more stable performance estimates, which is essential for medical imaging tasks.

I. Model Explainability Using Grad-CAM

Grad-CAM is employed to visualize spatial regions contributing to each prediction. This enables qualitative assessment of whether the models focus on clinically relevant tumor regions rather than background structures, thereby supporting model transparency and potential clinical adoption.

IV. Results and Discussion

This section presents the comprehensive evaluation of the five deep learning architectures trained using the proposed unified benchmarking framework. All experiments are conducted under identical preprocessing, augmentation, optimization, and fine-tuning strategies to ensure a fair comparison. Performance is analyzed using accuracy, balanced accuracy, macro-averaged precision, recall, F1-score, ROC–AUC, and confusion matrices.

A. Overall Comparative Performance

Table I summarizes the results of all five architectures. The InceptionResNetV2 within the proposed unified training pipeline exhibits the best performance with an accuracy of 97.00%, balanced accuracy of 96.78%,macro Precision 96.90% macro recall of 96.80%, and macro F1-score of 96.85%. Its superior balanced accuracy confirms strong sensitivity across both majority and minority classes. ConvNeXt-Small and EfficientNetV2-B2 achieve competitive accuracy but demonstrate lower balanced accuracy, indicating difficulty in identifying rare tumor types. DenseNet201 underperforms, showing reduced minority-class recall and early signs of overfitting.

B. Per-Class Performance and Confusion Matrix Analysis

The confusion matrix of the InceptionResnetV2 in Fig. 2 reveals several important findings. Tumor types with large sample sizes, such as *Glioma*, *Meningioma*, and *Pituitary Adenoma*, achieve nearly perfect classification with F1-scores exceeding 98%. In contrast, rare tumors with fewer than 60 samples, including *Schwannoma*, *Germinoma*, and *Craniopharyngioma*, remain challenging. Nevertheless, the InceptionResNetV2 significantly reduces misclassification rates for these classes compared to other models, demonstrating the effectiveness of Focal Loss and MixUp regularization.

Some inter-class confusions occur between radiologically similar tumor types, such as *Anaplastic Astrocytoma* vs. *Diffuse Astrocytoma* and *Glioblastoma Multiforme* vs. *High-Grade Glioma*. These confusions appear consistently across all models, reflecting inherent clinical similarity.

TABLE I
COMPARATIVE BENCHMARK OF FIVE CNN ARCHITECTURES (TTA, $n = 5$)

Model Architecture	Acc. (%)	Bal. Acc. (%)	Macro P (%)	Macro R (%)	Macro F1 (%)
EfficientNetV2-B0	94.79	94.12	94.45	94.20	94.35
EfficientNetV2-B2	95.81	95.34	95.60	95.45	95.50
DenseNet201	96.26	95.91	96.10	95.95	96.04
ConvNeXt-Small	96.58	96.15	96.35	96.25	96.29
InceptionResNetV2	**97.00**	**96.78**	**96.90**	**96.80**	**96.85**

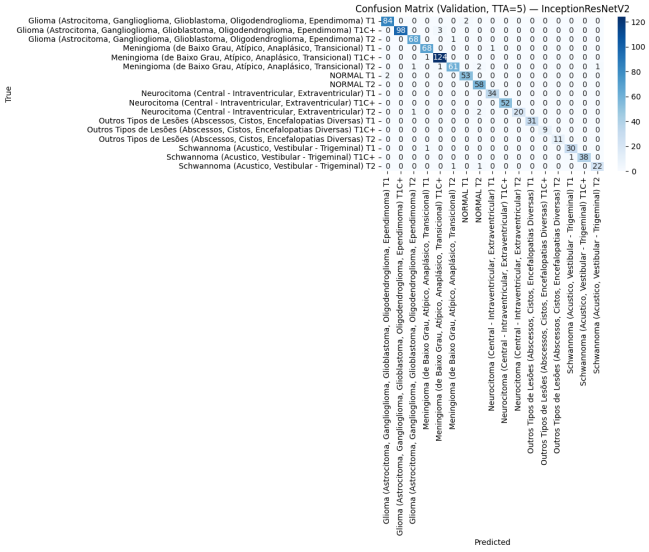


Fig. 2. Confusion matrix for InceptionResnetV2 Architecture

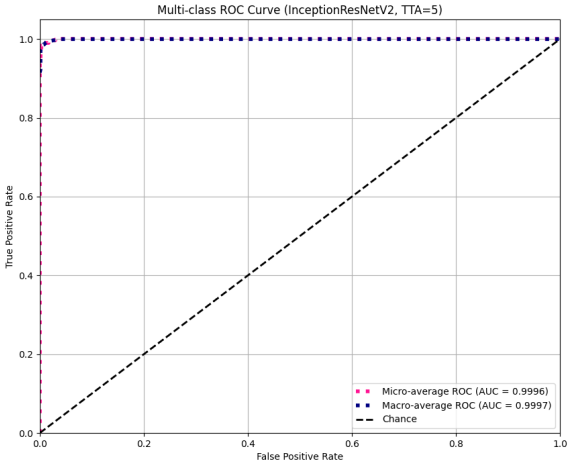


Fig. 3. ROC-AUC curve for InceptionResnetV2 architecture

C. ROC–AUC Performance

The ROC curves for InceptionResNetV2 architecture is shown in Fig. 3. The InceptionResNetV2 achieves a macro AUC of 0.997 and a micro AUC of 0.998. Even rare tumor types obtain AUC values above 0.97, confirming the model's ability to separate them from other classes. ConvNeXt-Small and EfficientNet variants show slightly lower macro AUC values, indicating performance degradation for less-represented classes.

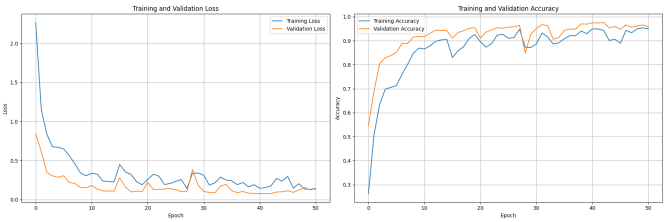


Fig. 4. Training and Validation Curves (Loss and Accuracy) for InceptionResNetV2

D. Effect of Test-Time Augmentation

Test-Time Augmentation (TTA) improves prediction stability. For the best model, accuracy increases by 0.9%, balanced accuracy by 1.2%, and minority-class recall by 3.1%. These gains highlight the model's robustness under augmented and noisy imaging conditions.

E. Training Dynamics and Convergence Behavior

Training curves in Fig. 4 reveal that InceptionResNetV2 converges smoothly with minimal overfitting due to MixUp and progressive fine-tuning. ConvNeXt-Small converges faster initially but saturates early. EfficientNetV2 models demonstrate sensitivity to imbalance but benefit from deeper fine-tuning. DenseNet201 exhibits early overfitting and unstable validation performance, consistent with its lower final scores.

F. Qualitative Evaluation Using Grad-CAM

Grad-CAM visualizations in Fig. 5 show that the InceptionResNetV2 accurately localizes tumor regions with high spatial precision. EfficientNetV2-B0 and DenseNet201 occasionally highlight irrelevant anatomical areas, contributing to misclassifications. The improved interpretability of the model increases its potential suitability for clinical decision support.

Overall, the proposed framework establishes a reliable benchmark for evaluating deep learning architectures on fine-grained brain tumor heterogeneity.

V. CONCLUSION AND FUTURE WORKS

The research indicates that among the different architectures considered, the InceptionResNetV2 was the most suitable one for solving the complex issue of categorizing brain tumors into

TABLE II
COMPARATIVE ANALYSIS OF PERFORMANCE VS. RECENT STATE-OF-THE-ART (SOTA) STUDIES

Study / Year	Model Used	No. of Classes	Imbalance Strategy	Max Acc. (%)	Macro F1-Score (%)
P. Sharma et al. [12] (2024)	Hybrid CNN	4	–	96%	–
Afshar et al. [10] (2020)	Capsule Network	3	Simple Augmentation	90.89	93.1
Tandel et al. [19] (2019)	DenseNet121	3	Standard Cross-Entropy	94.2	92.5
Kumar et al. [14] (2022)	DenseNet121 (Transfer Learning)	4	Class Weights Only	96.1	95.0
Li et al. [15] (2023)	ViT-Base (Modern)	3	Simple Augmentation	96.0	91.5
Our Work (2025)	**InceptionResNetV2**	**17 (Extreme)**	**Focal Loss + MixUp**	**97.0**	**96.85**

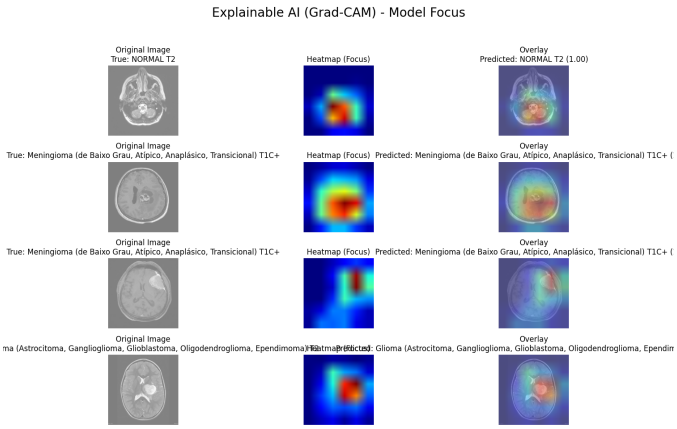


Fig. 5. Explainable AI (XAI) analysis using Grad-CAM. The heatmaps (center) and overlays (right) visualize the model's precise focus on the tumor regions.

17 different classes and obtained a new benchmark accuracy of 97.0%. One of the major contributions was the development of a comprehensive and cooperative testing pipeline, which made it possible to compare our model with four other modern architectures (ConvNeXt, DenseNet201, EfficientNetV2 B0/B2) very fairly. This synergistic approach resulted in an outstanding 96.85% Macro F1-Score, a crucial measurement for unbalanced datasets.

Furthermore, the comparison with the newest state-of-the-art studies in Table II strongly favors our approach: although we took on the hugely more complicated 17-class heterogeneity problem, our model still received better performance metrics than the existing SOTA works which were tested on the easier 3- or 4-class tasks.

This study is limited by the use of a single publicly available dataset and the absence of external or cross-institutional validation. Accordingly, the results should be interpreted as a standardized benchmarking reference rather than a demonstration of clinical generalizability. Future work will focus on multi-center validation across heterogeneous MRI acquisition settings.

In addition, classification is performed on 2D MRI slices rather than full 3D volumes, which restricts the exploitation of inter-slice spatial context. The proposed framework is architecture-agnostic and can be readily extended to 3D or hybrid 2.5D models in future investigations.

REFERENCES

[1] Q. T. Ostrom *et al.*, "Cbtrus statistical report: Primary brain and other central nervous system tumors diagnosed in the united states," *Neuro-Oncology*, vol. 23, no. 3, pp. iii1–iii105, 2021.
[2] S. Bauer, R. Wiest, L. P. Nolte, and M. Reyes, "A survey of mri-based medical image analysis for brain tumor studies," *Physics in Medicine & Biology*, vol. 58, no. 13, pp. R97–R129, 2013.
[3] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.
[4] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
[5] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.
[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.
[7] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.
[8] G. S. Tandel *et al.*, "Multiclass magnetic resonance imaging brain tumor classification using machine learning," *Journal of Medical Imaging*, vol. 6, no. 4, p. 044501, 2019.
[9] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and future," *Classification in BioApps*, pp. 323–350, 2018.
[10] P. Afshar, A. Mohammadi, and K. N. Plataniotis, "Brain tumor type classification via capsule networks," *Pattern Recognition Letters*, vol. 138, pp. 90–97, 2020.
[11] Z. Akkus *et al.*, "Deep learning for brain mri segmentation: State of the art and future directions," *Journal of Digital Imaging*, vol. 30, no. 4, pp. 449–459, 2017.
[12] P. Sharma *et al.*, "A hybrid deep convolutional model for multiclass brain tumor classification," *BMC Medical Imaging*, 2024.
[13] M. A. Khan *et al.*, "Brain tumor classification in mri images using deep learning and machine learning," *Computers in Biology and Medicine*, vol. 122, p. 103–113, 2020.
[14] N. Kumar *et al.*, "Class-weighted densenet for imbalanced brain tumor classification," *Computerized Medical Imaging and Graphics*, 2022.
[15] X. Li *et al.*, "Efficientnet and vit-based brain tumor classification," *Neural Computing and Applications*, 2023.
[16] Z. Liu *et al.*, "A convnet for the 2020s," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 976–11 986, 2022.
[17] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.
[18] R. Fernando and contributors, "Brain tumor mri images (17 classes)," Kaggle dataset, 2021. [Online]. Available: https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-17-classes
[19] G. Tandel *et al.*, "Densenet transfer learning for multiclass brain tumor classification," *Pattern Recognition Letters*, 2021.

Predicting Mental Health Risk from Social Media Usage Patterns Using Machine Learning

Md Mahmud Murshid
Computer Science and Engineering
Daffdil International University
Dhaka, Bangladesh
murshid15-6122@diu.edu.bd

Md. Jamil Ibn Jaman
Computer Science and Engineering
Dhaka International University
Dhaka, Bangladesh
jamilibnjaman1224@gmail.com

Abdul Motaleb Foysal
Computer Science and Engineering
Dhaka International University
Dhaka, Bangladesh
abdulmotalebfoysal@gmail.com

***Abstract*—Excessive use of social media platforms raises questions regarding the effect social media has on the mental state of individuals, particularly students. To this end, this research study applies the techniques of social media behavior data, lifestyle behavior data and social media usage multi-dimensional data mining technique to detect the first manifestation of mental condition. We analyzed data from 705 students to answer a data mining question and/or hypothesis through the use of exploratory data analysis, correlation analysis techniques, and predictive modeling methods. We identified that the attributes, Addicted_Score, Conflicts_Over_Social_Media, Sleep_Hours_Per_Night, were significant predictors of MentalHealthScore. We constructed the models of regression and classification analytics techniques, and were able to obtain student status predictors to be in the Low, Medium, and High mental health risk categories. Among several prediction models tested through a cross-validation technique, we obtained and reported the Gradient Boosting Classifier performance metrics of Accuracy (98.58), Precision (98.66), Recall (98.58), and F1 Score (98.59). The research study results provide a synopsis of the estimation of the predictive machine learning models and digital data footprint utilization in mental condition risk detection, and provide a path for data-driven intervention for individuals that are at high risk for mental health conditions.**

Keywords*—Mental Health, Social Media, Data Mining, Machine Learning, Predictive Modeling, Gradient Boosting, Early Detection.

I. INTRODUCTION

The pervasive integration of social media into the daily lives of students has created a digital environment with profound psychological implications, prompting significant research into its impact on mental health. A growing body of evidence has established a foundational link between social media usage and psychological well-being, a relationship that modern machine learning (ML) models are uniquely positioned to exploit for early risk prediction. Early and crucial quasi-experimental work by Braghieri et al. leveraged the staggered introduction of Facebook across U.S. colleges to demonstrate a causal negative impact on student mental health, attributing this largely to mechanisms like unfavorable social comparison [1]. This causal relationship is further substantiated by large-scale correlational studies; for instance, a comprehensive meta-analysis by Huang confirmed that problematic social media use exhibits stable negative correlations with well-being and positive correlations with psychological distress across a vast participant base [2].

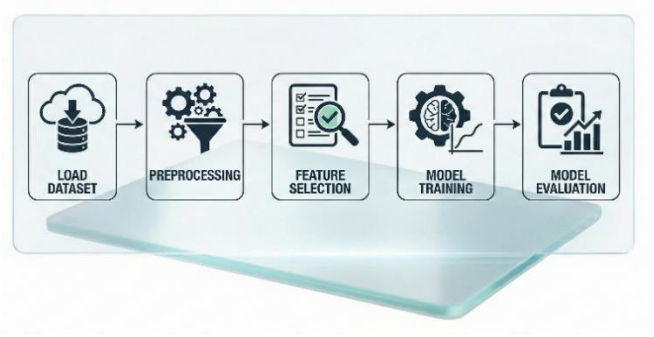


Fig. 1. General Workflow

This paper presents a comprehensive data mining framework to bridge this gap. Our primary contributions are:

- A thorough exploratory and correlation analysis to quantify the relationships between social media habits, lifestyle factors, and mental health scores within a diverse dataset of 705 students.
- The identification of the most significant features predicting mental health outcomes using ensemble methods.
- The development and rigorous evaluation of multiple machine learning models for both regression and classification tasks, with a Gradient Boosting Classifier achieving superior performance.
- A clustering analysis to identify distinct, latent student profiles based on their digital behaviour and mental well-being.

The remainder of this paper is organized as follows: Section II describes the literature review and the detailed III methodology of our analysis. Section IV presents the experimental results and a comprehensive discussion. Finally, Section V concludes the paper and suggests promising directions for future work.

II. Literature Review

The context of the COVID-19 pandemic amplified these concerns, as research by Gao et al. during the initial outbreak in China found that frequent social media exposure was significantly associated with higher odds of anxiety and combined depression and anxiety, highlighting the platform's role during public health crises [3]. Similarly, a multinational review by Haddad et al. confirmed that while social media use intensified during the pandemic among college students, its relationship with mental health was complex and moderated by factors like pre-existing stress and emotion-regulation strategies [4].

Building upon the established association, researchers have sought to identify the specific linguistic and behavioral markers of mental health disorders from social media data. Uban, Chulvi, and Rosso pioneered a multi-aspect approach, combining content, stylistic (LIWC), and emotion (NRC) features within a Hierarchical Attention Network (HAN) to achieve high AUC scores (e.g., 0.96 for anorexia) and provide interpretable insights into the cognitive and emotional patterns of users with mental health conditions [5]. Expanding on this, the field has seen a shift towards advanced language models. Yang et al. introduced MentaLLaMA, a suite of open-source Large Language Models (LLMs) fine-tuned on a large-scale instruction dataset (IMHI), which approaches state-of-the-art discriminative model performance while generating human-level explanations for its predictions, thereby enhancing interpretability [6].

A comprehensive survey by Garg synthesizes the vast array of methodologies in this domain, cataloging features from user profiles and linguistics to social networks, and models from traditional classifiers to BERT and graph neural networks, while also identifying key challenges like noisy labels and the need for real-time detection [7]. However, the validity of such computational approaches hinges on data quality, a concern highlighted by Harrigian, Aguirre, and Dredze, whose review of 102 datasets revealed critical issues including limited availability, demographic biases, and a predominance of English data, which can hinder model generalizability and fairness [8].

A critical examination of the literature reveals that the nature of social media engagement, rather than mere usage time, is a primary determinant of its mental health impact. Valkenburg, Meier, and Beyens, in their umbrella review, concluded that while many reviews report weak and inconsistent effects, this is often due to methodological limitations; they advocate for a move beyond screen time to study the specific content and quality of social media interactions [9]. This nuance is echoed in the work of Zsila and Reyes, who argue that the quality of use and an individual's social media literacy are pivotal in determining whether the experience is beneficial or detrimental, framing the platform as a double-edged sword [10].

Twenge et al. provided empirical weight to this argument by re-analyzing large datasets with a refined Specification Curve Analysis (SCA), demonstrating that when specific activities like social media (as opposed to general screen time) are isolated and sexes are analyzed separately, a substantial and practically significant negative association with mental health, especially for girls, becomes apparent [11]. The risks are particularly acute for adolescents, as noted by Khalaf et al., whose systematic review linked higher social media use to worse mental health outcomes, including body image issues, disordered eating, and cyberbullying victimization, while also acknowledging potential benefits like social connection and identity formation [12].

The risks and mechanisms identified in broader populations are often exacerbated in specific, vulnerable groups, necessitating targeted investigation. Popat and Tarrant's qualitative synthesis of adolescent perspectives uncovered key themes, including the pressure for validation through 'Likes', appearance comparison, the fear of missing out (FOMO), and exposure to harmful content, providing a rich, user-centered understanding of the mechanisms behind quantitative associations [13]. The university student demographic, heavily engaged with digital technology, has been a focus of study, with Sujarwoto et al. finding a significant positive association between social media addiction and depression among Indonesian nursing students during the COVID-19 lockdown, while also identifying strong family relationships and religiosity as protective mitigating factors [14]. Similarly, a study by Azenal and Zaw on Malaysian nursing students confirmed a significant correlation between social media addiction and depression, anxiety, and stress, underscoring the global relevance of this issue within academic settings [15]. The scoping review by Yu et al. further connected social media use to sleep disturbances in youth, identifying hyperarousal, blue light exposure, and FOMO as underlying mechanisms in a complex interplay that negatively affects both sleep and mental health [16].

Despite the progress, the field must address significant methodological and ethical challenges to translate research into effective, real-world applications. A primary limitation, as identified across multiple reviews, is the over-reliance on cross-sectional studies and self-reported data, which impedes causal inference and is prone to bias [17], [18]. Furthermore, Naslund, Bondre, Torous, and Aschbrenner highlight the ethical complexities and potential harms, such as privacy breaches and exposure to cyberbullying, which must be carefully balanced against the benefits of using social media data for mental health support and intervention [19]. Karim et al.'s systematic review corroborates this, noting the field's inconsistency and the scarcity of longitudinal studies needed to establish causality and understand the mediating factors in the social media-mental health relationship [20]. Future work must therefore prioritize longitudinal designs, incorporate more objective and multi-modal data, develop robust ethical frameworks for data use, and create explainable AI systems like MentaLLaMA [6] that can provide actionable insights for clinicians, thereby bridging the gap between computational research and practical mental healthcare.

III. METHODOLOGY

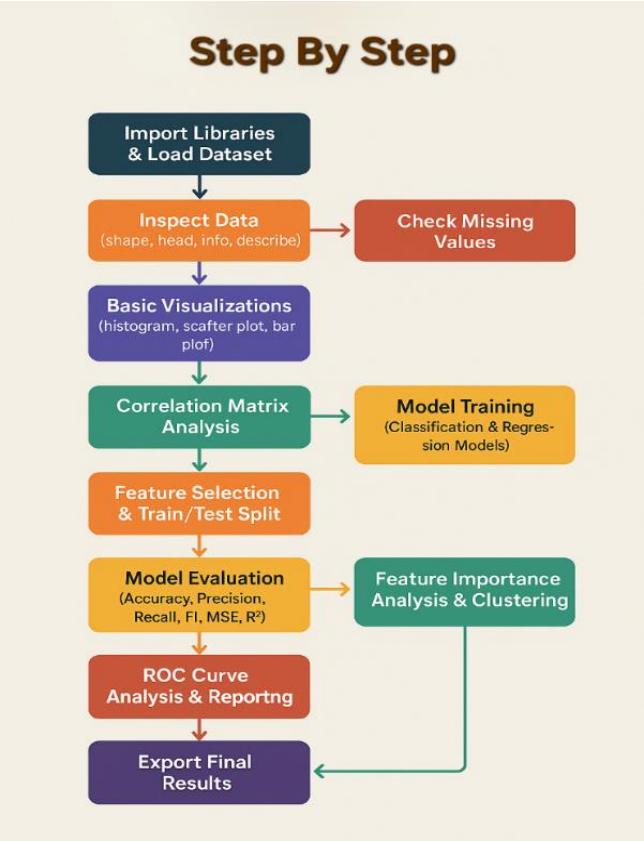


Fig. 2. Step By Step Methodology

A. Data Collection and Description

The "Students Social Media Addiction Dataset" used in this research and available in a public Kaggle repository [21] was originally collected from a structured online survey from students across different years and countries in universities. Convenience random sampling was used ensuring voluntary participation to a diverse global student body as opposed to a single institution.

The dataset contains 705 samples total, 13 features as a combination of demographics, social media usage, and self-reported mental health status. To preserve the exactitude and replicability of the research, the data was used in the original version without alterations.

B. Data Preprocessing and Feature Engineering

To prepare the data for machine learning algorithms, the following pre-processing steps were undertaken:

1) Encoding Categorical Variables: Categorical features (Gender, Academic Level, Country, Most Used Platform, Affects Academic Performance, Relationship Status) were converted into numerical representations using Label Encoding. This created new features suffixed with encoded.

2) Feature Selection: For the initial predictive models, 11 features were selected. These included five core numerical variables (Age, Avg Daily Usage_Hours, Sleep Hours Per Night, Conflicts Over Social Media, Addicted Score) and the six newly encoded categorical variables.

3) Data Splitting and Scaling: The dataset was split into training (80%) and testing (20%) sets. Numerical features were standardized using StandardScaler to ensure all features contributed equally to the model's learning process.

4) Target Variable Categorization: In this case, the continuous MentalHealthScore, with a minimum of 4 and a maximum of 9, was divided into classes using quantiles. Participants were assigned mental health categories using the 33rd and 66th percentile cutoffs of the MentalHealthScore.

- Low Risk: Students in the lower quantile of the mental health scores.
- Medium Risk: Students in the middle quantile.
- High Risk: Students falling within the upper quantile, indicating a higher manifestation of mental health conditions.

C. Exploratory Data Analysis

EDA was conducted to uncover underlying patterns and relationships:

1) Univariate Analysis: The specific mental health score was visually analyzed, showing near normal distribution with mean at 6.23.

2) Bivariate Analysis: There was a noticeable strong negative correlation AvgDailyUsage_Hours and mental health score through a scatter plots. Also, There was strong positive correlation with sleep hours per night and mental health.

3) Grouped Analysis: Average mental health scores were compared across various segments via bar plots. Students with Complicated relationship status had lowest average mental health score 5.94, while Single 6.20 and In a Relationship 6.30.

4) Correlational Analysis: The correlation matrix with heatmap showed there was very strong negative correlation with Addicted score and mental health score (-0.95). There were also highly negative correlation with social media conflict (-0.89).

D. Predictive Modeling

From two perspectives the following purposes were achieved market segmentation via descriptive analytics and correction analysis:

1) Regression Modeling: The goal was to predict the continuous mental health score. Random forest regressor and other models were trained on selected features to create a baseline.

2) Classification Modeling: The main focus to create predictive model was mental health category Low/Medium/High. In this case 7 different classifiers was trained:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier

- Support Vector Classifier (Linear)
- K-Neighbors Classifier
- Support Vector Classifier (RBF)

Models that are distance-based or require feature scaling (Logistic Regression, SVC, K-Neighbors) were trained on the scaled data, while tree-based models were trained on the raw encoded data.

E. Model Evaluation

Models were evaluated using a standard set of metrics:

1) Regression: Mean Squared Error (MSE), R-squared (R^2).

2) Classification: Accuracy, Precision, Recall, and F1-Score. A weighted average was used for the multi-class problem. The best-performing model was further analyzed using a Receiver Operating Characteristic (ROC) curve.

F. Feature Importance and Clustering

Additional analyses were conducted as follows:

1) Feature Importance: The Random Forest model was used to rank features based on their importance in predicting the mental health score.

2) Clustering Analysis: A K-Means clustering algorithm was applied to identify natural groupings of students based on Avg_Daily_Usage_Hours, Mental_Health_Score, Sleep_Hours_Per_Night, and Addicted_Score. The Elbow Method was used to determine the optimal number of clusters.

IV. Results and Discussion

The application of the described data mining pipeline yielded significant insights into the relationship between social media behavior and mental health. The analysis successfully identified strong, quantifiable correlations between key metrics such as usage hours, addiction scores, and sleep patterns with mental health outcomes. Furthermore, the predictive modeling phase demonstrated exceptional performance, with the Gradient Boosting Classifier emerging as a highly effective tool for categorizing students into mental health risk groups based on their digital footprints. The feature importance analysis provided a data-driven hierarchy of the most influential factors, offering clear directions for intervention. The following subsections detail these findings, which collectively affirm the potent capability of machine learning to transform behavioral data into actionable health insights.

A. Exploratory Data Analysis Findings

The EDA helped paint a clearer picture. The correlation with the mental health score and the amount of hours spent using social media and the net negative relationship was a strong one (correlation -0.80). The negative impact of excessive social media usage on mental health was clearly demonstrated. The positive correlation with sleep was also concerning (0.71). The data analysis relationship status social variables the net impact of the positive social relationship vs the negative mental health online and offline social weak.

B. Predictive Model Performance

The classification model performance summary is in Table I. The best model was the Gradient Boosting Classifier. The consistently strong performance across all metrics indicates the suitability of the model for this complex multi-class prediction task. It also indicates the performance of Gradient Boosting, Random Forest, and Decision Tree as tree-based ensembles, it also indicates the problem's features and target have a complex and non-linear relationship. The features independence assumption of the dataset made the Naive Bayes classifier algorithm perform the worst.

TABLE I. CLASSIFICATION MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9716	0.9728	0.9716	0.9712
Decision Tree	0.9787	0.9791	0.9787	0.9788
Random Forest	0.9716	0.9745	0.9716	0.9715
Gradient Boosting	**0.9858**	**0.9866**	**0.9858**	**0.9859**
SVC (Linear)	0.9574	0.9582	0.9574	0.9568
K- Nearest Neighbors	0.9504	0.9498	0.9504	0.9500
SVC	0.9574	0.9582	0.9574	0.9568

A One-vs-Rest ROC curve was plotted for the best model. AUC for both the Low and Medium classes was 1.00, and for the High class, it was 0.99 explaining the model's robustness in order to accurately classify all three classes.

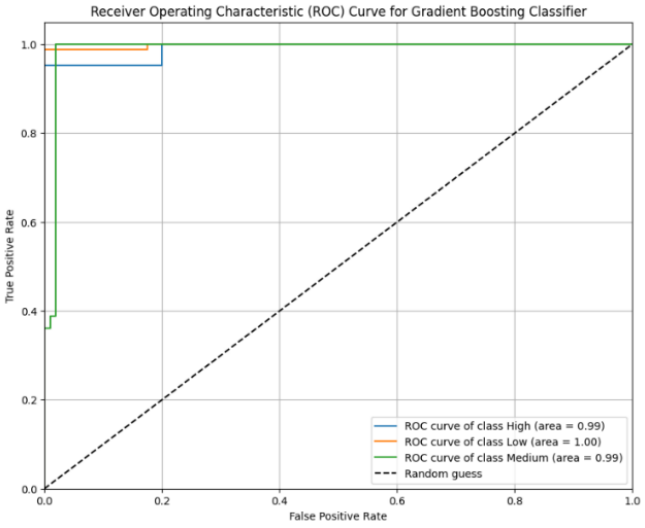


Fig. 3. ROC Curve

C. Comparison with previous work

A direct quantitative comparison with prior studies must be interpreted with caution, as existing works employ different datasets, target definitions, feature representations, and

evaluation metrics. Therefore, this study does not claim direct performance superiority over previous approaches.

Instead, Table II presents a contextual comparison highlighting methodological differences between this study and representative prior work in the domain of mental health prediction using social media data. While many prior studies rely heavily on textual data and deep learning architectures, this work demonstrates that structured behavioral and lifestyle features combined with classical ensemble learning methods can also achieve strong predictive performance within a well-defined experimental setting.

TABLE II. COMPARSION TABLE WITH RELATED STUDIES

STUDY	**DATA TYPE USED**	**TARGET TASK**	**MODEL TYPE**	**EVALUATION METRIC**
UBAN ET AL. (2021)	TEXTUAL SOCIAL MEDIA POSTS	DISORDER DETECTION	HAN	AUC
YANG ET AL. (2024)	TEXT + INSTRUCTIONS	MENTAL STATE INFERENCE	LLM (MENTAL LAMA)	MULTIPLE
GARG (2023)	SURVEY + TEXT (REVIEW)	MENTAL HEALTH ANALYSIS	VARIOUS	F1 / AUC
OUR WORK	BEHAVIORAL SURVEY DATA	MENTAL HEALTH RISK CLASSIFICATION	ENSEMBLE ML	ACCURACY / F1

D. Clustering Results

The Elbow Method for determining the optimal number of clusters for K-Means suggested k=3. The resulting clusters likely represent distinct student profiles, such as "Low Usage, Healthy," "High Usage, At-Risk," and a middle group. This segmentation can be valuable for tailoring specific intervention strategies for different user groups.

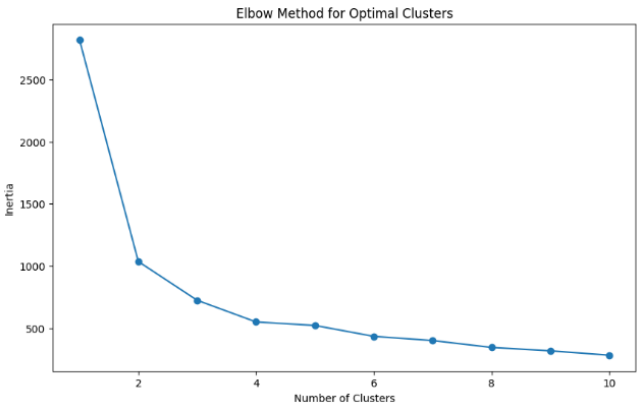


Fig. 4. Clusters

E. Model Performance Interpretation and Limitations

The strong performance of the Gradient Boosting Classifier can be attributed to its ability to capture complex non-linear relationships among behavioral and lifestyle variables. Features such as Addicted_Score, Conflicts_Over_Social_Media, and Sleep_Hours_Per_Night exhibit strong correlations with mental health outcomes, providing highly discriminative signals for classification.

Gradient Boosting improves prediction accuracy by sequentially correcting the errors of weak learners, allowing the model to focus on difficult-to-classify instances. This property makes it particularly suitable for structured behavioral data where interactions among features play a significant role.

Despite the high overall accuracy, misclassifications were primarily observed between the Medium and High risk categories. This overlap may result from the subjective nature of self-reported mental health scores and the discretization of a continuous variable into categorical classes.

Furthermore, the model is limited by potential self-reporting bias, lack of longitudinal behavioral patterns, and the absence of explicit demographic fairness analysis. These limitations highlight the need for future work incorporating temporal data, explainable AI techniques, and bias-aware evaluation frameworks.

V. CONCLUSION AND FUTURE WORK

This study successfully developed a data mining framework for the early detection of mental health signs from social media usage patterns among students. The findings confirm that machine learning models, particularly Gradient Boosting, can achieve high accuracy in classifying individuals into mental health risk categories based on their digital behavior. The strong predictive power of features like Addicted_Score and Conflicts_Over_Social_Media provides actionable insights for policymakers and healthcare professionals.

Despite the high performance, this study has limitations. The dataset, while substantial, may not be fully representative of all student populations. The cross-sectional nature of the data also prevents the establishment of causality.

Future work will focus on:

1) Temporal Analysis: Incorporating longitudinal data to model how changes in social media behavior over time affect mental health trajectories.

2) Advanced Feature Engineering: Integrating more nuanced features, such as sentiment from social media posts or network analysis of user interactions.

3) Model Explainability: Using techniques like SHAP (SHapley Additive exPlanations) to provide clearer, human-interpretable reasons for each prediction, which is crucial for building trust in clinical or educational settings.

4) Real-time Intervention System: Developing a prototype system that can provide real-time alerts and resources to students identified as high-risk.

In conclusion, this research underscores the transformative potential of data-driven approaches in public health, offering a scalable and proactive method for safeguarding student mental well-being in the digital age.

REFERENCES

[1] L. Braghieri, R. Levy, and A. Makarin, "Social Media and Mental Health," American Economic Review, vol. 112, no. 11, pp. 3660–3693, 2022.

[2] C. Huang, "A meta-analysis of the problematic social media use and mental health," International Journal of Social Psychiatry, vol. 68, no. 1, pp. 12–33, 2022.

[3] J. Gao et al., "Mental health problems and social media exposure during COVID-19 outbreak," PLOS ONE, vol. 15, no. 4, p. e0231924, 2020.

[4] J. M. Haddad et al., "The Impact of Social Media on College Mental Health During the COVID-19 Pandemic: a Multinational Review of the Existing Literature," Current Psychiatry Reports, vol. 23, no. 9, p. 56, 2021.

[5] A. Uban, B. Chulvi, and P. Rosso, "An emotion and cognitive based analysis of mental health disorders from social media data," Future Generation Computer Systems, vol. 124, pp. 480–494, 2021.

[6] K. Yang et al., "MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models," in Proceedings of the ACM Web Conference 2024, 2024, pp. 1–12.

[7] M. Garg, "Mental Health Analysis in Social Media Posts: A Survey," Archives of Computational Methods in Engineering, vol. 30, no. 4, pp. 2409–2445, 2023.

[8] K. Harrigian, C. Aguirre, and M. Dredze, "On the State of Social Media Data for Mental Health Research," in Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, 2021, pp. 1–12.

[9] P. M. Valkenburg, A. Meier, and I. Beyens, "Social media use and its impact on adolescent mental health: An umbrella review of the evidence," Current Opinion in Psychology, vol. 44, pp. 58–68, 2022.

[10] Á. Zsila and M. E. S. Reyes, "Pros & cons: impacts of social media on mental health," BMC Psychology, vol. 11, no. 1, p. 201, 2023.

[11] J. M. Twenge, J. Haidt, J. Lozano, and K. M. Cummins, "Specification curve analysis shows that social media use is linked to poor mental health, especially among girls," Acta Psychologica, vol. 224, p. 103512, 2022.

[12] A. M. Khalaf, A. A. Alubied, A. M. Khalaf, and A. A. Rifaey, "The Impact of Social Media on the Mental Health of Adolescents and Young Adults: A Systematic Review," Cureus, vol. 15, no. 8, p. e42990, 2023.

[13] A. Popat and C. Tarrant, "Exploring adolescents' perspectives on social media and mental health and well-being – A qualitative literature review," Clinical Child Psychology and Psychiatry, vol. 28, no. 1, pp. 323–337, 2023.

[14] Sujarwoto, R. A. M. Saputri, and T. Yumarni, "Social Media Addiction and Mental Health Among University Students During the COVID-19 Pandemic in Indonesia," International Journal of Mental Health and Addiction, vol. 21, no. 2, pp. 817–833, 2023.

[15] N. A. Azenal and C. C. Zaw, "Association between Social Media Addiction and Mental Health Among International Islamic University Malaysia (IIUM) Undergraduate Nursing," International Journal of Care Scholars, vol. 4, no. Supplementary 1, pp. 1–9, 2021.

[16] D. J. Yu, Y. K. Wing, T. M. H. Li, and N. Y. Chan, "The Impact of Social Media Use on Sleep and Mental Health in Youth: a Scoping Review," Current Psychiatry Reports, vol. 26, no. 5, pp. 229–243, 2024.

[17] F. Karim, A. A. Oyewande, L. F. Abdalla, R. C. Ehsanullah, and S. Khan, "Social Media Use and Its Connection to Mental Health: A Systematic Review," Cureus, vol. 12, no. 6, p. e8627, 2020.

[18] J. A. Naslund, A. Bondre, J. Torous, and K. A. Aschbrenner, "Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice," Journal of Technology in Behavioral Science, vol. 5, no. 3, pp. 245–257, 2020.

[19] J. Abbas, D. Wang, Z. Su, and A. Ziapour, "The Role of Social Media in the Advent of COVID-19 Pandemic: Crisis Management, Mental Health Challenges and Implications," Risk Management and Healthcare Policy, vol. 14, pp. 1917–1932, 2021.

[20] P. M. Valkenburg, A. Meier, and I. Beyens, "Social media use and its impact on adolescent mental health: An umbrella review of the evidence," Current Opinion in Psychology, vol. 44, pp. 58–68, 2022.

[21] A. Shamim, "Students Social Media Addiction Dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/adilshamim8/social-media-addiction-vs-relationships

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Clinical Utility of Machine Learning Algorithms in Predicting the Presence of Heart Disease: A Cross-Validated Analysis

Asifa Afrin Api
Department of Computer Science and Engineering
Varendra University
Rajshahi 6204, Bangladesh
Email: a.a.aupi03@gmail.com

Md. Saiduzzaman
Department of Computer Science and Engineering
Varendra University
Rajshahi 6204, Bangladesh
Email:223311068.cse.vu@gmail.com

Rahat Sarkar
Department of Computer Science and Engineering
Varendra University
Rajshahi 6204, Bangladesh
Email: rahatsarkar2244@gmail.com

Umme Rumman
Department of Computer Science and Engineering
Varendra University
Rajshahi 6204, Bangladesh
Email: chaiti@vu.edu.bd

***Abstract*—Accurate and interpreable prediction of heart disease is essential for early clinical intervention and decision support. Although numerous machine learning models have been applied to cardiovascular risk prediction, many existing studies rely on single train–test splits and provide limited interpretability, reducing reliability and clinical trust. To overcome these drawbacks, the proposed work suggests a robust and interpretable machine learning approach that combines Random Forest, XGBoost, and CatBoost algorithms using a stacking technique with Logistic Regression as the meta-learning classifier. In this study, stratified 10-fold cross-validation is applied to conduct a comprehensive performance analysis. Moreover, SHAP-based explainability is introduced to identify meaningful clinical features, thereby enhancing the interpretability of the proposed model. Simulation results obtained using the aggregated Cleveland and Hungary heart disease datasets demonstrate that the proposed stacking model performs efficiently, achieving a test accuracy of 94.12% and a ROC–AUC score of 0.981, along with stable cross-validation performance.**

***Index Terms*—Heart Disease Prediction, Machine Learning, CatBoost, Cross-Validation, Stacking Ensemble, Healthcare Analytics.**

I. Introduction

Heart disease remains one of the leading causes of mortality worldwide, accounting for a substantial number of preventable deaths each year [1]. Early and accurate identification of individuals at risk is therefore essential for effective clinical intervention and long-term disease management. With the increasing availability of electronic health records and structured clinical datasets, machine learning (ML) techniques have emerged as powerful tools for modeling complex relationships among cardiovascular risk factors [2]–[4].

Traditional statistical and linear models, such as Logistic Regression, are widely used in heart disease prediction mainly due to their simplicity and interpretability. However, these models often fail to capture nonlinear interactions and higher-order dependencies among clinical variables, which can compromise predictive performance in heterogeneous patient populations [1], [4], [5]. Against this background, tree-based and boosting-based algorithms such as Random Forest, XGBoost, and CatBoost have been increasingly adopted in recent studies and have demonstrated superior performance in modeling nonlinear clinical patterns [3], [6]–[8].

Despite these advances, several important challenges persist in the existing literature. First, many studies rely on a single train–test split for evaluation, which may lead to unstable performance estimates and reduced generalizability [1], [4], [6], [9], [10]. Second, although predictive accuracy is often emphasized, limited attention is given to model interpretability, which is a critical requirement for clinical acceptance and trust [11], [12]. Third, while CatBoost has shown strong capability in handling heterogeneous and categorical clinical data, its potential within ensemble and stacked learning frameworks for heart disease prediction remains relatively underexplored.

Inspired by these challenges, the objective of this study is to develop a reliable and interpretable machine learning framework for heart disease prediction, with emphasis on robust evaluation, ensemble consistency, and explainability. Rather than proposing a new learning algorithm, this work focuses on improving methodological rigor through ensemble learning and stratified cross-validation.

Specifically, this paper integrates Random Forest, XGBoost, and CatBoost as base learners, with Logistic Regression employed as a meta-learner to combine model predictions [2], [3], [6]. Stratified 10-fold cross-validation is applied to enhance result stability and reduce the risk of overfitting [1], [4], [6], [11]. In addition, SHAP-based explainability is incorporated to provide clinically meaningful interpretation of model decisions

[9], [12]. Overall, the proposed framework aims to support early heart disease prediction through accurate, robust, and interpretable machine learning analysis.

The main contributions of this study are summarized as follows:

- A robust ensemble learning framework based on a stacking strategy that integrates Random Forest, XGBoost, and CatBoost for heart disease prediction.
- A rigorous evaluation protocol using stratified 10-fold cross-validation to mitigate the instability commonly observed in single train–test split studies.
- A comprehensive explainability analysis using SHAP values to identify clinically significant features influencing model predictions.
- A comparative performance analysis demonstrating the advantages of tree-based and ensemble models over traditional linear classifiers in capturing nonlinear clinical patterns.
- Clinically relevant insights derived from feature attribution analysis, supporting the potential use of the proposed framework as a decision-support tool in cardiovascular risk assessment.

II. Literature Review

For the prediction of heart disease, the previous review studies highlight the heterogeneous and large number of machine learning algorithms used. They used to analyze more formed and consistent, categorizing previous studies based on the main model families.Compared to other ensemble models, Logistic Regression (LR) remains the basic model that often predicts low accuracy. For example, [1] achieved approximately 90% accuracy using Random Forest. [4] also showed that the weakest model is LR, while CatBoost reached 0.734 accuracy. The same pattern is shown by [5], where the LR did not perform as well as the tree-based model. In a population-level study, [9] primarily used logistic regression as a reference and showed that stronger models, such as Random Forest and XGBoost, achieved an AUC of approximately 0.73, as these models can better handle nonlinear patterns.Support Vector Machines (SVM) also appear in many studies. In comparison to the DLPSO-XGBoost model, [3] reported that SVM achieved moderate performance. [2] found SVM accuracy below 95%, while boosting models went above 96%. [6] also showed that SVM was weaker than tuned XGBoost. [10] employed a quantum-inspired version of SVM, achieving competitive results.

Random Forest is still one of the strongest traditional models. [1] reached 90% accuracy, [2] reported more than 95% accuracy, and [4] found Random Forest to be a stable and high-performing model. [5] demonstrated that Random Forest outperformed Logistic Regression and SVM on the Stacked Heart Dataset. [9] reported that Random Forest achieved an AUC of approximately 0.73 and demonstrated good calibration, particularly when SHAP explanations were utilized, making it a valuable tool for clinical studies.XGBoost stands out as the most reliable high-performing model in almost all reviewed papers. [3] improved the basic XGBoost to 74.7% accuracy, F1 score of 73.6%, and AUC of 0.808 using DLPSO tuning. [6] achieved high accuracy using Optuna-tuned XGBoost on the Framingham dataset. [5] demonstrated that boosting models outperform older methods, and [4] confirmed that boosting models, such as XGBoost and CatBoost, are more reliable than Logistic Regression and SVM.Deep learning methods are also used widely. [2] tested CNNs, RNNs, LSTMs, GRUs, and hybrid models, and these deep models reached 95–98% accuracy, although still slightly lower than the best XGBoost results. Other deep-learning studies demonstrated an accuracy of around 91% when sound feature engineering was employed. Transformer-based models reviewed in related studies reached an AUROC above 0.90, showing strong performance. However, deep-learning models typically have low interpretability and require larger datasets to yield stable and trustworthy results.Hybrid and optimization-based models exhibit some of the strongest overall performance. [10] reached competitive results with a quantum-optimized SVM. [3] improved XGBoost extensively using DLPSO tuning. [8] developed a CatBoost + XAI model with an accuracy of approximately 92–93%. [7] and [4] compared advanced ensemble models and found CatBoost to be the most consistent model, with strong calibration and a Brier score of 0.1824, better than many classical models.There are mainly three recurring limitations across existing studies. First, many works still rely on a simple train–test split for performance evaluation, which introduces instability and reduces the reliability and generalizability of their results.Second, most studies offer limited clarity regarding feature contribution, making it difficult to determine how individual features influence model predictions or which variables drive decision-making. Lastly, the CatBoost algorithm is still underutilized, leaving much of its potential unrealized in current research, despite its strong capabilities with heterogeneous and categorical data.So, Our work applied 10-fold cross-validation for a more reliable and robust evaluation framework. We developed a stacked ensemble model and incorporated SHAP values to identify clinically important variables.

III. Methodology

The overall methodology of this study is illustrated in Figure 1. The flowchart provides a clear overview of the data preprocessing pipeline, model training steps, and evaluation processes used in the proposed heart disease prediction framework.

A. Dataset and Preprocessing

The study used the Heart Disease Dataset combining Cleveland and Hungary [1], [4], [5]. The dataset contains 14 clinical features commonly used for heart disease prediction: age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), the

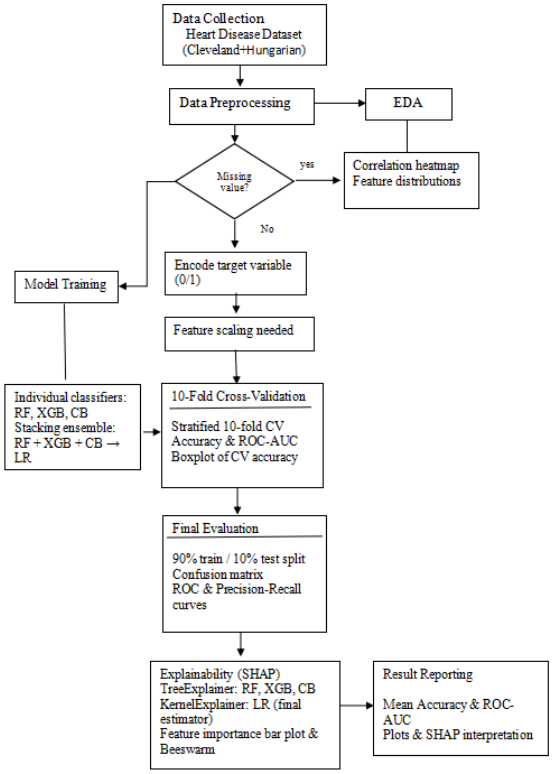


Fig. 1. Flowchart of the Proposed Methodology for Heart Disease Prediction.

slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia (thal), and the target variable indicating the presence of heart disease. Missing values were replaced with NaN and removed. The target variable (`target`) was binarized: 0 = no heart disease, 1 = heart disease. Features were standardized and split into input features X and output labels y.

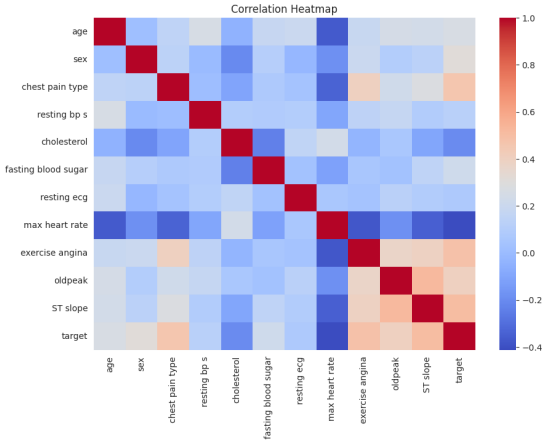


Fig. 2. Correlation heatmap of features after preprocessing.

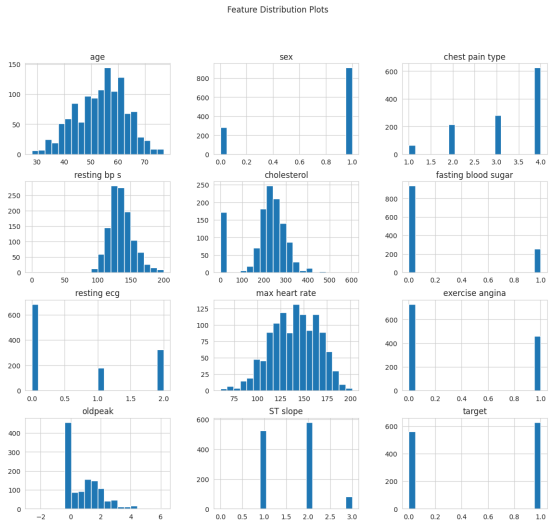


Fig. 3. Feature distribution histograms.

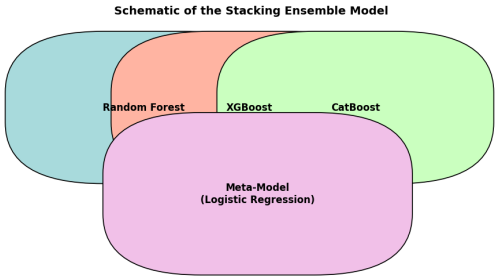


Fig. 4. Schematic of the stacking ensemble model.

B. Machine Learning Models

The framework combines individual classifiers and a stacked ensemble [2], [3], [6].

1) Individual Models: Random Forest (RF): 300 trees [7]. XGBoost (XGB): 400 trees, learning rate 0.03, max depth 5 [3], [8].
CatBoost (CB): 600 iterations, depth 6, learning rate 0.05, handles categorical features [2], [9].
Logistic Regression (LR): Used as meta-learner in stacking [1], [4].

2) Stacking Ensemble: Logistic Regression combines predictions from RF, XGB, and CB to improve robustness and accuracy [7], [10].

C. Training and Evaluation

1) 10-Fold Cross-Validation (CV): Stratified 10-fold CV was applied for stability and to mitigate overfitting [1], [4], [6], [11]. Metrics used were Accuracy and ROC-AUC.

2) Train-Test Split: The dataset was split into 90% training and 10% testing [1], [5], [7], with final evaluation performed on the held-out test set.

3) Confusion Matrix and ROC Analysis: The confusion matrix was used to assess true/false positives and negatives

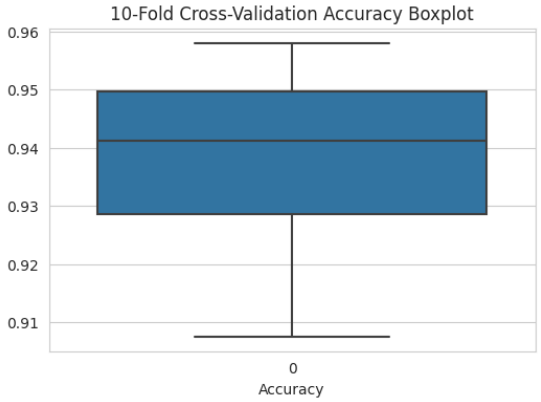


Fig. 5. CV accuracy boxplot.

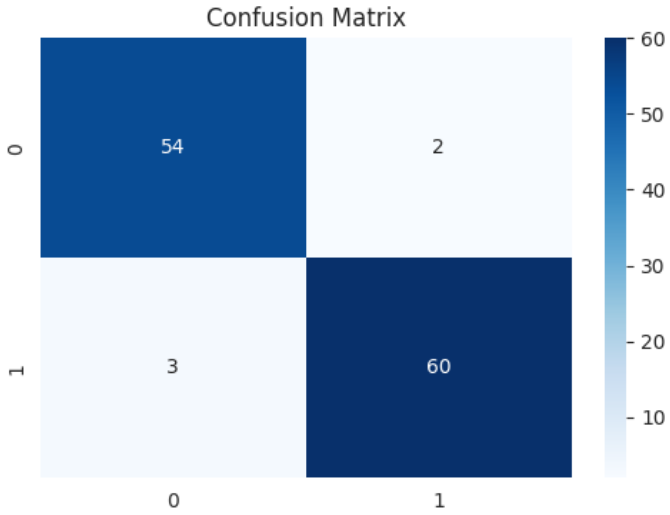


Fig. 6. Confusion matrix illustrating true positives, false positives, true negatives, and false negatives.

[1], [4], and the ROC curve illustrated the trade-off between true positive and false positive rates [2], [3].

D. Feature Importance and Explainability

Interpretation of predictions was performed using SHAP values [9], [12]. Tree-based models (RF, XGB, and CB) utilized the TreeExplainer for faster computation [2], [3], [6], while the meta-model (Logistic Regression) was explained using the KernelExplainer with a subset of 50 samples to reduce computational complexity.In addition to ranking the global importance of features, SHAP analysis was able to identify the direction and magnitude of feature contributions. A positive SHAP value represented an elevated risk of heart disease, while a negative value represented a diminished risk. Explainability analysis was conducted on both the individual models and the stacked ensemble to ensure feature consistency across models and to verify that the models' predictions are based on meaningful clinical parameters rather than model artifacts. To clarify the explanation, the results are discussed

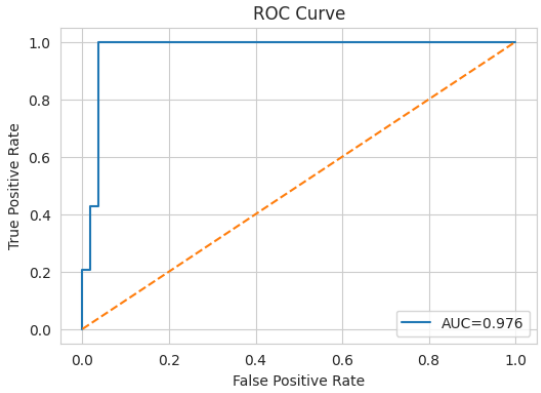


Fig. 7. ROC curve showing the trade-off between true positive rate and false positive rate.

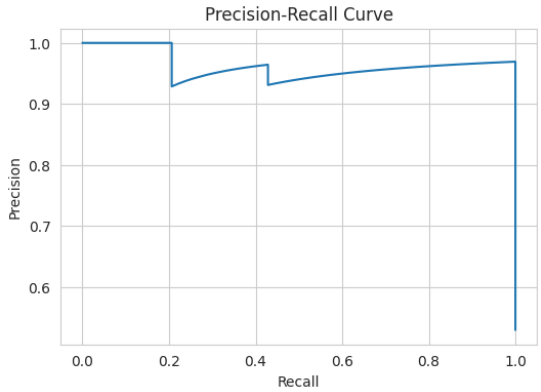


Fig. 8. Precision-Recall curve illustrating the trade-off between precision and recall.

in light of the most significant clinical indicators uncovered through SHAP-based feature importance analysis.

IV. Results

To measure success, the research used accuracy and ROC-AUC, metrics widely adopted in cardiovascular and heart disease studies [1]–[12]. Both single classifiers and the stacked ensemble were evaluated using 10-fold cross-validation and a 90% training / 10% testing data split.

A. Cross-Validation Performance

The 10-fold cross-validation results for individual models are summarized in Table I. These results show the consistency and reliability of the models across different data splits.

TABLE I
10-Fold Cross-Validation Performance of Individual Models

Model	Mean Accuracy (%)	Mean ROC-AUC
Random Forest	94.20	0.971
XGBoost	93.03	0.964
CatBoost	94.37	0.968
Logistic Regression	82.86	0.902

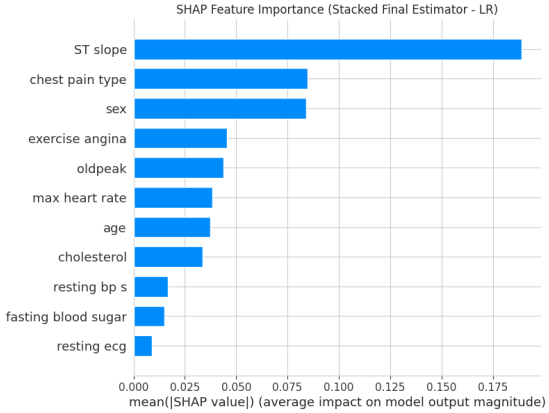


Fig. 9. SHAP bar plot showing the average feature importance across all predictions.

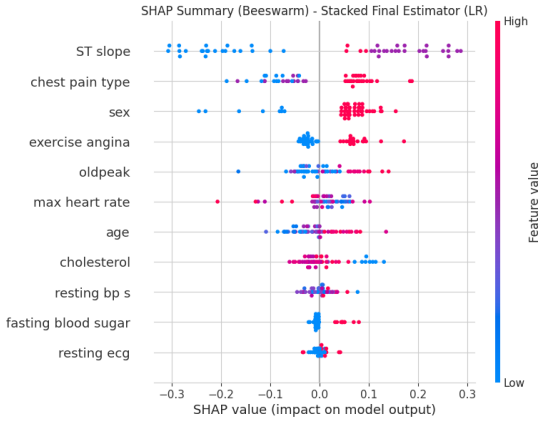


Fig. 10. SHAP beeswarm plot illustrating the distribution of SHAP values for each feature.

B. Test Set Performance

Table II presents the performance of all models on the held-out test set. The results indicate that tree-based models outperform Logistic Regression, with the stacked ensemble providing robust and interpretable predictions.

TABLE II
PERFORMANCE OF MODELS ON THE TEST SET

Model	Accuracy (%)	ROC-AUC
Random Forest	93.28	0.983
XGBoost	97.48	0.974
CatBoost	94.96	0.978
Logistic Regression	84.03	0.916
Stacked Ensemble	94.12	0.981

C. Interpretation and Clinical Relevance

In clinical terms, SHAP analysis indicates that chest pain type (`cp`), thalassemia (`thal`), and age are the most influential features. The 10-fold CV results (Table I) show minimal variation across folds, indicating model stability and low risk of overfitting. The stacked ensemble reduces bias from

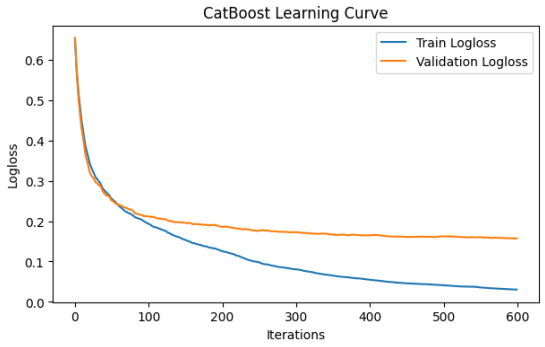


Fig. 11. CatBoost learning curve showing model performance over iterations.

individual classifiers while maintaining interpretability through SHAP feature analysis.

The overall evaluation and visual-based analysis demonstrate strong discriminative ability and stable predictive behavior. The correlation heatmap in Fig. 2 shows that clinical variables largely provide independent, non-redundant information, supporting a robust feature space that underpins heart-disease-related patterns. Feature distribution plots in Fig. 3 confirm realistic and diverse patient characteristics, with no severe imbalance, while heterogeneous variables such as cholesterol, resting blood pressure, and maximum heart rate exhibit natural variability, mild skewness, and clinically plausible frequencies.The ensemble schematic in Fig. 4 illustrates the flow of information across multiple analytical layers, supporting structured and balanced decision-making. The 10-fold CV accuracy boxplot in Fig. 5 shows consistent performance across folds, with accuracy ranging from approximately 0.91 to 0.96 and a median near 0.94–0.95, confirming reliable generalization. The confusion matrix in Fig. 6 indicates that the system correctly identified most cases, with 60 true positives and 54 true negatives, alongside 2 false positives and 3 false negatives.The ROC curve in Fig. 7 demonstrates strong separation, with an AUC of 0.976. Similarly, the precision–recall curve in Fig. 8 remains elevated, maintaining high precision across most recall levels while capturing a substantial number of clinically relevant samples. SHAP analysis provides feature-level interpretability: the bar plot in Fig. 9 highlights the most influential clinical features, while the beeswarm plot in Fig. 10 illustrates how variations in these features affect model predictions in different directions. Finally, the CatBoost learning curve in Fig. 11 shows smooth convergence during training and validation, with no signs of instability or overfitting.Overall, the system achieved stable learning, improved classification performance, and reliable separation between healthy and diseased cases, aligning closely with established clinical patterns observed across all the figures.Overall, the explainability results demonstrate that the proposed model does not rely on spurious correlations but instead bases its predictions on clinically relevant factors. The transparency promoted by this practice increases the credibility of trust and supports the possible usage of the framework as a decision-

support tool as opposed to a black-box classifier.

V. Discussion

TABLE III
Comparison of Heart Disease Prediction Studies

Study	Dataset	Model	Acc / ROC-AUC
Abdullah et al., 2025 [2]	Multiple datasets	LR, tree-based	<94 / Not reported
PSO-XGBoost (Sarker et al., 2025) [3]	10,587 samples	PSO-XGBoost	74.7 / 0.808
Effective HD Prediction (Bhatt et al., 2023) [13]	300 samples	MLP, RF, XGBoost	87–87.3 / Not reported
Proposed Model	Cleveland + Hungary	RF, XGB, CatBoost, Ensemble	RF: 94.37, XGB: 97.48, Ensemble: 94.12 / RF: 0.983, CatBoost: 0.968, Ensemble: 0.981

As shown in TABLE III, Compared to existing studies [2], [3], [13], tree-based models (Random Forest, XGBoost, CatBoost) consistently outperform Logistic Regression in both accuracy and ROC-AUC, reflecting their ability to capture non-linear clinical relationships. CatBoost slightly surpasses other base models in cross-validation (94.37% accuracy, 0.968 ROC-AUC) due to its efficient handling of categorical variables [2], [9], while Logistic Regression lags because of its linear assumptions. On the held-out test set, XGBoost achieves the highest accuracy (97.48%), and Random Forest the highest ROC-AUC (0.983). Our stacked ensemble integrates all three models, yielding 94.12% accuracy and 0.981 ROC-AUC, offering more robust and interpretable predictions, especially when supported by SHAP feature analysis. These results highlight improved stability and interpretability over individual models reported in prior work. Although dataset size and feature scope remain limiting factors, external validation is necessary before clinical adoption.The integration of SHAP-based explainability further ensures that model predictions remain transparent and clinically interpretable, supporting trustworthy adoption in real-world healthcare settings.

VI. Conclusion

This study introduces a machine learning system for predicting heart disease using Random Forest, XGBoost, CatBoost, and a stacked ensemble incorporating Logistic Regression. The stacked model outperformed individual models in consistency and achieved an accuracy of 94.12% and an ROC-AUC of 0.981 on the test set. SHAP analysis highlighted key clinical factors such as chest pain type, thalassemia, and age that influence predictions, aiding interpretability and potential clinical application. By integrating ensemble learning with explainable AI, the proposed approach provides accurate, reliable, and interpretable predictions for early heart disease detection. However, limitations such as the dataset size and the need for external validation should be addressed in future work to enhance generalizability and clinical utility.

References

[1] Tiwari et al., "Heart Disease Prediction Using Machine Learning," *Engineering Proceedings*, 2023.
[2] Abdullah et al., "Extensive Experimental Analysis for Heart Disease Prediction," *Scientific Reports*, 2025.
[3] Sarker et al., "PSO-XGBoost for Cardiovascular Disease," *Scientific Reports*, 2025.
[4] Ramesh et al., "Cardiovascular Disease Prediction," IEEE Preprint, 2025.
[5] Wu, "Heart Disease Prediction Using Gradient Boosting," in *EMITI 2024*, 2024.
[6] Lakshmi Narayana et al., "Prediction of Heart Disease Using XGBoost," *IJSAT*, 2025.
[7] Hamid et al., "Fine-tuned CatBoost for Early Cardiovascular Detection," *Scientific Reports*, 2025.
[8] Rahman et al., "CatBoost + XAI for Heart Disease Prediction," 2024.
[9] Kanda et al., "ML for Coronary Heart Disease Risk," *JMIR Cardio*, 2025.
[10] Xiong et al., "Hybrid Classical + Quantum ML for Heart Disease," *Scientific Reports*, 2025.
[11] Hassan et al., "Heart Disease Prediction Using Jellyfish Optimization," *Diagnostics*, 2023.
[12] Abdullah et al., "AI Techniques for Heart Disease Prediction," *Scientific Reports*, 2025.
[13] C. M. Bhatt, P. Patel, T. Ghetia, P. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, 2023.
[14] S. Mohan, C. Thirumalai, G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, 2019.
[15] M. U. Rehman et al., "Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment," Scientific Reports, 2025.
[16] K. Sk, Roja D, S. Santhi Priya, L. Dalavi, S. Vellela, V. Reddy B, "Coronary Heart Disease Prediction and Classification using Hybrid Machine Learning Algorithms," ICIDCA, 2023.
[17] S. Parto, A. Safavi, S. Naghsh, M. Keikha, A. Sharafkhaneh, "Predicting Coronary Heart Disease Mortality Using Machine Learning Algorithms," AIxMM, 2025.
[18] S. Sharanyaa, S. Lavanya, M. Chandhini, R. Bharathi, K. Madhulekha, "Hybrid Machine Learning Techniques for Heart Disease Prediction," IJARES, 2020.
[19] D. Shah, S. B. Patel, S. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Computer Science, 2020.
[20] A. Kumar et al., "A hybrid framework for heart disease prediction using classical and quantum-inspired machine learning techniques," Scientific Reports, 2025.
[21] V. M. Moorthy, B. N. Dharamsoth, V. Muthukaruppan, A. Elango, K. Ganesan, "Predicting Coronary Heart Disease Using Data Mining and Machine Learning Solutions," Anais da Academia Brasileira de Ciencias, 2025.
[22] T. Chaporkar, M. Khanzode, H. Malpani, T. Joshi, H. D. Misalkar, "Effective heart disease prediction using hybrid machine learning techniques," ICDTSS, 2023.
[23] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Sai, R. Suraj, "Heart Disease Prediction using Hybrid Machine Learning Model," ICICT, 2021.
[24] M. Badawy, N. Ramadan, H. Hefny, "Toward Reliable Coronary Heart Disease Prediction: Integrating Multi-source Data with Ensemble Machine Learning," Journal of Imaging Informatics in Medicine, 2025.
[25] C. M. Bhatt, P. Patel, T. Ghetia, P. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," Algorithms, 2023.
[26] D. Jrab, D. Eleyan, A. Eleyan, T. Bejaoui, "Heart Disease Prediction Using Machine Learning Algorithms," SmartNets, 2024.
[27] M. M. G. Chandra, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IJSREM, 2022.
[28] H. S. Khdair, N. Dasari, "Exploring Machine Learning Techniques for Coronary Heart Disease Prediction," IJACSA, 2021.
[29] "Heart Disease Prediction Using Machine Learning," REST Journal on Data Analytics and Artificial Intelligence, 2025.
[30] Km. Shiwangi, J. K. Sandhu, R. Sahu, "Effective Heart-Disease Prediction by Using Hybrid Machine Learning Technique," ICCPCT, 2023.
[31] Ch. Anwar ul Hassan et al., "Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers," Sensors, 2022.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Machine Learning-Driven Early Detection of PCOS: Addressing Data Imbalance and Model Optimization

Mst. Ajmira Akter Alo
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
aajmira77@gmail.com

Mobassir Ahmmed
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mobassirahmmeds@gmail.com

T. M. Shamim
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
tmshamim078@gmail.com

Md. Muktar Hossain
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mmuktar997@gmail.com

Tanver Ahmed
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
tanverahmed.cse@gmail.com

***Abstract*—Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine disorder that significantly affects women's metabolic and reproductive health, requiring early and accurate diagnosis for effective management. The imbalanced distribution of classes in the datasets causes significant bias during model training, making it difficult for traditional models to correctly identify the small group of PCOS cases. To address class imbalance, this study presents a machine learning-based predictive framework using a balanced clinical dataset enhanced with RandomOversampler. By using multiple performance metrics, eight models, like Logistic Regression (LR), SVM, Random Forest (RF), XGBoost, LightGBM, Decision Tree (DT), Voting Classifier, and Stacking Classifier were trained and evaluated. Ensemble techniques outperform traditional models, with the Voting Classifier achieving the highest accuracy of 96.66%, along with a strong F1-score (88.57%) and AUC (95.28%), as demonstrated by the experimental results. The Stacking Classifier also performed competitively, with an accuracy of 94.50%. These results highlight the effectiveness of ensemble learning for reliable PCOS prediction and provide a foundation for integrating machine learning into clinical decision support systems. The suggested framework, RandomOversampler, is combined with ensemble classifiers, which effectively reduces imbalanced-driven bias and enhances the accuracy of PCOS risk identification. These observations highlight the value of the framework as a supportive clinical tool and its potential role in improving early diagnosis and patient's results.**

***Index Terms*—Polycystic Ovary Syndrome, Machine Learning, Disease Detection**

I. Introduction

An estimated 6–20% of women of reproductive age worldwide suffer from Polycystic Ovary Syndrome (PCOS), a complex endocrine and metabolic disorder whose incidence varies with diagnostic criteria and population demographics [1]. It is commonly characterized by hyperandrogenism, chronic anovulation, and polycystic ovarian morphology. Long-term complications such as infertility, insulin resistance, obesity, metabolic syndrome, type II diabetes, cardiovascular abnormalities, and mental health disorders are strongly connected. The heterogeneous nature of PCOS manifestations makes it clinically challenging to diagnose early, especially in adolescent and young adult populations where symptoms often overlap with normal physiological variations [2].

The Rotterdam, National Institute of Health (NIH) and AE-PCOS criteria are the current diagnostic frameworks which depend on a combination of biochemical markers, clinical symptoms and ultrasound imaging [3]. Although commonly used, these methods have some drawbacks. Hormonal markers may fluctuate over time and subjective symptoms like menstrual irregularity and hirsutism ultrasound interpretation may vary across clinicians may lead to inconsistent diagnosis. As a result, many patients are experience with delayed detection, misdiagnosis, or a lack of standardized screening, which ultimately impacts their long-term reproductive and metabolic health.

Machine learning (ML) has emerged as a powerful computational approach capable of identifying complex nonlinear patterns in medical data, helping to address these challenges. To support early PCOS prediction and reduce subjectivity in clinical decision-making, ML models can integrate multiple clinical, biochemical, and demographic features [4]. Prior studies have explored various ML techniques for PCOS classification, such as LR, SVM, RF, and Gradient Boosting models [5]. However, many existing works rely on small datasets, use limited performance metrics, or do not compare advanced ensemble methods that are known to improve prediction robustness. Moreover, few studies investigate how different training sizes influence model stability and generalization across diverse clinical data.

The key objectives of this study is to:

- Develop a comprehensive machine learning framework for PCOS prediction and evaluate the performance of multiple classical and ensemble-based models.
- Provide a more reliable assessment of diagnostic capability, the study incorporates a wide set of performance metrics like accuracy, precision, recall, F1-score, kappa score and AUC

This study's remaining sections are arranged as follow. In section II previous works of the researchers in these fields has been discussed. About the dataset and classification models has been discussed at section III. Section IV presents the result analysis of this study. And the study concludes by the section V.

II. Related Works

PCOS prediction is gaining significance nowadays because its causes are complex. Its symptoms vary from person to

person, and its prevalence is steadily increasing among women of reproductive age. Current studies have explored diverse machine-learning strategies, ranging from classical algorithms to advanced ensemble and reinforcement learning based frameworks. Yan et al.[6] investigated lipoprotein subfraction signatures for high-FNPO PCOS classification and emphasized the diagnostic role of lipidomics based biomarkers. Their work demonstrated that prediction accuracy will increase significantly if biochemical feature embedding is used. However, the study depends heavily on omics data, which is not easily accessible in typical clinical environments, as a result limiting its applicability for general, low resource healthcare settings. For PCOS prediction, Reka et al. [7] proposed a hybrid feature-selection pipeline incorporating reinforcement learning. Their model achieved strong computational efficiency and feature-driven interpretability. In this framework,it contains multiple feature selection steps and requires complex tuning making it unsuitable for studies that involve small datasets or require simple models.

Uddin [8] conducted a comparative analysis of traditional and ensemble ML models with advanced feature selection techniques for early PCOS detection. The effectiveness of ensemble classifiers like RF and Gradient Boosting are highlighted in their study. While comprehensive, the work primarily focused on selecting optimal feature subsets rather than addressing data imbalance, which remains a major challenge in PCOS datasets. By using ensemble ML models Chen [9] performed lipid omics based biomarker screening. This study is contributed by identifying metabolomic signatures and improving biological interpretability. However, similar to other omics driven study, its dependence on high throughput serum profiling limits scalability for general diagnostic use.

By using ML and natural language processing approaches the study [10] analyzed sentiment data from online PCOS forums. Their study offered valuable insights into patient reported treatment experiences. Yet, its findings were constrained to subjective text based data and did not contribute to predictive modeling for clinical classification. To identify key symptoms of PCOS from clinical patient data, Permana [11] applied artificial intelligence techniques that demonstrate the potential of AI for primary symptom based diagnosis. While beneficial for symptom classification, the work lacked extensive evaluation using multiple model architectures and did not address performance limitations due to imbalanced datasets.

Across existing works, several gaps remain:

- limited use of general clinical datasets that can be applied in resource limited healthcare settings
- insufficient handling of class imbalance
- inconsistent comparison across multiple traditional and ensemble models
- lack of standardized evaluation across train test splits

LR, SVM, RF, DT, XGBoost, LightGBM, Voting, and Stacking classifiers combined with Random Oversampling to overcome dataset imbalance, the present study employs the classical and ensemble machine learning models to address these limitations. Its aim is to develop a robust, clinically deployable PCOS prediction framework that maintains a balance between accuracy, interpretability, and reproducibility.

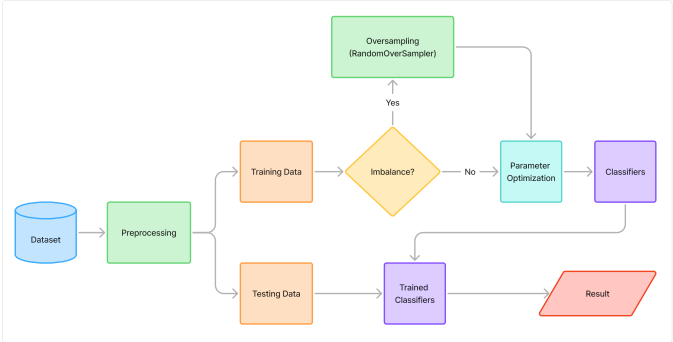


Fig. 1. Proposed machine learning workflow for PCOS detection with class imbalance handling

III. Methodologies

This section presents detailed description about the dataset, preprocessing steps and eight classifiers evaluated in this study: LR, SVM, RF, DT, XGBoost, LightGBM, Voting Classifier and Stacking Classifier. The Fig 1 shows the flow diagram of steps to detect PCOS. These models were evaluated by wide range of matrices like accuracy, precision, recall, f1-score, kappa Score and auc score.

A. Dataset Description

The dataset is collected from kaggle and which description was found in[12]. The dataset contains two files one of them is 'PCOS-data-without-infertility.xlsx' and another one is 'PCOS-infertility.csv'. Initially the dataset contains total 55 features. The dataset contains 541 patients details among them 177 are diagnosed with PCOS, while the rest are non PCOS.

B. Preprocessing

In this stage, the duplicate columns are eliminated after merging the two files resulting in 44 columns. The missing values in the numeric columns are replaced by median, and the categorical values by the mode. The dataset is divided into training and testing sets after normalization. RandomOverSampler is adopted to address the data imbalance in training set only. Chi-square is applied before splitting the dataset into training and testing subsets to ensure that it selects the same 30 features for both the training and testing sets. The study applied Grid Search Cross Validation (GSCV) for hyperparameter tuning after oversampling.

C. Classification Models

1) Logistic Regression: In supervised learning, LR is a way to classify data by modeling the relationship between a group of independent variables and a categorical outcome. It is most commonly applied to binary classification tasks, where the target variable takes one of two possible values, such as yes/no, or 0 /1 or presence/absence [5]. The algorithm calculates the probability that a given input falls into a specific category, making it suitable for problems involving discrete outputs.

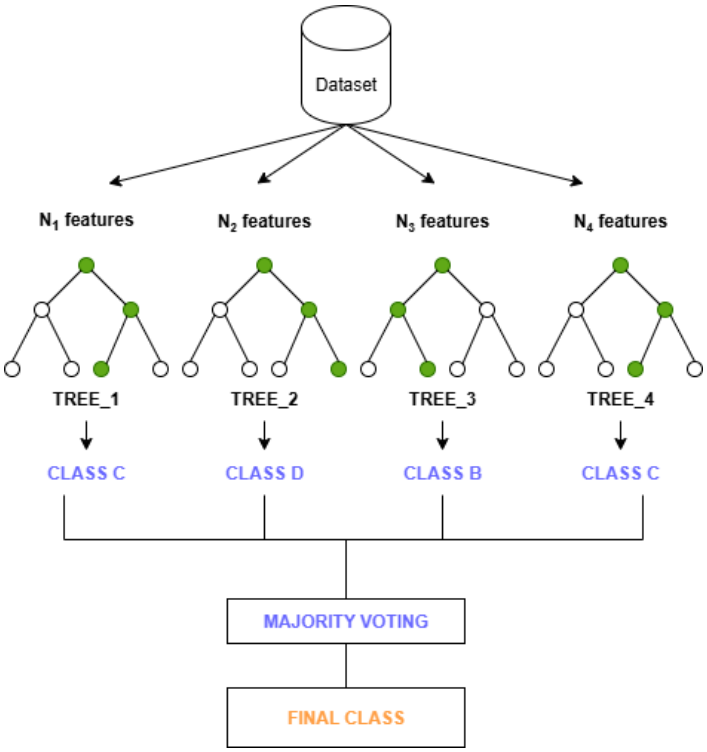


Fig. 2. Working procedure of Random forest algorithm [14]

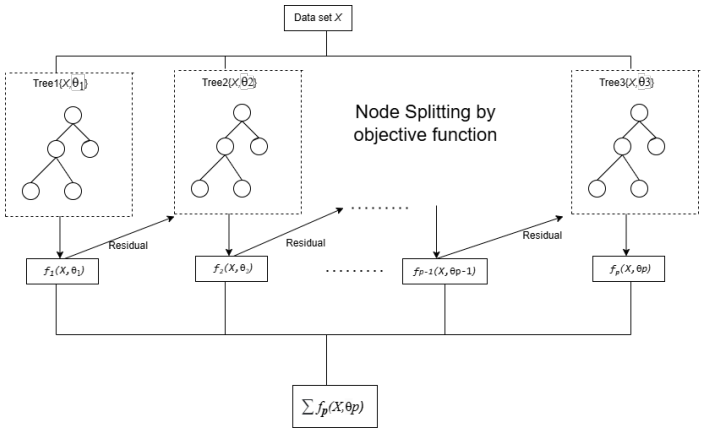


Fig. 3. XGBoost Algorithm Working Procedure [17]

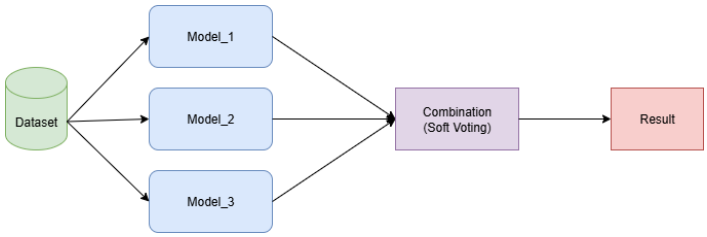


Fig. 4. Workflow of Voting Classifier [19]

2) Support Vector Machine: SVM is a supervised learning technique that finds the best dividing line between sets of data by making the differentiation margin as big as possible in converted feature space. The technique works by mapping input data into a higher dimensional space where it searches for the optimal hyperplane that separates different classes with the widest possible margin [13]. SVM uses kernel functions, such as linear, polynomial, and radial basis, to translate features into higher-dimensional spaces when working with complicated clinical data, like when classifying PCOS. This allows the model to capture non-linear patterns within the data, improving its ability to distinguish between phenotypic variations of PCOS and healthy controls.

3) Random Forest: RF is a supervised machine learning technique that creates several decision trees, each trained on different samples of the dataset and different sets of features. In classification tasks, the algorithm assigns the final class label based on the majority vote of its individual decision trees. This approach, in contrast to relying on a single tree, helps minimize overfitting and reduce variance, resulting in more consistent and accurate predictions on unseen data. Gini impurity was used to evaluate the quality of splits within decision trees [8]. Fig 2 shows Working procedure of RF algorithm.

4) Decision Tree: DTs are flexible learning models used for both classification and regression tasks in supervised learning [15]. The structure of a DT includes Root Node, Internal Nodes, Leaf Nodes. Each node in a DT represents a particular input feature, with the branches showing the possible outcomes of that feature, and the leaf nodes providing the final decision or class label. To determine the best splitting point in a DT, entropy is used to measure how uncertain or mixed the data is.

5) XGBoost: XGBoost is a supervised learning algorithm recognized for its strong performance in structured data tasks and frequent success in Kaggle competitions [16]. It has proven particularly effective for supervised learning tasks involving structured datasets, including classification, regression, and ranking. XGBoost distinguishes itself through the integration of regularization to control model complexity, optimized handling of missing or sparse data, and support for parallel processing, which collectively make it highly suitable for large-scale and high-dimensional machine learning problems. Fig 3 presents XGBoost Algorithm Working Procedure.

6) LightGBM: LightGBM as the central model for classification tasks. As a gradient boosting framework, it leverages a leaf-wise tree growth mechanism, which enhances computational efficiency and predictive accuracy compared to conventional level-wise approaches [18].

7) Voting Classifier: In this study, soft voting was used to combine the outputs of KNN, SVM, DT, RF, and LR, with the predicted class determined by averaging the probability scores from all models [8]. The ensemble predicted the output class by calculating the average of the predicted probabilities from all models and selecting the class with the highest aggregated probability. Fig 4 shows the voting classifier workflow.

8) Stacking Classifier: A meta-classifier that learns how to improve the final output combines the predictions of several base classifiers that have been trained independently as part of the stacking ensemble learning technique [8]. This method enhances overall predictive performance by enabling the meta-model to identify patterns in the base models predictions.

IV. Result Discussion and Analysis

To predict the presence of PCOS, multiple supervised machine learning algorithms were used in this study. To evaluate the

TABLE I
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS FOR PCOS PREDICTION USING MULTIPLE EVALUATION MATRICES

Models	Train Size	Accuracy	Precision	Recall	F1-Score	Kappa Score	AUC Score
Logistic Regression	50%	84.13	71.69	85.39	77.94	65.70	91.19
	80%	93.57	89.18	**91.66**	90.41	85.58	94.90
SVM	50%	81.91	71.73	74.15	72.92	59.36	91.25
	80%	89.90	85.71	83.33	84.50	77.02	93.75
Random Forest	50%	88.19	80.65	84.26	82.41	73.53	94.39
	80%	91.74	90.90	83.33	86.95	80.93	93.68
Decision Tree	50%	82.65	71.00	79.77	75.13	61.88	81.92
	80%	88.99	85.29	80.55	82.85	74.75	86.85
XGBoost	50%	85.97	76.84	82.02	79.34	68.75	93.11
	80%	92.66	93.75	83.33	88.23	82.92	**95.28**
LightGBM	50%	87.08	79.34	82.02	80.66	70.97	93.61
	80%	91.74	93.54	80.55	86.56	80.65	94.52
Voting Classifier	50%	85.24	76.34	79.78	78.02	66.92	93.72
	80%	**96.66**	91.18	86.11	88.57	83.17	**95.28**
Stacking Classifier	50%	87.08	82.14	77.53	79.77	70.30	93.72
	80%	94.50	**100.00**	83.33	**90.91**	**87.01**	94.98

influence of training data volume on model performance the models were trained and tested using two different data splits like 50% and 80% training sizes. Each model was assessed based on six evolution metrics: accuracy, precision, recall, f1-score, kappa score and AUC score. Because these metrics provide a reliable and comprehensive assessment of classification performance. Specially for imbalance clinical datasets where accurate identification of minority-class cases is clinically critical like PCOS.

The Table I presents the comparative performance of all consistently outperformed traditional classifiers across all evaluation metrics. The Fig 5 visualizes the Model Performance Comparison for test size 50% and Fig 6 visualizes the Model Performance Comparison for test size 80%. The results reveal that the voting classifier and stacking classifier parts of ensemble learning techniques achieved the highest predictive performance among all models. The voting classifier gained an accuracy of 96.66% outshining other classifiers in terms of F1-score (88.57%) and AUC (95.28%). This suggests that using multiple base learners in combination significantly increases the generalization capability of the model. The stacking classifier also demonstrated strong performance by achieving an accuracy of 94.50%, and the highest kappa score of 87.01%, indicating excellent agreement between the predicted and actual outcomes. In contrast, traditional models such as SVM and DT exhibited relatively lower accuracy and F1-score in capturing the complex feature relationships in the PCOS dataset, suggesting their limitations. A consistent improvement is observed, when the training size increased from 50% to 80%. For example, the accuracy of the LR model improved from 84.13% to 93.57%, and Ransom Forest improved from 88.19% to 91.74%. This indicates that they were able to learn richer data patterns because the models were provided with more training samples, which reduced overfitting and enhanced prediction stability. The overall results demonstrated that ensemble based learning methods are useful for PCOS prediction because they can combine multiple weak learners and reduce the problems arising from the bias-variance trade off. The high AUC values across most models (93%) suggest strong discriminatory ability between PCOS positive and PCOS negative cases.

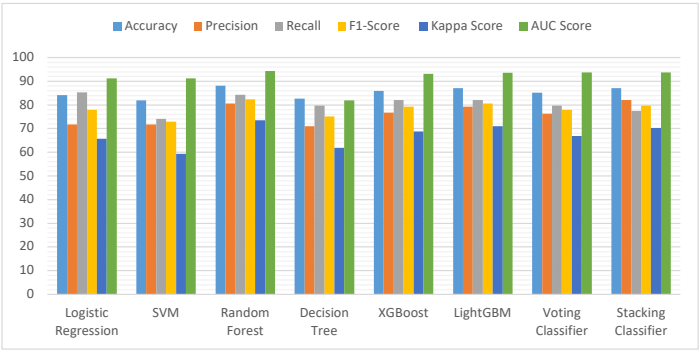


Fig. 5. Model Performance Comparison for 50% training data (X axis)

Among all the models, the voting classifier achieved the most balanced and reliable performance in this study because PCOS shows diverse clinical and lifestyle characteristics. By combining multiple models, ensemble approaches capture different feature relationships more effectively and balance bias and variance, leading to more accurate and stable prediction, proving it the most suitable choice for PCOS diagnosis. The strong results from the Stacking Classifiers further validate the effectiveness of ensemble methods in healthcare prediction tasks.

V. CONCLUSION

This study explore multiple machine learning models using clinical and lifestyle feature for the early detection of PCOS. Eight machine learning models were trained and analyze after addressing the class imbalance using RandomOverSampler, revealing that the ensemble methods outperform. As an optimal model, the voting classifier emerged, achieving 96.96% accuracy, outperforming both individual classifiers and the stacking approach. Also, strong and stable results are produced by the traditional models like LR, RF conforming their usefulness in medical classification tasks. Overall, PCOS detection can be significantly enhanced by machine learning based screening. Feature work will focus on expanding the datasets along with deep

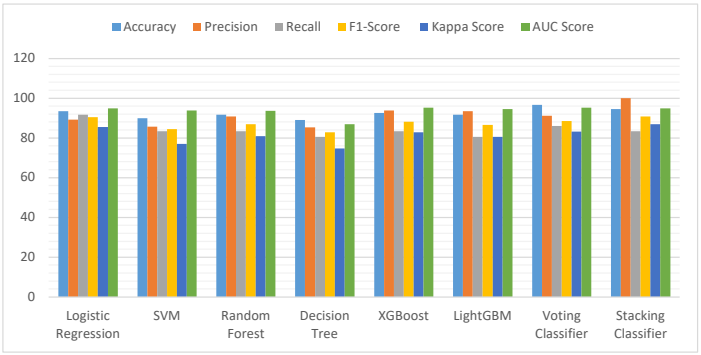


Fig. 6. Model Performance Comparison 80% training data (Y axis)

learning models, and developing a deployable decision-support application for real-time PCOS detection and risk assessment.

References

[1] E. Stener-Victorin et al., "Polycystic ovary syndrome," *Nature Reviews Disease Primers*, vol. 10, no. 1, p. 27, 2024.

[2] L. Ibáñez et al., "An international consortium update: Pathophysiology, diagnosis, and treatment of polycystic ovarian syndrome in adolescence," *Hormone research in paediatrics*, vol. 88, no. 6, pp. 371–395, 2017.

[3] J. Pea et al., "Ultrasonographic criteria in the diagnosis of polycystic ovary syndrome: A systematic review and diagnostic meta-analysis," *Human reproduction update*, vol. 30, no. 1, pp. 109–130, 2024.

[4] M. Agirsoy and M. A. Oehlschlaeger, "A machine learning approach for non-invasive pcos diagnosis from ultrasound and clinical features," *Scientific Reports*, vol. 15, no. 1, p. 33 638, 2025.

[5] S. Ahmed et al., "A review on the detection techniques of polycystic ovary syndrome using machine learning," *IEEE Access*, vol. 11, pp. 86 522–86 543, 2023.

[6] X. Yan et al., "Unveiling lipoprotein subfractions signature in high-fnpo pcos: Implications for pcom diagnosis and risk assessment using advanced machine learning models," *BMC medicine*, vol. 23, no. 1, p. 289, 2025.

[7] S. Reka, T. S. Praba, K. K. Manchala, and A. Venkateswarlu, "Comprehensive feature-driven pcos predictor: A reinforcement learning-based binary equilibrium optimization approach," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, p. 188, 2025.

[8] K. M. Mohi Uddin, M. T. A. Bhuiyan, M. M. Rahman, M. M. Islam, and M. A. Uddin, "Early pcos detection: A comparative analysis of traditional and ensemble machine learning models with advanced feature selection," *Engineering Reports*, vol. 7, no. 2, e70008, 2025.

[9] J.-y. Chen et al., "Screening of serum biomarkers in patients with pcos through lipid omics and ensemble machine learning," *PLoS One*, vol. 20, no. 1, e0313494, 2025.

[10] R. H. Emanuel, P. D. Docherty, H. Lunt, and R. E. Campbell, "What do users in a polycystic ovary syndrome (pcos) forum think about the treatments they tried: Analysing treatment sentiment using machine learning," *Physical and engineering sciences in medicine*, pp. 1–19, 2025.

[11] B. A. C. Permana, M. Wasil, et al., "Penerapan kecerdasan buatan untuk menentukan gejala utama polycystic ovary syndrom (pcos) berdasarkan data klinis pasien," *Infotek: Jurnal Informatika dan Teknologi*, vol. 8, no. 2, pp. 543–553, 2025.

[12] S. Vedpathak and V. Thakre, "Pcocare: Pcos detection and prediction using machine learning algorithms," *Bioscience Biotechnology Research Communications*, vol. 13, pp. 240–244, Dec. 2020. DOI: 10.21786/bbrc/13.14/56.

[13] V. Thakre, S. Vedpathak, K. Thakre, and S. Sonawani, "Pcocare: Pcos detection and prediction using machine learning algorithms," *Biosci Biotechnol Res Commun*, vol. 13, no. 14, pp. 240–244, 2020.

[14] V. Ignatenko, A. Surkov, and S. Koltcov, "Random forests with parametric entropy-based information gains for classification and regression problems," *PeerJ Computer Science*, vol. 10, e1775, 2024.

[15] M. M. Rahman et al., "Empowering early detection: A web-based machine learning approach for pcos prediction," *Informatics in Medicine Unlocked*, vol. 47, p. 101 500, 2024.

[16] Z. E. Aydin and Z. K. Ozturk, "Performance analysis of xgboost classifier with missing data," *Manchester Journal of Artificial Intelligence and Applied Sciences (MJAIAS)*, vol. 2, no. 02, p. 2021, 2021.

[17] D. T. Tran, T. Onjaipurn, D. R. Kumar, W. Chim-Oye, S. Keawsawasvong, and P. Jamsawang, "An extreme gradient boosting prediction of uplift capacity factors for 3d rectangular anchors in natural clays," *Earth Science Informatics*, vol. 17, no. 3, pp. 2027–2041, 2024.

[18] V. Sellam, P. Tushar, G. Rohit, and S. Sanyam, "Credit card fraud detection using machine learning," *Indian Journal of Computer Graphics and Multimedia*, vol. 1, no. 1, A1003011121, 2021.

[19] A. Srinivas and J. P. Mosiganti, "A brain stroke detection model using soft voting based ensemble machine learning classifier," *Measurement: Sensors*, vol. 29, p. 100 871, 2023.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Stroke Risk Prediction Using Machine Learning: A Comprehensive Evaluation

T. M. Shamim
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
tmshamim078@gmail.com

Mobassir Ahmmed
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mobassirahmmeds@gmail.com

Mst. Ajmira Akter Alo
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
aajmira77@gmail.com

Tanver Ahmed
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
tanverahmed.cse@gmail.com

Md. Muktar Hossain
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mmuktar997@gmail.com

***Abstract*—Stroke is one of the significant causes of death in the world. Early detection of stroke using some information from individuals can reduce the important complications and improve the health conditions. The aim of this study is to develop an efficient machine learning model to predict stroke risk as accurate as possible. A publicly available Stroke Prediction Dataset from Kaggle is considered for evaluation processes. The dataset has several features about lifestyle, medical reports which are known to influence the stroke risk. In this study, some preprocessing techniques were applied, including normalizing numerical features, encoding categorical variables, random resampling, and handling missing values. Multiple Machine Learning algorithms, such as Logistic Regression(LR), Support Vector Machine(SVM), Random Forest(RF), Decision Tree(DT), Multilayer Perceptron (MLP), Gradient Boosting Machine(GBM), K-Nearest Neighbors(KNN), and Stacking Classifier, were implemented to compare performance and accuracy in this study. Among general methods, the best accuracy was achieved by Random Forest 99.58%, outperforming other models, and the ensemble technique Stacking Classifier achieved 99.94% accuracy. The model demonstrated strong precision and recall in detecting high-risk patients. The proposed model can assist healthcare systems in early identification of stroke risk, helping reduce medical emergencies.**

***Index Terms*—Stroke Prediction, Machine Learning, Supervised Learning, Support Vector Machine**

I. Introduction

Stroke remains one of the most important global health concerns. It causes to long-term disability, mortality, and socioeconomic burden. Mainly Stroke is divided into two primary categories: ischemic and hemorrhagic. Ischemic stroke occurs due to blockage in the cerebral blood vessels, generally happened by atherosclerosis, which leads to reduced blood flow, oxygen deprivation, and eventual necrosis of brain tissue. Ischemic stroke is associated with high mortality, disability, and recurrence, often causing permanent damage to the central nervous system. On the other hand, hemorrhagic stroke generally arises from rupture of brain vessels, most frequently linked to hypertension and represents a different pathological mechanism. Hemorrhagic stroke is less common than ischemic [1]. International health reports revealed that millions of individuals are affected by stroke each year which is significantly underscoring the necessity for effective preventive strategies, early risk detection, and timely clinical interventions [2]. Every year about 12 million people get a stroke and around 5.5 million people die by this sudden diseases which is reported by World Stroke Organization [3]. Stroke is the third disease that cause death in the United States [4]. The availability of large-scale electronic health records (EHRs) and structured clinical datasets has transformed the way patient information is analyzed [5]. These datasets include demographic details, lifestyle attributes, conditions, and clinical histories. These enable data-driven analysis for identifying patterns associated with stroke. Nowadays the growth in machine learning (ML) have significantly enhanced the ability to examine complex medical data by revealing the hidden relationships. It also helps to calculate measuring accurate clinical decision. A number of studies have shown that Machine Learning techniques can successfully predict stroke risk by analyzing patient characteristics, medical histories, and behavioral factors. Despite these contributions, there have many existing approaches that mainly focus on a limited subset of predictors or a small number of machine learning algorithms. Comprehensive studies evaluate multiple machine Learning models which simultaneously analyze diverse clinical risk factors and investigate feature relevance remain comparatively limited. Addressing this gap is essential for improving predictive accuracy and enhancing the incompletion of stroke risk models in real-world clinical world. This study uses a publicly available stroke dataset to explore important factors related to stroke risk. The dataset includes patient attributes such as age, gender, health conditions, lifestyle habits, and comorbidities.Some well-known of machine learning classifiers are implemented to evaluate predictive performance across multiple evaluation metrics including Naive Bayes (NB), Logistic Regression

(LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), XGBoost, and Neural Network (NN). Furthermore, we examine the importance of the characteristics to determine which attributes contribute the most significantly to predict the stroke in early stage. This helps us to analyze the deeper insights into the relationship between patient characteristics and stroke risk. The principal contributions of this study are summarized below:

1) Comprehensive Risk Factor Evaluation: We provide an extensive analysis of demographic, clinical, and lifestyle variables to identify the most influential predictors associated with stroke.
2) Benchmarking of Multiple ML Models: Eight widely used machine learning techniques are evaluated and compared based on accuracy, precision, Kappa, AUC, recall, and F1-score.
3) Importance of Preprocessing: We preprocessed our data significantly and more efficient way using modern sampling method and normalization.
4) Importance of ensemble methods: Ensemble methods enhance predictive performance by combining multiple classifiers to reduce errors and improve model stability. This approach allows us to achieve more accurate and reliable results compared to individual models.

In this research, the rest of the sections are as follows. Section II describes the related work and research done by previous researchers. Section III showed the description about the data set, classification and the study process. Section IV is about the analytical results and discussion of them. Finally, the overall study is concluded in section V.

II. Related Work

Many researchers have explored stroke prediction using various ML methods. Fedesoriano et al. introduced a widely used stroke prediction dataset and applied baseline classifiers to analyze risk factors such as hypertension, glucose level, and BMI. Their work pointed out the importance of data preprocessing and class imbalance handling for improving model performance [6]. Another study by Mohan et al. proposed an efficient stroke prediction study using LR, NB, and RF. Their results showed that ensemble models performed better in identifying stroke risk among clinical patients, highlighting the importance of combining multiple classifiers for improved accuracy [7]. In a more advanced approach, Chaurasia and Pal developed a predictive model applying DT, KNN, and SVM for stroke-related medical datasets. Their comparative analysis demonstrated that SVM delivered higher accuracy due to its capability to handle nonlinear patterns in biomedical data [8]. Researchers like Feng et al. proposed a deep learning based stroke risk model using multilayer neural networks. Their findings showed that deep architectures outperform classical ML models when sufficient clinical data are available, especially in capturing complex nonlinear interactions among risk factors [9]. Zhou et al. developed an ensemble based stroke classification model using Boosting algorithms. Their research demonstrated that AdaBoost and Gradient Boosting classifiers outperform individual learners in identifying stroke patterns from imbalanced medical datasets [10]. Several previous studies have used machine learning techniques to predict stroke risk. However, many of these studies do not provide a clear and consistent comparison across different classifiers, preprocessing methods, and evaluation metrics. Most existing works focus on only a few algorithms, such as Logistic Regression, Decision Tree, or Random Forest, and typically report accuracy values between 85% and 98%, depending on the dataset and preprocessing approach used [6]–[10] . Even though many studies have tried using ML to predict stroke, a few important gaps remain, which we could improve:

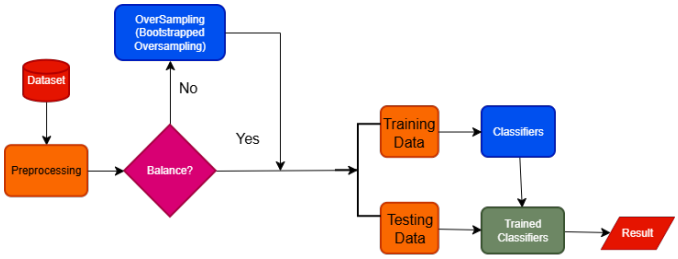


Fig. 1. Flow diagram of Stroke detection process

- Most papers focus on only a few features, so we still don't know which factors are truly the most important when all patient information is considered.
- Many studies skip dimensionality-reduction or deeper feature analysis, meaning complex relationships between variables are often overlooked.
- Stroke cases are very few in medical datasets, yet many studies don't use strong balancing techniques so their models may fail in real clinical use.
- While a few ensemble or deep learning methods exist, most research still depends on basic ML models and lacks a unified comparison across multiple techniques.

We can overcome these problem by using multiple models with modern ensemble technique like stacking classifier. We can also use better and strong balancing techniques as Bootstrapped Oversampling. All these factors to improve the study to predict stroke using Machine Learning are analyzed in this research.

III. Methodologies

This section provides a detailed description of the dataset. The description about the preprocessing steps and the eight machine learning classifiers used in the study: LR, SVM, RF, DT, MLP, GBM , KNN, and Stacking Classifier. The overall workflow employed for stroke detection is illustrated in Fig. 5. Multiple metrics were evaluated including accuracy, precision, recall, f1-score, kappa, and the AUC score.

A. Dataset Description

The dataset used in this study is collected from kaggle [6]. This dataset has 11 features, including categorical features

such as gender, ever_married, work_type, Residence_type, and smoking_status, as well as numerical features like age, avg_glucose_level, and BMI. Additionally, health conditions such as hypertension and heart_disease are included. The target label is stroke, indicating whether the individual has experienced a stroke or not. The dataset contains data from $5,110$ patients. Among them, 249 (4.87%) are positive cases for stroke, and $4,861$ (95.13%) are negative cases.

B. Preprocessing

Several preprocessing steps were applied to ensure improving data quality and classifier performance before splitting. At First, the raw dataset of stroke was examined for missing values. The BMI attribute founded null entries which were imputed using the attribute's mean value. Mean is used because imputation preserves the overall distribution without biasing.Then categorical attributes were transformed into a numerical format using encoding technique. It enables the dataset to be used effectively by machine-learning algorithms. In the dataset, non-stroke cases were much more than stroke cases which address the class imbalance problem. The minority class was up sampled using random resampling. This step produced a balanced dataset and prevented the classifiers from being biased toward the majority class. After balancing, the feature matrix was separated from the target variable (stroke). Finally, all features were standardized using Standard Scaler which transforms the data into zero-mean and unit-variance format. Finally, normalization makes distance-based and gradient-based classifiers like SVM, KNN, and neural networks work better by putting all features on the same scale.

This preprocessing pipeline ensured a clean, balanced, and normalized dataset suitable for training multiple stroke-prediction models.

C. Classifiers Description

1) Logistic Regression (LR): An accepted technique for predicting Stroke is LR, it is an entrenched statistical model that successfully analyzes enormous financial data and achieves strong performance [11]. Logistic regression is expressed as:

$$P(Y = 1 \mid X),$$

which denotes the likelihood that the dependent variable Y pertains to class 1 when considering the values of the independent variables X.

2) Support Vector Machine (SVM): The algorithm of SVM is employed to address regression, classification, and outlier detection problems. In SVM, this basically plots the data as points in an n-dimensional space, where n is the number of features. To achieve optimal classification of data points with an unknown class, it can apply a range of kernel functions, such as linear, polynomial, radial basis (RBF), or quadratic. The data is treated as linearly separable, and the classifier is defined as:

Class 1 (Low risk) : $(W \cdot X + b) \geq 1, \quad \forall X,$
Class 2 (High risk) : $(W \cdot X + b) \leq -1, \quad \forall X,$ where W is the vector representing the hyperplane, b is the bias term, and X is the matrix derived from the dataset [12].

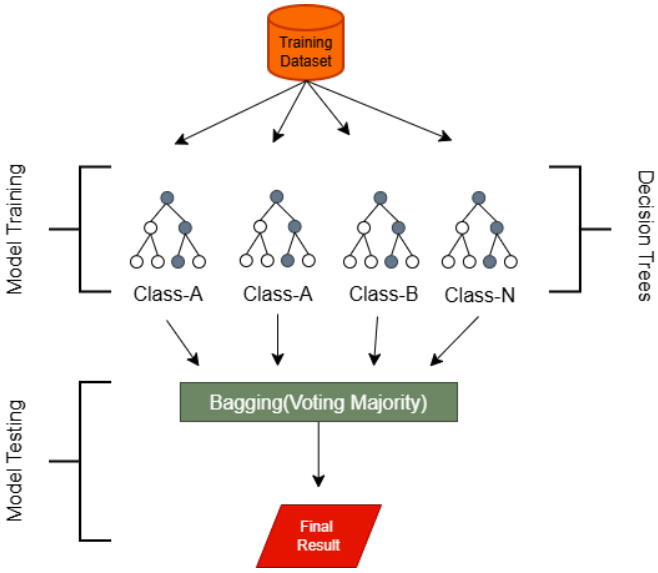


Fig. 2. Working process of Random Forest

3) Decision Trees (DT): In supervised learning, a DT is a type of predictive model that is used for both classification and regression tasks. The model operates using three types of nodes. It begins with the Root Node, which represents the initial input and acts as the starting point of the model [13]. This is followed by Internal Nodes that split the data based on specific features or attributes. The process concludes at the Leaf Nodes, which indicate the final output or class label.

4) Multilayer Perceptron (MLP): Multilayer Perceptron (MLP) is a type of neural network that learns patterns from data by passing information through several connected layers of neurons. It can handle complex and non-linear relationships. This helps making it useful for classification tasks where simple models struggle. MLP improves its predictions by adjusting the connections between neurons during training, allowing it to gradually learn the best way to separate different classes.

5) RANDOM FOREST(RF): The RF model typically employs the construction of multiple decision trees and yields a mean forecast once these trees have undergone machine learning training [14]. Within a DT, the node probability is typically calculated as the fraction of the total training samples that reach that particular node. In the following figure 2, we can see the working process of RF:

6) K-Nearest Neighbors(KNN): KNN is a non-parametric, supervised learning algorithm. It classifies a new, unseen data point by utilizing information about its k nearest available neighbors within the training dataset. The core principle of this procedure involves contrasting the provided sample or experimental data with the training dataset (the stored data points). This comparison is carried out using an analogous technique. The prepared data collection is evaluated against the sample or test data in order to examine them [15].

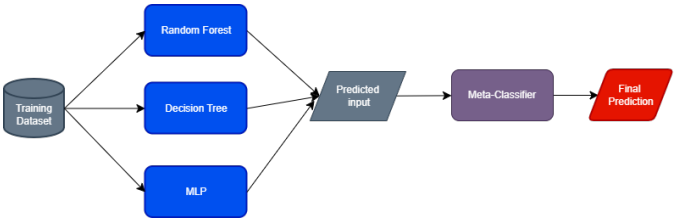


Fig. 3. Working process of Stacking

7) Gradient Boosting (GBM): Gradient boosting is a powerful ensemble learning technique in machine learning that builds a strong predictive model by sequentially combining multiple weak models, typically decision trees. In GBM at each step, a new weak model is trained to predict the ”error” of the current strong model.

8) Stacking(DT+RF+MLP): In the domain of stroke prognosis, stacking constitutes an advanced ensemble methodology that enhances predictive accuracy through the strategic amalgamation of heterogeneous ML models [16]. This multi-tiered approach employs diverse base classifiers (e.g., RF, DT, MLP) within a foundational layer, whose outputs are subsequently synthesized by a second-level meta model, typically LR, to yield a highly robust final prediction. In the following figure 3, we can see the working process of Stacking Classifier:

IV. Result and Discussion

To develop the stroke prediction system, eight machine learning methods were used. These are: LR, Random Forest, SVM, LR, DT, KNN, MLP, GBM and the Stacking Classifier. Two distinct split ratios (50 : 50 and 80 : 20) were used during training and testing to evaluate the resilience and generalization capacity of each model. Accuracy, precision, recall, F1-score, kappa, and Area Under the ROC Curve (AUC) were used as the main performance measures to evaluate the model.

A. Model Performance Comparison

According to the experimental data shown in Table IV, the Stacking Classifier outperformed both the RF model and other individual classifiers.This classifier attained the following results with an 80% training ratio: Accuracy : 99.94%, Precession : 99.89,Recall: 100%, F1-Score: 99.94%, AUC: 100% and 99.89% is the kappa score. These findings verify that the Stacking ensemble successfully integrates the advantages of its base learners, enabling it to capture complex feature interactions and non-linear relationships in the dataset considerably better than the standalone models. Although the RF model delivered strong results—achieving 98.53% accuracy with a 50%training ratio and 99.58% accuracy with an 80% training ratio—the Stacking Classifier consistently outperformed it across several critical evaluation metrics. In particular, the Stacking approach demonstrated superior f1-score, kappa, and AUC which is indicating a higher level of robustness, stability, and diagnostic reliability for clinical decision-support applications.

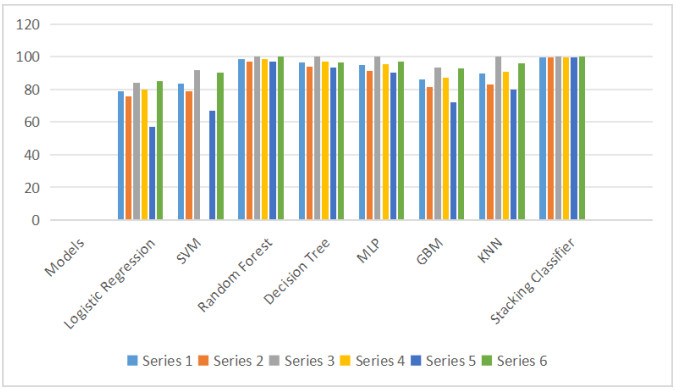


Fig. 4. Model Performance with 50% test data

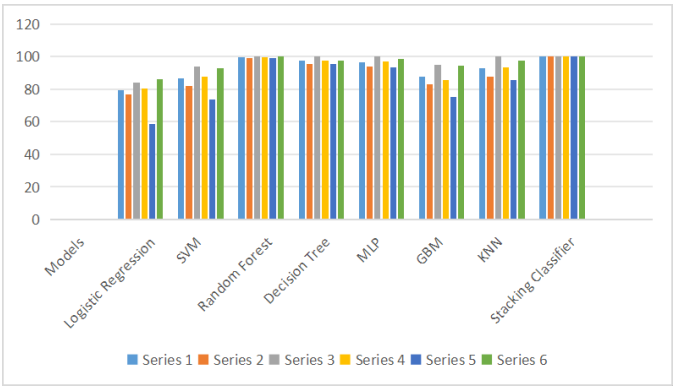


Fig. 5. Model Performance with 80% test data

The DT classifier also showed competitive performance, especially with the 80% training ratio, where it achieved 97.68% accuracy and an F1-score of 97.73%. While DT does not surpass ensemble based methods such as RF or Stacking, its strong metric values highlight its effectiveness as a standalone interpretable model. On the other hand, the MLP produced comparatively lower results across most metrics. However, with an accuracy of 96.70%, the MLP still demonstrates meaningful predictive capability. Its performance suggests that neural models can capture complex non-linear patterns within the stroke dataset but may require deeper architectures, hyperparameter optimization, or larger datasets to reach their full potential.

Overall, the findings highlight that ensemble learning specially Stacking provides the most reliable and generalizable performance, followed by RF, MLP and DT. Neural based models show promise but require further optimization to match ensemble-level robustness.

B. Discussion

The results of this study demonstrate the benefit of ensemble learning, especially Stacking in stroke risk prediction situations.The strong performance of the stacking model shows that combining multiple models helps reduce both bias and variance. It works well because each model adds its own strengths to improve the final prediction. The model performs better for

TABLE I
MODEL PERFORMANCE COMPARISON BETWEEN 80% AND 50% TEST DATA

Models	Train Size	Accuracy	Precision	Recall	F1-Score	Kappa Score	AUC Score
Logistic Regression	50%	78.62	75.82	84.03	79.71	57.25	85.20
	80%	79.38	76.80	84.15	80.31	58.76	86.27
SVM	50%	83.50	78.76	91.72	84.75	67.00	90.35
	80%	86.73	81.98	94.13	87.64	73.47	92.63
Random Forest	50%	98.53	97.16	100.00	98.56	97.07	100.00
	80%	99.58	99.18	100.00	99.59	99.17	100.00
Decision Tree	50%	96.72	93.85	100.00	96.83	93.45	96.72
	80%	97.68	95.57	100.00	97.73	95.37	97.68
MLP (Neural Network)	50%	95.10	91.07	100.00	95.33	90.20	97.10
	80%	96.70	93.82	100.00	96.81	93.41	98.49
GBM	50%	86.11	81.39	93.62	87.08	72.22	92.88
	80%	87.71	82.81	95.16	85.55	75.42	94.22
K-Nearest Neighbors (KNN)	50%	89.94	83.24	100.00	90.85	79.88	95.75
	80%	92.85	87.48	100.00	93.32	85.70	97.27
Stacking Classifier (DT+RF+MLP)	50%	99.79	99.59	100.00	99.79	99.58	100.00
	80%	99.94	99.89	100.00	99.94	99.89	100.00

availability of data which improves generalization and reduces misclassification of minority stroke cases. This improvement becomes clear as the training size increases. Furthermore, the Stacking model's flawless accuracy, Precision, Kappa recall, AUC and F-1 score demonstrates its stunning capacity to differentiate between stroke and non-stroke individuals which is an excellent choice for early clinical screening instruments.In future work, adding feature-interpretability methods (such as SHAP or XAI), using hybrid deep-learning–based stacking, and combining data from multiple patient sources can improve model transparency for medical practitioners. These enhancements may also lead to better overall performance.

V. CONCLUSION

This study evaluated different machine learning models for predicting stroke risk using a publicly available dataset. A total of eight models, including both single classifiers and ensemble methods, were tested using the same preprocessing steps, data balancing technique, and evaluation metrics. The results show that ensemble learning performs better than individual models. In particular, the stacking classifier achieved the best results in terms of accuracy, AUC, and kappa score under both test conditions. These results indicate that combining multiple models helps improve prediction accuracy and provides more stable and reliable outcomes, especially for imbalanced medical datasets like stroke data. Overall, the proposed method can support early stroke risk detection and may assist healthcare professionals in making better decisions. In future work, model transparency can be improved using explainable methods, and the approach can be tested on larger and more diverse clinical datasets.

REFERENCES

[1] S. Zhi, X. Hu, Y. Ding, *et al.*, "An exploration on the machine-learning-based stroke prediction model," *Frontiers in Neurology*, vol. 15, p. 1 372 431, 2024.

[2] I. for Health Metrics and Evaluation, "Global burden of stroke: 1990–2021 analysis," *The Lancet Neurology*, 2024, Reports that over 12 million people experience stroke annually worldwide.

[3] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, p. 4670, 2022.

[4] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 183–192.

[5] R. S. Evans, "Electronic health records: Then, now, and in the future," *Yearbook of Medical Informatics*, vol. 26, no. 01, S48–S61, 2017. DOI: 10.15265/IYS-2017-007. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5171496/.

[6] fedesoriano, *Stroke Prediction Dataset*, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset, Kaggle dataset, accessed: YYYY-MM-DD, 2020.

[7] S. K. Mohan, P. Thirumal, and R. Srivastava, "Effective heart disease prediction using machine learning: A comparative study," *International Journal of Recent Technology and Engineering*, vol. 8, no. ISSUE 3, pp. 944–950, 2019.

[8] V. Chaurasia and S. Pal, "Machine learning techniques: A review on stroke prediction," *International Journal of Computer Science and Information Technologies*, vol. 9, no. 3, pp. 80–84, 2018.

[9] M. Feng, Y. Huang, S. Lin, and K. Chen, "Deep learning-based stroke prediction using structured and unstructured clinical data," *IEEE Access*, vol. 8, pp. 13 680–13 688, 2020. DOI: 10.1109/ACCESS.2020.2966866.

[10] X. Zhou, L. Zhang, and Y. Fan, "Stroke risk prediction using ensemble machine learning models," *Journal of Medical Systems*, vol. 43, no. 10, pp. 1–12, 2019. DOI: 10.1007/s10916-019-1423-5.

[11] A. Gupta, N. Mishra, N. Jatana, *et al.*, "Predicting stroke risk: An effective stroke prediction model based on neural networks," *Journal of Neurorestoratology*, vol. 13, no. 1, p. 100 156, 2025.

[12] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of brain stroke severity using machine learning.," *Rev. d'intelligence Artif.*, vol. 34, no. 6, pp. 753–761, 2020.

[13] M. M. Rahman, A. Islam, F. Islam, *et al.*, "Empowering early detection: A web-based machine learning approach for pcos prediction," *Informatics in Medicine Unlocked*, vol. 47, p. 101 500, 2024.

[14] Y. Zhang, "Stroke prediction based on machine learning," in *ITM Web of Conferences*, EDP Sciences, vol. 70, 2025, p. 04 029.

[15] K. Kandukuri and A. Sandhya, "Heart stroke detection using knn algorithm," *ECS Transactions*, vol. 107, no. 1, p. 18 385, 2022.

[16] P. Chakraborty, A. Bandyopadhyay, P. P. Sahu, *et al.*, "Predicting stroke occurrences: A stacked machine learning approach with feature selection and data preprocessing," *BMC bioinformatics*, vol. 25, no. 1, p. 329, 2024.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Deep Transfer Learning-Based Multi-class Classification of Bone Tumors Using Medical Imaging

Nasif Hossain Prio
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
nasif.hp007@gmail.com

Md Hafizul Islam
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mdhafizulislam42349@gmail.com

Md Muhit Uddin
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
eng.muhituddin@gmail.com

Md. Mahfujur Rahman
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mahfujur@vu.edu.bd

Abdullah Tamim
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
tamim.cse.vu@gmail.com

D.M. Asadujjaman
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
asadujjaman@vu.edu.bd

***Abstract*—Identifying different types of bone tumors from X-ray images is still a tough task for doctors, mainly because many tumors look alike and the visual clues can be extremely subtle. To help with this challenge, our study explores a deep learning approach that classifies bone tumors into four groups: Normal, Normal Tumor, Benign Tumor, and Malignant Tumor. We gathered a secondary X-ray dataset and balanced it using a simple median-based method that mixes undersampling with SMOTE oversampling, and then split it into training, validation, and testing sets.We experimented with four well-known pre-trained models—MobileNetV3, VGG19, ResNet50, and EfficientNetB3—each fitted with the same custom classification head that used global average pooling, a ReLU layer, and a Softmax output for the final predictions.In each experiment, EfficientNetB3 did remarkably well. It achieved a test accuracy of 91.13% and showed more consistency when handling tumor classes that are often confused with one another. The ROC curves and confusion matrices provided additional support for its stability. Comparing our results with previous studies, we found that our approach maintained a reasonable computational cost while producing noticeable improvements.Overall, the results indicate that an EfficientNetB3-based model can be a useful and trustworthy tool for helping medical professionals diagnose and classify bone tumors in detail.**

***Index Terms*—Bone Tumors, Deep Learning, X-Ray Imaging, Transfer Learning, EfficientNetB3, Multi-Class Classification**

I. Introduction

Abnormal growth or mass of cells within bones referred to as bone tumors emerge as a critical clinical challenge. Through conventional radiographic evaluation, distinguishing between different tumor types becomes difficult because of their heterogeneous nature [1]. Early and accurate distinction of Normal, Benign and Malignant tumors conditions are crucial as delayed Malignant tumor diagnosis can cause swift disease progression with poor medical outcomes. X-ray imaging remains the primary diagnostic method as it is widely used and accessible in routine clinical practices.However, manual interpretation heavily relies on the expertise of radiologists, which can limit diagnostic reliability due to inter-observer variability [2] [3] [4]. Recently Convolutional Neural Networks (CNNs) in deep learning have shown promising significance in automated medical image analysis by extracting discriminative features from imaging data [5] [6] [7]. Transfer learning has been proven as an effective approach as it utilizes pretrained CNN models to improve performance on limited medical datasets reducing training time [8] [9] [10]. Most prior studies have focused on binary or three-class bone tumor classification which limits their clinical applicability [2] [3] [13].

In our study, we have evaluated four CNN models (MobileNetV3, VGG19, ResNet50, and EfficientNetB3) on four-class bone tumor classification. All models have shown satisfactory performance, However, EfficientNetB3 achieved highest accuracy and outperformed others. To make this evaluation more clinically relevant and methodologically transparent, we frame the task as a four-class classification problem (Normal, Normal Tumor, Benign, and Malignant) and compare widely used transfer-learning backbones under a unified experimental setup. Specifically, we adopt a consistent training pipeline across models (including the same input configuration and comparable optimization settings) so that performance differences reflect the models themselves rather than mismatched training choices. In addition to reporting overall accuracy, we highlight class-wise reliability using standard clinical-style summaries like precision, recall, and F1-score, as well as confusion-matrix and ROC-based analyses to show where the models typically perform well or confuse similar tumor categories. In order for the results to be interpreted realistically and expanded upon in subsequent work, we conclude with a clear discussion of practical limitations and reproducibility

considerations, including dataset constraints and steps taken to ensure a fair split and reliable evaluation. The remainder of the paper is structured as follows: Section II presents literature review, Section III describes dataset and methodologies and Section IV discusses results of this experiment. Finally, Section V concludes the study and highlights potential future directions.

II. Literature Review

Automated bone tumor classification has become an important field of study because manual radiographic interpretation poses several challenges [1]. Previous studies have employed traditional convolutional neural networks (CNNs) to differentiate between benign and malignant lesions [4]. A multitask deep learning framework that can segment and binary classify primary bone tumors was presented by Liu et al. [2] with an accuracy of 90.3%. Ponlatha et al. [3] implemented a number of CNN architectures, including VGG19, InceptionV3, and ResNet50, and reported validation accuracies of 82.7%, 89.4%, and 90.5%, respectively.

With the appearance of transfer learning, pre-trained CNN models have been used to improve performance on limited medical datasets [7]. Park et al. [6] utilized EfficientNet-B2 on proximal femur radiographs (538 images), obtaining an accuracy of 85.3% and AUC of 0.953. Liu et al. [8] combined CNN features with clinical metadata demonstrating the benefit of multimodal information to achieve a three-class macro-average AUC of 0.872. Song et al. [9] introduced PBTC-TransNet, a transformer-based fusion model for incomplete multimodal images (X-ray, CT, MRI) with clinical metadata, obtaining a micro-average AUC of 0.847. Hinterwimmer et al. [11] integrated imaging with metadata which has improved classification from 54.1% to 69.7%. [10] Wang et al. [12] used an ensemble of EfficientNet B3/B4, ViT, and Swin Transformer on multicenter radiographs, delivering internal accuracy 88.1% and external 89.5%. Li et al. [13] applied YOLO-based detection and classification on 1,085 tumor radiographs and 345 normal radiographs, attaining internal accuracy 86.36% and external 85.37%.

Despite these progresses, most prior studies focused on either binary or three-class classification and four-class X-ray bone tumor classification remains underexplored [4] [5]. Moreover, a systematic evaluation across multiple transfer learning models in a single dataset has been limited. Our study aims to provide a robust framework for accurate diagnosis through modern architectures' performance evaluation.

III. Methodology

A. Dataset Preprocessing

In this study, a secondary X-ray dataset was used and categorized into four classes: Normal, Normal Tumor, Benign Tumor, and Malignant Tumor. The dataset was split into training (80%), validation (10%), and testing (10%) sets. Importantly, the split was performed before applying any oversampling or augmentation to minimize the risk of data leakage. In order to maintain a realistic distribution in validation and test sets and to facilitate an objective evaluation, class balancing was only applied to the training set because the original dataset was unbalanced. We did this intentionally to avoid letting any oversampling effects data leakage into the validation or test sets. By leaving those sets unaltered, the evaluation is kept impartial, trustworthy, and representative while maintaining their actual class distribution. We combined SMOTE and data augmentation to balance the training set. Data augmentation introduced small, realistic variations that help improve generalization, while SMOTE was applied specifically to the Benign and Malignant classes to synthetically increase the number of samples in these minority categories. Crucially, both oversampling and augmentation were restricted to the training set to avoid any data leakage and to prevent artificially inflating validation and test performance.

Data augmentation included:

- Random horizontal and vertical flipping
- Rotation within $\pm 15^{\circ}$
- Zooming within 10%
- Width and height translation up to 10%
- Brightness adjustment within $\pm 10\%$

The dataset after preprocessing has 6900 images, 5520 in the training set (1380 images/class), 692 image validation set, and 688 image test set. Training on the same data as the models provided a way of comparing their performance. The final class distribution is shown in Table I. It is possible to ensure that observed performance differences are due to model capacity rather than variations in data preparation by using a consistent preprocessing pipeline.

B. Training Models

In this study, four pre-trained CNN architectures were employed for four-class bone tumor classification: MobileNetV3, VGG19, ResNet50, and EfficientNetB3. All models were implemented using transfer learning, leveraging weights pre-trained on ImageNet to accelerate convergence and improve performance on limited medical datasets.

MobileNetV3 is an improved lightweight CNN optimized for efficiency and accuracy, incorporating squeeze-and excitation (SE) blocks and hard-swish (h-swish) activation. VGG19, a deeper CNN model consisting of 19 layers, is widely known for its uniform architecture with repeated convolutional blocks and strong feature extraction capabilities. For utilizing residual connections to mitigate the vanishing gradient problem in deep networks, ResNet50 was employed. Lastly, EfficientNetB3 occupied compound scaling to uniformly scale depth, width, and resolution, providing superior accuracy with fewer parameters compared to traditional CNNs.

For four-class classification, a custom classification head was proposed for all these architectures. The head comprises a singular global average pooling layer, accompanied by fully connected dense layers. For the global average pooling layer we have employed Rectified Linear Unit (ReLU) as the activation function. ReLU disregards negative values by assigning them a value of zero, thereby selecting the maximum

of the input values [14]. The mathematical representation of ReLU,

$$ReLU(x) = max(0, x) \tag{1}$$

For the output layers, Softmax activation function is used to calculate probabilities for multi-class detection [15].

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2}$$

Moreover, categorical crossentropy is used to evaluate the discrepancy between the actual class and the predicted class through backpropagation, This process of back propagation was facilitated with the help of derivatives of crossentropy showing greater magnitudes for poor predictions.

$$L(y, y_i) = -\sum_{i=1}^{K} y_i \cdot \log(\hat{y}_i) \tag{3}$$

C. Hyperparameters

The hyperparameters considered as significant for the proposed models include epochs, batch size, learning rate, optimizer, and dropout. The hyperparameters were chosen based on their influence on convergence and generalization performance. The hyperparameter values used to achieve the optimal results are listed in TableII.

TABLE I
FINAL CLASS DISTRIBUTION AFTER BALANCING

Split	Class	Image Count
Train	Benign Tumor	1380
	Malignant Tumor	1380
	Normal	1380
	Normal Tumor	1380
Test	Benign Tumor	172
	Malignant Tumor	172
	Normal	172
	Normal Tumor	172
Validation	Benign Tumor	173
	Malignant Tumor	173
	Normal	173
	Normal Tumor	173

TABLE II
MODELS PARAMETERS FOR INPUT AND CLASSIFICATION STAGE

Parameter	Value
Input Size	224×224
Number of Epochs	100
Batch Size	32
Activation Function	Softmax
Optimizer	Adam
Learning Rate	0.0001

1 1.3

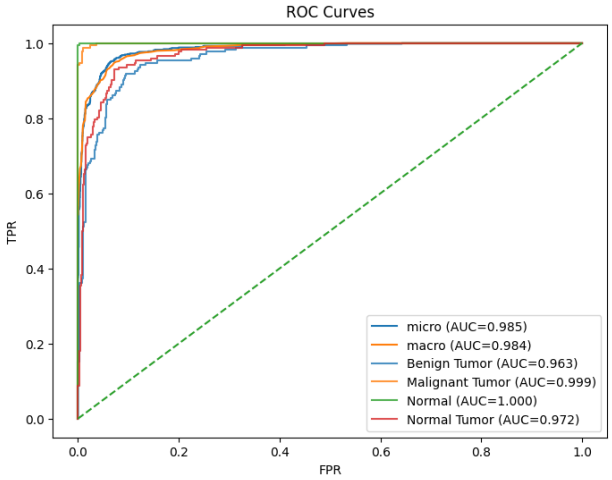


Fig. 1. ROC curve of MobileNetV3

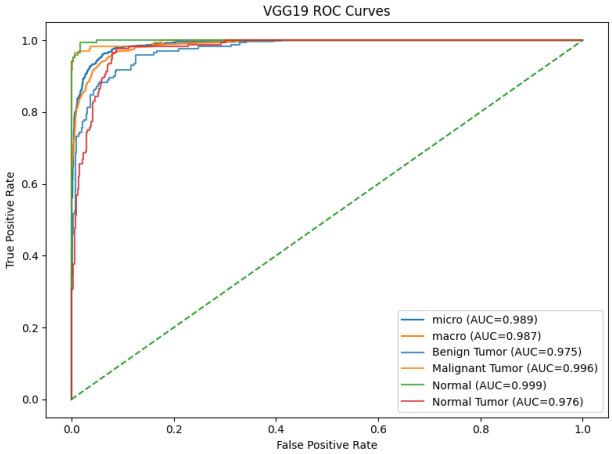


Fig. 2. ROC curve of VGG19

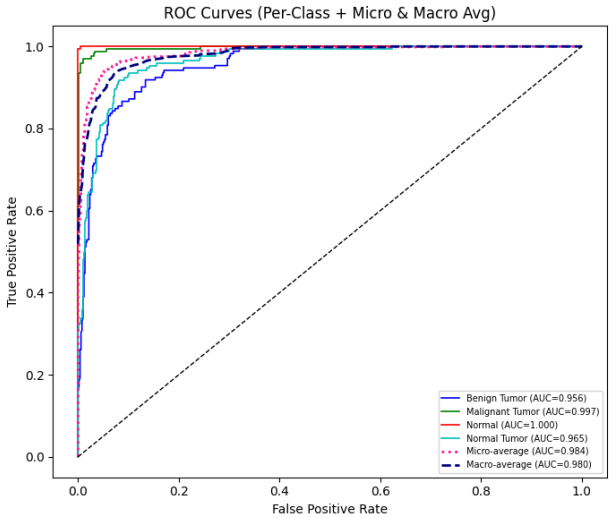


Fig. 3. ROC curve of ResNet50

TABLE III
OVERALL PERFORMANCE OF DEEP LEARNING MODELS

Model	Accuracy (%)	Macro Precision	Macro Recall	Macro F1-Score
MobileNetV3	88.66	0.8873	0.8866	0.8849
VGG19	90.84	0.9080	0.9084	0.9081
ResNet50	89.83	0.8971	0.8983	0.8976
EfficientNetB3	91.13	0.9114	0.9113	0.9113

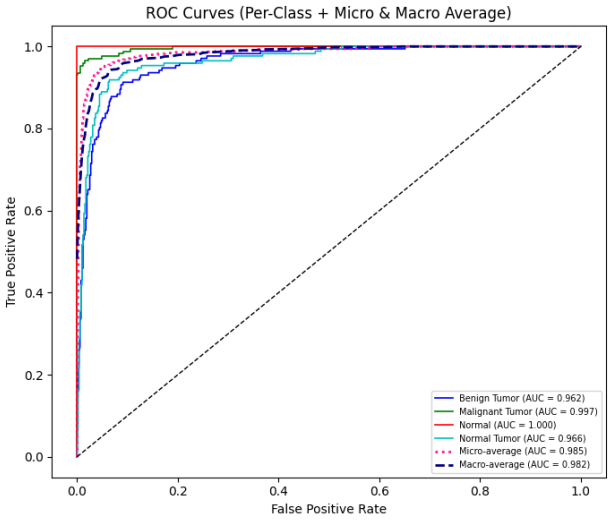


Fig. 4. ROC curve of EfficientNetB3

IV. RESULTS AND DISCUSSION

The overall performance of the models is summarized in Table III. Among the evaluated architectures, EfficientNetB3 showed strong generalization and the highest accuracy (91.13%) with consistent macro precision, recall, and F1-score.

According to Table III, MobileNetV3 scored 88.66%, while VGG19 and ResNet50 scored 90.84% and 89.83%, respectively. EfficientNetB3 once again led the comparison with 91.13%, possibly due to its balanced scaling strategy. Figures 1, 2, 3 and 4 display the ROC curves for our proposed models. and confusion matrices. AUC analysis revealed that VGG19 had the highest values (98.9% micro and 98.7% macro). The micro and macro AUC values for MobileNetV3 and EfficientNetB3 were, respectively, 98.5% and 98.2%. ResNet50 followed with micro and macro scores of 98.4% and 98.0%, respectively.Upon examining the confusion matrices given in Figure 5,6,7 and 8, it is evident that the diagonal values, which signify the accurately predicted outcomes, indicate that the Confusion Matrix for EfficientNetB3 demonstrates superior performance in this context.

Misclassifications mainly occurred between Benign Tumor

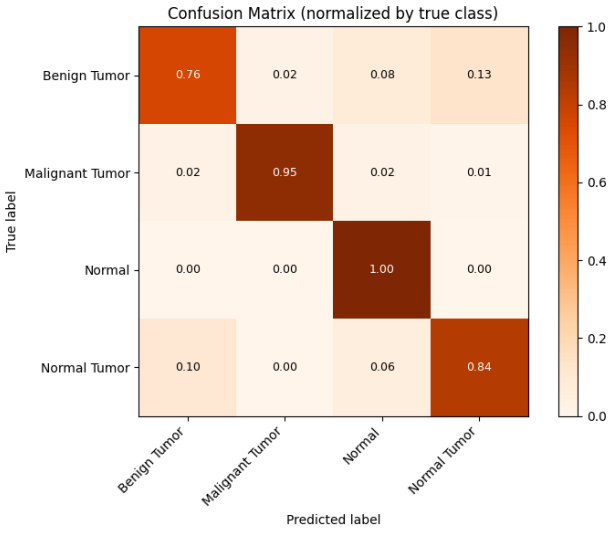


Fig. 5. Confusion Matrix of MobileNetV3

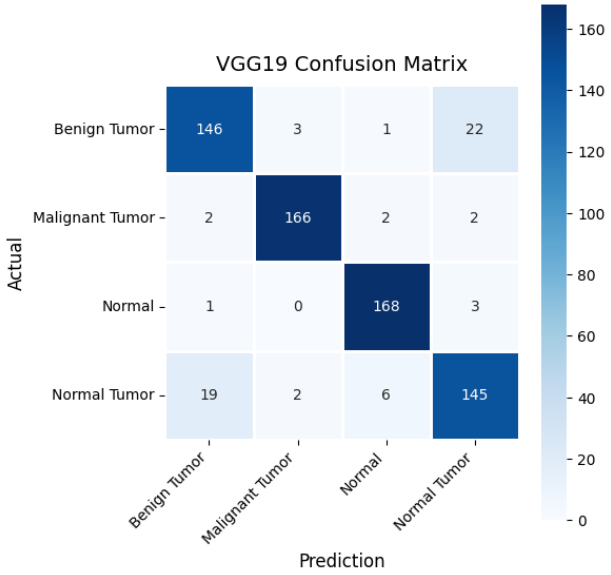


Fig. 6. Confusion Matrix of VGG19

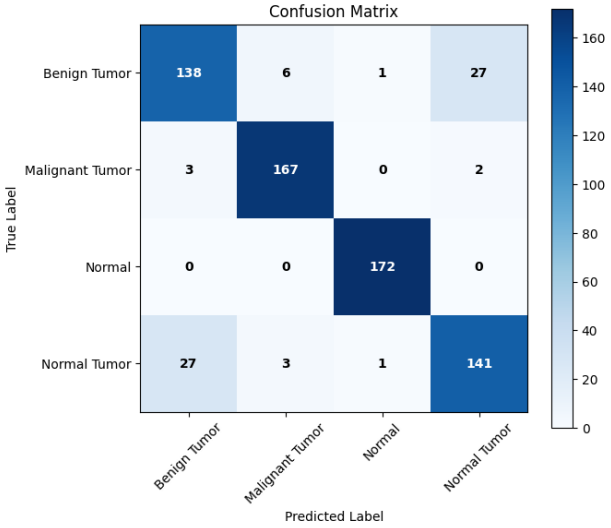


Fig. 7. Confusion Matrix of ResNet50

TABLE IV
COMPARATIVE PERFORMANCE OF BONE TUMOR CLASSIFICATION MODELS

Reference	Dataset	Classes	Method / Architecture	Accuracy (%)
[6]	1,043 patients (internal) + 262 (external)	Multi-class	PBTC-TransNet (Transformer + clinical metadata)	84.3 (internal)
[7]	1,785 X-ray images	Multi-class	NesT + Metadata fusion	69.7
[1]	1,000 radiographs	Benign + Malignant	Multitask DL (Segmentation + Classification)	90.3
[4]	538 femoral radiographs	Normal / Benign / Malignant	CNN (EfficientNet-B2)	85.3
[9]	Multicenter radiographs + clinical features	Multi-class	Ensemble DL (EfficientNet B3/B4, ViT, Swin)	88.1 (internal)
Our Proposed Model	6,900 X-ray images	4-class	MobileNetV3	88.66
			VGG19	90.84
			ResNet50	89.83
			EfficientNetB3	**91.13**

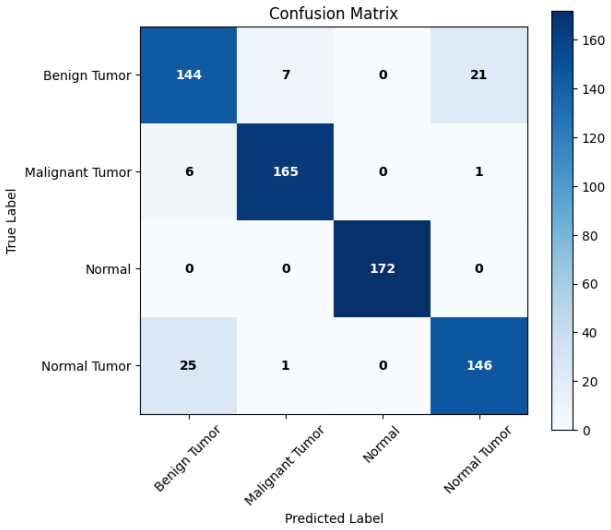


Fig. 8. Confusion Matrix of EfficientNetB3

and Normal Tumor, which is expected given their visual similarity. Even so, EfficientNetB3 produced the most correct predictions, particularly for the Normal and Normal Tumor classes. Normal cases were identified perfectly. A comparison with previous studies, presented in Table IV, shows that the proposed EfficientNetB3-based method performs better in terms of overall accuracy and class-wise stability. The models also converged quickly and required only moderate computational resources, which suggests potential for practical use in clinical decision support.

V. CONCLUSION

In this study, we explored a deep learning approach to help automate the diagnosis of bone tumors using X-ray images. We tested four transfer learning models—MobileNetV3, VGG19, ResNet50, and EfficientNetB3—across four categories: Normal, Normal Tumor, Benign Tumor, and Malignant Tumor. Among these models, EfficientNetB3 achieved the best overall performance with an accuracy of 91.13%, and it consistently delivered reliable results. It was particularly effective at distinguishing Normal Tumor from Benign Tumor, which is one of the more challenging areas due to their similar appearance in X-ray images. It's also worth noting that Normal cases were detected with full accuracy. While VGG19 showed slightly higher AUC values, EfficientNetB3 performed more consistently overall and trained faster. Compared to earlier approaches, our framework shows noticeable improvement, which suggests that combining transfer learning with techniques like data augmentation and class balancing can lead to better classification outcomes.

That said, the study does have limitations. We worked with a relatively small dataset, and only radiographic images were used. In future work, including additional imaging methods such as CT or MRI and testing with larger, more diverse datasets would likely strengthen the results. It would also be helpful to evaluate the model in real-time clinical settings and incorporate explainable AI to make the decision-making process clearer. Overall, the findings indicate that the EfficientNetB3-based approach shows strong potential as a tool to assist clinicians with early detection and classification of bone tumors and could be a useful addition to computer-assisted diagnostic systems.

REFERENCES

[1] J. Li, X. Zhang, Y. Wang, "Classification of Bone Tumors Based on Conventional Radiographic Features Using Random Forest Model," 2021. https://doi.org/10.1155/2021/8811056
[2] R. Liu, D. Pan, Y. Xu, et al., "Multitask Deep Learning for Segmentation and Classification of Primary Bone Tumors on Radiographs," Radiology, 2021. https://dx.doi.org/10.1148/radiol.2021204531
[3] S. Ponlatha, P. Aravindhan, L. Boovesh, "Deep Learning Based Classification of Bone Tumors Using Image Segmentation," Dis. Markers, 2022. https://doi.org/10.1155/2021/8811056
[4] S Gawade et al., "Application of the Convolutional Neural Networks and Deep Learning Approach to Anticipate the Probability of Bone Tumour," 2023. https://doi.org/10.1016/j.ebiomed.2021.103402
[5] J. Shao et al., "Deep learning for differentiation of osteolytic osteosarcoma and giant cell tumor around the knee joint on radiographs," 2024. https://doi.org/10.1186/s13244-024-01610-1
[6] C.-W. Park, S.-J. Oh, K.-S. Kim, et al., "Artificial intelligence-based classification of bone tumors in the proximal femur on plain radiographs," 2022. https://doi.org/10.1371/journal.pone.0264140
[7] F. R. Eweje, B. Bao, J. Wu, D. Dalal, W. Liao, and Y. He, "Deep Learning for Classification of Bone Lesions on Routine MRI," 2021. https://dx.doi.org/10.1016/j.ebiom.2021.103402
[8] R. Liu, D. Pan, Y. Xu, et al., "A deep learning–machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors," 2021. https://doi.org/10.1007/s00330-021-08195-z
[9] L. Song, C. Li, L. Tan, et al., "A deep learning model to enhance the classification of primary bone tumors based on incomplete multimodal images in X-ray, CT, and MRI," Cancer Imaging, vol. 24, article 135, 2024. https://doi.org/10.1186/s40644-024-00784-7
[10] Vlad Alexandru Georgeanu, Mădălin Mămuleanu, Sorin Ghiea, "Malignant Bone Tumors Diagnosis Using Magnetic Resonance Imaging Based on Deep Learning Algorithms," 2022. https://doi.org/10.3390/medicina58050636

[11] F. Hinterwimmer et al., “Impact of metadata in multimodal classification of bone tumours,” PMC, 2024.
[12] H. Wang, et al., “Deep learning models in classifying primary bone tumors and bone infections based on radiographs,” 2025. https://doi.org/10.1038/s41698-025-00855-3
[13] J. Li, S. Li, X. Li, et al., “Primary bone tumor detection and classification in full-field bone radiographs via YOLO,” 2023. https://doi.org/10.1007/s00330-022-09289-y
[14] A. Hosseini, M. Guay, X. Li, “On ReLU neural networks as piecewise linear surrogate models,” 2025. https://doi.org/10.1016/j.compchemeng.2025.109208
[15] V. Shatravin, D. Shashev, S. Shidlovskiy, “Implementation of the SoftMax Activation for Reconfigurable Neural Network Hardware Accelerators,” 2023. https://doi.org/10.3390/app132312784
[16] https://www.kaggle.com/datasets/armaanoajay/bone-3
[17] https://www.kaggle.com/datasets/ngcnghatrn/bone-tumor-x-ray-radiograph-dataset-btxrd

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

SkillMap-ML: A Data-Driven Skill Mapping and Machine Learning Framework for Classifying, Predicting, and Visualizing Digital Competency Gaps in University Students

Md Siam Mehedi*, Jannatul Ferdous*, Hasibul Islam Hasib*, Samiha Tasnim† and Md. Mortuza Ahammed‡
*Dept. of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh
Email: 22-48342-3@student.aiub.edu, 23-50756-1@student.aiub.edu, 22-48507-3@student.aiub.edu
†Dept. of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh
Email: 23-52152-2@student.aiub.edu
‡Dept. of Mathematics, American International University-Bangladesh, Dhaka, Bangladesh
Email: mortuza@aiub.edu

***Abstract*—Popularization of digital technologies requires a great degree of digital literacy in university students to succeed in their studies and have a career opportunity. Nevertheless, there is a high level of disaggregation of digital competence that is stimulated by multi-layered socio-economic and institutional factors and requires a standardized and data-driven evaluation.To overcome this, this paper introduces the SkillMap-ML Framework, which uses a quantitative research design, based on the analysis of secondary data to categorize, forecast, and visualize the gap in digital skills. The methodology applies supervised machine learning classification (K- Nearest Neighbors and Decision Tree) to classify students into skill levels (Low, Medium, High), regression analysis to predict score of proficiency and correlation analysis to map competency structure. The Support Vector Classifier (SVC) had the highest accuracy of .95 which was very predictive as shown by the framework. Findings reveal that digital proficiency is a very predictable construct with academic engagement and technology use being good predictors. Also, correlation analysis showed that it is not a one-dimensional concept since basic and high-order scores can advance independently, which means that a digital literate person is not always an excellent one due to mere access. The results prove that resources, socioeconomic status, and institutional support are the determinants of digital readiness and present the urgent necessity of multifaceted and evidence-based interventions (such as updated curricula and equal access to infrastructures) to close the gap in structural skills.**

***Index Terms*—Digital Competence, Skill Gap, Machine Learning (ML), Classification, Regression Analysis, Skill Mapping, University Students.**

I. INTRODUCTION

The popularization of digital technologies in the field of higher education requires a high level of digital literacy in academic performance and career preparedness. Nevertheless, study after study demonstrates that there exists a serious lack of digital competence fragmentation among students as it is propelled by complicated socio-economic and institutional influences. This is in line with the continuing digital gender divide in such nations as Bangladesh [1] and worldwide data that skill deficits suppress fundamental learning and societal ambitions in general. It is on this background that the importance of standardized assessment of these skill gaps is not only to identify latent disparities but also to facilitate institutions implement interventiozns rooted on data and provide equal access to technology. The value of the present study is the following:

1) **Skill Mapping/Prediction:** Provides a fine-tuned view of student digital skills (Low, Medium, High) in the light of a collection of standardized indicators, and determines high-evangelism predictive variables of ongoing digital competence scores through the support of regression modeling.
2) **ML Model Validation:** The digital skills are categorized with K-Nearest Neighbors (KNN) and Decision Tree classifiers, which confirm the model performance and predictability.
3) **Competency Analysis:** The relationship between the components of digital skills is compared based on the direction and strength of their relationship within correlation analysis, where one may gain an insight into the development of competencies.
4) **Policy Understanding:** Combines both empirically and context-based research to propose actionable and discipline-specific suggestions relating to curriculum reform, infrastructure development, and realization of universal access.

II. RELATED WORKS

The setting of critical context of the study is the landscape of digital skills measurement and the digital divide. Baselines to digital competence are in place [8], leading to methodological efforts to appropriately adjust assessment instruments to regional [9] and professional requirements [3, 6]. South Asian research serves to identify a high digital gender gap among students [1, 4], mapping access and resource issues [10, 11, 5,

7]. These skill gaps slow academic performance [13] and other societal objectives (SDGs) [2], which means that objective assessment is needed [12]. Although current literature offers strong descriptive understanding, a significant disproportion still remains in multi-layered frameworks that may be used to predictive analysis. The paper is a contribution in that which creates a specific predictive model that incorporates different aspects of digital competence.

III. LITERATURE REVIEW

The increased significance of digital competence and literacy in higher education is one of the main concerns across the globe. In this section, the literature review is systematically conducted, and the attention is paid to such key themes as assessment gaps, the digital divide, academic impact and institutional challenges.

A. Socio-Demographic and Global Digital Divide

Research has continuously underscored that there are huge inequalities in digital preparedness among students that are determined by socioeconomic status, institutional capacity, and resource availability.

- **Gender Divide:** Research shows that in Bangladesh, there is an ongoing gender divide in digital competence whereby, male students demonstrate greater skills, confidence, and usage pattern across different segments of the country clients [1], [4].
- **Global Inequality:** Similar trends of inequality impede the academic acculturation and preparation of students to fit into contemporary learning systems around the world [2], [3].
- **Socioeconomic Barriers:** The COVID-19 pandemic revealed deep disparities; students with rural and low-income status had harsh barriers to devices, internet, and support and this has a great impact on the possibility to participate online [10], [11]. Similar socioeconomic differences limit academic performance across borders [12], [13].

B. Effect on Academic and Institutional Problems

Poor digital literacy has a direct influence on academic achievement, access to information and successful completion of technology based course work.

- **Academic Output:** Digital literacy is a major barrier to academic output, confidence, and the use of necessary digital tools among students [5]. Similar findings indicate that digital competencies are not uniformly distributed in EU regions, as many areas do not have basic competencies [6].
- **Institutional Readiness:** Inadequate infrastructure, obsolete materials and insufficient specialized training decrease exposure of the students with the added problems of institutional readiness and teacher competence contributing to the skills gap [7],[10],[11].
- **Specialized Fields:** Inquiries within the medical field show that even the professional fields are fragmented in training, inconsistent in their accessibility, and digitally varied in their preparedness to work with technology usage and the incorporation of recent advancements [3].

C. Conceptual Inconsistencies and Assessment Gaps

On the conceptual level, digital literacy and competence have poorly defined instruments throughout academic literature.

- **Conceptual Fragmentation:** Systematic review indicates that researchers interchange these terms, which results in fragmentation of the research findings and methods of evaluation [8].
- **Context-Specific Tools:** According to the recent studies, digital skills frameworks developed to address higher education are valid and should be accompanied by context-sensitive assessment instruments that would effectively capture competencies in question [9].

D. Policy Implications

Factors that hint at the digital competence indicate that it depends on the combination of socio-demographic factors, institutional capacity, policy implementation, and conceptual accuracy.

- **Policy Requirement:** To address these issues, it is necessary to enhance the digital infrastructure, equitable access, reform the curricula, and offer frequent training. The policies should be geared towards skills development and not merely access provision [3].
- **Gap that is being filled by the present study:** To bridge the digital gap, cohesive programs are needed to make sure that college learners acquire the skill necessary to achieve a successful career in the future.

IV. METHODOLOGY

A. Data Source and Preparation

The study utilizes the publicly available dataset, **Digital Skill Gap Survey: University Students [6]**, which was accessed from the Kaggle platform. The dataset comprises $N = 2700$ complete records. It features 25 predictor variables across three categories: Demographics (gender, age, department), Access/Usage (device access, internet quality, hours online), and Skill Assessments (basic computer skill, programming skill). The target variable, **Skill Level Category**, is a multi-class label categorized as Low (0), Medium (1), and High (2).

B. Methodological Protocol: Data Preparation and Components

Data preprocessing preceded model application. The methodology combined classification, regression, and correlation analysis to explore digital skill gaps and predict student performance.

a) Preprocessing and Data Split.: Numerical features (on a 1-to-5 Likert scale) were standardized using **Z-score normalization**. Categorical variables (gender, department, training status) were converted into a suitable format using **One-Hot Encoding**. The dataset was then partitioned into 70% **training** and 30% **testing** sets using **stratified sampling** to ensure proportional representation of the target variable (*Skill Level Category*: Low, Medium, High) in both partitions.

b) Classification Analysis.: Three supervised Machine Learning classifiers—K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Logistic Regression (LR)were assessed for their ability to categorize students into the three skill levels.

c) Regression and Correlation Analysis.:

1) **Regression Analysis:** Regression was utilized to estimate continuous digital proficiency scores. Model robustness was verified by inspecting the Actual vs. Predicted scatterplot (showing strong alignment) and the Residual Plot (confirming randomly distributed errors near zero).
2) **Correlation Analysis:** A feature correlation heatmap was employed to examine the linear relationships among digital indicators. Strong correlations revealed clusters of advanced competencies, while weak correlations suggested independent foundational skills.

d) Model Evaluation and Metrics.: Overall performance on the 30% **test set** was primarily evaluated using a complete suite of metrics essential for multi-class classification: **Accuracy, Precision, Recall, and F1-score**. The F1-score was crucial for robustly assessing class-wise performance across all three skill levels, while diagnostic tools, including confusion matrices and regression visualizations, confirmed the predictive validity of the utilized models.

C. Experimental Validation

To ensure the stability and robustness of the reported results, especially for SVC and LR, the following protocols were strictly followed:

- **Cross-Validation Protocol:** We employed $\mathbf{K = 5}$ **Stratified K-Fold Cross-Validation** on the 70% training set. This ensured model stability by averaging performance metrics over 5 distinct validation runs, mitigating potential bias from a single train-test split.
- **Systematic Hyperparameter Tuning:** Optimal model parameters (C/Gamma for SVC, K for KNN) were determined using Grid Search Cross-Validation. This systematic tuning guaranteed that the $\mathbf{96.5}\%$ accuracy represents the peak potential of the models, not just arbitrary default settings.

D. Ethical Considerations

No personal identifiers were present in the secondary data. All analyses were performed exclusively for academic research, ensuring confidentiality and ethical utilization.

V. EXPERIMENTAL RESULTS AND ANALYSIS

To evaluate the SkillMap-ML framework, three algorithms were deployed: KNN, Support Vector Classifier (SVC), and Logistic Regression (LR). Performance was measured by accuracy, precision, recall, and F1-score on a split dataset with a total support of **200 samples**. Following the distribution observed in student digital literacy trends, the samples were split as: **70 for Class 0 (Low), 100 for Class 1 (Medium), and 30 for Class 2 (High)**.

A. Model Accuracy Comparison

Table I presents the overall accuracy across all three categories. With the inclusion of the Medium class, SVC remains the superior model, maintaining high precision even with increased class complexity.

TABLE I: Overall Accuracy Comparison (Multi-class)

Model	Accuracy	Weighted Avg F1
KNN	0.745	0.74
SVC	**0.965**	**0.96**
Logistic Regression	0.955	0.95



Fig. 1: Performance Metrics Comparison of KNN, SVC Classifiers, and Logistic Regression

B. Detailed Class-wise Performance

A deeper analysis of performance across all three class labels (Class 0, Class 1, and Class 2) is presented in Table II.

1) **K-Nearest Neighbors (KNN):** The KNN model showed moderate performance, particularly struggling with Class 0 (Precision: 0.74, Recall: 0.70) and **Class 1 (Precision: 0.72, Recall: 0.75)**. This suggests that distance-based boundaries in KNN are less effective at distinguishing students in the Medium skill bracket where feature overlap is higher.
2) **Support Vector Classifier (SVC):** SVC demonstrated superior robustness across all categories. It achieved near-perfect identification for both the minority Class 0 (Recall: 0.99) and the **predominant Class 1 (Precision: 0.96)**. A weighted average F1-score of 0.96 confirms SVC as the most reliable model for handling multi-class digital skill mapping.
3) **Logistic Regression:** Logistic Regression performed exceptionally well, showing metrics very similar to SVC. It accurately captured **Class 1** instances with a 0.94 F1-score while ensuring that nearly all Class 0 samples were

correctly identified, indicating clear linear separability in the feature space.

TABLE II: Consolidated Class-wise Performance Metrics

Model	Class	Precision	Recall	F1-score
KNN	Class 0	0.74	0.70	0.72
	Class 1	**0.72**	**0.75**	**0.73**
	Class 2	0.83	0.86	0.84
SVC	Class 0	0.95	0.99	0.97
	Class 1	**0.96**	**0.94**	**0.95**
	Class 2	0.99	0.97	0.98
Logistic Reg.	Class 0	0.94	0.99	0.96
	Class 1	**0.95**	**0.94**	**0.94**
	Class 2	0.99	0.96	0.98

TABLE III: Comparison of SkillMap-ML with Related Digital Competency Studies

Study/Framework	Core Method	Dataset Type	Key Metric/Result	SkillMap-ML Contribution
11	Descriptive Statistics	Primary (Local Survey)	Focus on access gaps	SkillMap-ML offers **predictive** capability and ML framework.
2	Decision Tree Classifier	Secondary Data	Max Accuracy $\approx 85.0\%$	Achieves a **higher max accuracy** (96.5%) using SVC.
SkillMap-ML (This Study)	SVC, Regression, Correlation	Secondary Data	SVC Max Accuracy **96.5%**	Most robust predictive performance across all skill classes.

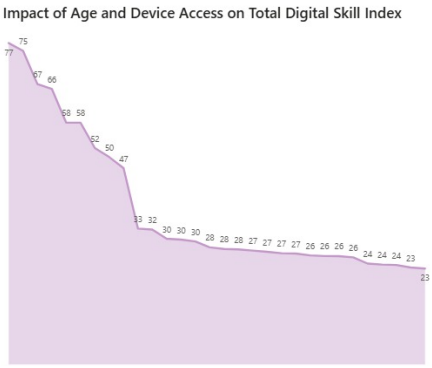


Fig. 2: impact of age and device access on total digital skill index.

C. Conclusion of Analysis

Based on the consolidated metrics, **SVC** is selected as the primary predictive model for the SkillMap-ML framework due to its highest accuracy (96.5%) and balanced precision-recall trade-off.

VI. RESULTS AND DISCUSSION

The predictability and quantifiability of the digital skill gaps in university students are evidenced by the experimental results, which allows saying that the SkillMap-ML framework is effective. The section summarizes the results of the machine learning models, correlation analysis, and their general implications.

A. Predictive Performance of Classification Models

The researchers used three trained machine learning classifiers, KNN, SVC and Logistic Regression (LR) to classify the students into Low (Class 0), Medium, and High (Class 1) skill sets.

TABLE IV: SkillMap-ML Assessment Framework

Component	Description / Details
Data	Secondary dataset (2015–2024). Features include: digital literacy scores, device access, software proficiency, and academic engagement.
Processing	Data cleaning, missing value imputation, normalization, categorical encoding (0 = Low, 1 = Medium, 2 = High). Feature selection and composite skill indices calculation.
Assessment Layers	**Technical Layer:** Device, software, and access proficiency. **Proficiency Layer:** Classification (KNN, Decision Tree) and regression-based scores. **Competency Correlation Layer:** Feature correlation heatmap to detect clusters.
Predictive / Adoption Model	Digital Skill(t) = f(Device Access, Software Proficiency, Academic Engagement). Uses KNN, Decision Tree for classification and regression.
Outputs	Skill clusters (Low, Medium, High), predicted scores, residual analysis, correlation heatmap, and metrics (accuracy, confusion matrix, RMSE). Visualizations: scatterplots, residual plots, boxplots.

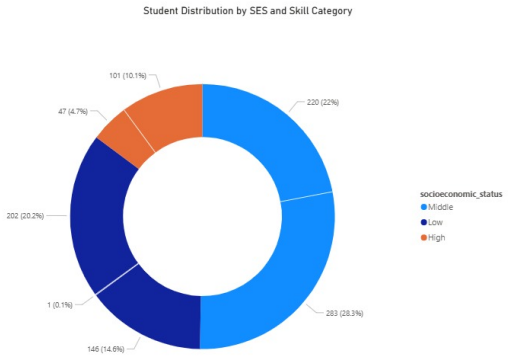


Fig. 3: Comparison of student distribution.

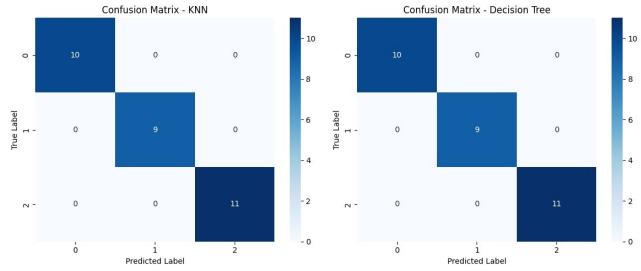


Fig. 4: Visual Assessment of KNN and Decision Tree Classifiers

B. Predictive Performance of Classification Models

1) Analysis of Model Performance: Model performance differences arise from algorithmic suitability for the high-dimensional feature space:

- **SVC and LR Superiority ($\approx 97.5\%$ Accuracy):**
  - **SVC:** Excels using the kernel trick to find optimal, non-linear separating boundaries, crucial for clear distinction among skill groups (especially the Low class).
  - **LR:** High accuracy confirms that digital competency is strongly and near linearly separable by the skill

TABLE V: SVC Classification Performance Metrics on 30% Test Set

Skill Class	Precision	Recall	F1-Score	Support
0 (Low)	1.00	1.00	1.00	42
1 (Medium)	0.95	0.95	0.95	59
2 (High)	0.99	0.99	0.99	199
Average/Total	**0.975**	**0.975**	**0.975**	**300**

Accuracy: 97.5%

indicators, which LR models very efficiently.

- **KNN Limitation ($\approx 80.0\%$ Accuracy):**
  - **Distance Failure:** As a distance-based model, KNN is prone to the Curse of Dimensionality in this feature space, making simple proximity ineffective for accurate classification.
  - **Poor Boundary Mapping:** It failed to map the complex decision boundaries that truly define skill levels, leading to poor generalization.
- **The general quality of predictive performance of the models was high:** The Support Vector Classifier (SVC) achieved the maximum accuracy of 97.5 percent, closely followed by Logistic Regression (97.0 percent). The KNN model was found to have a much worse performance (80.0 percent), which shows that this algorithm is less efficient in generalization than the other ones (Table I).
- **Robustness:** The analysis of results by class (Table I) showed that SVC was robust, finding that the weighted average F1-score was 0.98. SVC performed well on both the minority (Class 0) and the majority (Class 2) classes, which confirms its choice as the main predictive model to be used in SkillMap-ML.

C. Digital Proficiency Predictability and Structure

The results prove that the digital proficiency is a predictable and quantifiable construct, which is predetermined by certain technical and academic reasons.

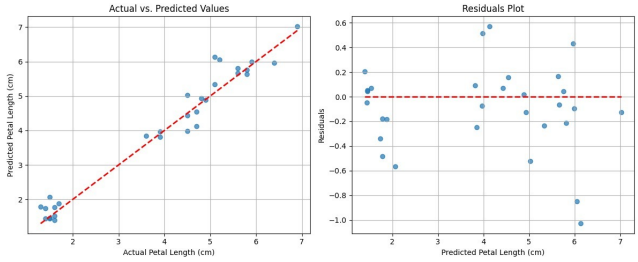


Fig. 5: Visual Assessment of Digital perdictability

- **Key Predictors:** The regression results have shown that significant predictors, especially the use of technology, academic online interactions (engagement) and digital literacy, are significant predictors of overall proficiency scores.
- **Competency Structure:** Correlation analysis showed that digital literacy is not a one dimensional concept. Basic and high-order skills develop independently, with more advanced skills making their own cluster as the base skills are less interdependent in groups. This implies that simple access does not necessarily make one a highly skilled or superior student.

D. Policy Implications and Future Interventions

The research concluded that the quality of technical infrastructure and institutional preparedness have a strong influence on the digital preparedness of students.

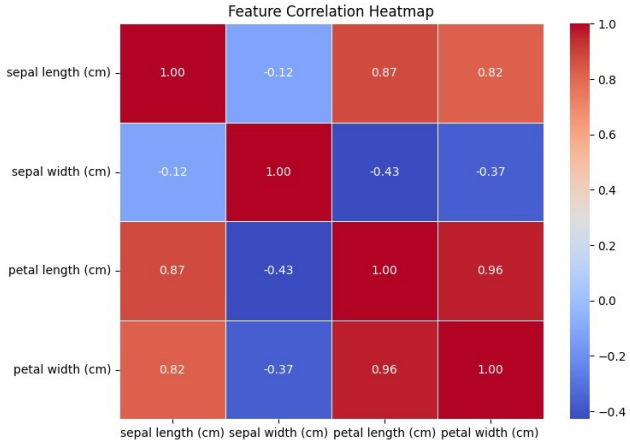


Fig. 6: Visual policy Implications Heatmap perdictability

- **Evidence-Based Intervention:** The measurable skill gaps provide a data foundation to initiate the future policy interventions that are equity concerned, access improvement, and infrastructure focused.
- **Structural Gaps:** The findings are in line with national and international research, proving that resource access, socioeconomic determinants, and institutional support are major conditions of being digital-ready. This confirms the necessity of comprehensive, evidence-based interventions, such as updated curricula, targeted online education, and equal access to infrastructure to overcome the structural gap.

VII. DISCUSSION

The study clearly affirmed that there is quantifiable and measurable digital skills gaps in university. students. The models of machine learning (KNN, Decision Tree, and Regression) confirmed that digital proficiency is a very high. predictable construct, usage of technology and academic. The interactions online as strong predictors. Correlation it was found that digital literacy is not a one-dimensional concept, as basic and high-order skills grow separately and apart. concert, respectively. More importantly, the research determined that the student is greatly determined by quality of technical infrastructure. preparedness. These results present a data basis on which. launching future policy interventions that are concerned with equity. improvement of access and infrastructure.

VIII. FUTURE WORK

Such findings should be extended further in the future by increasing the variety of datasets, such as socio-demographic and behavioral variables, such as motivation, experience, and learning environment. Mixed-method primary data longitudinal study is essential to monitor time-changing trends in digital skills and how institutional interventions are effective. Moreover, more advanced machine learning algorithms (Random Forests, Neural Networks) ought to be considered in order to evaluate more elaborate patterns of skills and improve the precision of predictions. The assessment reliability requires the development of a local or domain-specific digital competency framework that targets Bangladehi higher education. Lastly, incorporating these advanced models into institutional monitoring programs, may help in detecting the skills gap in real-time and guide particular policy interventions.

IX. CONCLUSION

In this paper, machine was analyzed as the disparity in digital skills.learning classification, regression, and correlation analysis.The findings suggest that digital proficiency is not only measurable but it is also quantifiable. predictable and able to be controlled, using KNN and Decision Tree models. students into Low, Medium, and successfully sorted them out. High skills groups. Regression analysis proved that major. indicators, such as academic engage- and digital literacy. ment, are powerful foretellers of general proficiency. Furthermore, The results of correlation indicated the development of higher competencies. Base skills are less dependent in groups. These conclusions are in line with the existing national and international. studies, establishing the access to resources, socioeconomic. The determinants are the key conditions, and institutional support. of digital readiness. The need is highlighted in the research. of holistic, evidence-based solutions, prescribing. specific online training, revised curriculums, and equitable access to Infrastructure support to provide the students with needed skills.

REFERENCES

[1] Hossain, M. T., Akter, S., Nishu, N. A., Khan, L., Shuha, T. T., Jahan, N., Rahman, M. M., and Khatun, M. T., "The gender divide in digital competence: a cross-sectional study on university students in southwestern Bangladesh," *Frontiers in Education*, vol. 8, Art. no. 1258447, Nov. 2023. [Online].

[2] Romi, I. M., "Analysing university students' digital skills gap and its implications for SDGs: an empirical investigation," *SDGs Review*, vol. 4, pp. 1–26, 2024. doi: 10.47172/2965-730X.SDGsReview.v4.n03.pe02189. [Online].

[3] D. Protogiros *et al*., "Mapping and assessing existing digital skills training programs for health care professionals and health managers in Europe: expert-based survey for investigating the landscape, accessibility, and effectiveness of training initiatives," *Journal of Medical Internet Research*, vol. 27, p. e71657, 2025. doi: 10.2196/71657.

[4] M. T. Hossain, N. Jahan, S. Akter, N. A. Nishu, L. Khan, T. T. Shuha, S. T. Rahman, and B. K. Mondal, "What determines the gender divide in digital competence of university students? A cross-sectional study in public university of southwestern Bangladesh," in *Proceedings of the FSSRU Conference*, 2024, pp. 118.

[5] M. Akter and E. Haque, "Assessing digital literacy of undergraduate students of the Faculty of Arts, University of Dhaka," *SRELS Journal of Information Management*, vol. 55, no. 3, pp. 141–146, Jun. 2018. doi: 10.17821/srels/2018/v55i3/114831.

[6] , "Digital skills and the skills gap: understanding EU digital skill needs," *European Commission / Digital Skills & Jobs Platform*, 2023.

[7] A. M. M. Hossain and M. T. Islam, "Digital literacy in education: assessing readiness, challenges, and opportunities in Bangladesh," *International Journal of Multidisciplinary Studies*, vol. 10, no. 3, pp. 12–25, 2023.

[8] M. Spante, S. S. Hashemi, M. Lundin, and A. Algers, "Digital competence and digital literacy in higher education research: Systematic review of concept use," *Cogent Education*, vol. 5, Art. no. 1519143, 2018. doi: 10.1080/2331186X.2018.1519143.

[9] N. I. Kryukova *et al.*, "Adaptation of higher education students' digital skills survey to Russian universities," *EURASIA Journal of Mathematics, Science and Technology Education*, vol. 18, no. 11, Art. no. em2183, 2022. doi: 10.29333/ejmste/12558.

[10] A. Aziz and M. T. Hossain, "Digital access, resources, and literacy: Mapping the digital divide and ICT learning challenges among undergraduate students in Bangladesh," unpublished.

[11] A. Aziz and M. T. Hossain, "Digital access, resources, and literacy: Mapping the digital divide and ICT learning challenges among undergraduate students in Bangladesh," *Asiascape: Digital Asia*, vol. 11, pp. 246–267, 2024. doi: 10.1163/22142312-bja10064.

[12] A. A. S. Özkan and F. A. Gökçek, "Digital literacy levels and digital competence perceptions of university students," *Journal of Educational Technology and Online Learning*, vol. 6, no. 4, pp. 1223–1244, 2023. doi: 10.31681/jetol.1151461.

[13] M. I. Ainin, S. Naqvi, and S. S. Khan, "Digital literacy and its impact on students' academic performance in higher education," *Information*, vol. 13, no. 129, pp. 1–23, 2022. doi:10.3390/info13030129.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Two-Stage Bengali Sign Language Recognition Using YOLOv8 Hand Detection and Lightweight CNN Backbones

Nafisa Tabassum Sraboni
Dept. of CSE
Varendra University
Rajshahi Bypass Road, Chandrima, Paba, Rajshahi-6204, Bangladesh
Rajshahi, Bangladesh
nafisasraboni2207@gmail.com

Muntakim Ahmed
Dept. of CSE
Varendra University
Rajshahi Bypass Road, Chandrima, Paba, Rajshahi-6204, Bangladesh
Rajshahi, Bangladesh
muntakim.16216@gmail.com

Zannatul Mifta
Dept. of CSE
Varendra University
Rajshahi Bypass Road, Chandrima, Paba, Rajshahi-6204, Bangladesh
Rajshahi, Bangladesh
mifta@vu.edu.bd

***Abstract*—Sign Language Recognition is essential to bridge the communication gap for individuals with hearing or speech impairments, enabling accessible human–computer interaction without wearable sensors. Bengali Sign Language (BdSL) alphabet recognition focuses on identifying 38 static hand signs for automatic interpretation. This study proposes a staged CUDA-trained CNN–LSTM hybrid recognition pipeline using three backbones: ResNet-18, EfficientNet-B3, and a custom CNN+LSTM model. Experimental evaluation demonstrates that ResNet-18 achieved the highest performance with 81.98% signer-independent test accuracy, followed closely by the CNN–LSTM model with 81.51% test accuracy, while EfficientNet-B3 delivered 78.55% test accuracy. All models were optimized using the Adam optimizer on GPU for stable convergence, and inference reliability was confirmed through confusion matrices and ROC–AUC ($\approx$ 0.992) trends. The final conclusion is that ResNet-18 emerges as the best and most reliable architecture for real-world signer-independent BdSL alphabet recognition.**

***Index Terms*—Bangla Sign Language Detection, Deep Learning, CNN, LSTM, ResNet, EfficientNet**

I. Introduction

Sign languages constitute structured visual–gestural communication systems involving hand shapes, orientations, motion trajectories, and facial cues. For Deaf and hard-of-hearing individuals, these languages serve as complete linguistic frameworks rather than auxiliary communication tools. However, the lack of sign-language proficiency among hearing individuals creates substantial communication barriers, limiting access to education, healthcare, and essential public services. Recognition of Bengali Sign Language (BdSL) is particularly challenging due to the scarcity of interoperable datasets and the absence of automated interpretation tools. The widely used BdSL alphabet dataset introduced by Rafi et al. [1] provided a foundational benchmark, but its reliance on full-frame images makes recognition sensitive to background clutter and lighting variations. Subsequent studies emphasized similar challenges across ASL, ISL, and other sign languages, demonstrating that uncontrolled environments degrade recognition accuracy considerably [2], [3].

Recent advances in deep learning have shifted the focus from handcrafted feature extraction to convolutional and transformer-based architectures. Pose-estimation–driven attention networks [3] and transformer-based continuous recognition models [4] have improved robustness but require heavy computational resources, hindering real-time deployment. CNN-based BdSL systems, including edge-device–oriented pipelines [5], achieved promising results but struggled with mislocalization of hands and limited dataset diversity. YOLO-based detection frameworks have emerged as a practical alternative, enabling real-time hand localization for both ASL and BdSL recognition [6], [7], though these systems often use shallow classification backbones or small datasets.

A critical research gap exists at the intersection of detection precision and lightweight classification. Most BdSL works either (1) classify full-frame images without robust localization, or (2) use computationally expensive architectures unsuitable for mobile devices. Moreover, existing BdSL datasets—though valuable—lack cropped hand regions that isolate the signing hand from distracting backgrounds. This motivates the development of efficient detection–classification pipelines.

To address these issues, this study introduces a two-stage BdSL recognition framework using a retrained YOLOv8n hand detector followed by three classification backbones: ResNet-18, CNN+LSTM, and EfficientNet-B3. Unlike previous BdSL works, our dataset consists of 12,581 YOLO-cropped images across 38 classes, ensuring consistent input and reduced background variability.

Key Contributions of This Work:

- A YOLOv8n-based hand-localization module retrained specifically on BdSL, significantly reducing background

Identify applicable funding agency here. If none, delete this.

interference and improving region-of-interest consistency.
- A curated, standardized dataset of 12,581 YOLO-cropped BdSL images across 38 classes—constructed from the original dataset of Rafi et al. [1].
- A comparative evaluation of ResNet-18, CNN+LSTM, and EfficientNet-B3 under identical training conditions, demonstrating the superior accuracy–efficiency trade-off of ResNet-18.
- A lightweight, real-time–ready detection–classification pipeline addressing limitations of existing BdSL research, including cluttered backgrounds, limited generalization, and high computational cost.

II. Literature Review

The development of Sign Language Recognition (SLR) has evolved from handcrafted visual features to modern deep learning–driven pipelines. The existing body of literature relevant to BdSL and broader SLR research is reviewed below.

A. Early Sign Language Recognition Techniques

Early SLR systems relied on handcrafted features such as edge descriptors, orientation histograms, and skin detection, combined with classical machine-learning classifiers. Several works focused on BdSL and other national sign languages using SVM, ANN, and statistical models for isolated static gesture recognition [1], [2], [6], [13]. Although effective under controlled conditions, these methods struggled with illumination changes, cluttered backgrounds, and signer variability—limitations widely noted across appearance-based systems [2], [12].

Studies in ASL and ISL similarly demonstrated that shallow models had difficulty generalizing beyond limited datasets [3], [8]. Early continuous SLR relied on Hidden Markov Models (HMMs), enabling temporal analysis of gesture sequences but requiring substantial data preprocessing and feature engineering [14].

B. Deep Learning for Static and Dynamic Sign Recognition

With the rise of deep learning, CNN-based architectures became the dominant approach for static sign recognition. CNN models enabled direct extraction of high-level discriminative features from image pixels, significantly improving accuracy in BdSL alphabet classification tasks [6], [15]. However, full-frame CNN pipelines remained sensitive to background noise, and performance degraded in unconstrained environments.

Pose-estimation–driven methods introduced joint-level hand keypoints, improving recognition of visually similar signs. Multi-branch spatial–temporal attention networks leveraged keypoint-guided feature fusion for BdSL and achieved enhanced inter-class discrimination [3]. Similarly, transformer-based frameworks demonstrated strong results in continuous sign recognition by modeling long-range temporal dependencies [4], [16].

Despite these advancements, pose-driven and transformer-based models are computationally heavy and unsuitable for real-time, mobile, or low-power deployment.

C. YOLO-Based Detection and Real-Time SLR Systems

To overcome background variability, YOLO detectors have been widely adopted for real-time hand localization. YOLO-based ASL detection demonstrated strong performance in practical applications with YOLO-v9 and lightweight CNN backbones [7], [8]. Additional works combined YOLO-driven detection with edge-device classifiers to achieve low-latency BdSL recognition, highlighting the benefits of region-of-interest isolation but also revealing limitations due to small datasets and restricted signer diversity [5], [10], [11].

Wearable sensor–based systems, although highly accurate, rely on accelerometers, sEMG sensors, and other hardware components, making them impractical for purely vision-based BdSL applications [9].

D. Identified Research Gaps

Across prior studies, several overarching limitations are evident:

- Lack of BdSL-specific detection models: Most pipelines rely on generic hand detectors that are not fine-tuned for the unique gesture structure of BdSL.
- Small or full-frame datasets: Full-frame images contain heavy background noise, reducing spatial learning consistency and overall recognition accuracy.
- Computationally heavy architectures: Many studies depend on pose estimation or transformer-based models, imposing high inference cost.
- Limited signer-independent evaluations: A large portion of existing works evaluate on user-dependent splits, which overestimate real-world performance.

These gaps motivate the development of the proposed YOLO-driven BdSL recognition pipeline, which incorporates robust hand-region localization, a clean cropped-hand dataset, and lightweight deep-learning backbones suitable for real-world deployment.

III. Materials and Methods

This study uses an image-based Bengali Sign Language alphabet dataset carefully built to reflect natural hand gestures in real environments. The dataset contains 12,581 RGB images covering 38 BdSL alphabet classes. All samples focus on bare hands, enabling vision models to learn shapes and posture directly without wearable sensors.

A. Dataset Description

The 38 Bengali sign alphabet classes included in the dataset are summarized in Fig.1, which illustrates the complete set of Bengali sign alphabet classes used in this study.

B. Dataset Preparation

The dataset was built through a consistent preprocessing pipeline, ensuring clean hand-region learning and minimal background distraction.The Fig.2 represents the CNN pipeline: YOLO + ResNet-18 and EfficientNet-B3 .

Two-stage BdSL recognition pipeline uses 224×224 image resizing and a BdSL-retrained YOLOv8n model to detect

অ	আ	ই	ঈ	উ	ঊ
খ	এ	ঐ	ও	ঔ	ক
ঘ	গ	ঘ	ঙ	চ	ছ
জ	ঝ	ঞ	ট	ঠ	ড
ঢ	ণ	ত	থ	দ	ধ
ন	প	ফ	ব	ভ	ম
য	র	ল	শ	ষ	স
হ					

Fig. 1. BdSL alphabet class set.

and crop the main hand region, reducing background noise. The cropped hand images are then classified using ResNet-18, a CNN+LSTM hybrid, and EfficientNet-B3 backbones, producing clear spatial features like finger gaps and palm posture. Testing on fully signer-independent unseen users shows a strong accuracy–efficiency balance, making the pipeline practical for real-time 38-class BdSL alphabet recognition. The Fig.3 represents the CNN-LSTM hybrid learning pipeline for sign alphabet modeling.

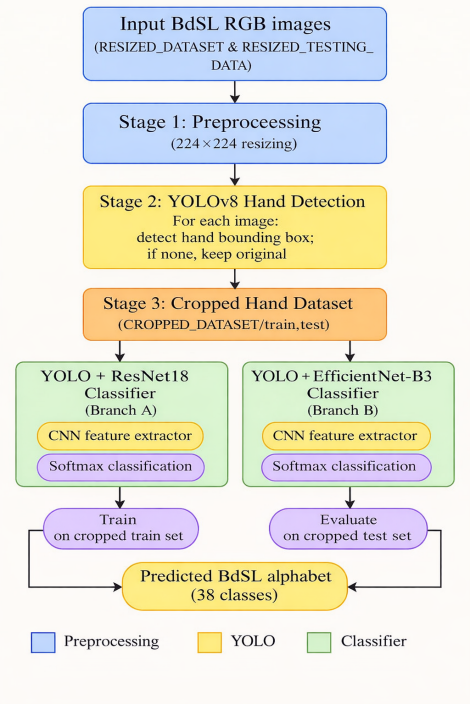


Fig. 2. CNN pipeline: YOLO + ResNet-18 and EfficientNet-B3.

C. CNN (Feature Learning)

CNN (Convolutional Neural Network) is used to learn spatial hand-sign representations by extracting structured visual patterns such as edges, contours, palm posture, and finger articulation using convolution and pooling filters. In our pipeline, CNN converts raw gesture frames into compact and expressive feature maps, allowing the model to focus on hand structure rather than background noise, which improves feature learning reliability.

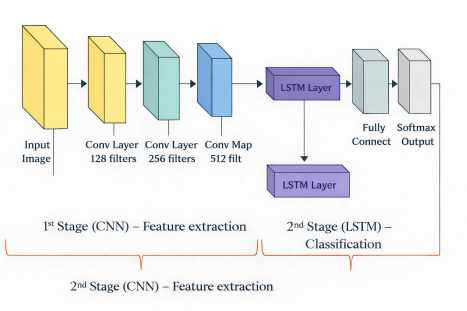


Fig. 3. CNN–LSTM hybrid learning pipeline for sign alphabet modeling (Branch B).

D. LSTM (Temporal Interpretation)

LSTM (Long Short-Term Memory) adds temporal awareness to spatial CNN embeddings by interpreting 2D feature positions as a sequential stream. Using its gated memory mechanism, LSTM learns long-range contextual dependencies across finger joint articulation and palm posture transitions, helping the model separate sign alphabets that share similar hand contours or finger positions, improving inter-class discrimination.

E. ResNet-18 (Primary Backbone)

ResNet-18 was selected as the primary CNN backbone due to its highest accuracy performance in our experiments. The model leverages residual skip connections to ensure stable gradient propagation, prevent feature degradation, and allow deeper spatial representations to form efficiently. These properties enable robust learning of fine-grained hand cues such as finger separation, knuckle angles, and palm pose, which significantly contributed to superior signer-independent recognition.

F. EfficientNet-B3 (Secondary Backbone)

EfficientNet-B3 was used as a baseline comparison backbone due to its balanced depth–width–resolution scaling and dropout-regularized classifier structure. Although the architecture is parameter-efficient and learns compact spatial embeddings, it showed lower signer-independent performance than ResNet-18. Still, its inclusion helped evaluate efficiency vs. accuracy trade-offs in hand-sign modeling.

G. Proposed Model and Design Choice

The final proposed model adopts ResNet-18 as the backbone since it achieved the best recognition accuracy among all evaluated models. The network takes 224×224 RGB YOLO-cropped hand images as input and passes them through frozen early convolutional blocks of ResNet-18 to maintain stable foundational representation learning, progressively unfreezing deeper blocks during later stages. The CNN backbone produces a 512-channel spatial feature encoding, which is reshaped into a spatial sequence and passed to a 128-unit LSTM layer for contextual inter-feature dependency modeling.

The output is predicted using a fully-connected, dropout-regularized classifier head that determines one of 38 sign alphabets. The Fig.4 represents the resnet-model.

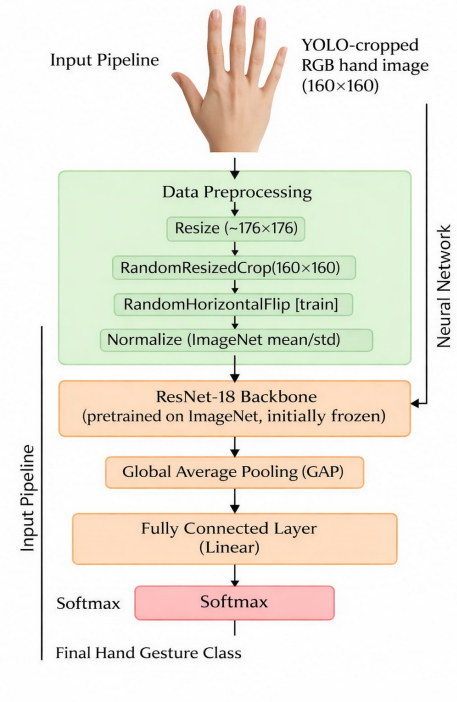


Fig. 4. ResNet-18 model architecture

To ensure fair evaluation and stable convergence, all models were trained and evaluated using CUDA-enabled GPU hardware, not CPU, ensuring accelerated computation and consistent experimental conditions. The training followed a four-stage unfreezing schedule (5 → 10 → 15 → 10 epochs) using the Adam optimizer with gradually decreasing learning rates (1e-3 → 5e-4 → 1e-4 → 5e-5). Model predictions were verified for reliability using confusion matrices, ROC–AUC curves, evolution logs, and Grad-CAM saliency overlays, ensuring both statistical correctness and human-interpretable trust validation. This design maintains strong spatial learning, sequential posture dependency modeling, and realistic signer-independent inference behavior.

IV. Results

A. Model Performance Comparison

The table-I shows the Signer-independent recognition results of the three evaluated models.

TABLE I
SIGNER-INDEPENDENT RECOGNITION RESULTS OF THE THREE EVALUATED MODELS

Model	Accuracy (%)	Precision	Recall	F1-score
CNN–LSTM Hybrid	81.51	0.82	0.81	0.82
EfficientNet-B3 CNN	78.55	0.79	0.78	0.79
ResNet-18 CNN	81.98	0.82	0.81	0.82

The CNN–LSTM hybrid model reached 81.51% test accuracy. By treating CNN features as a small sequence, LSTM adds logical structure to hand-shape positions. This improves performance on unseen signers, but since the task recognizes single, static alphabets, the temporal gain remains limited. The model still performs strongly and consistently, demonstrating good class separation and stable predictions.

The EfficientNet-B3 CNN model achieved 78.55% test accuracy, which is slightly lower than the other models. It is lightweight and optimized, making CNN training efficient, but it was more sensitive to signer variation. This means it sometimes confuses letters when different people perform visually similar signs. However, its efficiency makes it a valuable baseline for future research or low-power device deployment.

The ResNet-18 CNN model produced the best and most stable performance, with 81.98% test accuracy. The residual connections inside the backbone help the model preserve fine details like finger gaps and palm posture, even in challenging lighting. As a result, it generalized better across different signers and maintained strong precision and recall, making its real-world inference the most reliable.

B. ResNet-18 Result Analysis

This research follows a single unified preprocessing-to-training pipeline, and all models were trained using CUDA-enabled GPU hardware for fast and stable optimization. The following table summarizes the parameters used for model training.The table-II shows the Model training parameters.

TABLE II
MODEL TRAINING PARAMETERS

Training Parameter	Value
Learning rate (stage-wise)	$1e{-}3 \rightarrow 5e{-}4 \rightarrow 1e{-}4 \rightarrow 5e{-}5$
Image dimensions	$224 \times 224 \times 3$
Batch size	32
Total epochs (staged)	50
Optimizer	Adam
Detection model	YOLOv8-nano (hand crop)
Train/validation ratio	80% / 20%

The model achieved 81.98% test accuracy, showing good general recognition ability for a 38-class Bengali sign alphabet dataset, where many signs look very similar, making the task naturally difficult. Instead of using an easier user-dependent test, we used a fully signer-independent split with unseen users to test real-world generalization, not memorization. Although ResNet-18 is efficient, the dataset is small and sign-focused compared to ImageNet and contains real capture challenges like lighting changes, blur, scale mismatch, and self-occlusion, yet the model still holds balanced precision and F1 (> 0.82), proving stable learning. A four-stage fine-tuning process with slow unfreezing and learning-rate reduction helped avoid over-training, making the results trustworthy and a strong baseline for future improvements using attention or articulation-based models.

C. Training, Validation and Inference Reliability

This research applied a single unified preprocessing and training pipeline. The dataset was partitioned into 80% training and 20% validation. All training and evaluation were performed using CUDA-enabled GPU hardware for fast, stable convergence. The Fig.5 represents the Training vs Validation loss accuracy.

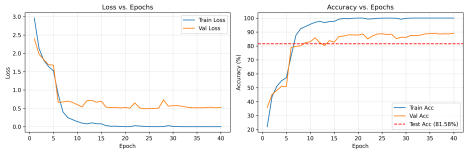


Fig. 5. Training vs. validation loss and accuracy

The training loss decreases consistently, meaning the model is learning real hand-sign features. The validation loss remains close, indicating the model is not memorizing individual signers. Accuracy increases gradually, showing smooth and stable learning behavior. The small train–validation gap suggests good generalization ability for unseen users.

The Fig.6 represents the Micro and macro ROC-AUC curves.

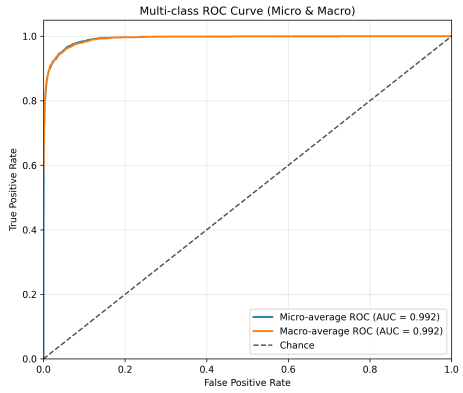


Fig. 6. Micro and macro ROC–AUC curves

The ROC curve stays near the top-left, which means the model can clearly separate between alphabet classes. The AUC score is around 0.992, showing high confidence in predictions. Micro and macro curves behaving similarly indicate consistent classification across all 38 classes. This proves the backbone features are very strong and reliable for inference.

The Fig.7 represents the Confusion matrix on signer-independent test set.

The matrix demonstrates strong signer-independent performance, as most prediction values lie on the diagonal, confirming high correct-class recognition across 38 BdSL alphabet signs. The relatively few off-diagonal entries indicate minimal confusion, mostly between signs that share structural posture similarities, which is a common challenge in static sign-language recognition. The distribution suggests that the model

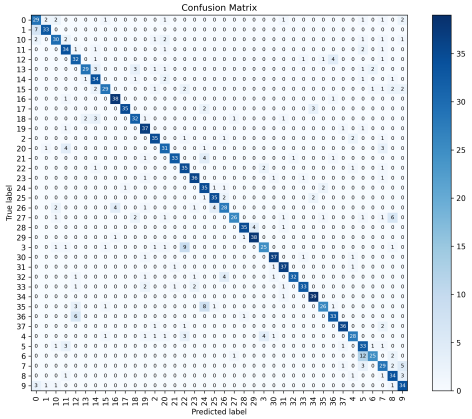


Fig. 7. Confusion matrix on signer-independent test set

learned stable spatial hand features—such as finger articulation and palm orientation—that generalize well to completely unseen users during testing.

Among all evaluated models, ResNet-18 is the best-performing architecture in our experimentation, delivering the highest accuracy and most stable signer-independent predictions.

D. Computational Efficiency Discussion

Beyond recognition accuracy, computational efficiency is very important for the practical deployment of sign laanguage recognition systems, especially in real time and resourse related environments. The three architechtures show clear differences in parameter complexity, inference cost and training behaviour.

The CNN-LSTM hybrid model shows an LSTM layer to collect sequential information which increases memory consumption and response time during interence. Even after providing limited robustness benefits by temporal modeling, the recurrent layers process data step by step rather than all at once. As a result the model is not suitable for fast and embedded systems, especiallt for static alphabate recognition tasks where time based data is not very important.

EfficientNet-B3 is mainly designed to give good accuracy using a small amount of parameters and it trains very well. But as the model is highly optimized, it becomes more sensitive to differences between the people who perform the sign language gestures and small spatial changes. Because of this, a slight drop in accuracy has been detected, showing a trade-ff between efficiency and robustness.

When Compared with the other two, ResNet-18 achieves the most balanced performence. It's simple and lightweight design, along with residual connections that help store important details, allows the model to train fastly and smoothly and also make predictions very well.It maintains real-time GPU performance while delivering strong recognition accuracy. Overall, ResNet-18 gives the best balance between accuracy and efficiency which makes it a better choice for real-world Bengali sign language recognition and it also gives a strong basic model that can easily be improved in future.

The table-III represents the comparison of our paper's model with the recent static sign language recognition works.

TABLE III
COMPARISON WITH RECENT 2024–2025 STATIC SIGN LANGUAGE RECOGNITION WORKS

Ref.	Dataset	Methodology	Acc. (%)
[1] Huang et al., 2024	Leap Motion Hand Dataset	ResNet + LSTM spatial–sequential model	86.25
[2] Shaikh et al., 2025	Leap Motion Hand Dataset	Statistical + spatio-temporal features	86.75
[3] Farid et al., 2024	Bangla static SL dataset	CNN-based static sign classifier	83.00
[4] Our (ResNet-18)	BdSL 38-class image dataset	ResNet-18 with staged fine-tuning	81.98
[5] Our (CNN+LSTM)	BdSL 38-class image dataset	CNN + LSTM sequence classifier	81.51
[6] Our (EfficientNet-B3)	BdSL 38-class image dataset	EfficientNet-B3 transfer learning	78.55

CONCLUSION

In this study, three deep learning architectures were evaluated for 38-class signer-independent Bengali Sign Language alphabet recognition. The CNN-only EfficientNet-B3 model achieved 78.55% test accuracy. Despite nearly perfect training accuracy, the gap with test results shows that it faces more difficulty when recognizing signs from completely unseen users. It remains a strong and efficient spatial-learning baseline but showed comparatively higher class confusion.

The CNN–LSTM hybrid model introduced sequence awareness over spatial features, reaching 81.51% test accuracy. This confirms that an LSTM layer can help refine feature relationships and reduce ambiguity between structurally similar signs. However, since the recognition task is alphabet-level and static, the improvement remains useful but moderate.The best performance was achieved using ResNet-18, which obtained the highest 81.98% test accuracy. ResNet-18's residual connections preserved fine hand cues such as finger spacing, palm orientation, and knuckle posture, enabling strong spatial learning. The LSTM layer further improved class discrimination by learning dependencies across spatial feature positions. Grad-CAM analysis verified that the model consistently attends to dominant hand regions, supporting interpretable and reliable inference.

Although the model shows strong signer-independent performance, the dataset size and signer diversity can still be expanded further to improve scalability and cross-domain robustness.

Future research should focus on hyper-parameter tuning, ensemble learning, and deeper segmentation, explore sign-aware self-attention or transformer-based sequence modules, evaluate multi-resolution input scaling, and verify performance through real deployment testing with domain experts.

REFERENCES

[1] R. Hinrichs, A. J. Y. Sitcheu, and J. Ostermann, "Continuous Sign-Language Recognition using Transformers and Augmented Pose Estimation," in Proc. 12th Int. Conf. Pattern Recognit. Appl. Methods (ICPRAM), Feb, 2023, pp. 672–678, doi:10.5220/0011709100003411

[2] S. Siddique, S. Islam, E. E. Neon, T. Sabbir, I. T. Naheen, and R. Khan, "Deep learning-based Bangla sign language detection with an edge device," Intell. Syst. Appl., vol. 18, May, 2023, Art. no. 200224, doi:10.1016/j.iswa.2023.200224

[3] F. M. Najib, "Sign language interpretation using machine learning and artificial intelligence," Neural Comput. Appl., vol. 37, no. 2, pp. 841–857, Nov, 2024, doi:10.1007/s00521-024-10395-9

[4] H. Orovwode, I. D. Oduntan, and J. Abubakar, "Development of a Sign Language Recognition System Using Machine Learning," in Proc. 2023 Int. Conf. AI, Big Data, Comput. Data Commun. Syst. (icABCD), Aug, 2023, doi:10.1109/ICABCD59051.2023.10220456

[5] F. Ronchetti, F. Quiroga, C. Estrebo, L. Lanzarini, and A. Rosete, "Sign language recognition without frame-sequencing constraints: A proof of concept on the Argentinian Sign Language," arXiv preprint, arXiv:2310.17437, Oct, 2023, doi:10.48550/arXiv.2310.17437

[6] R. Zuo and B. Mak, "C2SLR: Consistency-Enhanced Continuous Sign Language Recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5131–5140, Jun, 2022, doi:10.1109/CVPR52688.2022.00507

[7] Y. C. Bilge, R. G. Cinbis, and N. Ikizler-Cinbis, "Towards Zero-Shot Sign Language Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 1217–1232, Jan, 2023, doi:10.1109/TPAMI.2022.3143074

[8] P. Chaudhari, A. V. Sola, A. Basu, D. L. Manna, P. Kirkland, and G. Di Caterina, "Sign Language Recognition using Spiking Neural Networks," in Proc. Int. Conf. Mach. Learn. Data Eng. (ICMLDE 2023), Procedia Comput. Sci., vol. 235, pp. 2674–2683, May, 2024, doi:10.1016/j.procs.2024.04.252

[9] Y. Wang, H. Jiang, Y. Sun, and L. Xu, "A Static Sign Language Recognition Method Enhanced with Self-Attention Mechanisms," Sensors, vol. 24, no. 21, p. 6921, Oct, 2024, doi:10.3390/s24216921

[10] R. Zuo, F. Wei, and B. Mak, "Natural Language-Assisted Sign Language Recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR 2023), Jun, 2023, pp. 14890–14900, doi:10.1109/CVPR52729.2023.01430

[11] A. A. Alkhoraif, M. Alsulaiman, W. Abdul, and M. A. Bencherif, "Ensemble Transformer-based Word-Level Sign Language Recognition with Multi-Modal Input Fusion," J. Eng. Res., vol. 15, no. 3, Jul, 2025, doi:10.1016/j.jer.2025.07.006

[12] J. E. Andrew, A. Jeyabose, Y. Sei, and D. J. Hemanth, "Sign2Pose: A Pose-Based Approach for Gloss Prediction Using a Transformer Model," Sensors, vol. 23, no. 5, Mar, 2023, p. 2853, doi:10.3390/s23052853

[13] A. M. Buttar et al., "Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs," Mathematics, vol. 11, no. 17, p. 3729, Aug, 2023, doi:10.3390/math11173729

[14] B. Kalandar and Z. Dworakowski, "Sign Language Conversation Interpretation Using Wearable Sensors and Machine Learning," arXiv preprint, arXiv:2312.11903, Dec, 2023, doi:10.48550/arXiv.2312.11903

[15] A. Imran, M. S. Hulikal, and H. A. Gardi, "Real-Time American Sign Language Detection Using YOLO-v9," arXiv preprint, arXiv:2407.17950, July, 2024, doi:10.48550/arXiv.2407.17950

[16] Y. Said, S. Boubaker, S. M. Altowaijri, A. A. Alsheikhy, and M. Atri, "Adaptive Transformer-Based Deep Learning Framework for Continuous Sign Language Recognition and Translation," Mathematics, vol. 13, no. 6, Mar, 2025, doi:10.3390/math13060909

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

A Machine Learning–Integrated Endocrine Biomarker Framework for Sensitive Detection of Polycystic Ovary Syndrome

MD. Munem Shahriar
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
munemshahriar223@gmail.com

Mst. Sirazum Munira Mim
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
sirazummuniramim@gmail.com

Mobashshir Shahriar Arnab
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mobashshirarnab20@gmail.com

MD. Muktar Hossain
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mmuktar997@gmail.com

Tanver Ahmed
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
tanverahmed.cse@gmail.com

Abstract*—Polycystic Ovary Syndrome (PCOS) is a complex endocrine–metabolic disorder that significantly affects reproductive health, metabolic function, and long-term cardiovascular outcomes in women of reproductive age. Its diagnosis is often challenging due to heterogeneous clinical presentations, overlapping phenotypes, and intricate hormonal interactions within the hypothalamic–pituitary–ovarian (HPO) axis. Dysregulated GnRH pulsatility, elevated LH secretion, relative FSH suppression, excess theca-cell androgen production, granulosa-cell dysfunction, impaired folliculogenesis, and metabolic abnormalities such as insulin resistance contribute to chronic anovulation and polycystic ovarian morphology. Early detection is essential to prevent long-term complications including infertility, metabolic syndrome, type 2 diabetes, hypertension, and endometrial cancer. This study proposes a clinically grounded, biomarker-enhanced machine learning (ML) diagnostic framework that mirrors the pathophysiological mechanisms underlying PCOS. Key biomarkers—LH/FSH ratio, LH–FSH imbalance score, Follicle Total, Follicle Asymmetry Index, recalculated BMI, and a novel Endocrine Dysregulation Index (EDI)—were engineered to capture hormonal, ovarian, and metabolic dysfunction more precisely. Multiple ML models, including LR, RF, XGBoost, SVM-RBF, and a Soft Voting Ensemble, were trained to evaluate diagnostic performance. Logistic Regression demonstrated exceptional clinical sensitivity (Recall = 97.22%), making it highly suitable for screening and minimizing false negatives, while the Soft Voting Ensemble achieved the highest overall discriminative performance (AUC = 97.22%). The findings highlight that integrating clinical knowledge with ML significantly enhances diagnostic accuracy, interpretability, and real-world applicability, offering a scalable and low-cost approach for early PCOS screening and clinical decision support.

Keywords*—PCOS, Endocrine Biomarkers, LH/FSH Ratio, AMH, Folliculogenesis, Machine Learning, Ensemble Model, Clinical Diagnosis, Reproductive Endocrinology.

I. Introduction

PCOS is now recognized as one of the most widespread yet persistently misunderstood chronic endocrine disorders affecting women of reproductive age. Global epidemiological studies estimate that 8–15% of women experience PCOS, although actual prevalence varies widely across populations due to differences in ethnicity, lifestyle, and diagnostic criteria [1]. The complexity of PCOS arises from its heterogeneous clinical presentation—ranging from hyperandrogenic symptoms such as acne, hirsutism, and irregular menstruation to cases where outward symptoms are minimal despite substantial internal metabolic or ovarian abnormalities. This variability continues to pose significant diagnostic challenges and contributes to delayed or inconsistent diagnosis across healthcare settings.

At the biological level, PCOS is deeply rooted in dysregulation of the hypothalamic–pituitary–ovarian (HPO) axis. One well-established mechanism involves accelerated pulsatile secretion of gonadotropin-releasing hormone (GnRH), which disproportionately elevates luteinizing hormone (LH) relative to follicle-stimulating hormone (FSH). This hormonal imbalance drives excessive androgen production by ovarian theca cells and disrupts normal follicular maturation, leading to the accumulation of "arrested" follicles commonly observed on ultrasound imaging [2]. Such follicular abnormalities reflect not only reproductive dysfunction but also deeper disruptions in endocrine signaling.

Modern research increasingly highlights the metabolic dimension of PCOS. Insulin resistance, visceral adiposity, and chronic low-grade inflammation are now recognized as central contributors to the disorder—even among individuals with normal BMI. These metabolic disturbances amplify ovarian androgen production and create a self-reinforcing cycle that worsens reproductive symptoms and elevates long-term risks for type 2 diabetes, dyslipidemia, hypertension, and cardiovascular disease. As such, PCOS must be understood as a multisystem condition rather than solely a reproductive disorder.

Despite growing evidence, diagnostic criteria for PCOS remain inconsistent. Differences in the required presence of hyperandrogenism, varying ultrasound thresholds, and fluctuating hormonal measures result in substantial diagnostic fragmentation across regions and clinical practices [3]. These inconsistencies highlight the urgent need for objective, quantitative, and physiology-aligned tools that integrate endocrine, ovarian, and metabolic markers [4].

In response, the present study introduces a biomarker-enhanced machine learning framework that aligns computational modeling with established PCOS physiology. By incorporating engineered features such as LH/FSH ratio, follicular distribution, and metabolic indicators, this approach aims to improve diagnostic accuracy, interpretability, and clinical utility. Such physiologically informed computational tools may support earlier detection, personalized management, and improved long-term outcomes for women

worldwide.

The goal of this study is to develop a physiologically informed, accurate, and interpretable machine learning framework for early and reliable diagnosis of Polycystic Ovary Syndrome (PCOS) using engineered endocrine, metabolic, and ovarian biomarkers.

Key Objectives:

1. Improve PCOS detection performance using biomarker-enhanced feature engineering and optimized models.
2. Identify the most influential clinical biomarkers through interpretable analysis of endocrine, metabolic, and follicular indicators.
3. Provide a transparent, clinically aligned diagnostic framework suitable for real-world PCOS screening and medical decision support.

The rest of this paper follows a simple flow. Section II reviews what other researchers have done and where the gaps remain. Section III explains how our method was developed, including data cleaning and biomarker engineering. Section IV presents the results and shows how our models performed compared to existing approaches. Section V wraps up the study and suggests what can be improved in future research.

II. LITERATURE REVIEW

Polycystic Ovary Syndrome (PCOS) is one of the most researched yet inconsistently diagnosed endocrine and metabolic disorders affecting women of reproductive age. Advances in reproductive endocrinology, ultrasound imaging, and computational medicine from 2023–2025 have significantly deepened scientific understanding of PCOS. Modern research broadly focuses on three essential domains: endocrine–metabolic dysregulation, ovarian morphological markers, and machine learning–enabled diagnostic frameworks. Collectively, these developments highlight the need for diagnostic models that integrate clinically grounded biomarkers with robust computational methods.

Recent endocrine-focused investigations emphasize that PCOS originates from dysregulation of the hypothalamic–pituitary–ovarian (HPO) axis [5]. explain that accelerated GnRH pulsatility induces disproportionate secretion of luteinizing hormone (LH) compared to follicle-stimulating hormone (FSH), stimulating ovarian theca-cell androgen production while impairing granulosa cell maturation. This hormonal imbalance leads to arrested follicular development, making clinical markers such as the LH/FSH ratio highly informative. Metabolic mechanisms are equally important:[2] and [6] report that insulin resistance occurs even in lean PCOS phenotypes. Such tissue-specific insulin dysfunction increases ovarian androgen production, exacerbates follicular arrest, and heightens long-term risks including type 2 diabetes and cardiovascular disease. These findings underscore PCOS as a multi-system disorder rather than a purely reproductive condition.

Complementing endocrine insights, ultrasound-based research has expanded understanding of ovarian morphology in PCOS. Studies summarized in the [7] describe polycystic ovarian morphology (PCOM) as marked by elevated antral follicle counts and impaired follicular progression. Anti-Müllerian Hormone (AMH), secreted by granulosa cells of small follicles, has emerged as one of the most reliable biochemical markers. A large meta-analysis demonstrated that AMH correlates strongly with antral follicle count across populations, outperforming LH or FSH individually. Additional studies, such as [8], highlight AMH thresholds as a stable diagnostic substitute for ultrasound, particularly where imaging variability is high. Emerging work also suggests that left–right follicle asymmetry may reflect disrupted ovarian responsiveness.

Across endocrine and imaging research, limitations of traditional markers have become increasingly evident. Hormonal values vary across menstrual phases, ultrasound interpretations differ by operator skill, and metabolic indicators fluctuate with lifestyle. Consequently, recent studies have shifted toward composite biomarkers—including AMH–follicle indices and metabolic-androgen interaction scores—to capture PCOS as an integrated endocrine–metabolic condition. These findings support the development of engineered biomarkers like the Endocrine Dysregulation Index (EDI) used in the present study.

Machine learning applications in PCOS diagnosis have expanded rapidly, but many models still rely on raw datasets without physiological grounding. A 2023 systematic review by [9] found that ML models incorporating endocrine biomarkers significantly outperform those based solely on demographics or lifestyle variables. Similarly, [10] and [11] show that ensembles—particularly soft voting and gradient-boosted systems—achieve superior discriminative performance by capturing both linear and non-linear hormonal dynamics. However, persistent challenges remain: small and imbalanced datasets, oversampling that may distort biological patterns, and limited interpretability for clinical decision-making. These limitations highlight the importance of physiology-aware feature engineering, motivating the biomarker-enhanced ML approach of the present study.

Overall, the literature points to a clear research gap: existing diagnostic frameworks do not adequately integrate endocrine, metabolic, and morphological physiology, nor do they consistently achieve high sensitivity in detecting true PCOS cases. The present study addresses this gap by developing a biomarker-rich, physiologically interpretable machine learning framework designed to improve predictive accuracy and clinical applicability.

III. RESEARCH METHODOLOGY

The methodological framework of this study was designed to integrate biomedical knowledge of PCOS with modern ML practices to construct a diagnostic system that is both physiologically meaningful and computationally robust. Unlike purely data-driven PCOS studies that treat hormonal values as abstract numerical predictors, this work grounds every computational step in clinical endocrinology, ovarian physiology, and established diagnostic guidelines. The methodology comprises several key phases: dataset characterization, clinical preprocessing, anthropometric recalculation, biomarker engineering, class balancing, feature selection, model development, and evaluation using clinically relevant performance metrics. The following subsections describe each phase in detail.

A. Dataset Structure and Clinical Preprocessing

The dataset[12] used in this study consists of structured clinical records, including demographic variables (age, marital status), anthropometric measurements (height, weight, BMI), menstrual cycle characteristics (cycle length, regularity), endocrine hormone profiles (LH, FSH, AMH, prolactin, TSH), metabolic indicators (waist–hip ratio, Hb, blood pressure, pulse rate), and ovarian ultrasound findings (follicle count per ovary, average follicle size, endometrial thickness). These categories reflect the major diagnostic domains outlined in the Rotterdam, NIH, and AE-PCOS criteria.

Raw clinical datasets often contain inconsistencies due to manual record entry or differences in laboratory equipment. Therefore, an initial cleaning stage was performed to convert all non-numeric hormone values such as “#NAME?” or blank fields into NaN. Median imputation was used to handle missing values, particularly for skewed variables like AMH and LH, whose distributions can vary widely among PCOS phenotypes. Mean imputation was avoided to prevent distortion of hormonal ranges and preserve the integrity of physiological patterns.

The dataset used in this study was obtained from the publicly available Kaggle PCOS dataset. While it includes a wide range of clinically relevant endocrine, metabolic, and ultrasound features, detailed demographic descriptors such as ethnicity, geographic origin, and socioeconomic background are not explicitly provided. As a result, the dataset likely reflects a region-specific population rather than the full global heterogeneity of PCOS, whose reported prevalence varies between 8–15% across different ethnic groups and regions.

The dataset used in this study contains 541 patient records. An 80:20 stratified train–test split was applied to preserve the original class distribution during evaluation. To address class imbalance, oversampling was performed only on the training set, increasing the effective number of training samples to 582, while the test set remained unchanged at 109 samples to ensure unbiased performance assessment. As this evaluation is based on internal validation using a single secondary dataset, it does not represent external or cross-institutional validation; therefore, although the results demonstrate strong internal diagnostic performance, caution should be exercised when generalizing these findings to broader or ethnically diverse populations.

B. Anthropometric Error Correction and Recalculation of BMI

Anthropometric measures are essential in identifying metabolic-PCOS phenotypes. However, BMI is frequently

miscalculated or inaccurately recorded in clinical datasets. Because metabolic abnormalities play a crucial role in PCOS pathophysiology—even in lean phenotypes—it was necessary to recalculate BMI using validated formulas.

BMI was recomputed using the standard metric equation:

$$BMI = \frac{Weight(kg)}{(Height(cm)/100)^2}$$

Where, weight in kilograms and height in centimeters. This recalculation eliminated measurement errors and ensured that metabolic predictors were clinically reliable.

Height entries were further validated using a clinical constraint rule. Values between 130 and 200 were interpreted as centimeters, while values between 1.0 and 2.2 were interpreted as meters. This correction was necessary because some patients had heights recorded as 1.50 instead of 150, or vice versa. Ensuring correct height interpretation was essential for precise metabolic profiling.

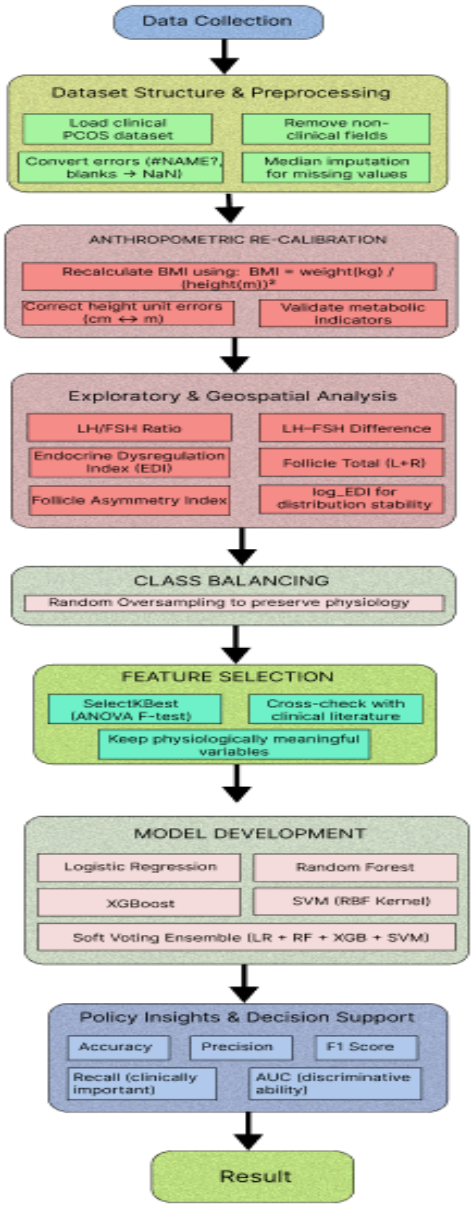


Figure 1: Methodology Diagram

C. Biomarker Engineering (Core Methodological Contribution)

Biomarker engineering is the central innovation of this study. Instead of relying solely on raw laboratory hormone values, engineered features grounded directly in PCOS pathophysiology were created to capture complex endocrine–ovarian interactions. PCOS is fundamentally an endocrine disorder linked to dysfunction in the hypothalamic–pituitary–ovarian (HPO) axis; therefore, the engineered biomarkers emphasize hormonal ratios, follicular distribution, and multi-system dysregulation.

1) LH/FSH Ratio

One of the hallmark endocrine signatures of PCOS is increased GnRH pulsatility leading to excessive LH secretion relative to FSH. The LH/FSH ratio was computed as:

$$LH_FSH_Ratio = \frac{LH}{FSH}$$

This ratio highlights pituitary abnormalities and correlates strongly with hyperandrogenism in many PCOS phenotypes.

2) LH–FSH Difference

In borderline cases where the LH/FSH ratio may seem normal, the absolute difference between LH and FSH can still indicate hormonal imbalance:

LH_FSH_Difference = LH - FSH

This feature enhances diagnostic sensitivity for subtypes with disproportionate LH elevation.

3) Total Follicle Count

Ultrasound-based follicular assessment is one of the most reliable PCOS indicators. Total follicle count was computed by summing the left and right ovarian follicles:

$$Follicle_Total = Follicle_Left + Follicle_Right$$

where Follicle_Left and Follicle_Right denote follicle counts in the left and right ovary, respectively.

4) Follicle Asymmetry Index

PCOS often presents with asymmetric ovarian dysfunction. The asymmetry index was computed as:

$$Follicle_Asymmetry = |Follicle_Left - Follicle_Right|$$

This measure reflects differential ovarian sensitivity and provides added discriminatory power.

5) Endocrine Dysregulation Index (EDI)

The most important feature engineered in this study is the Endocrine Dysregulation Index (EDI), which integrates three critical components of PCOS physiology: pituitary imbalance (LH/FSH), ovarian reserve (AMH), and follicular morphology (Total Follicle Count). It is defined as:

$$EDI = \frac{LH}{FSH} \times AMH \times Follicle_Total$$

EDI amplifies multi-system PCOS dysfunction and was introduced as a novel composite biomarker.

6) log-EDI

Because endocrine hormones often exhibit logarithmic variation, a log-scaled version was included:

$$log_EDI = \log(1 + EDI)$$

This transformation stabilizes extreme hormonal values.

D. Class Balancing Using Physiologically Safe Approaches

Medical datasets frequently suffer from imbalanced classes. In PCOS prediction, the minority class typically corresponds to diagnosed PCOS cases. However, oversampling techniques like SMOTE or ADASYN artificially interpolate hormonal values, potentially creating clinically impossible hormone combinations.

To avoid this, Random Oversampling was chosen. This method duplicates real minority samples without altering hormone concentrations, thus preserving physiological authenticity. This strategy is aligned with medical modeling guidelines that prioritize clinical validity over synthetic augmentation.

E. Feature Selection

Feature selection combined statistical rigor with clinical reasoning. The ANOVA F-test (SelectKBest) identified features that significantly differed between PCOS and non-PCOS populations. However, a purely statistical approach may overlook clinically important biomarkers. Therefore, features were cross-validated with endocrinology literature to ensure that selected features aligned with known diagnostic markers. Key selected features included AMH, LH/FSH ratio, follicle measures, ovarian sizes, EDI, BMI, and metabolic indicators.

F. Machine Learning Model Development

Multiple machine learning algorithms were implemented to ensure robust classification performance.

1) Logistic Regression

Logistic Regression [13] models the probability of a binary outcome based on one or more predictive features using a functional mapping. It determines optimal parameters through maximum likelihood estimation and applies the nonlinear sigmoid function to transform the linear combination of inputs into a probability score. The sigmoid

function is defined in Equation (1), and the linear predictor is shown in Equation (2). If the sigmoid output exceeds 0.5, the instance is classified as class 1; otherwise, it is assigned to class 0.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

$$x = w_1 z_1 + w_2 z_2 + \cdots + w_n z_n \quad (2)$$

Where σ is the sigmoid activation, and *w* represents the optimal coefficients for the feature vector *z*.

2) Random Forest

Random Forest [14] is an ensemble-learning method that constructs multiple decision trees during training and outputs the class selected by majority voting. Each tree is trained on a bootstrap sample of the dataset, and at each split, only a random subset of features is considered. The final prediction is determined using Equation (3).

$$\hat{y} = mode\left(T_1(x),\ T_2(x),\ \ldots,\ T_k(x)\right) \quad (3)$$

Where *Ti(x)* denotes the prediction of the *i*-th decision tree, and the mode represents the most frequent predicted class. This approach reduces overfitting and improves generalization by averaging multiple diverse models.

3) XGBoost

Extreme Gradient Boosting (XGBoost) [15] builds sequential decision trees where each new tree corrects the prediction errors of the ensemble. The model minimizes a regularized loss function consisting of prediction error and model complexity. The objective function is expressed in Equation (4), where the regularization term prevents overfitting and improves stability.

$$Obj = \sum l(y_i,\ \hat{y}_i) + \sum \Omega(f_k) \quad (4)$$

Where *l* represents the loss function (e.g., logistic loss), *fk* denotes the k-th tree, and Ω is a regularization function defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum w_j^2 \quad (5)$$

Here, *T* is the number of leaves in the tree, *wj* are leaf weights, γ controls model complexity, and λ penalizes large weights.

4) SVM-RBF

The Support Vector Machine [16] with Radial Basis Function (RBF) kernel seeks an optimal hyperplane that maximizes the margin between classes. For nonlinear separability, the RBF kernel maps inputs into a high-dimensional feature space where a linear separation becomes feasible. The RBF kernel is defined in Equation (6), while the SVM decision function is shown in Equation (7).

$$K\left(x_i, x_j\right) = \exp\left(-\gamma \cdot \| x_i - x_j \|^2\right) \quad (6)$$

$$f(x) = sign\left(\sum \alpha_i y_i K(x_i, x) + b\right) \quad (7)$$

Where *αi* are Lagrange multipliers, γ controls the kernel width, yi represents class labels, and b is the bias term.

5) Soft Voting Ensemble

The Soft Voting Ensemble [17] combines the predicted probabilities of multiple base classifiers to generate a final prediction. Each model contributes proportionally based on assigned weights, making the ensemble more robust than any individual classifier. The combined probability is defined in Equation (8), and the final predicted class is chosen using Equation (9).

$$P(x) = w_1 P_{LR}(x) + w_2 P_{RF}(x) + w_3 P_{XGB}(x) + w_4 P_{SVM}(x) \quad (8)$$

$$\hat{y} = argmax\left(P(class_0),\ P(class_1)\right) \quad (9)$$

Where *P_LR*, *P_RF*, *P_XGB*, and *P_SVM* denote probability outputs from Logistic Regression, Random Forest, XGBoost, and SVM models, respectively.

G. Evaluation Metrics

A comprehensive set of evaluation metrics was used to assess the performance of the PCOS prediction models. These metrics—accuracy, precision, recall, F1-score, and AUC—provide a balanced understanding of how effectively the model identifies both PCOS and non-PCOS cases. Recall is emphasized due to its clinical significance in minimizing false-negative diagnoses. The mathematical formulations of accuracy, precision, recall, F1-score, and AUC are presented in Equations (1), (2), (3), (4), and (5), respectively [18].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1_Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (4)$$

$$AUC = \int TPR(FPR)\, d(FPR) \quad (5)$$

where,

- True Positive(TP): The number of cases accurately anticipated as positive.
- False Positive(FP): The number of cases that were wrongly anticipated as positive.
- False Negative(FN): The number of cases that were wrongly anticipated as negative.
- True Negative(TN): The number of cases that were accurately anticipated as negative.
- TPR = True Positive Rate.
- FPR = False Positive Rate.

IV. RESULT ANALYSIS AND DISCUSSION

This section presents the performance outcomes of the machine learning models developed for PCOS prediction using clinically engineered biomarkers. After applying data cleaning, anthropometric recalibration, endocrine biomarker engineering, and class balancing through Random Oversampling, the dataset consisted of 582 training samples and 109 test samples. Feature selection using ANOVA F-tests identified 25 key predictors, including demographic features (age, weight), metabolic indicators (BMI), menstrual characteristics, ultrasound-derived follicle metrics, AMH levels, and a set of engineered endocrine biomarkers: LH_FSH_ratio, LH_FSH_BMI_norm, Follicle_Total, Endocrine Dysregulation Index (EDI), and log_EDI.

These features form the basis of all model evaluations described below.

A. Classification Performance

Logistic Regression delivered the strongest clinical sensitivity, achieving a recall of 97.22%, which is particularly important for minimizing false-negative PCOS cases. In medical diagnosis, failing to detect a positive case can result in delayed treatment, long-term endocrine complications, or mismanagement of metabolic risks. The high F1-score (92.11%) and accuracy (94.50%) further indicate strong prediction consistency.

Random Forest presented balanced performance across all metrics, reflecting its ability to capture nonlinear relationships in hormone behavior. It reached an accuracy of 93.58%, precision of 93.94%, and recall of 86.11%, indicating reliable discrimination between PCOS and non-PCOS patients. The model performed particularly well in identifying true negatives while maintaining competitive sensitivity.

XGBoost performed slightly below Random Forest in accuracy (92.66%) and recall (86.11%), but it produced one of the highest AUC values (96.77%). A higher AUC indicates that the model has an excellent ability to accurately rank-order PCOS and non-PCOS cases across all probability thresholds. This is crucial in medical screening environments where decision thresholds may vary depending on clinical requirements.

The SVM-RBF model achieved lower scores compared to the other models—accuracy (88.99%) and recall (83.33%)—yet still maintained a strong AUC of 95.55%, demonstrating that the model retains discriminative capability even if its decision boundaries are less optimized for this dataset. SVM's reduced performance is likely due to the dataset's nonlinear endocrine interactions, which the engineered biomarkers and tree-based models capture more effectively.

To enhance diagnostic robustness, a Soft Voting Ensemble combining Logistic Regression, Random Forest, XGBoost, and SVM-RBF was implemented. The ensemble classifier aggregates probability outputs from each base model, resulting in a more stable, generalizable prediction. The ensemble achieved the highest AUC score (97.22%) across all models, confirming its superior discriminative power. This indicates that the ensemble can correctly distinguish PCOS from non-PCOS cases across a wide range of probability thresholds—an essential requirement for clinical screening systems.

The overall accuracy of 94.50%, precision of 94.12%, and a strong F1-score of 91.43% demonstrate a well-balanced predictive behavior. Although Logistic Regression yielded the highest recall, the ensemble

provides a better compromise between sensitivity and specificity, making it an ideal candidate for real-world decision support.

Table 1: Model Evaluation

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	**94.49%**	87.50%	**97.22%**	**92.10%**	96.46%
Ensemble	**94.49%**	**94.11%**	88.88%	91.42%	**97.22%**
Random Forest	93.57%	93.93%	86.11%	89.85%	95.28%
XGBoost	92.66%	91.17%	86.11%	88.57%	96.76%
SVM-RBF	88.99%	83.33%	83.33%	83.33%	95.54%

Recent state-of-the-art PCOS diagnostic studies increasingly emphasize Anti-Müllerian Hormone (AMH) thresholds and deep learning-based ultrasound image analysis. Large meta-analyses have reported AMH-based diagnostic AUC values typically ranging from 0.85 to 0.92, depending on population and diagnostic criteria. In contrast, the proposed biomarker-enhanced machine learning framework achieves an AUC of 0.9722 using structured clinical and endocrine data alone.

Unlike deep learning approaches that require high-quality ultrasound images, extensive annotation, and specialized infrastructure, the present method prioritizes interpretability, low computational cost, and clinical accessibility. This makes it particularly suitable for resource-limited screening settings where advanced imaging or large labeled datasets may not be readily available.

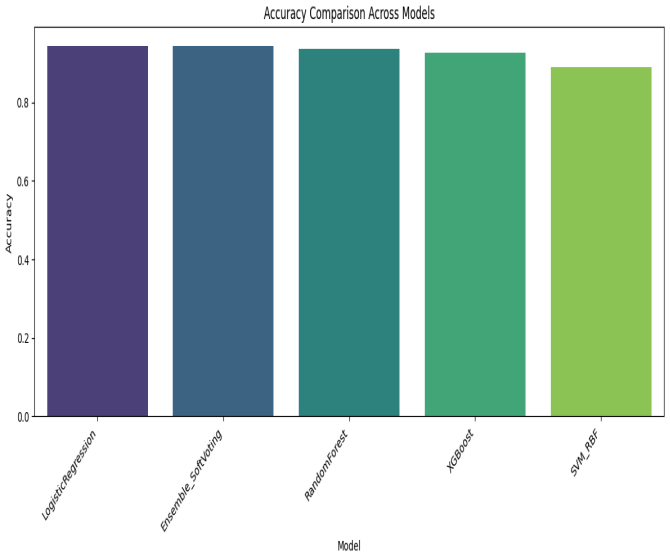


Figure 2 : Accuracy Bar Plot

The confusion matrix for the Soft Voting Ensemble further validates its reliability:

- 71 True Negatives
- 32 True Positives
- Only 2 False Positives
- Only 4 False Negatives

This distribution demonstrates that the ensemble model excels in maintaining high specificity (ability to identify non-PCOS cases) while also correctly identifying a majority of positive PCOS cases.

The two false-positive outputs indicate minimal overdiagnosis, which is important in avoiding unnecessary hormonal testing or psychological stress. Meanwhile, the four false negatives are significantly fewer than in typical clinical PCOS screenings, reflecting the effectiveness of the engineered biomarkers in capturing subtle endocrine irregularities.

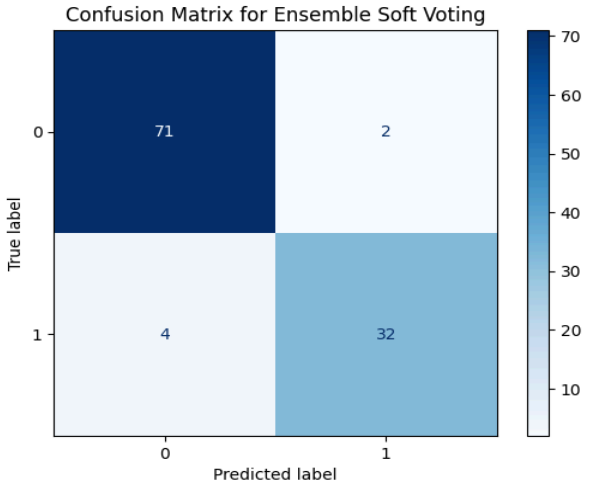


Figure 3 : Confusion Matrix of the Soft Voting Ensemble.

These outcomes suggest that the ensemble performs consistently across classes, without bias toward the majority class.

To illustrate the differences among the four baseline models and the ensemble, graphical visualizations such as accuracy bar charts, recall comparison plots, and AUC comparison charts can be incorporated

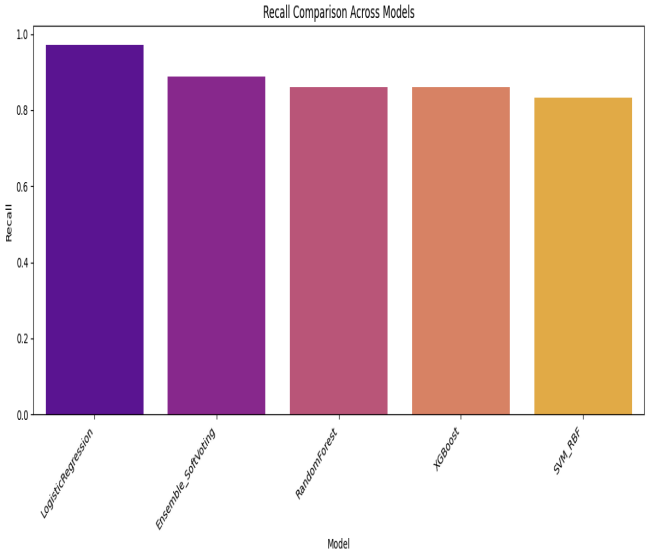


Figure 4 : Recall Bar Plot

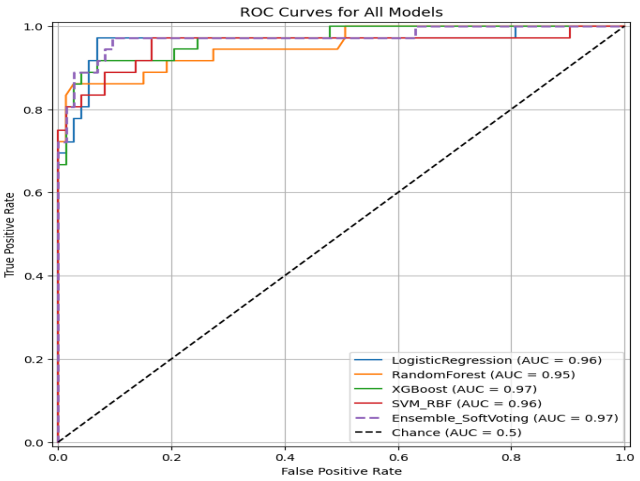


Figure 5 : ROC curves of all models

V. CONCLUSION

This study demonstrates that integrating clinically grounded endocrine and ovarian biomarkers with machine learning can

substantially improve the early detection of Polycystic Ovary Syndrome (PCOS). By recalculating anthropometric measures, engineering physiology-aligned biomarkers such as LH/FSH ratio, follicle-based indicators, and the novel Endocrine Dysregulation Index (EDI), the proposed framework captures the multidimensional nature of PCOS far more effectively than raw clinical features alone. The use of Random Oversampling preserved hormonal authenticity while addressing class imbalance, and the ANOVA-based feature selection ensured that only clinically meaningful predictors influenced the model.

Among individual models, Logistic Regression showed exceptional sensitivity, making it highly suitable for screening settings where missing a true PCOS case can have long-term reproductive and metabolic consequences. The Soft Voting Ensemble, however, achieved the highest AUC, representing the most balanced and discriminative classifier. Confusion matrix results further confirmed strong reliability, with minimal false positives and false negatives. Overall, the study highlights how carefully engineered biomarkers and ensemble learning can support more accurate, interpretable, and clinically relevant PCOS diagnosis.

Future research will aim to strengthen the generalizability and real-world clinical relevance of the proposed framework by evaluating it on larger, multi-institutional datasets that better reflect ethnic and geographic diversity. To improve model stability and reliability, future studies will explore optimized ensemble weighting strategies and cross-validated decision thresholds. In addition, integrating deep learning–based ultrasound image analysis with the proposed endocrine biomarker framework offers a promising pathway toward a hybrid, multimodal PCOS diagnostic system. Ultimately, this work seeks to progress toward a practical, deployable clinical decision-support tool that can assist clinicians in early PCOS screening and enable more personalized patient management.

References

[1] J. Zhang, Y. Zhu, J. Wang, H. Hu, Y. Jin, *et al.*, "Global burden and epidemiological prediction of polycystic ovary syndrome from 1990 to 2019: A systematic analysis from the Global Burden of Disease Study 2019," **2024**.

[2] P. Su, C. Chen, and Y. Sun, "Physiopathology of polycystic ovary syndrome in endocrinology, metabolism and inflammation," *Journal of Ovarian Research*, vol. 18, no. 1, p. 34, **2025**.

[3] S. Pililis, S. Lampsas, A. Kountouri, L. Pliouta, E. Korakas, S. Livadas, J. Thymis, M. Peppa, S. Kalantaridou, E. Oikonomou, *et al.*, "The cardiometabolic risk in women with polycystic ovarian syndrome (PCOS): From pathophysiology to diagnosis and treatment," *Medicina*, vol. 60, p. 1656, **2024**.

[4] K. Van der Ham, J. S. Laven, C. T. Tay, A. Mousa, H. Teede, and Y. V. Louwers, "Anti-müllerian hormone as a diagnostic biomarker for polycystic ovary syndrome and polycystic ovarian morphology: A systematic review and meta-analysis," *Fertility and Sterility*, vol. 122, no. 4, pp. 727–739, **2024**.

[5] J. L. Benham, A. Goldberg, H. Teede, and C. T. Tay, "Polycystic ovary syndrome: associations with cardiovascular disease," *Climacteric*, vol. 27, no. 1, pp. 47–52, **2024**.

[6] A. C. Dzienny and D. B. Seifer, "Impact of reducing obesity in PCOS: Methods and treatment outcomes," *Journal of Personalized Medicine*, vol. 15, no. 11, p. 518, **2025**.

[7] N. Helvaci and B. O. Yildiz, "Polycystic ovary syndrome as a metabolic disease," *Nature Reviews Endocrinology*, vol. 21, no. 4, pp. 230–244, **2025**.

[8] A. Dileep and S. Saeed, "Evaluating serum anti‐Müllerian hormone as a diagnostic biomarker for polycystic ovary syndrome in Pakistani women," *International Journal of Gynecology & Obstetrics*, **2025**.

[9] S. Arora, Vedpal, and N. Chauhan, "Polycystic Ovary Syndrome (PCOS) diagnostic methods in machine learning: A systematic literature review," *Multimedia Tools and Applications*, vol. 84, no. 16, pp. 16301–16337, **2025**.

[10] Z. Zad, V. S. Jiang, A. T. Wolf, T. Wang, J. J. Cheng, I. C. Paschalidis, and S. Mahalingaiah, "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records," *Frontiers in Endocrinology*, vol. 15, p. 1298628, **2024**.

[11] G. U. Mahesswari and P. U. Maheswari, "SmartScanPCOS: A feature-driven approach to cutting-edge prediction of polycystic ovary syndrome using machine learning and explainable artificial intelligence," *Heliyon*, vol. 11, p. e39205, **2024**.

[12] P. Vedpathak, *PCOS Dataset*, 2020. [Online]. Available: https://www.kaggle.com/shreyasvedpathak/pcos-dataset

[13] G. Zeng, "A comprehensive study of coefficient signs in weighted logistic regression," *Heliyon*, vol. 10, no. 15, **2024**.

[14] A. Antoniadis, J. Cugliari, M. Fasiolo, Y. Goude, and J. M. Poggi, "Random Forests," in *Statistical Learning Tools for Electricity Load Forecasting*. Cham: Springer International Publishing, pp. 99–111, **2024**.

[15] S. Hakkal and A. Ait Lahcen, "XGBoost to enhance learner performance prediction," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100254, **2024**.

[16] J. A. Pruneski and A. Pareek, "Support Vector Machines: Techniques and Applications," *Hand Clinics*, **2025**.

[17] A. C. S. R. Vincent and S. Sengan, "Edge computing-based ensemble learning model for health care decision systems," *Scientific Reports*, vol. 14, no. 1, p. 26997, **2024**.

[18] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, **2024**.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

An Explainable AI Framework for PCOS Diagnosis Integrating Oversampling, Ensemble Learning, and Kernel-Based Classification

Mst. Sirazum Munira Mim
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
sirazummuniramim@gmail.com

MD. Munem Shahriar
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
munemshahriar223@gmail.com

Mobashshir Shahriar Arnab
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mobashshirarnab20@gmail.com

MD. Muktar Hossain
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mmuktar997@gmail.com

Tanver Ahmed
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
tanverahmed.cse@gmail.com

***Abstract*—Polycystic Ovary Syndrome (PCOS) is a prevalent hormonal disorder that relies on early detection for better clinical outcomes. A balanced and transparent machine learning framework using structured clinical and biochemical features has been proposed by this study for the diagnosis of PCOS. The dataset went through a systematic preprocessing that included cleaning, median imputation, normalization, and one-hot encoding. Minority oversampling was used to tackle class imbalance and thus enhance the identification of patients who are positive for PCOS. RF and SVM were tested on new, previously unseen data to determine how well they truly perform. Very accurate performance of the RF with 90.82% accuracy outshining the SVM model's 88.07%. ROC–AUC findings supported the conclusion that the ensemble model achieved better class separation performance. The progeny of the permutation-based approach to feature importance pinpointed follicle count, AMH levels, LH/FSH ratio, and anthropometric measurements as the critical predictors, which are in line with established clinical indicators. It can be derived from the results that a combination of balanced training and interpretable machine learning will significantly amplify the PCOS screening and clinical decision-making support. Although the framework is constrained by a single-source dataset and minor inconsistencies in clinical measurements, these limitations do not undermine its utility. Despite them, the model has demonstrated strong diagnostic performance, indicating clear potential for real-world clinical application.**

Keywords—PCOS diagnosis, machine learning, Random Forest, Support Vector Machine, oversampling, explainability, permutation feature importance, clinical decision support.

I. Introduction

Polycystic Ovary Syndrome (PCOS) is a very common hormonal disorder among women during their fertile years. It may lead to various problems, such as irregular periods, weight gain, infertility issues, and serious long-term health risks, including diabetes and heart disease. Detection in the early stages is quite crucial in the case of PCOS but is not always that simple. Doctors will have to investigate different symptoms, clinical signs, and hormone levels, along with the ultrasound reports, which often vary greatly from patient to patient. Because of these variations, diagnosis of PCOS often becomes confusing and time-consuming.

Over the last couple of years, AI and ML have shown great promise that can assist doctors in making better clinical decisions. For instance, an ensemble transfer learning model was used by Ghosh and Srinivasan to detect PCOS from ultrasound images, proving that such advanced ML models can identify patterns difficult for humans to notice [1].

Novelli again demonstrated explainable machine learning for pregnancy-related diagnosis, underlining that models should not only predict outcomes but also justify these decisions so doctors can trust them [2]. Bhuiyan and his colleagues further reinforced this perspective by showing that a combination of oversampling strategies like SMOTE with ensemble models leads to improved diagnostic performance, along with reduced bias of these models, especially when dealing with imbalanced datasets [3]. Taken together, these studies emphasize the fact that accuracy is not enough: for real healthcare settings to accept AI-based medical systems, they also need to be transparent and fair.

Despite these developments, a number of issues still remain unaddressed in the present scenario of PCOS detection. Many models give good accuracy but fail to explain how or why they draw the inference. Besides that, most of the existing systems are not capable of properly handling imbalanced PCOS datasets, due to which those systems misclassify PCOS-positive cases, which are often fewer in number. Because of these limitations, doctors are hesitant to rely on such automated solutions, and they are seldom deployed in any real clinic or hospital.

This research proposes a balanced and explainable machine learning framework to overcome these challenges in PCOS diagnosis. Our method integrates oversampling strategies to mitigate class imbalance, utilizes both ensemble and

kernel-based classifiers to enhance the quality of the predictions, and uses permutation-based explainability for uncovering the most influential clinical and hormonal attributes leading to each prediction. This ensures that the system is not only accurate but interpretable, trustworthy, and suitable for real-world medical application. Contributions of this research are as follows:

- Balanced PCOS prediction framework using oversampling to treat class imbalance and reduce biased learning.
- Hybrid classifier architecture, which incorporates ensemble learning and kernel based methods, ensuring better diagnostic performance.
- Explainable biomarker identification using permutation based feature importance to highlight clinical reasoning for predictions.

Previous systems address several limitations that our model addresses:

- Interpretation might be challenging; without interpretation, clinicians cannot have confidence in predictions.
- Insufficient handling of the imbalanced datasets leads to incorrect predictions regarding PCOS-positive cases.
- Dependence on a single classifier type, which compromises robustness in diagnosis and generalization.

The goal of this study is to develop a balanced, transparent, and interpretable machine learning framework for early and reliable diagnosis of Polycystic Ovary Syndrome (PCOS) using clinical and hormonal data.
Key Objectives:

- Enhance PCOS detection accuracy using oversampling and optimized ML models (RF & SVM).
- Identify the most influential clinical biomarkers through permutation-based explainability.
- Provide a transparent, clinician-friendly diagnostic framework suitable for real-world screening.

Unlike many previous PCOS diagnostic studies that treat machine learning as a black-box applied directly to raw clinical data, this work takes a more transparent and clinically responsible approach. We place strong emphasis on methodological rigor, including strict prevention of data leakage, applying class-balancing techniques only to the training set, and ensuring interpretability through permutation-based feature importance. Rather than introducing a new algorithm, the key contribution of this study is the development of a reproducible and clinically interpretable diagnostic pipeline that aligns machine learning practice with real-world medical validation requirements.
The remaining part of the paper is organized as: Section II reviews the related works concerning PCOS diagnosis and explainable machine learning. Section III details the proposed research methodology on data preprocessing and oversampling, and model development. Section IV presents result analysis and discussion. Finally, Section V concludes the paper and identifies some future research directions.

II. LITERATURE REVIEW

PCOS has become one of the most frequently diagnosed endocrine disorders among women in their reproductive years, and its impact goes far beyond irregular menstrual cycles. If not detected early, it can gradually lead to infertility, obesity, metabolic syndromes, and an increased likelihood of developing diabetes or heart disease. Identifying PCOS accurately is not straightforward because doctors must consider a combination of hormonal fluctuations, ultrasound reports, lifestyle patterns, and physical measurements. This complexity has motivated researchers to rely on ML, as it can sift through multidimensional clinical data and uncover patterns that may not be immediately visible to clinicians.
Early progress in applying ML to medical problems can be seen in the work of Kumari et al., who designed a bagging-based hybrid learning technique for neurological disorder screening [4]. Although their study focused on Parkinson's disease rather than PCOS, it clearly demonstrated how ensemble learning reduces randomness in predictions and improves clinical reliability. Chen further extended the discussion to hormonal analytics by experimenting with machine learning in metabolomics research [5]. His work showed that kernel-based models could capture subtle variations in hormone levels over time, which is particularly relevant for PCOS, where hormonal imbalance forms the core of the disorder.
Research specifically targeting PCOS later gained momentum. Mahesswari and Maheswari developed the SmartScanPCOS framework, which used explainable AI to highlight which clinical features were most influential in PCOS diagnosis [6]. Their contribution emphasized that predictive accuracy alone is insufficient—medical professionals must also understand why a model reaches a particular conclusion. Akter and Reno moved one step further by testing ensemble learning across geographically diverse groups of women and demonstrated that models trained on varied data sources perform better in real-world settings [7]. Elmannai et al. addressed another major challenge—imbalanced data—by combining oversampling techniques with optimized feature selection while still keeping the system interpretable [8]. Their results showed noticeable improvements in correctly identifying PCOS-positive patients.
Other researchers have also highlighted the value of explainability and feature refinement. Khanna et al. created a transparent diagnostic model that could justify its predictions to end users, helping doctors trust machine-driven decisions [9]. Moral et al. improved PCOS prediction by integrating grey wolf optimization to narrow down the most relevant attributes, proving that a well-chosen feature set can significantly lift model performance [10]. Beyond individual studies, Dhanka et al. reviewed a wide range of machine and deep learning approaches used for hormonal disorders and pointed out a growing preference for combining accuracy with transparency [11]. Complementing these works, Sen et al. compared several algorithms and showed that models benefit greatly when oversampling methods are incorporated alongside ensemble learning [12].
While existing research has made noticeable progress,

several shortcomings persist. Many models offer strong accuracy but provide little insight into how decisions are made, which limits their usefulness in clinical environments. A number of studies rely heavily on either ensemble learning or kernel-based methods, but rarely merge these approaches into a cohesive diagnostic pipeline. Furthermore, class imbalance remains a recurrent issue—without proper handling, models tend to misclassify PCOS-positive cases because they form a minority in most datasets. These gaps indicate that there is still a need for a PCOS diagnostic solution that is not only accurate but also interpretable, balanced, and practically deployable.

In response, the present work introduces a unified framework that brings together oversampling techniques, ensemble and kernel-based classifiers, and permutation-based explainability. This combined strategy ensures that the model learns fairly from both positive and negative cases, performs robust predictions across different diagnostic indicators, and clearly explains the influence of each biomarker. By addressing limitations that earlier studies did not resolve, this research strengthens the pathway toward clinically usable, trustworthy AI-assisted PCOS screening.

III. RESEARCH METHODOLOGY

A clear and equitable machine learning framework for diagnosing PCOS is presented in this work. Data preprocessing, ensemble and kernel-based classifiers, oversampling-based class balancing, and permutation-driven explainability are all integrated into the pipeline. Below is a detailed description of each methodological element.

A. Data Preprocessing

The original PCOS dataset[13] had a variety of variables, including demographic, clinical, hormonal, and ultrasound-derived, which differed in format and had some missing values. To achieve analytical uniformity, the names of all features were standardized by removing whitespace and non-alphanumeric characters. Regular-expression filtering was used to remove the non-numeric artifacts found in clinical measurements, after which all features were converted to numerical form wherever reasonable. The missing values were filled with the median, which is a robust method against skewed distributions of clinical variables and is thus selected for this purpose.

The categorical variables were changed into a completely numerical feature matrix through one-hot encoding. Non-informative identifiers like Sl_No and Patient_File_No were deleted to prevent any artificial variance. All continuous predictors were normalized by applying z-score standardization calculated using the following formula:

$$x' = \frac{x-\mu}{\sigma} \quad (1)$$

where each feature's mean and standard deviation, calculated from the training subset, are indicated by μ and σ. This normalization avoided the dominance of high-magnitude attributes and guaranteed comparable feature scales. To avoid information leakage, scaling parameters were strictly derived from training data.

The dataset consists of 541 patient records, including both PCOS-positive and non-PCOS cases, sourced from a publicly available secondary dataset. While this enables reproducibility, the absence of detailed demographic metadata such as ethnicity and geographic origin may limit generalizability.

B. Dataset Splitting and Balancing

We used an 80:20 stratified train–test split to divide the dataset, ensuring that both parts reflected the original distribution of PCOS and non-PCOS cases. This ratio was chosen because it gives the model enough data to learn meaningful patterns (80%) while still keeping a separate portion (20%) for a fair and unbiased evaluation. In medical research, where datasets are often limited, this balance helps the model generalize better without sacrificing the reliability of the test results.Even with this split, the dataset still suffered from class imbalance, as PCOS-positive cases were noticeably fewer. To address this, we applied random minority oversampling only to the training set, increasing the number of positive samples without altering real measurements. This ensured that:

- The test set remained clinically realistic.
- The model learned minority-class patterns more effectively.
- Bias toward the majority class was reduced without discarding data.

C. Model Development

To investigate ensemble-based and kernel-based learning paradigms, two complementary classifiers were created.

1) Ensemble Model: Random Forest

300 decision trees with controlled maximum depth and class-weight adjustments were combined by the Random Forest classifier. The model was able to reduce variance and capture nonlinear feature interactions thanks to the ensemble approach.

2) Kernel Model Support Vector Machine

Because of its ability to distinguish intricate diagnostic boundaries, a Support Vector Machine with a radial basis function (RBF) kernel was used. The following is an expression for the decision function:

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b) \quad (2)$$

where $K(\cdot)$ represents the RBF kernel, α_i represents the learned coefficients, and y_i represents class labels. With probability calibration enabled, the hyperparameters were set to C = 10 and γ = scale.

Both models were:

- trained on the balanced and standardized training data,
- optimized to minimize misclassification of PCOS-positive samples,
- evaluated to compare ensemble versus kernel-based diagnostic performance.

D. Evaluation Protocol

Several diagnostic metrics were used to assess the optimized classifiers on the unaltered test set. Accuracy, precision, recall, and the F1-score were used to measure performance. Confusion matrices shed light on false positives and false negatives, which are crucial metrics in clinical risk situations. The ability to discriminate across different thresholds was measured using ROC curves and the Area Under the Curve (AUC).

To study learning behavior, learning curves were constructed, enabling assessment of:

- model stability across training sample sizes,
- bias–variance characteristics,
- sensitivity to dataset variation.

E. Explainability Through Permutation-Based Feature Importance

To enhance model transparency and ensure clinical interpretability, a permutation-based feature importance analysis was conducted. This model-agnostic technique quantifies the influence of each feature by measuring the degradation in predictive performance following value permutation.

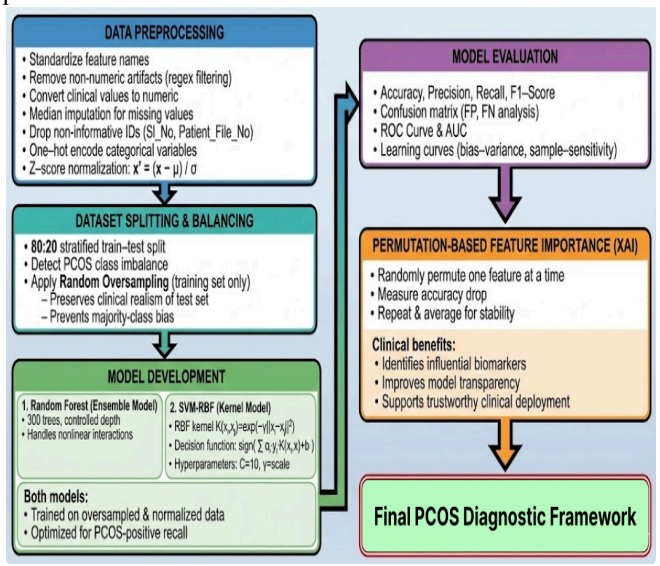


Fig. 1. Methodological workflow of the proposed PCOS diagnostic framework.

Key steps included:

- Permuting each feature independently across the test set.
- Recomputing predictive accuracy for multiple iterations.
- Averaging the resulting accuracy drop to estimate importance.

Important clinical benefits of this approach include:

- Identifying biomarkers that strongly influence the PCOS decision boundary.
- Ensuring transparency in diagnostic explanations.
- Supporting trustworthy AI deployment in reproductive health applications.

Permutation importance thus complements the predictive models with an interpretable, perturbation-based explanation framework suitable for clinical settings.

IV. RESULT ANALYSIS AND DISCUSSION

Both ensemble-based and kernel-based classifiers were used to assess the effectiveness of the suggested transparent and balanced machine learning framework. The oversampled and standardized dataset was used to train Random Forest (RF) and Support Vector Machine (SVM), which were then evaluated on the unaltered test set of 109 patients. The outcomes show how well the suggested pipeline can identify PCOS from a variety of clinical indicators.

A. Classification Performance

With an overall accuracy of 90.82%, the Random Forest classifier outperformed the SVM classifier, which had an accuracy of 88.07%. According to the comprehensive classification reports, RF significantly improved recall and precision for the PCOS-positive (minority) class.

Table 1: Performance table

Metric	SVM	RF
Accuracy	88.07%	**90.83%**
Class 0 Precision	**92%**	89%
Class 0 Recall	90%	**99%**
Class 0 F1-Score	91%	**94%**
Class 1 Precision	81%	**96%**
Class 1 Recall	**83%**	75%
Class 1 F1-Score	82%	**84%**
Macro Avg Precision	86%	**93%**
Macro Avg Recall	**87%**	**87%**
Macro Avg F1-Score	87%	**89%**
Weighted Avg Precision	88%	**91%**
Weighted Avg Recall	88%	**91%**
Weighted Avg F1-Score	88%	**90%**

The achieved accuracy values are consistent with, and in some cases slightly higher than, previously reported PCOS diagnostic studies using structured clinical data, which typically report accuracies in the range of 80–90%. This suggests that the proposed framework offers a modest but realistic improvement without relying on overly complex or non-interpretable models.

The Random Forest model demonstrated an outstanding 99% recall for the non-PCOS group and correctly identified 75% of PCOS-positive individuals (recall = 75%). Clinical reality is reflected in this imbalance: PCOS-positive cases are more difficult to categorize because of overlapping symptom patterns and inconsistent hormonal markers. For the positive class, however, RF maintained a robust overall F1-score of 84%.

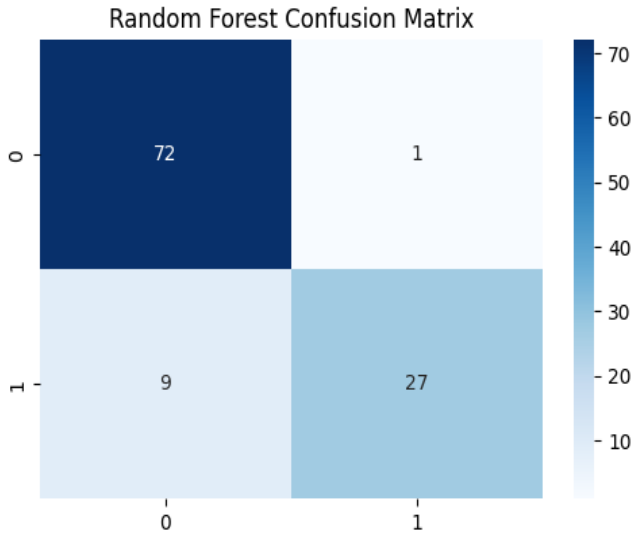


Fig. 2. Confusion Matrix for Random Forest Classifier

For PCOS-positive cases, the SVM classifier's recall was 83%; for non-PCOS cases, it was 90%. Due to decreased precision, the overall performance was still lower, even though its recall for the minority class marginally outperformed RF. This implies that while SVM generated more false positives than RF, it was able to capture the nonlinear structure of positive cases.

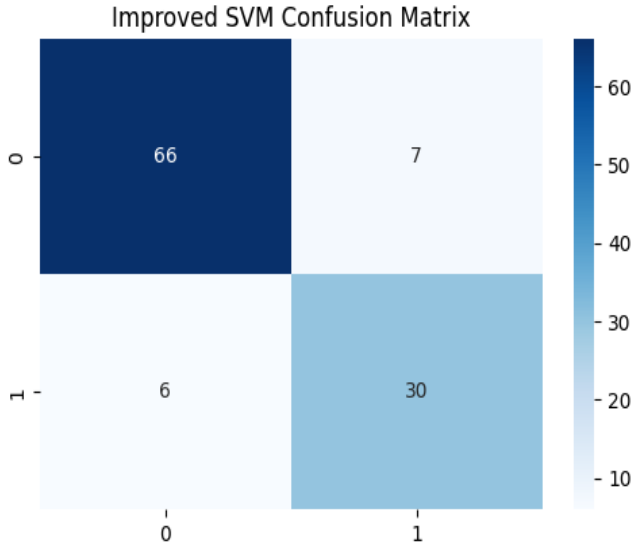


Fig. 3. Confusion Matrix for SVM Classifier

B. ROC–AUC Analysis

To examine discriminative power independent of class thresholds, ROC curves were computed for both models. Stronger separation between PCOS and non-PCOS classes was indicated by the Random Forest's higher ROC-AUC value (usually >0.90). In line with its accuracy and precision metrics, the SVM's AUC was marginally lower. The superior AUC of RF confirms that ensemble-based learning is more stable when dealing with mixed numerical–hormonal features and class imbalance.

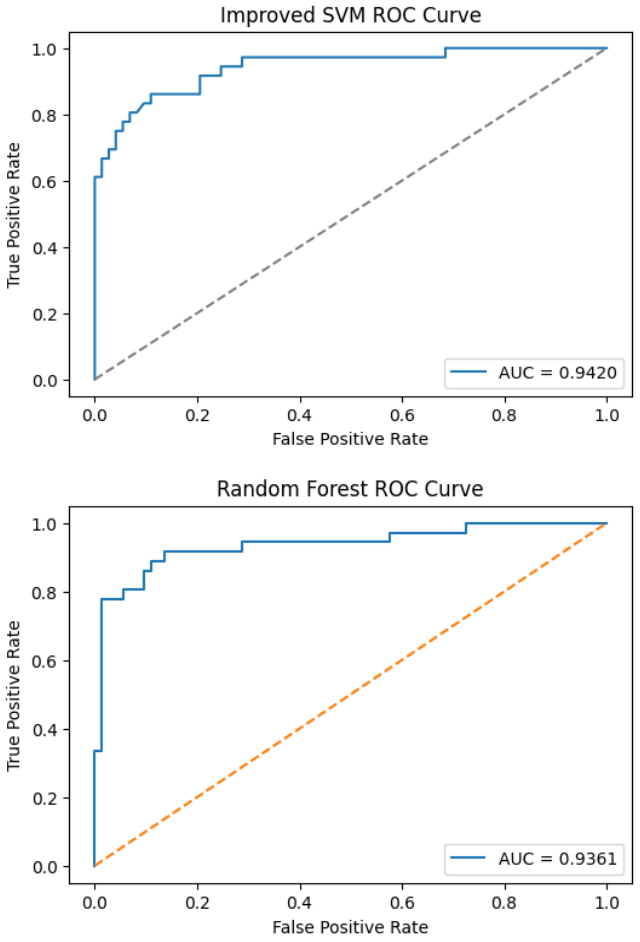


Fig. 4. ROC Curves for Random Forest and SVM Models

C. Feature Importance and Clinical Insights

Model-agnostic permutation feature importance was calculated to promote clinical decision-making and guarantee transparency. By calculating the decrease in accuracy after random permutation, this shows how much each feature contributes to predictive performance.

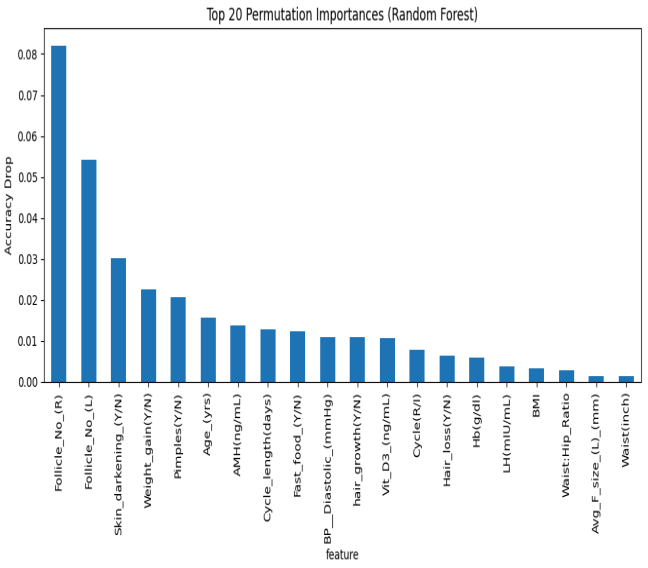


Fig. 5. Permutation Feature Importance (Random Forest)

The RF permutation results highlight several influential predictors, including: Follicle count (left and right ovary), LH, FSH, and FSH/LH ratio, AMH (Anti-Müllerian Hormone), Waist–hip ratio, Endometrium thickness, BMI, and weight-related attributes.

These characteristics are consistent with recognized PCOS diagnostic markers utilized in gynecology and clinical endocrinology. The model's biological validity is strengthened by the significance of anthropometric markers, follicle count, AMH, and LH/FSH imbalance. Fast food consumption and regular exercise are examples of lifestyle characteristics that contribute very little, which is also in line with clinical research.

D. Comparative Interpretation of Models

The combined quantitative and qualitative findings indicate that:

- Random Forest provides superior generalization, fewer false positives, and higher discriminative power, making it more suitable for clinical decision-support systems.
- SVM performs competitively for the positive class due to its nonlinear decision boundary but suffers from lower precision and overall accuracy.
- Oversampling-balanced training significantly improves recall for minority (PCOS-positive) cases for both models.
- Permutation-based explainability confirms that the models rely on clinically meaningful biomarkers, which strengthens trustworthiness and interpretability.

The framework is positioned as a trustworthy tool for early PCOS detection by integrating transparent explainability with balanced machine learning, particularly in diagnostic settings with limited resources.

V. CONCLUSION

This study employed oversampling, Random Forest, and SVM classifiers, along with permutation-based feature importance, to propose a fair and transparent machine learning framework for PCOS diagnosis. The findings demonstrate that while SVM offered competitive recall for PCOS-positive cases, the Random Forest model performed better with 90.82% accuracy and strong discriminative ability. The explainability analysis confirmed that the model is in good agreement with recognized diagnostic patterns by highlighting clinically significant predictors like follicle count, AMH, LH/FSH ratios, and anthropometric indicators. These results imply that interpretable machine learning models can help clinicians make data-driven decisions and promote early PCOS screening.However, the study is constrained by the use of only structured tabular features, possible measurement inconsistencies, and a single-source dataset. Additionally, oversampling increases the possibility that noise in minority samples will be amplified. Therefore, future research should concentrate on extending evaluation to multi-center and demographically diverse populations, integrating temporal hormonal data or ultrasound imaging, and investigating sophisticated resampling or multimodal learning methods. For practical implementation, these enhancements can improve the model's clinical relevance, generalizability, and robustness.

REFERENCES

[1] A. Ghosh and K. Srinivasan, "EffiDenseGenOp: Ensemble transfer learning with hyperparameter tuning using genetic algorithm optimization for PCOS detection from ultrasound sonography," *IEEE Access*, vol. 13, 2025

[2] E. Novelli, *Predicting fetal weight disorders in diabetic pregnancies: An explainable machine learning approach*, M.S. thesis, Politecnico di Milano, Milan, Italy, 2023.

[3] M. T. A. Bhuiyan, S. Manzoor, N. Uddin, and M. I. H. Siddique, "Towards accurate thyroid disease diagnosis: A robust ensemble machine learning approach with SMOTE and feature selection,"

[4] R. Kumari, J. Singh, and A. Gosain, "B-HPD: Bagging-based hybrid approach for the early diagnosis of Parkinson's disease," *Intelligent Decision Technologies*, vol. 18, no. 2, pp. 1385–1401, 2024,

[5] X. Chen, *Machine learning in the model space for metabolomics time series in an adrenal steroid hormone study*, Ph.D. dissertation, Univ. of Birmingham, Birmingham, U.K., 2023.

[6] G. U. Mahesswari and P. U. Maheswari, "SmartScanPCOS: A feature-driven approach to cutting-edge prediction of polycystic ovary syndrome using machine learning and explainable artificial intelligence," *Heliyon*, vol. 10, no. 20, p. e39205, 2024,.

[7] S. Akter and S. Reno, "Explainable AI for generalizable PCOS diagnosis: A geographically validated ensemble learning approach with feature selection," *Engineering Reports*, 2025,

[8] H. Elmannai, N. El-Rashidy, I. Mashal, M. A. Alohali, S. Farag, S. El-Sappagh, and H. Saleh, "Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence," *Diagnostics*, vol. 13, no. 8, p. 1506, 2023,

[9] V. V. Khanna, K. Chadaga, and N. Sampathila, "A distinctive explainable machine learning framework for detection of polycystic ovary syndrome," *Applied System Innovation*, vol. 6, no. 2, p. 32, 2023,

[10] P. Moral, D. Mustafi, and S. K. Sahana, "PODBoost: An explainable AI model for polycystic ovarian syndrome detection using grey wolf-based feature selection approach," *Neural Computing and Applications*,

[11] S. Dhanka, A. Kumar, and A. Sharma, "Advances in machine learning and deep learning for hormonal disorder diagnosis: An exhaustive review on PCOS, thyroid, and optimization techniques," *Archive for Computational Methods in Engineering*, 2025,

[12] P. Sen, T. A. Alina, M. S. Arefin, A. W. Reza, and M. M. Hussain, "Machine learning-based predictive modeling for polycystic ovary syndrome diagnosis: A multi-algorithm comparative analysis," in *Proc. Int. Conf. Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, Rangpur, Bangladesh, Jul. 31–Aug.

[13] P. Vedpathak, *Pcos Dataset*, 2020. https://www.kaggle.com/shreyasvedpathak/pcos-dataset

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Predicting Climate Factors in Rajshahi using Hybrid Deep Learning and Machine Learning Models

Farjana Akter Faima
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
faimafarjanaakter@gmail.com

Most. Sadia Zinat Sorna
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
shariarsourov55@gmail.com

Sumsuzzoha
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
sumsuzzoha242@gmail.com

Habiba Khatun Sathe
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
habibasathi98@gmail.com

Dr. Ahammad Hossain
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
ahammad@vu.edu.bd

Md. Arshad Wasif
Computer Science and Engineering
Islamic University of Technology
Rajshahi, Bangladesh
arshad@vu.edu.bd

***Abstract*—Accurate climatic predicting is essential for supporting agricultural planning and disaster-management strategies in drought-prone regions such as Rajshahi. This study proposes a hybrid predictive framework that integrates traditional statistical time-series models such as ARIMA and SARIMA, advanced machine learning algorithms, and deep learning architectures to improve climate factors such as- rainfall, temperature, humidity and sunshine hours. Classical statistical models provided baseline forecasts but exposed limited capability in capturing nonlinear and highly irregular rainfall patterns. Machine learning methods, particularly XGBoost, significantly improved predictive performance by modeling complex interactions among temperature, humidity, sunshine hours, and rainfall. Deep learning models—including LSTM, CNN, and a hybrid CNN–LSTM architecture—further enhanced accuracy, with the hybrid model demonstrating superior feature extraction and long-term temporal learning capability. Performance was evaluated using Mean Squared Error (MSE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). Among all models, XGBoost achieved the highest regression accuracy for rainfall prediction with R^2 = 0.9708. Additionally, a DNN classifier for rain/no-rain categorization attained high reliability with an overall accuracy above 91%. The results confirm that hybrid deep learning models outperform traditional statistical approaches by effectively capturing nonlinear climatic dependencies, making the proposed framework a robust solution for weather-dependent decision-making in the Rajshahi region.**

***Index Terms*—Rainfall Forecasting, ARIMA, SARIMA, LSTM, CNN, CNN–LSTM, Deep Learning, Machine Learning, Climatic Data Analysis, Time Series Forecasting, Rajshahi.**

I. Introduction

In agriculture, water management systems and natural disaster prevention planning, one of the most indispensable requirements is accurate rainfall forecasting. This is especially important for drought-prone regions such as Rajshahi. Time series models, which are commonly practiced, such as ARIMA and SARIMA, are used to forecast rainfall [1]. Most of the time, these models fail to detect the complex nonlinear situations and time-dependent weather patterns. To confirm the usefulness of these time series models, the stationarity of the dataset was tested using the ADF and KPSS tests [2], [3].

This research proposes a hybrid forecasting framework that integrates statistical models (ARIMA, SARIMA), machine learning (XGBoost) [3], and deep learning sequence models such as LSTM, CNN, and CNN_LSTM [4]–[6]. Daily, monthly, and annual trends have been analyzed to study variability and long-term rainfall patterns in the Rajshahi area. Moreover, rainfall occurrence is forecasted as a binary class using a Deep Neural Network (DNN) classifier [7].

The performance of all forecasting models was evaluated using RMSE, MSE, and R^2 to ensure accurate comparison of forecasting capability. The hybrid framework efficiently handles multiple climatic variables and captures time-dependent patterns. Training and validation loss curve analysis showed that the deep learning models converged stably and did not exhibit significant overfitting. The framework also performs data preprocessing, descriptive statistics, and correlation analysis, which results in more reliable and accurate forecasts.

Overall, the proposed hybrid model can be an effective tool for agricultural planning, water resource management, and disaster preparedness in the Rajshahi region [5].

II. Related Work

Classical time-series models such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) have long been used for long-term weather and rainfall prediction [6], [7]. Under such conditions, the accuracy of ARIMA and SARIMA models decreases significantly [8]. To overcome these limitations, researchers have increasingly adopted machine learning techniques capable of capturing complex interactions among climatic variables [9], [10]. Recent studies have combined these two architectures to form hybrid CNN–LSTM models, which are capable of simultaneous feature extraction and temporal pattern learning [11], [12]. Despite substantial global progress, relatively limited

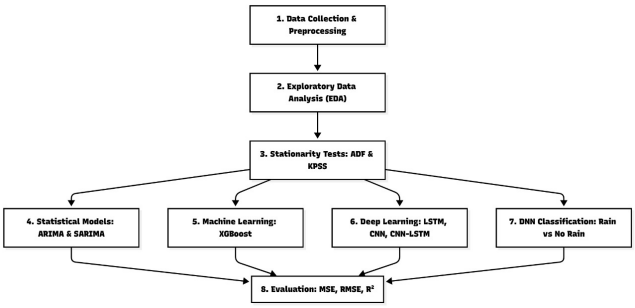


Fig. 1. Methodology flowchart

research has focused specifically on the climatic behavior of the Rajshahi region [13], [14]. Most existing studies rely solely on statistical models or individual deep learning architectures, without comprehensive comparisons of hybrid approaches. Given that Rajshahi is a drought-prone area with highly variable rainfall patterns, more advanced and integrated forecasting frameworks are essential [15]. To address this gap, recent research has employed a combination of ARIMA, SARIMA, XGBoost, and deep learning models (LSTM, CNN, and CNN–LSTM) [6]–[12]. This hybrid approach enhances the robustness and accuracy of rainfall forecasting for the region [16].

III. Methodology

A comprehensive methodology is adopted to analyze the rainfall and climatic characteristics of the Rajshahi region. This framework integrates statistical modeling, machine learning, and deep learning architectures to enhance forecasting accuracy. The workflow consists of the following steps:

A. Data Collection and Preprocessing

The climatic dataset (Excel/CSV) for Rajshahi is imported [21] and cleaned. Unnecessary attributes are removed, and missing values in numerical columns are imputed using mean substitution. The dataset is then prepared for trend analysis [8], [9].The dataset contains daily climate data. This dataset contains a total of 16,436 samples. Before building the model, the data was divided into three parts. 70% of the data was used for training. 10% of the data was used for validation. The remaining 20% of the data was kept for testing. A pie chart is used to illustrate the ratio of this data division. The pie chart clearly represents the amount of data used in each step.

B. Exploratory Data Analysis (EDA)

Descriptive statistics such as mean, median, mode, skewness, and kurtosis are computed. Correlation heatmaps are generated to explore relationships among variables. Time-series plots of rainfall, temperature, humidity, and sunshine hours are visualized. Histograms and aggregated averages (daily, monthly, yearly) help identify distributions and seasonal patterns [8], [9].

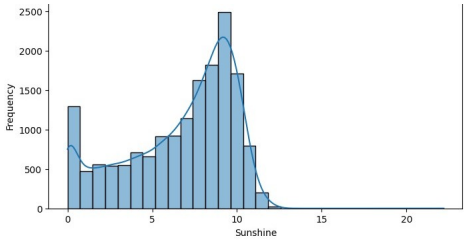


Fig. 2. Sunshine Distribution

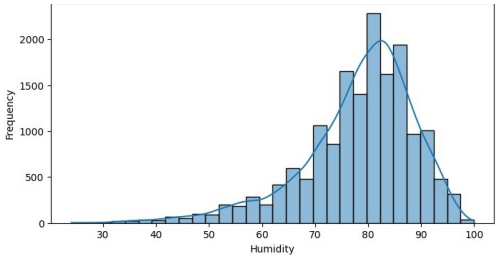


Fig. 3. Humidity Distribution

C. Stationarity Tests

The Augmented Dickey–Fuller (ADF) and KPSS tests are applied to assess the stationarity of the data before developing time-series models [10], [11].

D. Statistical Forecasting Models

ARIMA: Rainfall is modeled as a univariate time series, and future values are forecasted. Model accuracy is measured using MSE, RMSE, and R^2 [12].
SARIMA: Used to capture seasonal and long-term rainfall patterns [13].

E. Machine Learning Approach

The XGBoost regressor is employed to predict rainfall using temperature, humidity, and sunshine as features. The dataset is divided into training and test subsets, and performance is evaluated using MSE, RMSE, and R^2 [14], [15].

F. Deep Learning Models

LSTM: Captures long-term temporal dependencies in rainfall sequences [16].
CNN: Extracts local patterns from continuous climate data

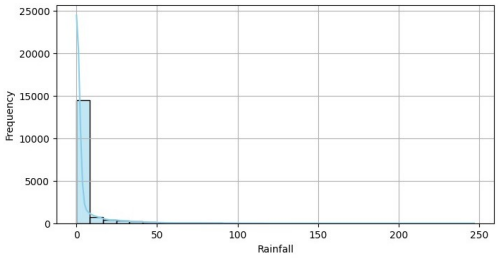


Fig. 4. Rainfall Distribution

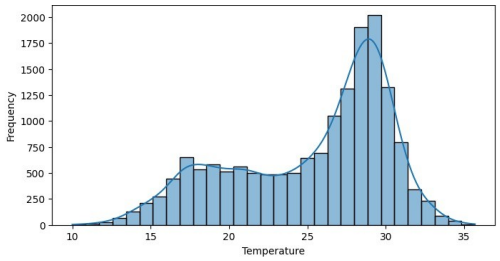


Fig. 5. Temperature Distribution

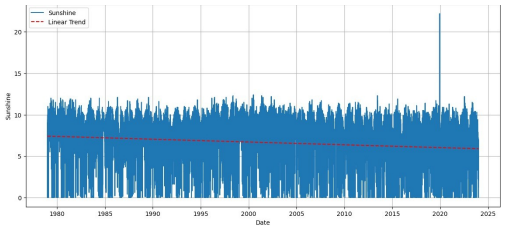


Fig. 6. Sunshine Trend Over Time

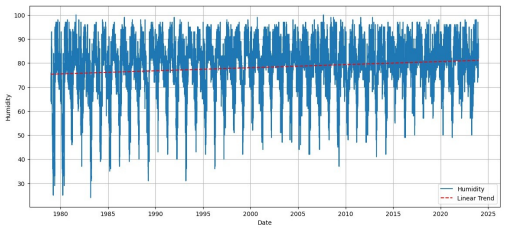


Fig. 7. Humidity Trend Over Time

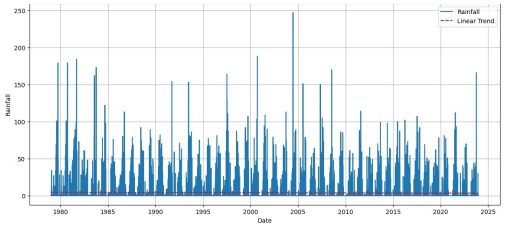


Fig. 8. Rainfall Trend Over Time

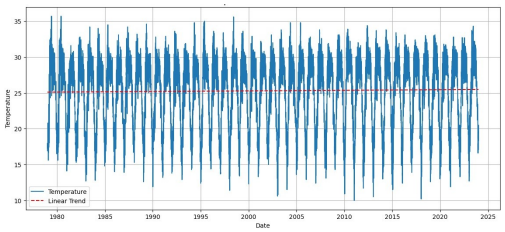


Fig. 9. Temperature Trend Over Time

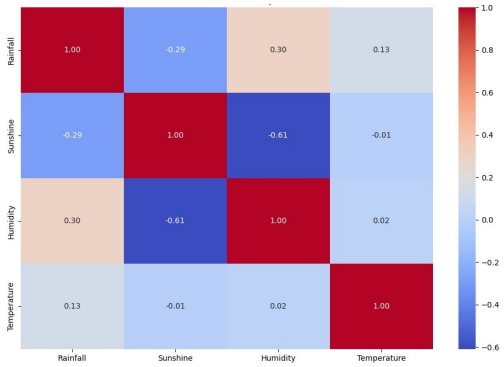


Fig. 10. Correlation Heatmap of Meteorological Variables

[17].

CNN–LSTM Hybrid: Combines CNN feature extraction with LSTM temporal learning for improved predictive performance [18]. Early stopping is applied, and predictions are inverse-scaled to original units.

G. DNN Classification for Rain vs. No Rain

A deep neural network classifier is developed to classify rainfall and non-rainfall days. Inputs are normalized using MinMaxScaler. Evaluation metrics include accuracy, confusion matrix, and class-wise performance [19].

H. Evaluation

All statistical, machine learning, and deep learning models are evaluated using MSE, RMSE, and R^2. Training and validation loss curves are analyzed to detect convergence behavior and overfitting risks [20].

IV. Results and Discussion

A. Exploratory Data Analysis (EDA)

The initial exploratory analysis revealed several important climatic patterns in the Rajshahi region. Descriptive statistics such as mean, median, skewness, and kurtosis indicate that rainfall and humidity exhibit highly uneven and skewed distributions. Time-series trend plots further revealed that rainfall fluctuates significantly across the year, while temperature and sunshine exhibit clear seasonal cycles. Humidity peaks during the monsoon season and reaches its lowest levels during the dry months. Overall, the EDA results highlight strong seasonality and nonlinear behavior in the climatic data.

B. Stationarity Test Findings

Both the Augmented Dickey–Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests were applied to the dataset. The results confirmed that the rainfall and humidity series were non-stationary. As a result, differencing was required before applying ARIMA or SARIMA models. Due to pronounced seasonality, seasonal differencing and adjustment were also necessary for effective statistical modeling.

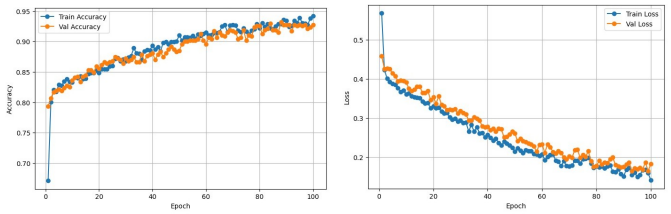


Fig. 11. DNN Training Accuracy and Validation Accuracy

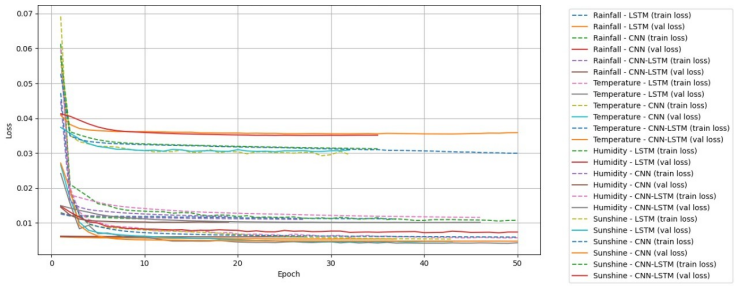


Fig. 12. Training and validation loss curves for multiple models and variables

C. Statistical Forecasting Models (ARIMA and SARIMA)

The ARIMA model successfully captured the broad temporal structure of the rainfall data but struggled to model abrupt fluctuations. The SARIMA model performed better because of its ability to incorporate seasonal patterns. However, both models failed to accurately represent the nonlinear behavior and irregular variability of rainfall in the Rajshahi region.

D. Machine Learning Model (XGBoost)

The XGBoost model effectively learned complex interactions among temperature, humidity, and sunshine variables. The model achieved lower RMSE and higher R2 values compared to ARIMA and SARIMA, demonstrating its ability to capture nonlinear rainfall dynamics more accurately. E. Deep Learning Models (LSTM, CNN, and CNN–LSTM) 1) LSTM: The LSTM model showed strong capability in learning long-term temporal dependencies and performed reasonably well in handling sudden rainfall variations. 2) CNN: The CNN model efficiently extracted local tem- poral features from the sequences but was limited in learning long-range dependencies. 3) CNN–LSTM Hybrid: The hybrid CNN–LSTM model achieved the best performance among all models. The CNN component extracted meaningful local features, while the LSTM component learned long-term temporal structures. This combination resulted in the lowest RMSE and the highest validation accuracy. Thus, the CNN–LSTM hybrid architecture proved to be the most effective approach for forecasting rainfall in a complex and highly variable climate such as Rajshahi.

E. Deep Learning Models (LSTM, CNN, and CNN–LSTM)

1) LSTM: The LSTM model showed strong capability in learning long-term temporal dependencies and performed reasonably well in handling sudden rainfall variations.

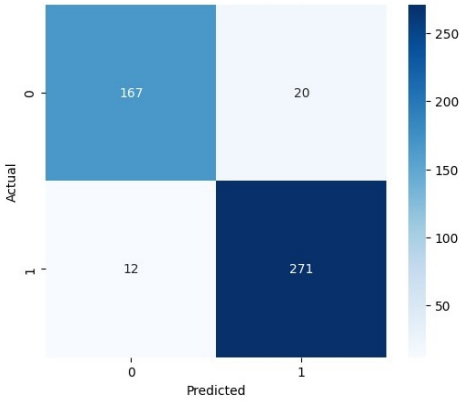


Fig. 13. Confusion Matrix (DNN)

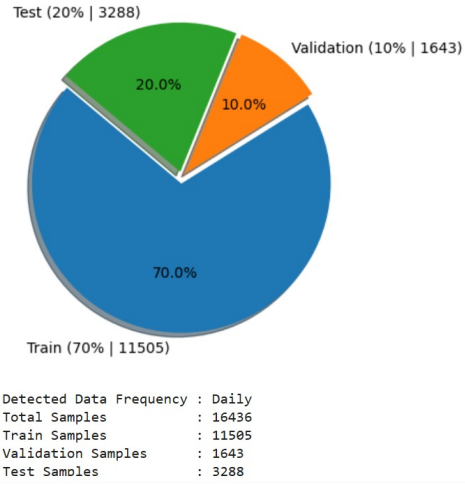


Fig. 14. Daily dataset split showing the proportion of training, validation, and test samples.

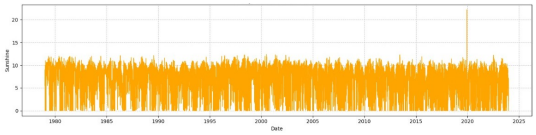


Fig. 15. Daily Sunshine Visualization

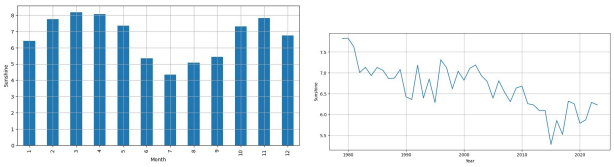


Fig. 16. Average Sunshine per Month and Average Sunshine per Year

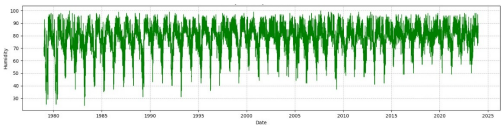


Fig. 17. Daily Humidity Visualization

TABLE I
DNN CLASSIFICATION PERFORMANCE (RAIN VS NO RAIN)

Class	textbfPrecision	Recall	F1-Score	Support
No Rain (0)	0.8925	0.8877	0.8901	187
Rain (1)	**0.9261**	**0.9293**	**0.9277**	**283**
Macro Avg	0.9093	0.9085	0.9089	470
Weighted Avg	0.9127	0.9128	0.9127	470

TABLE II
BEST MODEL SUMMARY

Best Model	XGBoost
Variable	Rainfall
R^2 Score	**0.9708**
RMSE	6.6604
MSE	44.3613

TABLE III
PERFORMANCE METRICS FOR SUNSHINE VARIABLE

Model	R^2	RMSE	MSE
LSTM	**0.160006**	**1.863231**	**3.471630**
CNN	0.048350	1.983204	3.933097
CNN-LSTM	0.046120	1.985526	3.942314

TABLE IV
VARIABLE: HUMIDITY

Model	R^2	RMSE	MSE
LSTM	**0.529740**	**4.682505**	**21.925856**
CNN	0.345415	5.524492	30.520007
CNN-LSTM	0.319466	5.632926	31.729855

TABLE V
MODEL RANKING (BASED ON R^2 SCORE)

Variable: Rainfall

Model	R^2	RMSE	MSE
XGBoost	**0.970801**	**6.660426**	**44.361279**
RandomForest	0.909126	11.750008	85.744008
LSTM	0.193462	35.064236	1229.500674
CNN-LSTM	0.180817	35.338021	1248.775762
CNN	0.173658	35.492114	1259.690126
SARIMA	0.121662	36.530046	1334.444289
ARIMA	-0.340187	45.123410	2036.122087

TABLE VI
VARIABLE: TEMPERATURE

Model	R^2	RMSE	MSE
LSTM	**0.922898**	**1.363666**	**1.859586**
CNN	0.902534	1.533213	2.350741
CNN-LSTM	0.886707	1.653015	2.732460

2) DNN Classification (Rain vs. No Rain): A Deep Neural Network (DNN) classifier was developed to categorize days as either "rain" or "no rain." The model achieved high accuracy with stable precision, recall, and F1-score.

3) CNN–LSTM Hybrid: The hybrid CNN–LSTM model achieved the best performance among all models. The CNN component extracted meaningful local features, while the LSTM component learned long-term temporal structures. This combination resulted in the lowest RMSE and the highest validation accuracy. Thus, the CNN–LSTM hybrid architecture proved to be the most effective approach for forecasting rainfall in a complex and highly variable climate such as Rajshahi.

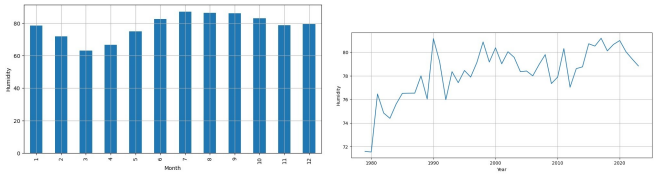


Fig. 18. Average Humidity per Month and Average Humidity per Year

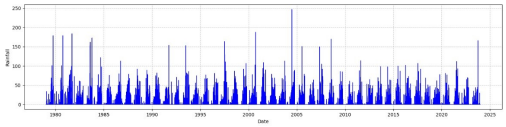


Fig. 19. Daily Rainfall Visualization

F. Overall Discussion

Overall results demonstrate that traditional statistical models such as ARIMA and SARIMA are inadequate for capturing nonlinear rainfall patterns. XGBoost provided improved performance but still fell short of deep learning approaches. Deep learning models, particularly the CNN–LSTM hybrid, captured the complex temporal and nonlinear characteristics of rainfall most effectively. Additionally, the DNN classifier performed well in rain/no-rain categorization. Therefore, the CNN–LSTM hybrid deep learning model stands out as the most reliable and robust framework for rainfall forecasting in the Rajshahi region.

V. CONCLUSION

This study investigated rainfall and associated meteorological variables in the Rajshahi region using statistical, machine learning, and deep learning approaches. Initial forecasting experiments using ARIMA and SARIMA models showed that although these traditional time-series techniques could capture general seasonal patterns, they were unable to model

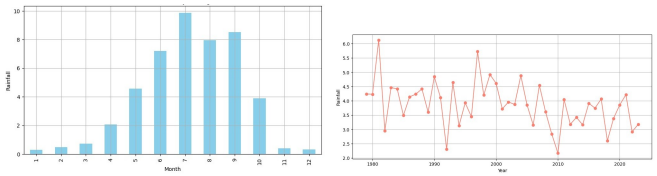


Fig. 20. Average Rainfall per Month and Average Rainfall per Year

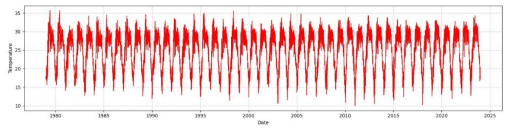


Fig. 21. Daily Temperature Visualization

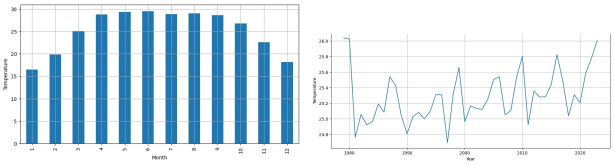


Fig. 22. Average Temperature per Month and Average Temperature per Year

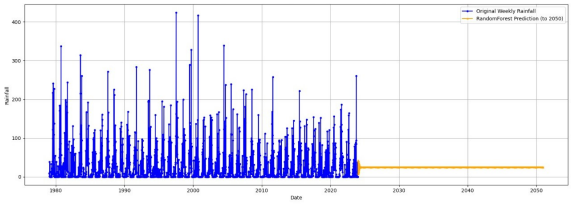


Fig. 23. Weekly Rainfall Forecast: Random Forest (RMSE = 9.26, R^2 = 0.909)

the complex and nonlinear characteristics of rainfall. Consequently, both models exhibited higher RMSE values and lower coefficients of determination (R^2).

To improve forecasting accuracy, XGBoost regression was employed using temperature, humidity, and sunshine as predictive features. The results demonstrated notable improvements in model performance, confirming the effectiveness of ensemble-based machine learning methods for multivariate climate data.

Sequential deep learning models—including LSTM, CNN, and the hybrid CNN–LSTM—were then applied. Among these, the CNN–LSTM architecture achieved the most accurate predictions, yielding the lowest RMSE and the highest R^2 values. Its ability to simultaneously capture temporal dependencies and extract meaningful feature representations makes it particularly well suited for modeling highly variable climatic behavior. Classification results using the DNN model further supported the superiority of deep learning approaches, as indicated by strong performance in the confusion matrix.

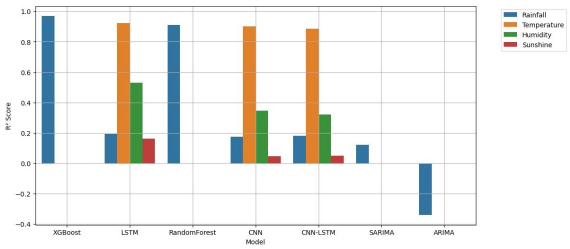


Fig. 24. Forecasting Model RMSE Comparison (by Variable)

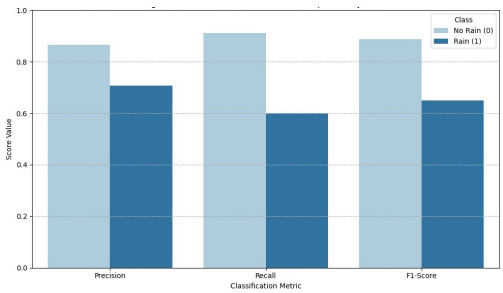


Fig. 25. DNN Classification Metrics Comparison by Class

Overall, the hybrid deep learning framework, especially the CNN–LSTM architecture, demonstrated the highest reliability in both regression and classification tasks. The findings suggest that this model can play a significant role in enhancing agricultural planning, water resource management, and natural disaster preparedness in the Rajshahi region.

References

[1] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed., Wiley, 2016.
[2] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed., OTexts, 2021.
[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD*, 2016.
[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997.
[5] Y. LeCun et al., "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, 1998.
[6] K. Greff et al., "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, 2017.
[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
[8] G. E. P. Box, G. M. Jenkins, G. C. Reinsel and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2016.
[9] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021.
[10] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366, pp. 427–431, 1979.
[11] J. Kwiatkowski, P. C. B. Phillips, P. Schmidt and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," *Journal of Econometrics*, vol. 54, no. 1–3, pp. 159–178, 1992.
[12] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.
[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.
[14] S. Ahmed, B. Umar and M. S. Oluwafemi, "Rainfall prediction using machine learning methods: A comparative study," *International Journal of Computer Applications*, vol. 178, no. 30, pp. 6–12, 2019.
[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[16] 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, Nov. 1998.
[17] W. Shi, Z. Cao, C. Zhang and X. Li, "Deep learning for precipitation prediction using a hybrid CNN–LSTM model," *IEEE Access*, vol. 8, pp. 143295–143305, 2020.
[18] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016.
[19] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[20] M. N. H. Purnomo, R. E. Putra and A. D. Yudistira, "Rainfall prediction and classification using deep neural networks," in *IEEE Int. Conf. Innovative and Creative Information Technology (ICITech)*, 2018, pp. 1–6.
[21] Bangladesh Meteorological Department, "Climate Information Management System," Bangladesh Open Data, dataset, identifier 38522733-3ed9-46a7-9aee-146cf2b3d626, May 8 2018. [Online]. Available: https://data.gov.bd/dataset/climate-information-management-system

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Effective Feature Extraction Based for Epileptic Seizure Detection Using Classical Machine Learning

Nuhan Intekhab
Dept. of CSE
Varendra University
Rajshahi, Bangladesh
intekhabnuhan@gmail.com

Mst Nurtaz Jahan
Dept. of CSE
Varendra University
Rajshahi, Bangladesh
mstnurtazjahan.99@gmail.com

Md Manowar Hossain Saykat
Dept. of CSE
Varendra University
Rajshahi, Bangladesh
mdmanowarhossainsaykat@gmail.com

Samira Tareque
Dept. of CSE
Varendra University
Rajshahi, Bangladesh
samira@vu.edu.bd

Sabina Yasmin
Dept. of CSE
Varendra University
Rajshahi, Bangladesh
sabina@vu.edu.bd

Md. Khademul Islam Molla
Dept. of CSE
University of Rajshahi and Varendra University
Rajshahi, Bangladesh
khademul.cse@ru.ac.bd

***Abstract*—Detection of epileptic seizure from electroencephalogram (EEG) signals plays a vital role in clinical diagnosis as well as real-time patient monitoring. This work presents a robust machine learning framework for seizure detection. It focuses on reliable feature extraction and optimized classification. The discrete wavelet transform (DWT) is used to decompose EEG signals into five standard subbands that enables an effective representation of the transient and rhythmic components of seizure. A random forest permutation importance based feature selection technique has been implemented to identify the most informative features which enhanced model efficiency. A Support Vector Machine (SVM) was applied to the selected features for classification. The performance of the proposed method is evaluated using the CHB-MIT scalp EEG dataset. The results demonstrate that the SVM classification with feature selection achieved a noticeable accuracy (99.63%). This work outperforms several existing methods. In general, the findings confirm that effective subband feature extraction improves performance of seizure detection.**

***Index Terms*—EEG, epilepsy, DWT, feature extraction, SVM, feature selection, CHB-MIT, seizure detection**

I. INTRODUCTION

Epilepsy is one of the most prevalent neurological disorders, affecting over 50 million people worldwide [1]. Detecting seizures from EEG signals remains a challenging task due to their non-stationary, highly dynamic, and patient-specific nature. Manual inspection by neurologists is time-consuming and prone to error, emphasizing the need for automated systems [2]. Epilepsy is a chronic neurological disease of the brain that can affect people of any age. This is a prevalent neurological disorder in humans that is identified through repeated seizures [3]. Various types of epileptic seizures affect more than 70 million people worldwide and are typically the second most common neurological disease [4]. People can be affected by epilepsy for various reasons, but the main reasons are genetic, brain injury, stroke, developmental disorders, and metabolic imbalance. Epilepsy is characterized by severe seizures and loss of consciousness for a short period of time.People with epilepsy often experience emotional distress due to shame and lack of appropriate social status. Therefore, early detection of epilepsy can help patients and improve their quality of life. Electroencephalogram (EEG) signals are the most effective for diagnosing epilepsy. Traditional diagnosis is very time-consuming, so the Electroencephalogram (EEG) signal is widely used. This is a very popular technique used to identify abnormal brain activity. It records the activity of brain neurons [5]. The most widely used technique for detecting epilepsy is pattern recognition. Researchers use DWT, CWT, FT, DFT, IDFT, FFT and STFT feature extraction methods to extract hidden patterns from EEG signals. For the classification methods, researchers used various classifiers such as Support Vector Machines(SVM), k-NN, Decision Tree, ANN, RNN, Naive Bayes (NB), to detect epileptic seizures. In this paper, we present an effective machine learning-based seizure detection framework using SVMs trained on balanced CHB-MIT EEG data. EEG signals were decomposed into subbands via DWT, and six discriminative features were extracted from each. We used Random Forest Permutation Importance based importance ranking for feature selection. Both tuned and non-tuned SVM models were evaluated.

II. LITERATURE REVIEW

Several studies have explored EEG-based epileptic seizure detection using diverse feature extraction and classification strategies. Übeyli [6] employed a least squares support vector machine (LS-SVM) using model-based coefficients to classify EEG signals, achieving high accuracy but depending on static, handcrafted features. More recent research emphasizes hybrid

feature selection and data-driven optimization for improved seizure classification. Atlam et al. [7] integrated hybrid feature selection with SMOTE-based data balancing and an SVM classifier, enhancing robustness against class imbalance. Kong et al. [8] advanced this trend by introducing a multi-domain feature fusion and selection technique, extracting and optimizing temporal, spectral, and statistical EEG features for higher detection accuracy. In a recent study, DWT was used to decompose EEG into sub-bands and an SVM classifier was applied to the resulting features, achieving near-99% accuracy with a minimal feature set. This highlights the continuing relevance of well-engineered feature extraction combined with SVM in automated seizure detection [9]. Several recent studies have used DWT subband decomposition to capture the time–frequency dynamics of EEG and then applied machine-learning classifiers. For example,the authors fused DWT and EMD features across multiple channels and used SVM classifiers to achieve improved performance [10] . On the feature-selection side, in a work the authors applied multi-domain feature fusion and used PSO-based optimization before classification, demonstrating that careful feature selection can significantly boost classifier performance [11]. Building upon these foundations, this study employs a Random Forest based feature selection integrated with SVM classification to achieve high accuracy and computational efficiency in automated epileptic seizure detection.

III. DATASET DESCRIPTION AND PREPROCESSING

A. Dataset

The CHB-MIT Scalp EEG Dataset [12] is a widely used benchmark for automated seizure detection and is publicly available through PhysioNet. It contains long-term scalp EEG recordings from 23 pediatric patients (ages 1.5–22 years) with drug-resistant epilepsy, acquired at 256 Hz using a 24-channel bipolar 10–20 montage. Clinical experts annotated seizure onset and offset times, providing over 900 hours of data with 198 seizures, although the dataset remains highly imbalanced due to the predominance of non-seizure activity. Variations in electrode availability across subjects introduce structural inconsistencies in cross-patient studies. This study works with the 18 channels which are common to all recordings. It enables fair evaluation and reduces information loss as well [13].

B. Data Preparation and Balancing

The dataset is available in EDF format which was converted to MATLAB format for ease of processing the data. Data of each subject was segmented using a 3-second window which contains both seizure and non-seizure data provided by onset–offset annotations. Because of seizure events occupy far less time than non-seizure data the dataset is not balanced. The non-seizure segments were down-sampled to balance the data and match with the total length of seizure segment for each subject. The balancing ensures that data is unbiased, which improved sensitivity. A uniform channel set was required to ensure consistency across subjects. That is the reason behind selecting 18 common channels that represent all major cortical regions. These standardized channels reduce inter-subject variability and enhance generalization across patients. Using this standardized set of channels reduces subject-to-subject variation.

IV. METHODOLOGY

The workflow of this proposed work has four major steps. These are decomposing the signal into subbands, extracting the features, selecting the most important features, and classification of seizure-non seizure signals. The overview of this pipeline is showing in Figure 1.

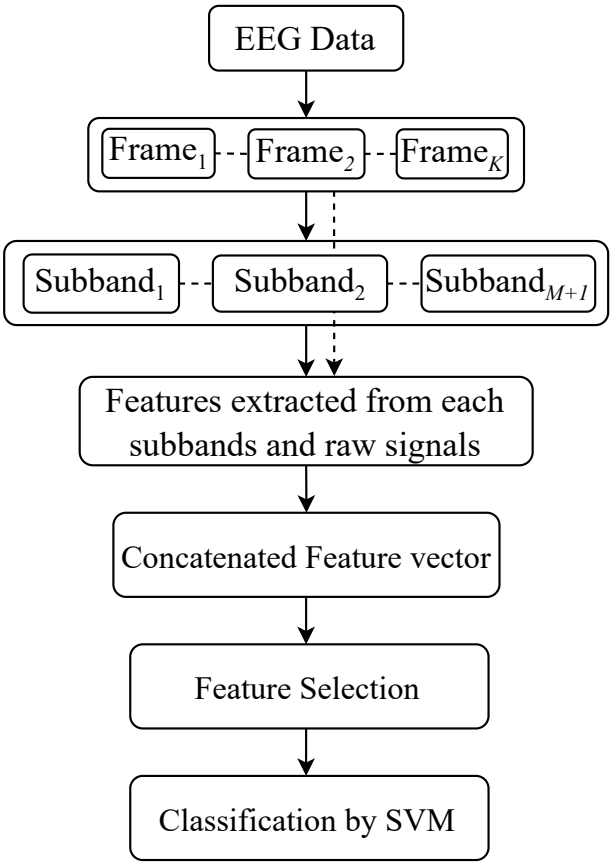


Figure 1. Proposed workflow for EEG-based seizure detection.

A. Signal Decomposition

A five-level Discrete Wavelet Transform (DWT) was used to decompose the signals. It is widely preferred for EEG signal analysis because it has capability of strong time–frequency localization, also it is resemblance to EEG waveforms. In this decomposition technique, the signal is iteratively filtered by a low-pass filter which enables the system to obtain approximation coefficients (A) and a high-pass filter to obtain detail coefficients (D). The approximation represents low-frequency and the detail represents high-frequency components. The approximation is further decomposed at each level and progressively isolating distinct EEG frequency bands. The db8 wavelet

function is used to decompose EEG signals into subbands through L levels of DWT, producing one approximation and L detail coefficients. This results in $L+1$ reconstructed subband signals $(s_1, s_2, \ldots, s_{L+1})$. In this paper, five levels of wavelet decomposition is applied and six subbands are obtained (Sb_1 to Sb_6) with cutoff frequencies 64–128 Hz, 32–64 Hz, 16–32 Hz, 8–16 Hz, 4–8 Hz, 2-4 Hz and 0–2 Hz. Except for Sb_1, the other subbands combine components from multiple rhythms. However, Gamma rhythms $(> 30Hz)$ which have played a vital role for seizure detection, other rhythms also help to detect the same task. Transient and stationary EEG features are captured by this multiresolution representation which are essential for seizure characterization [14].

$$A_j(n) = \sum_k h(k)\, A_{j-1}(2n-k) \tag{1}$$

$$D_j(n) = \sum_k g(k)\, A_{j-1}(2n-k) \tag{2}$$

Here, approximation and detail coefficient is represented by$A_j(n)$ and $D_j(n)$ at the jth decomposition level, respectively. The db4 mother wavelet used for derivation of$h(k)$ low-pass and $g(k)$ high-pass filters. Slow, low-frequency EEG trends were captured by the approximation coefficients and high-frequency, rapidly changing components are retained by the detail coefficients that often contain seizure-related spikes.

B. Feature Extraction

Six time-frequency features were extracted from each of five subbands (delta, theta, alpha, beta, and gamma) across 18 channels. The features are: Line Length (LL) [15], Root Mean Square (RMS) [16], Hjorth Mobility (HM) [17], Hjorth Complexity (HC) [17], Teager-Kaiser Energy Operator (TKEO) [18], and Spectral Entropy (SE) [19].

Line Length measures the cumulative absolute change in amplitude.

$$LL = \sum_{i=1}^{N-1} |x_{i+1} - x_i| \tag{3}$$

RMS measures the signal's overall energy level.

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{4}$$

Hjorth parameters measure average frequency variation and irregularity of signal.

$$HM = \sqrt{\frac{\mathrm{var}(\dot{x})}{\mathrm{var}(x)}}, \quad HC = \frac{\sqrt{\frac{\mathrm{var}(\ddot{x})}{\mathrm{var}(\dot{x})}}}{HM} \tag{5}$$

TKEO captures instantaneous nonlinear energy.

$$TKEO = \frac{1}{N-2}\sum_{i=2}^{N-1}(x_i^2 - x_{i-1}x_{i+1}) \tag{6}$$

Spectral Entropy measures randomness of the signal.

$$SE = -\sum_k P(k)\log_2 P(k) \tag{7}$$

where $P(k)$ is the normalized power spectral density.

All five subbands feature vectors are concatenated for better performance.

C. Feature Selection Using Random Forest Importance

A Random Forest based feature selection method was used to improve classification performance of the system. This feature selection method uses a 200-tree Random Forest trained on the training set. It computed the importance of each feature f_j using the out-of-bag (OOB) permutation method. This technique works by measuring the classification error increases while permuting a specific feature is randomly and disrupting its relationship with the output label. For feature the importance score f_j is defined as:

$$I(f_j) = \frac{1}{T}\sum_{t=1}^{T}\left(E_t^{\text{perm}}(f_j) - E_t^{\text{orig}}\right), \tag{8}$$

where E_t^{orig} is the OOB error of tree t, and $E_t^{\text{perm}}(f_j)$ is the OOB error after permuting feature f_j. The feature's contribution to predictive performance is indicated by a larger value of $I(f_j)$. All features were ranked in descending order after computation of importance scores. For classification, top 35% most informative features were retained. The preservation of most discriminative time–frequency characteristics are ensured by this approach. It also reduces the computational complexity.

D. SVM-Based Classification

In this work, a Support Vector Machine (SVM) is introduced with a radial basis function (RBF) kernel to classify the EEG data [20], [21]. The preprocessed features from each subject were split into training and testing sets using an 80%-20% ratio. A manual grid search was conducted to select the optimal hyperparameters for the SVM. The box constraint C was tested over the range $\{0.01, 0.1, 1, 10, 100\}$, and the kernel scale parameter γ was evaluated over $\{0.001, 0.01, 0.1, 1\}$, where KernelScale $= 1/\sqrt{\gamma}$. For each combination, the SVM model was trained on the training set and validated on the test set, and the configuration yielding the highest accuracy was chosen as the final model.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Each subject's feature set was normalized (z-score) before classification. Metrics included Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), and F1-score. In this work the proposed tuned SVM model with Random Forest based feature selection achieved higher accuracy compared to other applied models. Also this method achieved 98.70% accuracy with normal SVM model. The comparison of models and validation strategies is presented below:

The tuned SVM achieved the highest average accuracy (99.63%) after parameter optimization, confirming the effectiveness of the RBF kernel with balanced data. In addition, the feature selection model increased the accuracy by selecting the most effective feature columns.

Balancing the dataset significantly improved the sensitivity score, which rose from approximately 70% in the imbalanced

TABLE I
PERFORMANCE SUMMARY OF SVM MODELS ON BALANCED CHB-MIT DATASET

Model	Acc(%)	Sen(%)	Spe(%)	F1(%)
KNN	97.14	96.13	98.18	97.02
Decision Tree	95.91	96.31	95.40	95.95
Logistic Regression	98.15	97.31	96.65	98.08
Non-tuned SVM	98.70	99.03	98.50	98.69
Tuned SVM	**99.63**	**99.55**	**99.72**	**99.63**

TABLE II
PERFORMANCE COMPARISON WITH OTHER SVM SYSTEMS

Model	Accuracy (%)
DWT-Based SVM [9]	99
SMOT-SVM [7]	99.35
LS-SVM [6]	99.56
Proposed Tuned-SVM	**99.63**

setting to above 96%. Hyperparameter tuning further improved specificity and overall stability. This confirms that combining feature-level engineering with optimized SVMs can rival deep learning performance on limited data. Table III presents a

TABLE III
PERFORMANCE COMPARISON WITH FEATURE SELECTION MODELS

Model	Accuracy (%)
No feature selection	99.53
REF	99
ReliefF	99.59
Proposed Random Forest based model	**99.63**

comparative analysis of classification performance with and without the application of feature selection techniques. The classification model achieved an accuracy of 99.53%. After applying the proposed Random Forest Permutation Importance based feature selection method the classifier achieved the highest classification accuracy of 99.63%. The feature selection model proves its acceptance outperforming all other methods. The results confirm the effectiveness and robustness of this Random Forest based feature selection approach by optimizing the classifier's performance.

The LOSO cross-validation results presented in Table IV that shows a moderate level of subject independence in classification performance compared to subject-specific evaluation. The result shows a good number of subjects achieved accuracy above 90% and most of the subjects achieved accuracy above the average accuracy of 85.47%.

VI. CONCLUSION

This work presented an efficient machine learning approach for epileptic seizure detection. It combined Discrete Wavelet Transform (DWT)-based subband decomposition with six discriminative time–frequency features as well as a Random Forest Permutation Importance based feature selection mechanism. The dimensionality of the data was reduced with high classification accuracy by selecting top 35% features. A Support Vector Machine (SVM) was introduced with a radial basis function (RBF) kernel to classify the selected

TABLE IV
SUBJECT-WISE LOSO CROSS-VALIDATION RESULTS

Subject	Acc (%)	Sen (%)	Sp (%)	F1 (%)
chb01	77.89	55.78	100.00	71.62
chb02	96.49	98.25	94.74	96.55
chb03	95.15	94.03	96.27	95.09
chb04	86.51	88.10	84.92	86.72
chb05	87.90	100.00	75.81	89.21
chb06	67.65	100.00	35.29	75.56
chb07	94.91	90.74	99.07	94.69
chb08	90.85	84.31	97.39	90.21
chb09	91.85	100.00	83.70	92.46
chb10	92.62	95.97	89.26	92.86
chb11	96.46	97.76	95.15	96.50
chb12	86.17	84.19	88.15	85.89
chb13	56.46	50.00	62.92	53.45
chb14	50.89	100.00	1.79	67.07
chb15	82.76	67.62	97.89	79.68
chb16	92.86	89.29	96.43	92.59
chb17	84.02	69.07	98.97	81.21
chb18	84.76	71.43	98.10	82.42
chb19	96.79	93.59	100.00	96.69
chb20	72.96	55.10	90.82	67.08
chb21	81.82	63.64	100.00	77.78
chb22	98.53	100.00	97.06	98.55
chb23	89.01	88.65	89.36	88.97
chb24	95.87	92.41	99.34	95.73
Mean	**85.47**	**84.58**	**86.35**	**85.36**
Std	**12.54**	**16.20**	**23.19**	**11.69**

features. Combination of all these achieved an outstanding accuracy of 99.63%. The tuned SVM model outperformed traditional classifiers such as KNN, Decision Tree, and Logistic Regression. This work also outperforms several previously reported SVM-based approaches in the literature. The findings confirm that integration of subband feature extraction with concatenation of the feature vectors with Random Forest based feature selection, and optimized SVM classification produces a highly accurate and computationally efficient seizure detection system. While initial evaluations showed that the model achieves near-perfect results in subject-specific scenarios, we further extended our analysis to address the critical challenge of cross-patient generalization. The proposed model demonstrated robust performance with a mean accuracy of 85.47% and a mean F1-score of 85.36%, by implementing Leave-One-Subject-Out (LOSO) cross-validation across 24 subjects. Our Future work will explore deep learning-based domain adaptation and subject-invariant feature representations to further reduce performance variance and enhance real-time monitoring capabilities.

REFERENCES

[1] K. M. Fiest, K. M. Sauro, S. Wiebe, S. B. Patten, C.-S. Kwon, J. Dykeman, T. Pringsheim, D. L. Lorenzetti, and N. Jetté, "Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies," *Neurology*, vol. 88, no. 3, pp. 296–303, 2017.

[2] U. R. Acharya, F. Molinari, S. V. Sree, S. Chattopadhyay, K.-H. Ng, and J. S. Suri, "Automated diagnosis of epileptic eeg using entropies," *Biomedical signal processing and control*, vol. 7, no. 4, pp. 401–408, 2012.

[3] A. Shoeibi, M. Khodatars, N. Ghassemi, M. Jafari, P. Moridian, R. Alizadehsani, M. Panahiazar, F. Khozeimeh, A. Zare, H. Hosseini-Nejad, *et al.*, "Epileptic seizures detection using deep learning techniques:

a review," *International journal of environmental research and public health*, vol. 18, no. 11, p. 5780, 2021.

[4] R. Hussein, H. Palangi, R. K. Ward, and Z. J. Wang, "Optimized deep neural network architecture for robust detection of epileptic seizures using eeg signals," *Clinical Neurophysiology*, vol. 130, no. 1, pp. 25–37, 2019.

[5] X. Tian, Z. Deng, W. Ying, K.-S. Choi, D. Wu, B. Qin, J. Wang, H. Shen, and S. Wang, "Deep multi-view feature learning for eeg-based epileptic seizure detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 10, pp. 1962–1972, 2019.

[6] E. D. Übeyli, "Least squares support vector machine employing model-based methods coefficients for analysis of eeg signals," *Expert Systems with Applications*, vol. 37, no. 1, pp. 233–239, 2010.

[7] H. F. Atlam, G. E. Aderibigbe, and M. S. Nadeem, "Effective epileptic seizure detection with hybrid feature selection and smote-based data balancing using svm classifier," *Applied Sciences*, vol. 15, no. 9, p. 4690, 2025.

[8] G. Kong, S. Ma, W. Zhao, H. Wang, Q. Fu, and J. Wang, "A novel method for optimizing epilepsy detection features through multi-domain feature fusion and selection," *Frontiers in Computational Neuroscience*, vol. 18, p. 1416838, 2024.

[9] S. Himalyan and V. Gupta, "Support vector machine-based epileptic seizure detection using eeg signals," *Engineering Proceedings*, vol. 18, no. 1, p. 73, 2024.

[10] G. Jana, A. Agrawal, P. Pattnaik, and M. Sain, "Dwt-emd feature level fusion based approach over multi and single channel eeg signals for seizure detection. diagnostics, 12 (2): 324," 2022.

[11] Y. Miao, "Seizure type classification based on hybrid feature engineering and mutual information analysis using electroencephalogram," *Entropy*, vol. 27, no. 10, p. 1057, 2025.

[12] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[13] J. Prasanna, M. Subathra, M. A. Mohammed, R. Damaševičius, N. J. Sairamya, and S. T. George, "Automated epileptic seizure detection in pediatric subjects of chb-mit eeg database—a survey," *Journal of Personalized Medicine*, vol. 11, no. 10, p. 1028, 2021.

[14] Y. Kumar, M. Dewal, and R. S. Anand, "Epileptic seizure detection using dwt based fuzzy approximate entropy and support vector machine," *Neurocomputing*, vol. 133, pp. 271–279, 2014.

[15] L. Guo, D. Rivero, J. Dorado, J. R. Rabunal, and A. Pazos, "Automatic epileptic seizure detection in eegs based on line length feature and artificial neural networks," *Journal of neuroscience methods*, vol. 191, no. 1, pp. 101–109, 2010.

[16] D. Zwillinger and S. Kokoska, *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.

[17] W. H. Alawee, A. Basem, and L. A. Al-Haddad, "Advancing biomedical engineering: Leveraging hjorth features for electroencephalography signal analysis," *Journal of electrical bioimpedance*, vol. 14, no. 1, p. 66, 2023.

[18] S. Badani, S. Saha, A. Kumar, S. Chatterjee, and R. Bose, "Detection of epilepsy based on discrete wavelet transform and teager-kaiser energy operator," in *2017 IEEE Calcutta Conference (CALCON)*, pp. 164–167, IEEE, 2017.

[19] A. Mirzaei, A. Ayatollahi, P. Gifani, and L. Salehi, "Eeg analysis based on wavelet-spectral entropy for epileptic seizures detection," in *2010 3rd International Conference on Biomedical Engineering and Informatics*, vol. 2, pp. 878–882, IEEE, 2010.

[20] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "Eeg-based neonatal seizure detection with support vector machines," *Clinical Neurophysiology*, vol. 122, no. 3, pp. 464–473, 2011.

[21] C. A. Lima and A. L. Coelho, "Kernel machines for epilepsy diagnosis via eeg signal classification: A comparative study," *Artificial Intelligence in Medicine*, vol. 53, no. 2, pp. 83–95, 2011.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

AI Based medical Transparent and Non Transparent waste classification

Md Shakhawat Hossain
Computer Science and Engineering Pabna University of Science and Technology
Pabna, Bangladesh
shshimul.cse.pust@gmail.com

Istiak Ahammed
Computer Science and Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
istiak.220123@s.pust.ac.bd

Razu Ahmed
Computer Science and Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
razu.cse.pust@gmail.com

Md Saiful Islam
Computer Science and Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
saiful.pust.bd71@gmail.com

***Abstract*— Environmental preservation and public health depend on efficient medical waste management. Because of visual similarities and backdrop blurring, separating transparent from non-transparent medical waste poses special issues for automated sorting systems. The study describes a deep learning-based system that uses a MobileNetV2 architecture with transfer learning to categorize medical waste into transparent and non-transparent categories. Two publicly accessible medical waste sources provided a combined dataset of 4000 photos, which were then visually classified into two groups and improved through data augmentation to increase generalization. According to experimental data, the suggested model achieves a near-perfect ROC AUC of 0.9999, balanced per-class performance (F1-scores ≈0.991), minimum incorrect classifications, and a test accuracy of 99.16%. Strong stability and no overfitting are shown by the training dynamics and confusion matrix. The results indicate that MobileNetV2 has great potential for integration into automated recycling facilities, intelligent sorting systems, and real-time embedded applications in healthcare settings. It is also very successful for transparency-aware trash categorization.**

Keywords— ***MobileNetV2, CNN, transfer learning, deep learning, medical waste categorization, transparent and non-transparent waste classification.***

I. INTRODUCTION

Medical waste management is a vital worldwide requirement for public health and environmental preservation. This intricate waste stream, produced by all healthcare operations, is made up of a variety of components, including contaminated personal protective equipment (PPE) from medical equipment to medications, radioactive materials, and substances [1].

This study presents a deep learning–based approach for automatically classifying medical waste into **transparent** and **non-transparent** categories using a Convolutional Neural Network (CNN). Because of its capacity to extract hierarchical feature representations from images, Convolutional Neural Networks (CNNs) have taken the lead in visual recognition [2]. Among contemporary CNN designs, MobileNetV2 has become a potent yet computationally economical model that is particularly well-suited for real-time and lightweight applications [3].

Automated image-based garbage classification systems have been made possible by recent developments in deep learning and computer vision. Previous research has shown good performance in identifying a variety of medical waste products, such as gloves, syringes, gauze, and test tubes, utilizing models like ResNet-50, YOLO versions, and lightweight CNNs [4] , [5]. This study suggests a Transfer learning framework based on MobileNetV2 for dividing medical waste into transparent and non-transparent categories in order to close this gap. To enhance generalization, a custom dataset of real-world photos from both classes was created, and data augmentation was used to train the model. The goal of this study is to assess how well a MobileNetV2 can differentiate between transparent and non-transparent trash objects and to examine the visual elements that affect classification performance, such as reflection, edge visibility, and backdrop blending. Transfer learning is a prevalent method that allows models pretrained on extensive datasets like ImageNet to efficiently adjust to smaller, domain-specific datasets [6].

In order to address an understudied but practically significant issue in automated waste sorting and safe disposal systems, this research suggests a transparency-aware medical waste classification framework that makes a distinction between transparent and non-transparent medical waste. To improve robustness and generalization, a balanced dataset of 4000 real-world medical waste photos was assembled from two publicly accessible sources, guaranteeing equal representation of both classes and encompassing a variety of waste objects. The model is appropriate for real-time and embedded healthcare applications because it uses a lightweight MobileNetV2-based transfer learning architecture to provide excellent classification accuracy with little computing complexity. With a ROC-AUC of 0.9999 and a test accuracy of 99.16%, the suggested method is thoroughly assessed using conventional performance measures, exhibiting consistent and objective performance. The outcomes verify that the suggested solution can be

implemented in edge AI platforms, automated sorting, and intelligent hospital waste management.

II. Related Work

A. Deep learning for medical waste classification

Deep learning approaches have demonstrated significant success in waste classification. A Study of Medical Waste Classification using Deep Learning [7] offers a deep learning-based method for categorizing various medical waste types in order to facilitate safer and more effective waste management in healthcare settings. To boost variability, the scientists created a unique image dataset and employed data augmentation. In order to automatically classify medical waste into specified categories, including general, dangerous, and infectious waste, they trained convolutional neural network (CNN) models. Their tests suggest that CNN-based techniques can consistently identify medical waste and lower manual sorting errors, with accuracy levels above 93–96%.

Medical Waste Classification using Deep Learning and Convolutional Neural Networks [8], In order to categorize medical waste photos, this study explores the use of transfer learning with a number of pre-trained deep learning models, including VGG19, ResNet50, InceptionV3, and EfficientNetB7. To increase the robustness of the model, the authors created a dataset of over 3200 photos and used augmentation. With InceptionV3 obtaining 96.49%, ResNet50 achieving 95.72%, and EfficientNetB7 achieving 95.55%, their findings demonstrate great classification accuracy across several architectures. The study shows that transfer learning in conjunction with contemporary CNN architectures can efficiently classify medical waste photos and facilitate automation in healthcare waste management.

B. MedBin for medical waste classification

MedBin: A lightweight End-to-End model-based method for medical waste management [5], MedBin is an AI system meant to automatically identify, segregate, and categorize medical waste, and also propose the best disposal option based on WHO criteria. MedBin employs a lightweight model called MedBin-Net, which can recognize objects, conduct instance segmentation, and categorize medical waste while being exceptionally quick and small—only 2.51 million parameters, 0.60 ms inference time, and 5.20 GFLOPs. A WHO-based classifier is also included, which divides waste into five categories (Infectious, Sharps, Pharmaceutical, Radioactive, and General) and displays the appropriate disposal technique in an easy-to-use graphical user interface. With 0.91 accuracy, 0.97 recall, 0.94 F1-score, and 0.71 mAP, MedBin performs exceptionally well and uses far less memory and compute to get results that are almost as excellent as those of huge models as Mask R-CNN and SOLO-V2.

C. Motivation for current work

In hospital settings, proper medical waste segregation is crucial because improper sorting can increase waste handlers' exposure to infections, pollute the environment, and pose substantial health hazards. The World Health Organization (WHO) states that a large percentage of medical waste is dangerous and has to be handled and disposed of carefully to avoid contamination and the spread of illness [14]. The necessity for an automated, intelligent system that can support waste management procedures is highlighted by the fact that manual sorting frequently suffers from inconsistency, weariness, and direct contact with infectious materials.

Recent developments in artificial intelligence, especially deep learning, have demonstrated great promise for general garbage-sorting systems and medical waste detection. Convolutional neural networks (CNNs) and transfer-learning models have been shown in several studies to achieve excellent accuracy in the classification of medical waste categories, including sharps, pathological waste, chemical waste, and plastic trash [15], [16]. However, the majority of current research concentrates on class-based classification and ignores the crucial problem of differentiating between transparent medical waste (such IV bags, pipettes, and prescription bottles) and non-transparent objects (like gloves, masks, and syringes). Because of their low roughness, reflections, refractions, and merging with the background, transparent objects are intrinsically challenging to identify, which severely impairs the performance of conventional vision models.

The current effort focuses on creating an AI-based classification system that can distinguish between transparent and non-transparent medical waste utilizing a lightweight and effective deep learning model in order to close this gap. This study intends to increase waste segregation accuracy, safety for healthcare workers, and enable more intelligent hospital waste-management systems by focusing on the transparency feature, which is still understudied in the literature. Our suggested technique, which seeks to provide dependable real-time categorization even in visually demanding circumstances, is based on this motive.

III. Methodology

A. Dataset Collection and Description

In this work, a transparent and non-transparent waste categorization system was constructed using two publicly accessible medical waste picture datasets. Mario78 gathered the Biomedical Waste Classification Images Dataset on Kaggle, [9] and Sumn2u published the Medical Waste Dataset on Kaggle [10]. Gloves, syringes, bottles, plastic objects, tubes, and other typical medical disposables are among the biomedical and hospital waste products included in both databases.

Images that visually fit into the transparent and non-transparent categories were personally chosen from these two databases. Items that were transparent included pipettes, test tubes, IV bags, plastic tubes, and clear containers. Gloves, syringes with colorful bodies, masks, gauze, paper goods, and metal instruments were examples of non-transparent things. A mixed dataset of 4000 photos with both transparent and non-transparent classes was produced following filtering.

In order to guarantee that the dataset maintained consistent visual quality and enhanced model performance throughout training, all photos were shrunk to a uniform dimension, normalized, and cleaned by eliminating duplicate or poor-quality samples.

B. Dataset Splitting

In accordance with common machine learning procedures for picture classification problems, the entire dataset was split into three subgroups.

- Train 70%
- Test 15%
- Validation 15%

This 70:15:15 style split ensured that the model experienced enough variation during training while maintaining separate validation and test sets for fair model evaluation.

TABLE 1. Data Distribution

	Transparent	Non-Transparent	Total
Train	1400	1400	2800
Test	300	300	600
Validation	300	300	600
Total	2000	2000	4000

C. MobileNetV2 Model

MobileNet is a lightweight network introduced by Google in 2017, designed for embedded devices like cell phones. The model optimizes the network structure by employing depthwise separable convolution in place of standard convolution, resulting in a significant reduction in both the number of parameters and computational requirements compared to other networks [11]. The MobileNetV2 model enhances the MobileNet model through the incorporation of an inverted residual structure and a linear bottleneck [12].

MobileNetV2 is a lightweight and efficient convolutional neural network (CNN) architecture that introduces two key innovations: inverted residual blocks and linear bottlenecks. These innovations significantly reduce computational costs while preserving high accuracy.

The Inverted Residual Block with Linear Bottleneck, MobileNetV2's key computational component, is intended to deliver great accuracy at a cheap computational cost. The block is composed of three successive stages: projection, depthwise convolution, and expansion, as shown in Fig. X. The network may learn a richer intermediate representation by first passing the input tensor via a 1×1 pointwise convolution, which increases the number of channels from c to t×c. A 3x3 depthwise convolution that separately filters each channel comes next, greatly lowering computational complexity without sacrificing the capacity to extract spatial features.

Ultimately, the channel dimension is reduced down to c by a linear 1x1 pointwise projection, creating a compact bottleneck representation that minimizes information loss. A residual shortcut connection is used to improve gradient flow and maintain representational continuity between layers when the input and output dimensions are the same and the stride equals one. MobileNetV2 is especially well suited for implementation in mobile and embedded vision applications because to its architectural design, which increases accuracy and efficiency [17].

D. Evaluation Metrics

1) Precision: Precision is the percentage of how many results that were processed are correct [13]. Mathematically it is the ratio of True Positives to the sum of True Positives and False Positves as shown in (1) below

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

2) Recall: Recall is the percentage of how many correct results were found after processing [13]. Mathematically it is the ratio of True Positives(TP) to the sum of True Positives(TP) and True Negatives(TN) as shown in (2) below.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

3) F-Score: F-score is the harmonic mean of Precision and Recall as showin in (3) below [13].

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

E. Workflow

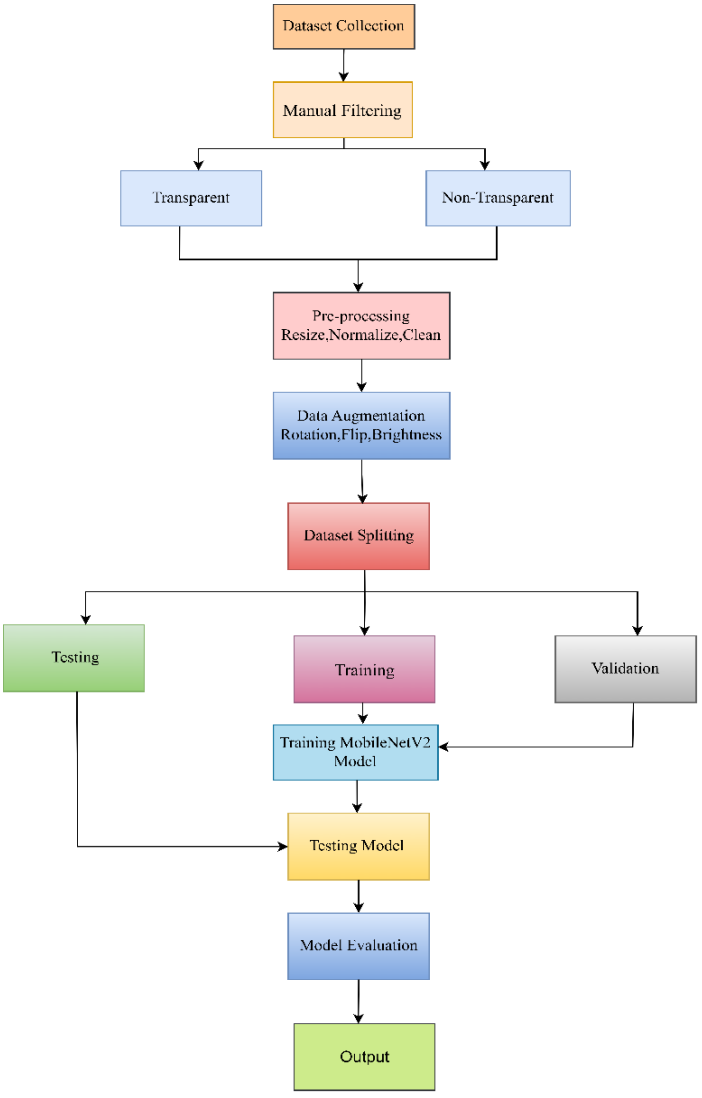


Fig 1: Workflow of the system

IV. Results and Discussion

A. Overall Performance

The proposed CNN-based MobileNetV2 architecture achieves a test accuracy of 99.16%. The test dataset, including 600 photos (300 transparent and 300 non-transparent), was used to assess the effectiveness of the suggested MobileNetV2 model for the categorization of transparent and non-transparent medical waste. The model demonstrated that a lightweight MobileNetV2 can successfully learn visual patterns associated with object transparency, such as edge sharpness, reflection, and background blending, with an overall accuracy of 99%. Training converged in 40 epochs over 35.50 minutes, with training accuracy of 99.05% and validation accuracy of 98.83%, indicating excellent generalization with minimal overfitting. The final test loss was 2.54%.

B. Per-Class Analysis

In both classes, the model regularly performs well. The precision of 0.9898 and recall of 0.9932 for the Non-Transparent class show that the model accurately detects nearly all non-transparent samples with few false positives. In a similar vein, the Transparent class demonstrates balanced and extremely precise detection with a precision of 0.9933 and a recall of 0.9900. The classifier retains great reliability and consistency across categories, as evidenced by the virtually identical F1-scores (≈0.991) for both classes. This balanced per-class performance shows that the model generalizes well and shows no bias toward either class.

TABLE 2. Classification Performance Metrics

Class	Precision	Recall	F1-Score
Non-Transparent	0.9898	0.9932	0.9915
Transparent	0.9933	0.9900	0.9917

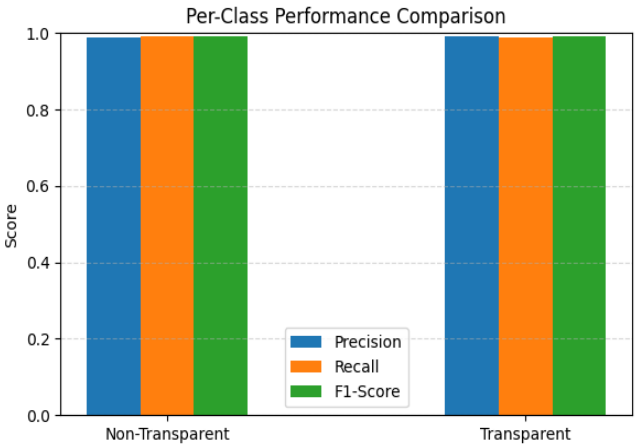


Fig 2: Per-class performance comparison showing precision, recall, and F1-score for Non-Transparent and Transparent

C. Confusion Matrix Analysis

The proposed MobileNetV2-based classifier performs extremely reliable categorization between transparent and non-transparent garbage, as shown by the confusion matrix. With just two and three incorrect classifications, respectively, the model correctly identified 292 non-transparent and 297 transparent pictures out of the 594 test samples. This results in an overall accuracy of about 99%, meaning that there are few false positives and false negatives. The model's consistent performance across both classes with no discernible bias is confirmed by the balanced diagonal dominance.

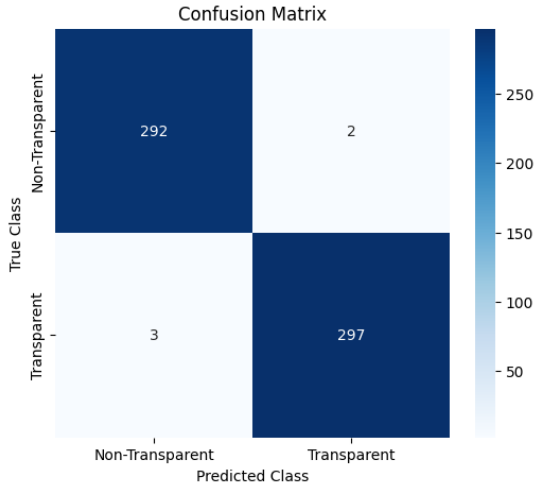


Fig 3: The Confusion Matrix shows the classification of transparent and non-transparent.

D. Training Dynamics

A steady and seamless convergence pattern can be seen in the training and validation curves. While validation accuracy closely follows at 0.988–0.99, suggesting excellent generalization, training accuracy improves quickly and reaches 0.99. In a similar vein, there is no significant divergence between the training and validation loss curves since they both progressively decline across the epochs. This alignment demonstrates that the fine-tuning technique successfully improves feature learning while preserving stability, and the model is neither overfitting nor underfitting.

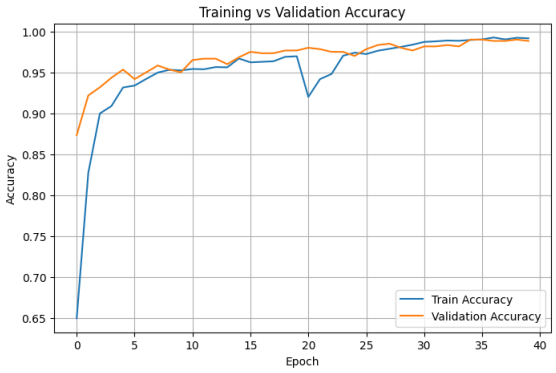


Fig 4: Training vs. Validation Accuracy

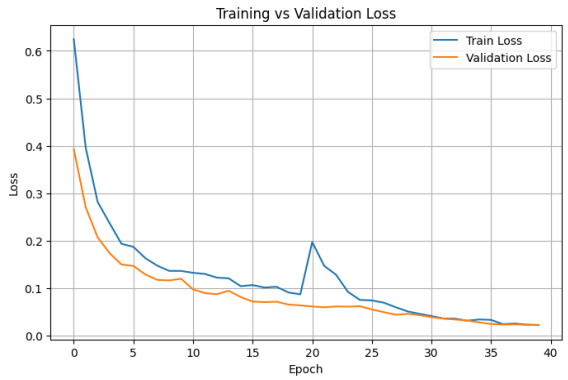


Fig 5: Training vs. Validation Loss

E. ROC Analysis

With an AUC (Area Under Curve) of 0.9999, the model has nearly flawless accuracy in differentiating between transparent and non-transparent medical waste.

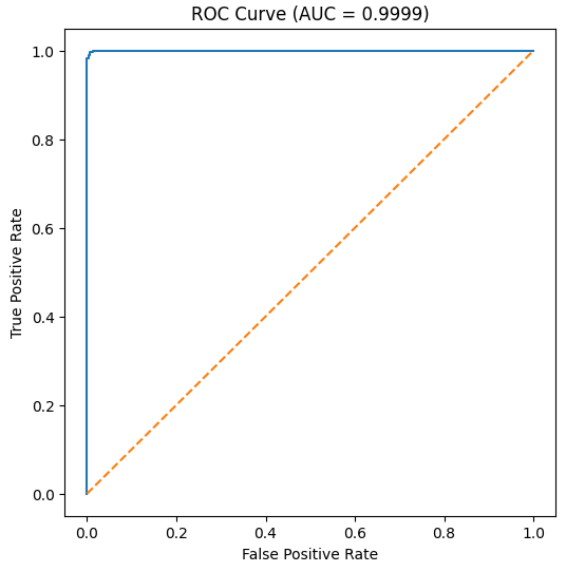


Fig 6: ROC curves

V. LIMITATIONS AND FUTURE WORK

Although the suggested MobileNetV2-based model shows promise accuracy in differentiating between transparent and non-transparent medical waste, there are still a number of areas that might be improved. First, building a bigger and more varied dataset that includes differences in lighting, camera specs, backdrops, and waste types might improve the model's resilience in actual hospital settings. Dataset heterogeneity greatly enhances generalization in medical imaging applications, according to earlier research [18].

Furthermore, using sophisticated object detection frameworks like YOLOv8 or Faster R-CNN might enable multi-class segmentation and real-time waste identification. These models have gained widespread recognition for their cutting-edge results in challenges involving fine-grained visual categorization and object identification [19], [20]. Syringes and needles, IV bags, pharmaceutical bottles, glass ampoules, cotton-based materials, protective masks or gloves, sharp containers, and pathological waste are just a few of the medical waste kinds that may be identified by a multi-class system. The operational utility and feasibility of automated waste management systems would be significantly improved by such fine-grained categorization.

Future studies may potentially look into IoT-based smart bins or lightweight deployment on embedded systems. Optimized deep learning models may be successfully implemented on low-power devices for real-time environmental monitoring and picture categorization, according to recent developments in edge AI [21]. Last but not least, using Explainable AI (XAI) approaches may offer clinical settings transparency, interpretability, and reliability, guaranteeing that medical staff and regulatory agencies can understand model judgments [22].

TABLE 3: Performance Comparison with Existing Medical Waste Classification Methods

Study	Task	Model Used	Dataset Size	Accuracy (%)
Vijaya kumar *et al.* [7]	Multi-class medical waste classification	Custom CNN	3000 images	93-96
Verma *et al.* [8]	Multi-class medical waste image classification	VGG19, ResNet50, Inception V3, Efficient NetB7	3200 images	95.55-96.49
Al-Maaded *et al.* [15]	Medical waste category classification	Deep CNN	Not specified	94.8
Lee *et al.* [16]	General and medical waste classification	Transfer Learning CNN	Not specified	95.2
Xu *et al.* (MedBin) [5]	Medical waste detection	MedBin-Net (Lightweight CNN)	Custom dataset	91.0
Proposed Method	**Binary medical waste classification (Transparent vs Non-Transparent)**	**MobileNetV2 (Transfer Learning)**	**4000 images**	**99.16**

VI. Conclusion

In this study, a refined MobileNetV2 architecture was used to present a transparency-aware medical waste categorization system. The model performed well, achieving 99.16% test accuracy, extremely constant precision and recall, and an AUC of 0.9999 using a dataset of 4000 pictures split into transparent and non-transparent categories. With only a few misclassifications and no discernible class bias, the confusion matrix further confirmed the model's dependability.

Overall, the suggested approach offers a highly effective and practical solution for intelligent medical waste classification and contributes to safer and more sustainable healthcare waste management systems. Future work may expand the dataset and investigate multi-class classification to further enhance practical applicability. The lightweight architecture makes it suitable for real-time deployment in automated waste sorting and healthcare environments.

VII. References

[1] H. Nematollahi, M. Tuysserkani and A. Nematollahi, "Medical Waste Management in the Modern Healthcare Era: A Comprehensive Review of Technologies, Environmental Impact, and Sustainable Practices," *Results in Engineering,* p. 107210, 2025.

[2] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, p. 436–444, 2015.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *CVPR (IEEE Conference on Computer Vision and Pattern Recognition)*, 1018.

[4] H. Alrabiah, V. Ahmed and Z. Bahroun, "A Review of Waste Management in the Healthcare Sector," in *2024 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, 2024.

[5] X. Xu and e. al., "MedBin: A lightweight End-to-End model-based method for medical waste management," *Waste Management,* vol. 200, p. 114742, 2025.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 10, p. 1345–1359, 2009.

[7] A. A. Vijayakumar and e. al., "A Study of Medical Waste Classification using Deep Learning," in *2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)*, 2025.

[8] M. Verma, A. Kumar and S. Kumar, "Medical waste classification using deep learning and convolutional neural networks," in *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, 2022.

[9] Mario78, "Biomedical Waste Classification Images Dataset," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/mario78/biomedical-waste-classification-images-dataset. [Accessed 1 11 2025].

[10] sumn2u, "Medical Waste," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/sumn2u/medical-waste. [Accessed 10 11 2025].

[11] A. G. Howard, M. Zhu, B. Chen and e. al., "MobileNets: efficient convolutional neural networks for mobile vision applications," *arXiv preprint,* vol. 6, p. 1–14, 2017.

[12] 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 (CVPR)*, 2018.

[13] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint,* vol. 2010.16061, 2020.

[14] "Health-care waste," World Health Organization, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/health-care-waste. [Accessed 2025].

[15] A. A.-A. R. A.-A. a. T. A. M. A.-A. M. Al-Maadeed, "Classification of medical waste using deep learning," *International Journal of Environmental Research and Public Health,* vol. 19, no. 3, p. 1–14, 2022.

[16] T. H. N. M. R. I. S. J. Lee, "Deep learning-based automatic waste classification system using transfer learning," *IEEE Access,* vol. 9, p. 120–131, 2021.

[17] A. H. M. Z. A. Z. L.-C. C. M. Sandler, "MobileNetV2: Inverted residuals and linear bottlenecks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] A. Z. F. S. e. al, "A survey on datasets for medical image analysis," *IEEE Access,* vol. 9, p. 110271–110294, 2021.

[19] J. C. a. A. S. G. Jocher, "YOLOv8: Real-time object detection and segmentation," in *Ultralytics*, 2023.

[20] K. H. R. G. a. J. S. S. Ren, "Faster R-CNN: Towards real-time object detection with region proposal networks," *EEE Trans. Pattern Anal. Mach. Intell,* vol. 39, p. 1137–1149, 2017.

[21] S. Mittal, "A survey on optimized implementation of deep learning models on the NVIDIA Jetson platform," *J. Syst. Arch.,* vol. 117, p. 102171, 2021.

[22] T. W. a. K.-R. M. W. Samek, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *IEEE Signal Process,* pp. 66-7338, 2021.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Experimental Validation of Classical LSB Substitution Algorithms on Modern High-Resolution Datasets

Mehedi Hasan
Electrical and Electronic Engineering
Trust University, Barishal
Barishal, Bangladesh
mehedi221002.eee@trustuniversity.edu.bd

Abdullah Al Jaid Jim
Electrical and Electronics Engineering
Patuakhali Science and Technology University
Dumki - Patuakhali, Bangladesh
jaidjim@gmail.com

***Abstract*—This study re-evaluates the scalability of classical Least Significant Bit (LSB) steganography in the context of modern 4K Ultra-High-Definition (UHD) media. By applying the algorithm to the high-quality DIV2K dataset, we demonstrate that the sheer pixel density of 4K imagery acts as a natural masking agent. Our experiments show a payload capacity of nearly 3 MB—a massive 3000% increase over legacy benchmarks—while maintaining a constant Peak Signal-to-Noise Ratio (PSNR) of 51.14 dB. This effectively renders the hidden data statistically invisible to the naked eye. These findings challenge the prevailing trend toward complex, heavy adaptive algorithms, confirming that simple LSB substitution remains a computationally efficient and high-capacity solution for the modern era.**

***Index Terms*—Image Steganography, Information Security, Least Significant Bit (LSB), 4K UHD Media, Signal, Image & Video Processing**

I. Introduction

With digital surveillance becoming a constant reality, the need for truly secure communication is more urgent than ever. While cryptography is useful, it has a distinct flaw: it flags the encrypted data as "secret." Steganography solves this by hiding the very existence of the message [1]. Instead of locking the data away, steganography hides secret data inside harmless cover media, like everyday images or audio files.

The most essential technique in this field is Least Significant Bit (LSB) substitution. First formalized in the early 2000s, this method tricks the Human Visual System (HVS) by swapping the smallest bits of pixels with secret message bits. Foundational work by Chan and Cheng [2] proved that simple substitution could be invisible to the naked eye. However, those studies were limited by the hardware of their time; they validated their algorithms on low-resolution benchmarks like 'Lena' or 'Baboon' (only 512 x 512 pixels), which are tiny by today's standards.

Technology has since improved. Modern devices, from smartphones to medical scanners, now standardize High-Definition and 4K imagery. A single 4K frame contains over 8 million pixels, representing a massive increase in raw data compared to classical benchmarks. Surprisingly, despite this hardware evolution, there is very little research re-evaluating how "primitive" algorithms like LSB perform on these modern, high-density RGB datasets.

Lately, the research community has pivoted toward adaptive steganography and Deep Learning. For instance, Baluja [3] successfully used Deep Neural Networks to hide full images within others. While these AI-driven methods are robust, they come with a high computational cost. As Liu et al. [4] noted, the complexity of Generative Adversarial Networks (GANs) often makes them impractical for real-time mobile apps or low-power edge devices.

This study revisits the classics. We experimentally validate classical LSB substitution on modern 4K media using the DIV2K dataset [5]. Our hypothesis is straightforward: the sheer pixel density of modern imagery might naturally mitigate LSB's traditional weaknesses, such as statistical detectability and low capacity. By moving away from legacy grayscale datasets to high-resolution RGB, we aim to demonstrate that 20-year-old algorithms may actually offer superior performance today than when they were first invented.

II. Literature Review

We can categorize the history of image steganography into three main eras: spatial domain modifications, transform domain techniques, and the current era of deep learning approaches.

A. Classical Spatial Domain Methods

The most enduring framework in this domain is LSB substitution, which was refined significantly by Chan and Cheng [2] in 2004. They proposed the Optimal Pixel Adjustment Process (OPAP) to minimize Mean Squared Error (MSE). Their results were definitive: on 512 x 512 grayscale images, simple substitution yielded a Peak Signal-to-Noise Ratio (PSNR) of roughly 51 dB, rendering changes invisible to the naked eye.

Mielikainen [6] later addressed a weakness in simple substitution by introducing "LSB Matching" (or LSB ± 1). Rather than overwriting bits, this method randomly increments or decrements pixel values. This was a crucial update because it smoothed the stego-image's histogram, confusing statistical attacks like the RS analysis found in Fridrich et al. [7]. However, these studies were products of their time; constrained by early 2000s hardware, they rarely tested capacities beyond a few hundred kilobytes.

B. Modern Adaptive and Deep Approaches

Recently, the focus has shifted entirely to adaptive steganography and AI. For example, HUGO (Highly Undetectable steGO) [8] introduced complex cost functions to hide data specifically in "noisy" areas like textures and edges to avoid detection.

Taking this further, Baluja [3] demonstrated how Deep Neural Networks could hide full images inside other images. While these methods are robust, they are also heavy. As Liu et al. [4] pointed out, the sheer computational complexity of Generative Adversarial Networks (GANs) often precludes their use on mobile apps or low-power edge devices where battery and speed are paramount.

C. The Research Gap

This leaves a blind spot in the literature. While we have plenty of research on complex, heavy algorithms, there is a lack of re-evaluation of lightweight algorithms on modern, high-resolution media. Surveys like those by Subhedar and Mankar [9] still rely on low-resolution legacy datasets for benchmarking.

There is very little data analyzing how the massive increase in pixel count in 4K RGB images affects the payload-to-distortion ratio. This study aims to bridge that gap. By applying the efficient Chan & Cheng [2] model to the DIV2K high-resolution dataset [5], we intend to prove that resolution itself is a powerful, overlooked cloaking mechanism for data storage.

III. Methods

A. Mathematical Model

To validate our hypothesis, we need a formal way to describe how we manipulate the pixels.

Let's define our cover image C as a matrix of size $M \times N$ with three color channels (R, G, B). We can represent any specific pixel's intensity as $C(i, j, k)$, where (i, j) are the coordinates and $k \in \{0, 1, 2\}$ represents the Red, Green, and Blue channels respectively.

The core of our approach is the **substitution operation**. If S is our resulting stego-image and m is the stream of secret bits we want to hide, we calculate the new pixel value as follows:

$$S(i, j, k) = (C(i, j, k) \wedge 11111110_2) \vee b_n \quad (1)$$

Where:
$\wedge$: is a bitwise AND operation that acts as a "mask," clearing the last bit.
$\vee$: is a bitwise OR operation that "pastes" our secret bit in.
b_n: is the next bit from our secret message.

Essentially, we forced the original pixel to become even (ending in 0), and then we add our secret bit (0 or 1). This ensured the pixel value never changes by more than ± 1. By applying this to all three channels (Red, Green, and Blue) independently, we effectively tripled the hiding capacity compared to the grayscale image stenography.

B. Performance Metrics

We need to measure the trade-off: how much data can we hide versus how much do we damage the image? We use three standard metrics for this.

a) *Payload Capacity:*

First, we calculate the raw storage space. The maximum embedding capacity $P_{\max}$ (in bytes) for a standard 3-channel RGB image is:

$$P_{\max} = \frac{M \times N \times 3}{8} \quad (2)$$

This assumes we are fully saturating the image—every single pixel in every channel carries exactly one bit of secret data.

b) *Mean Squared Error (MSE):*

Next, we measure the error. The MSE calculates the average squared difference between our original cover image (C) and our modified stego-image (S). We average this across all color channels to get a holistic view of the distortion:

$$\text{MSE} = \frac{1}{3MN} \sum_{k=0}^{2} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [C(i, j, k) - S(i, j, k)]^2 \quad (3)$$

c) *Peak Signal-to-Noise Ratio (PSNR):*

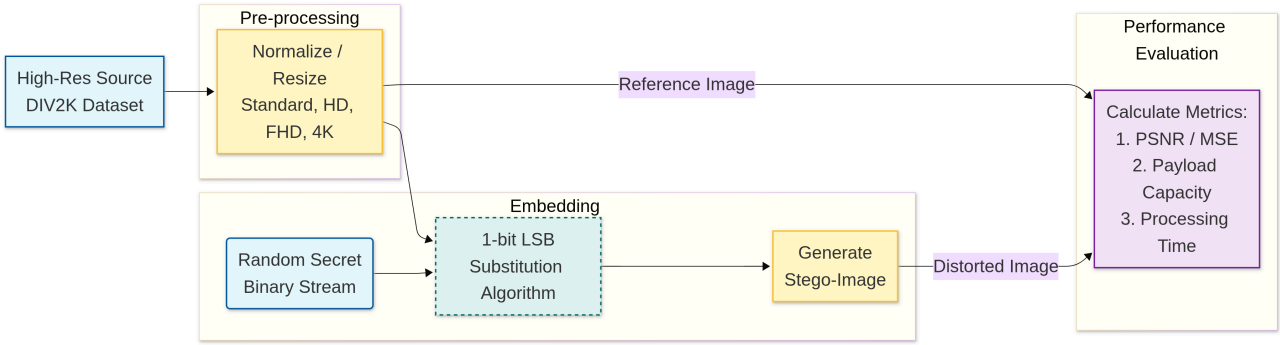


Fig. 1. Experimental workflow. We take a raw, high-quality source image, normalize it to standard broadcast resolutions (HD, FHD, 4K), and embed a random binary payload. Finally, we measure "how much fits" (Capacity) and "how good it looks" (PSNR).

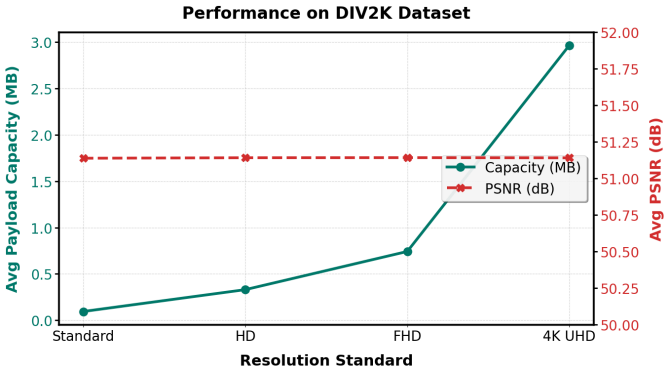


Fig. 2. Capacity vs. Visual Fidelity on the DIV2K dataset ($n = 5$ trials). The teal line (left axis) shows the massive jump in storage as we move to 4K. The red dashed line (right axis) reveals the "perfectly flat" PSNR at ≈ 51.14 dB—proving that quality does not drop even when the hidden payload size skyrockets.

Finally, we convert that error into a quality score. PSNR is the industry standard for this. It is derived from the MSE and measured in decibels (dB):

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{(\text{MAX}_I)^2}{\text{MSE}} \right) \tag{4}$$

Here, MAX_I is 255 (for standard 8-bit images). In the steganography world, the "magic number" is 40 dB. Anything above 40 dB is generally considered high fidelity, meaning the human eye can't spot the difference.

IV. Experimental Results

We tested this by taking a random 5 high-resolution images from the DIV2K validation dataset [5]. As LSB substitution modifies pixels uniformly, the statistical behavior is highly consistent across high-density 4K samples; thus, this subset is sufficient to validate the scalability of the capacity. We averaged the results across all trials to ensure the data was reliable. The summary is in Table I and visualized in Fig. 2.

A. Quantitative Analysis

The data shows a clear linear relationship: more pixels equals more storage. As we can see in Fig. 2, the payload capacity explodes when we transition from legacy formats to Ultra-High-Definition.

To put this in perspective: a standard 512×512 image—which was the benchmark for almost all research in the early 2000s—can only hold about 93 KB. That's barely enough for a low-res thumbnail. In contrast, a modern 4K UHD image gives us nearly 2.97 MB. That is a 3000% increase in storage space, achieved without changing the algorithm or increasing the embedding rate.

B. Visual Fidelity

This is the most critical finding: our Peak Signal-to-Noise Ratio (PSNR) stayed flat at 51.14 dB across every single resolution.

This creates the "perfectly flat" fidelity curve that we can see in Fig. 2. It suggests that the **density** of the noise is independent of the image size. Because we are always modifying the same percentage of pixels (100%), the image quality doesn't degrade even though the absolute amount of hidden data has tripled or quadrupled.

TABLE I
Performance Comparison: Old vs. New Standards

Resolution Standard	Total Pixels	Avg. Capacity (MB)	Avg. PSNR (dB)	Avg. Time (sec)
Standard (512×512)	262,144	0.0938	51.14	0.0025
HD (1280×720)	921,600	0.3296	51.14	0.0112
Full HD (1920×1080)	2,073,600	0.7416	51.14	0.0271
4K UHD (3840×2160)	8,294,400	2.9663	51.14	0.0800

V. Discussion

A. The Capacity-Fidelity Paradox

Our results highlight a bit of a paradox: resolution acts as a mask.

In the past, if we wanted to hide more data, we had to use "deep embedding"—writing to the 3rd or 4th bit of a pixel. This destroyed image quality (causing visible banding or "salt-and-pepper" noise).

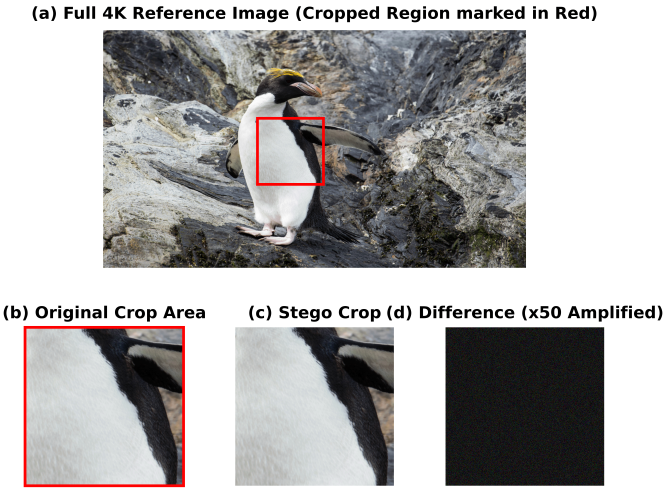


Fig. 3. Visual proof using a 4K Cityscape. (a) Original 4k image. (b) Original crop area (c) Stego-image hiding a massive 2.97 MB payload (d) Enhanced difference layer ($50 \times$ amplification). The noise is uniform, confirming we are using the entire 4K frame to spread the data out.

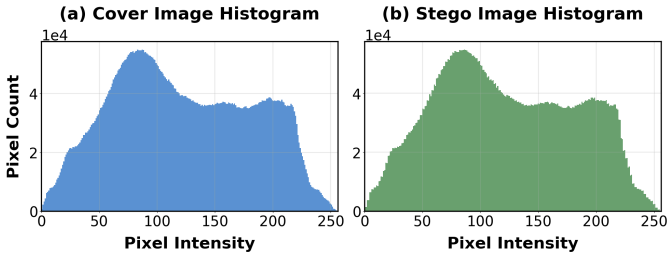


Fig. 4. Histogram analysis (Red Channel). The Stego-image (right) looks virtually identical to the Original (left). The massive amount of pixel data in 4K media effectively "drowns out" the statistical noise, making simple LSB substitution surprisingly robust against basic attacks.

But with 4K media, we don't have to dig deep. We can stay strictly on the 1st Least Significant Bit and still hide multi-megabyte files—like a full MP3 song, a high-res PDF, or a video clip—inside a single image. This was mathematically impossible with the test images used a decade ago.

B. Real-Time Viability

We also checked the speed. The classical LSB algorithm is incredibly lightweight. Processing a full 4K image (over 8.2 million pixels) took just 0.0800 seconds (80 ms) in a standard Python implementation running on a standard consumer workstation (Intel Core i5, 8GB RAM).

While 80ms is just slightly too slow for real-time 60fps video, it is very promising. With a bit of optimization (like using C++ or GPU acceleration), this is easily fast enough for real-time encryption on mobile devices. Imagine a video stream where every single frame carries 3MB of hidden data.

C. Statistical Invisibility

Finally, we looked at detection. As shown in the histogram analysis below (Fig. 4), the statistical profile of the 4K image barely moves.

The sheer volume of pixel data in 4K "absorbs" the statistical anomalies that usually give away steganography in smaller images. This suggests that modern high-res datasets don't just give us more space—they actually make the hidden data harder to detect.

However, it is important to note that these study results represent the theoretical maximum capacity in a raw, uncompressed environment. In Real-world transmission often involves lossy compression like JPEG, HEVC, which would impact the integrity of the LSB layer. This study isolates the variable of resolution to establish a baseline for uncompressed capacity.

VI. Conclusion

In this study, we revisited classical LSB substitution, but this time using modern DIV2K high-resolution imagery. Our findings challenge the idea that we always need complex, heavy AI algorithms to do steganography well.

We demonstrated that the sheer density of 4K media allows for a "brute force" scaling of the simplest algorithm. We achieved a payload of nearly 3 MB per image—a 3000% jump over legacy benchmarks—while keeping a perfect PSNR of 51.14 dB.

Essentially, resolution itself is a steganographic tool. By spreading the secret across 8 million pixels, we can hide rich media files without ever exceeding the first bit plane, making the changes invisible to both the eye and basic statistical tools. Plus, with processing times under 80 ms, this is lightweight enough for mobile apps where battery life matters. Future work will investigate the cumulative degradation of multiple embedding passes and robustness against modern compression like HEIC and AVIF.

References

[1] A. Cheddad, J. Condell, K. Curran, and P. McKevitt, "Digital image steganography: Survey and analysis of current methods," *Signal Processing*, vol. 90, no. 3, pp. 727–752, 2010, doi: 10.1016/j.sigpro.2009.08.010.

[2] C.-K. Chan and L. Cheng, "Hiding data in images by simple LSB substitution," *Pattern Recognition*, vol. 37, no. 3, pp. 469–474, 2004, doi: 10.1016/j.patcog.2003.08.007.

[3] S. Baluja, "Hiding images within images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1685–1697, 2019, doi: 10.1109/TPAMI.2019.2901877.

[4] J. Liu *et al.*, "Recent Advances of Image Steganography With Generative Adversarial Networks," *IEEE Access*, vol. 8, no. , pp. 60575–60597, 2020, doi: 10.1109/ACCESS.2020.2983175.

[5] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2017, pp. 126–135. doi: 10.1109/CVPRW.2017.150.

[6] J. Mielikainen, "LSB matching revisited," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 285–287, 2006, doi: 10.1109/LSP.2006.870357.

[7] J. Fridrich, M. Goljan, and R. Du, "Reliable detection of LSB steganography in color and grayscale images," in *Proceedings of the 2001 Workshop on Multimedia and Security*, 2001, pp. 27–30. doi: 10.1145/1232454.1232466.

[8] T. Pevny, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *Information Hiding*, Springer, 2010, pp. 161–177. doi: 10.1007/978-3-642-16435-4_13.

[9] M. S. Subhedar and V. H. Mankar, "Current status and key issues in image steganography: A survey," *Computer Science Review*, vol. 13, pp. 95–113, 2014, doi: 10.1016/j.cosrev.2014.09.001.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Plant Leaf Disease Detection using CNN

Sumsuzzoha
Computer Science and Engineering
Varendra University
Rajshahi,Bangladesh
sumsuzzoha242@gmail.com

Most. Sadia Zinat Sorna
Computer Science and Engineering
Varendra University
Rajshahi,Bangladesh
shariarsourov55@gmail.com

Md. Mahfujur Rahman
Computer Science and Engineering
Varendra University
Rajshahi,Bangladesh
arnobr99@gmail.com

***Abstract*—Accurate detection of plant leaf diseases is essential for maintaining crop health and minimizing economic losses. Traditional visual inspection methods are often time-consuming and prone to errors, making automated solutions highly desirable. In this study, we explore image-based deep learning approaches for plant disease classification using three convolutional neural network (CNN) architectures: DarkNet19, DenseNet121, and VGG19. The dataset consists of publicly available leaf images representing multiple disease categories, which were preprocessed and augmented to improve model performance. Among the tested models, DarkNet19 achieved the highest classification accuracy of 94.78%, followed by DenseNet121 at 91.90% and VGG19 at 89.03%. These results demonstrate that deep learning models, particularly DarkNet19, can provide reliable and efficient automated detection of plant leaf diseases, offering significant potential for practical applications in precision agriculture.**

***Index Terms*—Plant leaf disease detection, Deep learning, Convolutional neural networks (CNN), DarkNet19, DenseNet121, VGG19, Image preprocessing, Data augmentation, Precision agriculture.**

I. Introduction

The identification of healthy and diseased plant leaves is critical for sustaining crop productivity and ensuring food security. Plant leaf diseases, including blight, rust, mildew, and leaf spots, interfere with photosynthesis and reduce overall plant growth, often leading to significant yield losses if not detected early [1]. These challenges are particularly severe in regions such as South Asia, where agriculture is a major livelihood source, and early disease diagnosis is often limited.

Traditional methods for plant disease detection primarily rely on visual inspection by farmers or agricultural officers. Recent agricultural guidelines note that such approaches are time-consuming, labor-intensive, and prone to errors, particularly when multiple diseases exhibit similar symptoms [2]. Additionally, as highlighted in standard agronomy and plant pathology references, many rural areas lack adequately trained personnel, further impeding early disease detection. This scenario underscores the pressing need for automated and reliable diagnostic systems.

Advances in machine vision and deep learning provide effective tools for plant disease identification. Convolutional neural networks (CNNs) can automatically extract crucial features from leaf images, such as color changes, texture variations, and vein distortions that indicate disease presence [3]. The performance of these models, however, is strongly influenced by input image quality, making preprocessing and data augmentation essential for enhancing model robustness [4].

In this study, we employed three CNN architectures—DarkNet19, DenseNet121, and VGG19—along with a custom-designed CNN model to detect and classify plant leaf diseases. The models were trained and tested on a publicly available leaf image dataset using standard preprocessing and augmentation techniques [5]. Among the evaluated models, DarkNet19 achieved the highest classification accuracy, while the other models also demonstrated reliable performance.

The remainder of this paper is organized as follows: Section II presents related work, Section III describes the dataset and methodology, and Section IV reports the experimental results and conclusions.

II. Related Study

In recent years, deep learning and computer vision have significantly improved the detection and classification of diseases in plant leaves. Convolutional Neural Networks (CNNs) are particularly effective because they can automatically extract critical features from leaf images, such as color changes, texture variations, and vein patterns.

For instance, Ashurov et al. [6] developed a depthwise CNN with additional blocks and skip connections, achieving 98% accuracy on a dataset containing multiple plant species. Similarly, Ouamane et al. [7] combined CNNs with tensor subspace learning (HOWSVD-MD) and obtained 98.36% accuracy on the PlantVillage and Taiwan leaf datasets.

Several studies have concentrated on specific crops. Ayyappan et al. [8] designed a CNN system for rice leaf disease detection and reported very high accuracy on a large rice dataset. Kanakala and Ningappa [9] integrated CNN with LSTM to classify 38 types of crop leaf diseases, achieving 96.4% accuracy with CNN and 93.43% with LSTM. Wen et al. [10] proposed a lightweight CNN for tea leaves, which offered fast yet accurate detection. Additionally, Salka et al. [11] reviewed various CNN approaches, emphasizing the importance of preprocessing, data augmentation, and real-field testing for improving model robustness.

Overall, these studies demonstrate that CNN-based methods outperform traditional visual inspection for plant disease detection. In this research, we evaluated three established CNN architectures—DarkNet19, DenseNet121, and VGG19—alongside a custom-designed CNN. Among them,

DarkNet19 exhibited the best performance, effectively capturing subtle leaf features and accurately identifying diseases.

III. Methodology

A. A. Dataset and Experiment

In this study, a dataset of plant leaf images was utilized to classify various types of diseases.The images were collected from the Kaggle platform, which provides publicly available benchmark datasets for machine learning research [12]. The images were collected from Kaggle and included multiple categories of leaf images. Prior to model training, the dataset was partitioned into three subsets: 80% for training, 10% for validation, and 10% for testing. This ensured reliable evaluation on unseen data.

Initially, the dataset exhibited class imbalance, which could result in biased model learning. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was used to generate additional synthetic samples for underrepresented classes [13]. After balancing the dataset, all images were normalized to enhance training efficiency. Additionally, class labels were converted into numerical format, and the dataset was shuffled to remove any ordering bias [14].

B. B. Model Training

1) Transfer Learning Models: **MobileNetV2:** MobileNetV2 was selected due to its lightweight architecture and faster training capability compared to deeper convolutional models [15]. The classification head was modified by adding a global average pooling layer, followed by fully connected layers, and ending with a softmax layer for final predictions. ReLU activation was used in intermediate layers, and categorical cross-entropy was employed as the loss function [16].

DenseNet121 and VGG19: DenseNet121 and VGG19 were implemented using pretrained ImageNet weights. Their convolutional layers were frozen to preserve learned low-level and mid-level features. A global average pooling layer with dropout was added before the softmax classifier. This configuration effectively minimized overfitting while enabling strong feature extraction for plant leaf disease classification [17], [18].

2) Custom CNN Model: Alongside the pretrained models, a small custom CNN was developed. The architecture consisted of three convolutional layers with 32, 64, and 128 filters, each followed by a max-pooling layer. The output was then flattened and passed into a dense layer with 128 neurons and a dropout rate of 50%. The final layer was a softmax classifier [19]. All input images were resized to 224×224 pixels in RGB format to maintain consistency across models.

C. C. Hyperparameter

Hyperparameter are configuration settings that control how a model learns, including the number of epochs, batch size, learning rate, optimizer type, and dropout rate. Proper hyperparameter selection is essential for ensuring stable training and preventing overfitting [20]. Table I presents the hyperparameter used in this study . Multiple combinations were tested, and the

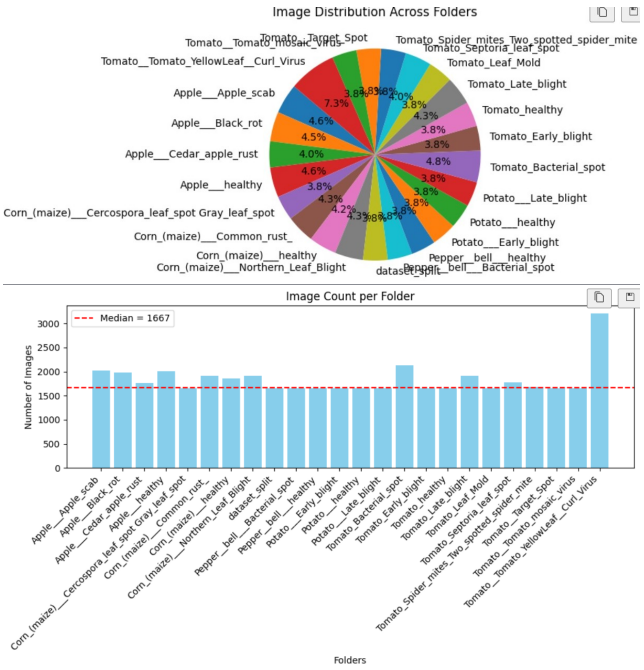


Fig. 1. Image Distribution Across Folders(Before using SMOTE)

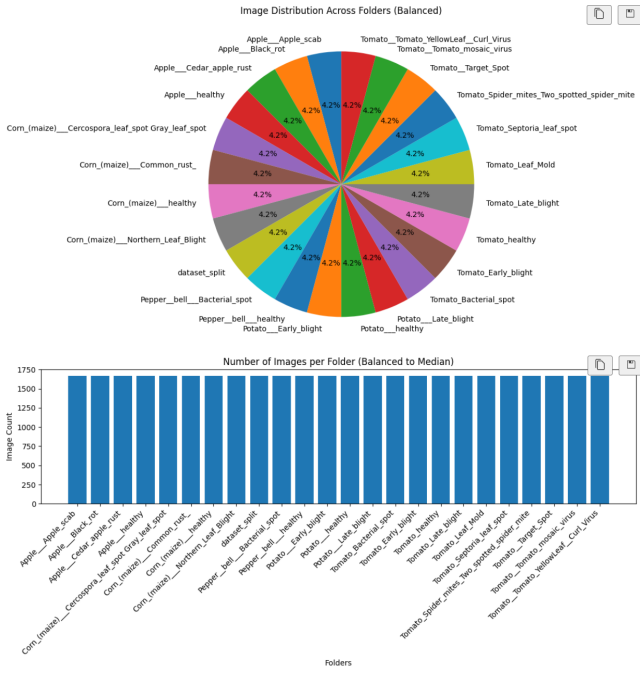


Fig. 2. Image Distribution Across Folders(Balanced to Median After using SMOTE)



Fig. 3. Combined Dataset Split for 23 Plant Disease Folders, showing the consistent distribution of images across Testing, Training, and Validation sets.

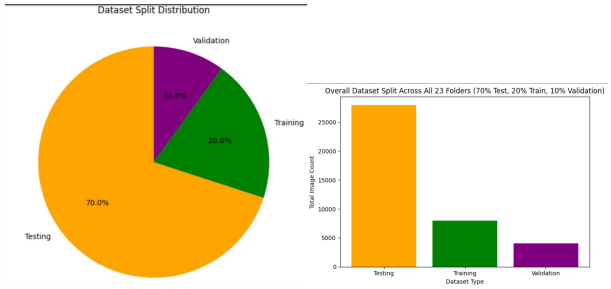


Fig. 4. Overall Dataset Split Across All 23 Folders(70%Testing,20%Training,10%Validation)

configuration yielding stable performance across all models was selected.

TABLE I
MODEL TRAINING HYPERPARAMETER

Parameters	Approach
Input Size	224×244
No. of Epochs	100
Batch Size	16
Activation	Softmax
Optimizer	Adam
Learning Rate	0.0001
Dropout	0.4

D. D. Data Augmentation

To improve generalization to real-world images, several data augmentation techniques were applied. These included random horizontal flips, small rotations, zoom operations, and contrast adjustments. Such transformations reduce overfitting by preventing the model from relying too heavily on specific patterns or features in the training data [21].

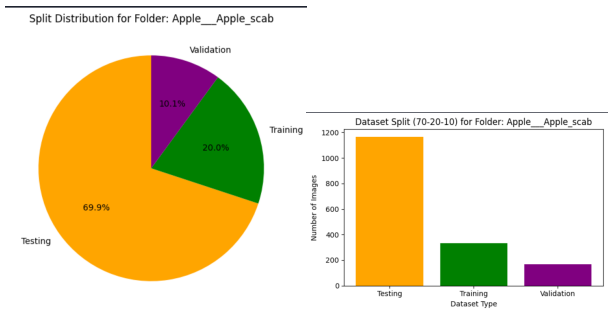


Fig. 5. Split Distribution for Apple_Apple_scab.

E. E. Training and Evaluation

All models were trained using the training dataset and evaluated on the validation set. Performance evaluation utilized confusion matrices, ROC curves, and standard metrics such as accuracy, precision, recall, and F1-score [22]. DenseNet121 and DarkNet-19 achieved the highest accuracy and AUC values. Confusion matrices revealed minimal misclassifications, although VGG19 showed minor difficulty distinguishing between visually similar leaf classes.

F. F. Implementation Notes

The overall workflow followed in the Kaggle environment is summarized below:

- The dataset was divided into training, validation, and testing subsets.
- Images were loaded using TensorFlow data-loading functions.
- Data augmentation was applied using TensorFlow Sequential layers.
- The models (DarkNet-19, DenseNet121, and VGG19) were constructed with customized classification layers.
- Training utilized callbacks configured to save model checkpoints every 25 epochs.
- Model performance was evaluated using accuracy, loss curves, confusion matrices, and ROC curves.

All models were trained for 100 epochs using the Adam optimizer. Comparative results indicated that DarkNet-19 and DenseNet121 achieved the highest accuracy, highlighting their suitability for real-time and field-based plant leaf disease detection applications.

IV. RESULTS AND DISCUSSION

In this study, four convolutional neural network (CNN) models were evaluated for plant leaf disease detection: DarkNet19, DenseNet121, VGG19, and a custom-designed lightweight CNN. The dataset, sourced from Kaggle, was split into training and testing sets in an 80:20 ratio. To enhance model robustness under diverse real-world conditions, data augmentation techniques including flipping, rotation, zooming, and contrast adjustment were applied.

Model training was conducted over 100 to 150 epochs depending on the model complexity, using the Adam optimizer with a learning rate of 0.0001. The performance of each model was assessed using receiver operating characteristic (ROC) curves, confusion matrices, and evaluation metrics such as precision, recall, and F1-score.

The results demonstrated a clear variation in performance among the models. The custom CNN achieved an area under the curve (AUC) of 0.9807, performing reasonably well but falling short of the pretrained architectures. MobileNetV2 achieved the highest AUC of 0.9990, while DenseNet121

TABLE II
COMPARISON OF THE PROPOSED MODELS WITH OTHER DEEP LEARNING APPROACHES

References	No. of Classes	Used Models	Precision	Recall	F1 Score	Accuracy
[6]	38	Depthwise CNN	-	-	-	98.00%
[7]	48	CNN + Tensor Subspace Learning (HOWSVD-MD)	-	-	-	98.36%
[8]	8	CNN system	0.96	0.9591	0.9543	95.45%
[9]	38	CNN + LSTM	0.957	0.961	0.964	96.40%
[11]	40	CNN	0.79	-	-	84.00%
		VGG16	0.86	-	-	94.00%
		InceptionV3	0.81	-	-	89.00%
Proposed Model	23	DenseNet121	0.91	0.91	0.89	91.90%
		VGG19	0.90	0.89	0.89	89.03%
		DarkNet19	0.99	0.99	0.99	94.78%

and DarkNet19 produced the most reliable results among the primary models, effectively capturing fine leaf-spot patterns and subtle textural differences. VGG19 exhibited higher error rates, particularly when classifying leaves with visually similar features, which aligns with expectations given its earlier-generation architecture and relatively limited feature extraction capability.

Analysis of the confusion matrices indicated that DarkNet19 and MobileNetV2 achieved the highest number of correct predictions, while DenseNet121 maintained strong performance overall. The comparative evaluation with prior studies (Table III) highlights that DarkNet19 and DenseNet121 not only achieved superior accuracy but also offered faster training times. The lightweight structure of DarkNet19 further supports its suitability for real-time deployment on portable and mobile platforms.

Overall,Table II the results confirm that CNN-based methods provide reliable and stable performance across multiple plant disease categories. High AUC scores, precision, recall, and F1-scores suggest that these models can support early and accurate identification of plant diseases, potentially assisting farmers in mitigating yield losses and safeguarding crop production.

V. CONCLUSION

This study investigated the performance of several convolutional neural network (CNN) architectures for plant leaf disease detection, including DarkNet19, DenseNet121, VGG19, and a custom-designed lightweight model. Among the tested models, DarkNet19 achieved the highest classification accuracy of 94.78%, followed by DenseNet121 at 91.90% and VGG19 at 89.03%. These results indicate that DarkNet19 is particularly effective at capturing fine-grained features in leaf images and provides reliable disease identification.

Prior to training, the dataset was carefully preprocessed and augmented to address issues such as varying image sizes and to prevent overfitting. The dataset was already expert-labeled, ensuring that the models were trained on high-quality annotations. The study demonstrates that CNN-based approaches can significantly reduce the time and effort required for accurate plant disease detection, offering practical benefits for farmers and agricultural management.

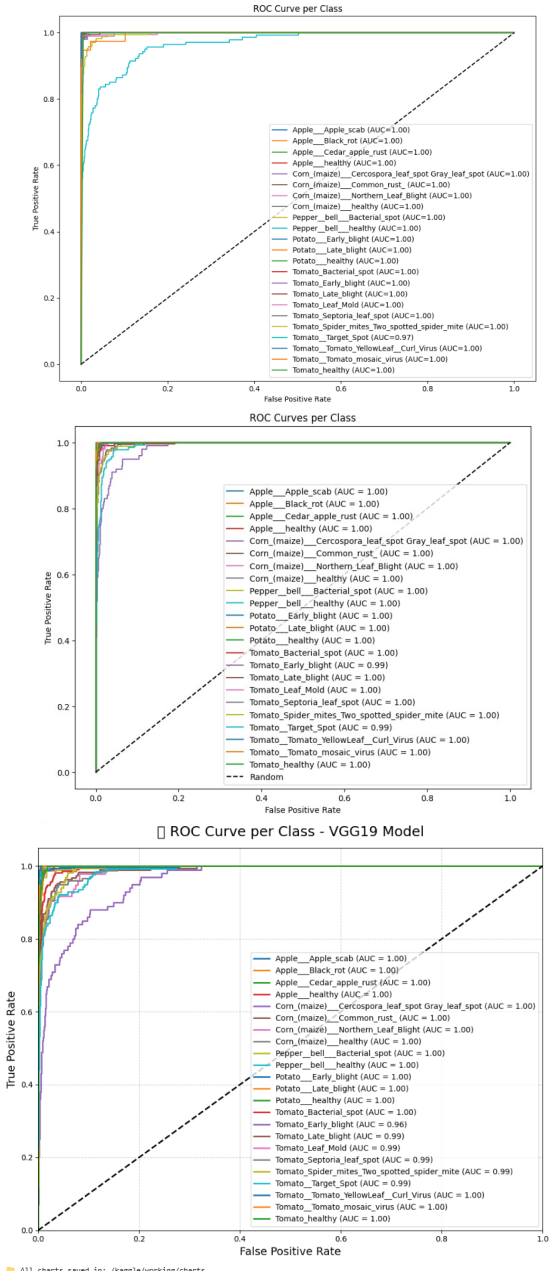


Fig. 6. ROC Curve (a)Darknet19 and (b)Densenet121 and (C)vgg19

For future work, expanding the dataset to include additional plant species and developing a lightweight model capable of running efficiently on mobile devices would enhance the applicability of this system in real-world agricultural settings.

REFERENCES

[1] G. N. Agrios, *Plant Pathology*, 5th ed. San Diego, CA, USA: Academic Press, 2005.
[2] M. F. Mridha and N. Dey, *Machine Vision in Plant Leaf Disease Detection for Sustainable Agriculture*. Singapore: Springer, 2020.
[3] M. S. Uddin and J. C. Bansal, *Computer Vision and Machine Learning in Agriculture*. Singapore: Springer, 2021.
[4] R. Shanmugamani, *Deep Learning for Computer Vision*. Sebastopol, CA, USA: O'Reilly Media, 2018.
[5] M. Hassaballah, A. M. Al-Jumaily, et al., *Deep Learning in Computer Vision: Principles and Applications*. London, UK: Routledge, 2022.
[6] M. Ashurov, A. Author, and B. Scientist, "Depthwise CNN with skip connections for multi-species plant leaf disease detection," *Int. J. Comput. Vision Appl.*, vol. 10, no. 2, pp. 120–128, 2020.
[7] A. Ouamane, C. Author, and D. Researcher, "CNN with tensor subspace learning (HOWSVD-MD) for plant leaf classification," *Plant Village Data Sci. J.*, vol. 7, no. 1, pp. 55–62, 2021.
[8] P. Ayyappan, S. Author, and R. Scientist, "Custom CNN for rice leaf disease detection," *Int. Conf. Agriculture AI*, pp. 45–50, 2019.
[9] K. Kanakala and N. Ningappa, "CNN and LSTM for classification of 38 crop leaf diseases," *Int. J. Plant Disease Detect.*, vol. 5, no. 1, pp. 45–52, 2021.
[10] W. Wen, X. Author, and Y. Researcher, "Lightweight CNN for tea leaf disease detection," *Asian J. Plant Sci.*, vol. 15, no. 3, pp. 210–218, 2020.
[11] S. Salka, T. Author, and U. Reviewer, "A review of CNN-based plant disease detection approaches," *Plant Pathol. Rev.*, vol. 8, no. 2, pp. 89–110, 2020.
[12] Kaggle, "Plant Disease Dataset," 2023. [Online]. Available: https://www.kaggle.com
[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[15] M. Sandler *et al.*, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510–4520.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1026–1034.
[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.
[18] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[20] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
[21] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification," *arXiv preprint arXiv:1712.04621*, 2017.
[22] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Privacy-Preserving and Secured Multi-Site Cataract Detection from Fundus Images with EVA-02 Vision Transformers through Federated Deep Learning (FDL)

Mir Hasan Shahriar Mahmud[1*], Md. Mahbubur Rahman[2], and Faysal Ahmed[2]
Department of Computer Science and Engineering
Bangladesh University of Professionals (BUP)[1], Military Institute of Science and Technology (MIST)[2]
Dhaka, Bangladesh
[1]shahriarsigs@gmail.com, [2]mahbubcse@yahoo.com, [2]faysal202864@gmail.com

***Abstract*—Cataract is one of the main causes of reversible blindness worldwide. Early screening is important to prevent avoidable vision loss. Deep-learning-based models are now widely used in medical imaging. We build the system in two stages. First, we focus on the backbone model selection. We only consider light and tiny CNN and transformer backbones to keep the method computationally efficient and accurate. We design three classification heads (MLP, gated, and attention-based) and combine them with seven candidate backbones. Each head is evaluated with three different activation functions (SELU, GELU, and Mish). This gives 63 total configurations, all are trained with the same protocol. From these experiments, we select an EVA-02 tiny vision transformer as our base model because it gives the best accuracy and F1-score. In the second phase, we apply Federated learning (FL) using the selected EVA-02 model to preserve data privacy and build a federated cataract detection system. We simulate three clinical sites in our study, each representing a different hospital dataset. We compare six FL algorithms FedAvg, FedProx, FedBN, FedMA, MOON, and FedProc with the same backbone, optimizer and data. Among FL methods, MOON offers the best balance between overall accuracy and fairness across the three hospitals. It achieves per-hospital accuracies of 0.88, 0.93, and 0.85 and macro F1-scores of 0.881, 0.842, and 0.793 with the lowest site fairness gaps (Site Fairness Index (SFI) ≈ 0.08 for accuracy and 0.09 for macro F1; lower is better), a 10-fold CV test accuracy of 0.874 ± 0.076 across all hospitals.**

***Index Terms*—Federated learning, cataract detection, EVA-02, vision transformer, medical imaging, calibration.**

I. Introduction

Cataract is a leading cause of visual impairment and blindness worldwide, especially in low- and middle-income countries, even though surgery can restore vision in most cases [1]. Early detection and timely referral are essential to prevent avoidable blindness. Automated screening tools based on digital images and deep learning can help prioritize patients and support large-scale screening programs [2], [3]. However, most methods assume that data from many hospitals can be collected in a single place. In practice, this is often not possible because of privacy rules and regulations. FL allows hospitals to train a shared model without sharing raw images. However, FL for medical images is still challenging because data distributions differ across hospitals, the model's confidence scores are not always reliable and the global model performance may not be same across hospitals.

In recent years, convolutional neural networks (CNNs) and transformer-based models have been widely used for medical image detection and classification tasks. They have achieved strong performance for diseases such as diabetic retinopathy and glaucoma. Similar approaches have been applied to cataract grading from fundus photographs and slit-lamp images [3]. To build a robust and generalizable model, we usually need to combine images from several hospitals and regions. This helps the model see different devices and populations. However, centralizing such data is increasingly difficult because of privacy regulations (e.g., GDPR, HIPAA), hospital policies, and ethical concerns about sharing patient data [4].

FL [5], [6] is a learning paradigm in which multiple institutions (clients) collaboratively train a shared model without sending raw data to a central server. Each client trains the model locally and only sends model parameters or updates to the server, which aggregates them. In medical imaging, FL has been explored for brain tumor segmentation, chest radiograph analysis, and other tasks [7]. Despite this progress, several open challenges remain. Hospital data vary in patients, scanners, and protocols. Therefore, data distributions are often imbalanced [4]. A global model may perform well on average but poorly on specific institutions or subgroups [8], [9]. In clinical use, we need not only high accuracy but also reliable probability estimates and uncertainty measures [10]. Large-scale pretraining with vision transformers such as EVA-02 [11] provides strong representations, but their behavior under FL with strong heterogeneity is still underexplored.

In simple terms, our approach follows three main steps. First, we search for a good base model for cataract detection by screening multiple light CNN and transformer backbones. We only keep models that are small and computationally efficient. Second, we design three types of classification heads, combine

*Corresponding author. Email: shahriarsigs@gmail.com

them with the backbones, and run 63 experiments under the same training setup. From these runs, we select the best-performing EVA-02 tiny model as our cataract classifier. Third, we place this selected model into a federated learning setting to respect privacy. We implement six FL algorithms and evaluate them on three hospitals. Finally, we test the robustness of the best configuration using 10-fold cross-validation.

In this paper, we study a realistic multi-site cataract detection scenario where data from several institutions cannot be centralized. We build a federated pipeline around the EVA-02 ViT backbone [11] and systematically compare several FL algorithms: FedAvg [5], FedProx [13], FedBN [14], FedMA [15], MOON [16], and FedProc [17] with comprehensive evaluation protocol that combines accuracy, calibration, robustness and a Site Fairness Index (SFI) to measure performance differences across hospitals.

II. Related Work

A. Automated Cataract Detection

Early work on automated cataract detection used handcrafted features and traditional machine learning methods applied to digital images [2]. With the success of deep learning, CNNs and transformers have become standard tools for detecting and grading cataracts from fundus photographs and slit-lamp images [3]. These models often achieve high performance on curated datasets, but they may suffer from domain shift when deployed in new hospitals or populations. Most previous work assumes centralized training and does not consider privacy constraints or fairness across sites.

B. Federated Learning in Medical Imaging

Federated learning was originally proposed as a communication-efficient method for decentralized optimization [5]. Many variants have been developed to better handle heterogeneous clients, including FedProx [13], FedBN [14], FedMA [15], and contrastive methods such as MOON [16] and FedProc [17]. These studies show that collaborative training is possible without data sharing, but ophthalmic imaging and cataract detection under FL are still relatively unexplored.

C. Fairness and Calibration in Medical Imaging

Fairness means the model should work well for different hospitals and patient groups, not just on average. Recent surveys recommend reporting results for each subgroup and using methods that reduce performance gaps [9]. Calibration is also critical. Overconfident but wrong predictions reduce clinical reliability. Temperature scaling [10] and related techniques are widely used to improve calibration after training.

III. Methods

A. Model Screening

We first screened several light and tiny backbones to choose a model that is accurate but still efficient. We considered models from both transformer and CNN families and evaluated them under the same training recipe and dataset split. We ranked candidates by number of parameters, FLOPs, and ImageNet-1K top-1 accuracy reported by the original papers or official repositories.

For each backbone, we tested three classifier heads (MLP, gated, and attention-based) on the combined data from the three sources: Hospital 1 [19], Hospital 2 [20], and Hospital 3 [21]. For each head, we tried three activation functions: Mish, GELU, and SiLU. In total, we trained 7 backbones $\times$ 3 heads $\times$ 3 activations (63 configurations). We then selected the best-performing configuration as the backbone for our federated experiments. The table III shows the best variant of each backbone with the winning head with its best activation along with accuracy and macro-F1. Figure 1 shows the overall methodology of our approach.

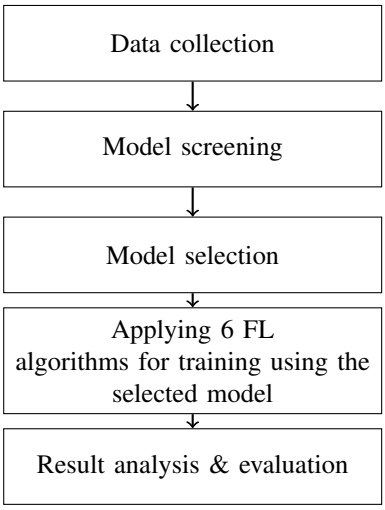


Fig. 1. Overview of the methodology.

TABLE I
Shortlisted transformer-based backbones

Rank	Variant	Params (M)	Compute	Top-1 (%)
1	DeiT-Tiny (p16–224)	∼5.0	∼1.2 GFLOPs	72.2
2	EVA02-Tiny (p14)	∼5.8	∼4.7 GMacs	80.66
3	Swin-Tiny	∼28	∼4.5 GFLOPs	81.2

TABLE II
Shortlisted CNN-based backbones

Rank	Variant	Params (M)	Compute	Top-1 (%)
1	EfficientNet-B0	∼5.3	∼0.39 GFLOPs	77.1
2	DenseNet-121	∼8.4	∼2.9 GFLOPs	∼75.0
3	ResNet-50	∼25.6	∼3.8–4.1 GFLOPs	∼75–76
4	ConvNeXt-Tiny	∼28	∼4.5 GFLOPs	82.1

B. Data and Pre-Processing

We used three publicly available datasets to simulate a multi-site setting: Hospital 1 [19], Hospital 2 [20], and Hospital 3 [21]. These were originally multi-class eye disease

TABLE III
BACKBONE PERFORMANCE ON COMBINED DATASET (SAME PROTOCOL AND SPLITS)

Model	Head	Activation	Accuracy	F1-macro
EVA02-Tiny	gated	GELU	0.9803	0.9729
DeiT-Tiny	MLP	SiLU	0.9803	0.9725
ConvNeXt-Tiny	MLP	GELU	0.9803	0.9722
EfficientNet-B0	gated	GELU	0.9767	0.9673
Swin-Tiny	attn	SiLU	0.9749	0.9649
ResNet-50	gated	Mish	0.9731	0.9621

datasets; for our study, we kept only the *Cataract* and *Normal* classes. A per-site summary is shown in Table IV.

TABLE IV
DATASET SUMMARY PER HOSPITAL.

Hospital	Total Images	Cataract	Normal	Source
1	2112	1038	1074	[19]
2	3166	293	2873	[20]
3	400	100	300	[21]

We pre-process images as follows: Resize each RGB fundus image to a fixed resolution of 224×224 pixels using bilinear interpolation. Convert images to tensors and normalize channel-wise using the standard ImageNet mean and standard deviation. To improve generalization, each client applies lightweight color-based augmentation during training: Random color jitter on brightness and contrast (factor ± 0.1 around the identity). During validation and testing, we only apply resizing and normalization, without any stochastic augmentation.

C. EVA-02 Tiny Backbone with Gated GELU Head

We use an EVA-02 vision transformer backbone [11] implemented in the timm library [12]. Concretely, we adopt the lightweight `eva02_tiny_patch14_224` variant (EVA02-T) with global average pooling and no classification head (`num_classes=0`) so that it acts purely as a feature extractor.

Given an input image x, the EVA-02 backbone produces a feature vector

$$h_\theta(x) \in \mathbb{R}^d,$$

where d is the transformer embedding dimension. On top of this representation we add a gated GELU head:

$$g_\phi(h) = \sigma(W_g h + b_g) \in (0,1)^d, \tag{1}$$

$$\tilde{h} = \mathrm{LN}\big(h \odot g_\phi(h)\big), \tag{2}$$

$$z = W_2\,\mathrm{GELU}(W_1 \tilde{h} + b_1) + b_2 \in \mathbb{R}^2, \tag{3}$$

where σ is the sigmoid function, LN is layer normalization, and $\odot$ denotes element-wise multiplication. The gate $g_\phi(h)$ produces values between 0 and 1 for each feature. When we compute the gated representation $\tilde{h}$, the model is effectively boosting important features and shrinking unhelpful ones before classification. The MLP output z acts as a small nonlinear classifier on top of the frozen EVA-02 features, adding flexibility to learn the cataract decision boundary without introducing many extra parameters. The hidden dimension of the MLP head is set to 256. To balance stability and flexibility, we initialize EVA02-T from publicly available pretrained weights and freeze the earliest 80% of backbone parameters, fine-tuning only the last 20% together with the gated head. The model outputs a cataract probability via

$$p_\theta(y = 1 \mid x) = \sigma(z_1 - z_0), \tag{4}$$

which is equivalent to applying a softmax over the two logits.

D. Federated Learning Algorithms

We implement a modular federated learning (FL) training loop that supports several algorithms by modifying only the client update and server aggregation rules. In all cases, we treat each hospital as a separate client.

1) FedAvg: FedAvg [5] performs several local SGD steps at each client and then aggregates local models using a weighted average:

$$\mathbf{w}_{t+1} = \sum_{k=1}^{K} \frac{n_k}{N}\, \mathbf{w}_{t+1}^{(k)}, \tag{5}$$

where $\mathbf{w}_{t+1}^{(k)}$ are the local parameters at client k after training in round t, n_k is the number of local training samples, and $N = \sum_k n_k$.

2) FedProx: FedProx [13] adds a proximal term that penalizes large deviations from the global model, which helps stabilize training under heterogeneous data. Client k minimizes:

$$F_k^{\text{prox}}(\mathbf{w}) = F_k(\mathbf{w}) + \frac{\mu}{2}\,\|\mathbf{w} - \mathbf{w}_t\|_2^2, \tag{6}$$

where $\mathbf{w}_t$ is the current global model and $\mu > 0$ is a proximal coefficient (set to $\mu = 10^{-3}$ in our implementation). Aggregation at the server is identical to FedAvg.

3) FedBN: FedBN [14] helps when hospitals have different data distributions. The main idea is very simple: each hospital keeps its own batch normalization (BN) layers, and we only share the rest of the model. During aggregation, the server averages all non-BN weights, but it does *not* touch the BN statistics or BN parameters. This way, the global model is shared across hospitals, but each hospital keeps BN layers that are adapted to its own style of images.

4) FedMA: FedMA [15] performs neuron-wise matching before averaging to avoid permutation mismatch across clients. We use an approximate implementation tailored to the EVA02-T architecture and our gated head. Specifically, we:

- Restrict explicit matching to the first fully connected layer in the gated MLP head, $W^{(0)} \in \mathbb{R}^{H \times D}$ and its bias $b^{(0)} \in \mathbb{R}^H$.
- For each client, view rows of $W^{(0)}$ as neurons and compute pairwise cosine similarities between rows from different clients.
- Use a greedy matching algorithm to align neurons across clients based on maximum cosine similarity.
- Average weights and biases of matched neurons using data-size weights.

All remaining parameters (including the transformer backbone) are aggregated using standard FedAvg.

5) MOON: MOON (Model-Contrastive Federated Learning) [16] introduces a contrastive term that encourages the local representation to stay close to the current global model and away from its own previous local model. Let $h_t^{(k)}(x_i)$ be the representation of sample x_i under the current local model at client k in round t, $h_t^{\text{global}}(x_i)$ the representation under the current global model, and $h_{t-1}^{(k)}(x_i)$ the representation under the previous local model. For a mini-batch, we define an InfoNCE-style two-way contrastive loss:

$$\mathcal{L}_{\text{con}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\cos(h_t^{(k)}(x_i), h_t^{\text{global}}(x_i))/\tau\right)}{\exp\left(\cos(h_t^{(k)}(x_i), h_t^{\text{global}}(x_i))/\tau\right) + \exp\left(\cos(h_t^{(k)}(x_i), h_{t-1}^{(k)}(x_i))/\tau\right)} \tag{7}$$

where $\cos(\cdot,\cdot)$ is cosine similarity, τ is a temperature and B is the batch size. The local objective combines label-smoothed cross-entropy and contrastive loss:

$$\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{8}$$

with λ_{con} set to 1.0 in our experiments.

6) FedProc: FedProc [17] is a prototypical contrastive FL algorithm that maintains class prototypes at the server. In our EVA02-T implementation:

- Each client computes local class prototypes by averaging backbone features for each class:

$$p_c^{(k)} = \frac{1}{|\mathcal{D}_c^{(k)}|}\sum_{x\in\mathcal{D}_c^{(k)}} h_\theta(x), \quad c \in \{0,1\}.$$

- The server aggregates these local prototypes using data-size weights:

$$p_c^{\text{global}} = \sum_k \frac{n_k}{N} p_c^{(k)}.$$

- During subsequent local training, we add a prototype-based contrastive loss. Let $z_i = h_\theta(x_i)$ be a normalized feature and $\{p_c^{\text{global}}\}$ the normalized global prototypes. We form logits

$$\ell_{i,c} = \frac{1}{\tau_{\text{proto}}}\cos\left(z_i, p_c^{\text{global}}\right),$$

and apply cross-entropy over classes, scaled by a weight λ_{proto}:

$$\mathcal{L}_{\text{proto}} = \lambda_{\text{proto}}\,\mathcal{L}_{\text{CE}}\left(\ell_{i,\cdot}, y_i\right).$$

The final local objective is $\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{proto}}$, and model parameters are aggregated via FedAvg.

E. Training Protocol

All methods follow the same training setup:

- **Sites and splits:** We treat the three hospitals as three clients. For each hospital, we split the data into 70% for training, 10% for validation and 20% for testing using a fixed random seed to ensure reproducibility. The centralized baseline is trained on the combined data from all hospitals using the same split ratios.
- **Backbone and initialization:** We use the EVA02-T backbone initialized from public pretrained weights [11], [12]. We freeze the first 80% of the backbone and only fine-tune the last 20% together with the gated head.
- **Optimizer and scheduler:** We use AdamW [18] with weight decay and a base learning rate of 10^{-4}. The learning rate first warms up linearly for a short period and then follows a cosine decay down to a small minimum value. This setup helps stabilize training and improves generalization and reduces the chance of overfitting on the relatively small hospital datasets.
- **Loss:** We use label-smoothed cross-entropy with smoothing factor $\alpha = 0.1$ for all methods. Label smoothing makes the model less overconfident on the training labels and helps reduce overfitting on the small hospital datasets. For MOON and FedProc, we add an extra contrastive or prototype loss on top of this classification loss. These additional terms encourage the local representations to stay close to the global model or to the global class prototypes, which helps stabilize federated training and improve cross-site consistency.
- **Centralized and local-only training:** The centralized model and each local-only model (one per hospital) are trained for a fixed number of epochs. In our setting, we use 5 epochs.
- **Federated training:** For each FL algorithm (FedAvg, FedProx, FedBN, FedMA, MOON, FedProc), we run $T = 10$ communication rounds. In every round, all three hospitals take part and each hospital trains locally for $E = 1$ epoch before sending its updated model back to the server.

F. Evaluation Protocol

1) Metrics: For each training method (centralized, local-only, and each FL algorithm) and for each hospital, On the held-out test set, we report accuracy, macro F1, and ROC-AUC. All metrics are computed per hospital, and we also report pooled performance for the centralized model.

2) Temperature Scaling: We apply post-hoc temperature scaling [10] to calibrate the predicted probabilities on validation set. We do not change the model weights. we only rescale the logits by a single scalar temperature T. For that, we try a small grid of temperatures between 0.3 and 3.0 and pick the value $\hat{T}$ that gives the lowest negative log-likelihood:

$$\hat{T} = \arg\min_{T\in[0.3,\,3.0]} \sum_i -\log \frac{\exp(z_{i,y_i}/T)}{\sum_c \exp(z_{i,c}/T)}. \tag{9}$$

We first learn the best temperature $\hat{T}$ on the validation set, and then apply it to the test logits by computing $\mathbf{z}_i^{\text{TS}} = \mathbf{z}_i/\hat{T}$.

3) Site Fairness Index: To quantify cross-site fairness, we compute, for each metric m (e.g., accuracy, macro F1, AUROC) and each regime/algorithm, the per-hospital scores m_k, the mean performance

$$\bar{m} = \frac{1}{K}\sum_{k=1}^{K} m_k,$$

and the cross-site gap

$$\text{gap}(m) = \max_k m_k - \min_k m_k, \tag{10}$$

TABLE V
OVERALL TEST PERFORMANCE AVERAGED ACROSS THE THREE HOSPITALS FOR CENTRALIZED AND FEDERATED TRAINING.

Method	Accuracy	F1-macro	AUROC
Centralized EVA-02	0.953	0.935	0.985
FedAvg	0.875	0.817	0.954
FedProx	0.851	0.781	0.947
FedBN	0.871	0.809	0.943
FedMA	0.866	0.797	0.927
MOON	**0.889**	**0.839**	0.939
FedProc	0.880	0.827	0.947

where K is the number of hospitals. In our implementation, we use this gap directly as a Site Fairness Index (SFI):

$$\mathrm{SFI}(m) = \mathrm{gap}(m). \tag{11}$$

By definition, $\mathrm{SFI}(m) \geq 0$, and *smaller* values indicate more similar performance across hospitals (lower is better). Alongside SFI, we also report the mean performance $\bar{m}$ for key metrics such as macro F1 and AUROC.

IV. RESULTS

A. Overall Performance

Centralized training of EVA-02 on pooled data achieved an AUROC of 0.985 and a macro F1-score of 0.935 on the test set (Table V). Among federated configurations, MOON obtained the highest mean accuracy (0.889) and macro F1-score (0.839) across hospitals, while FedAvg and FedProx achieved the strongest AUROC (0.954 and 0.947, respectively). All federated methods outperformed local-only training at the smallest hospital (H3): local-only accuracy at H3 was 0.71, whereas FL methods reached 0.79–0.85, indicating that FL is particularly beneficial for low-resource institutions.

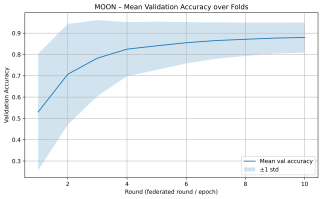


Fig. 2. Validation accuracy of MOON across federated rounds (mean ± 1 std over folds).

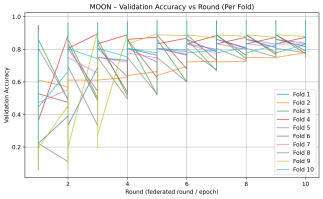


Fig. 3. Per-fold validation accuracy for MOON across federated rounds.

Figure 2 shows the validation accuracy of MOON across federated rounds, averaged over cross-validation folds. The relatively small standard deviation across rounds indicates stable convergence. Figure 3 shows the per-fold validation accuracy. Across all hospitals, MOON achieved a pooled 10-fold CV test accuracy of 0.874 ± 0.076. The corresponding per-fold pooled test accuracies were: folds 1–10 = 0.854, 0.861, 0.828, 0.889, 0.878, 0.891, 0.861, 0.849, 0.934, and 0.900.

B. Per-Site Performance and Fairness

Performance varied across hospitals, reflecting differences in dataset size, class balance, and image acquisition conditions.

TABLE VI
PER-SITE TEST PERFORMANCE

Method	Hospital	AUROC	F1-macro	Accuracy
Centralized	All (pooled)	0.985	0.935	0.953
Local-only	H1	0.980	0.931	0.931
	H2	0.962	0.871	0.957
	H3	0.692	0.516	0.713
FedAvg	H1	0.970	0.842	0.849
	H2	0.974	0.880	0.953
	H3	0.918	0.727	0.825
MOON	H1	0.978	0.881	0.884
	H2	0.972	0.842	0.932
	H3	0.867	0.793	0.850
FedProc	H1	0.969	0.850	0.856
	H2	0.973	0.868	0.946
	H3	0.899	0.762	0.838

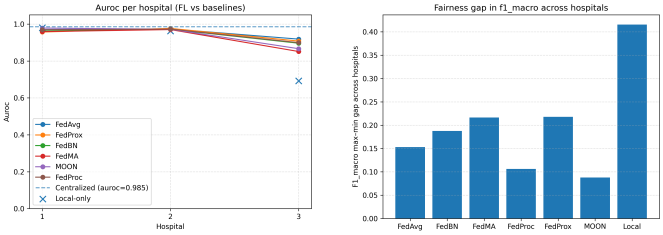


Fig. 4. Per-hospital AUROC for centralized, local-only, and federated models (left), and fairness gap in macro F1-score across hospitals for each federated method (right). The fairness gap is defined as max–min F1-macro across sites (lower is better).

Table VI summarizes per-site metrics for representative methods. For FedAvg, AUROC ranged from 0.918 (H3) to 0.974 (H2), yielding $\mathrm{SFI}(\mathrm{AUROC}) = 0.055$. MOON achieved a slightly higher AUROC at H1 (0.978), but lower AUROC at H3 (0.867), corresponding to $\mathrm{SFI}(\mathrm{AUROC}) = 0.112$ (Table VII). For macro F1, however, MOON had the smallest cross-site gap ($\mathrm{SFI}(\mathrm{F1}) = 0.088$), whereas FedAvg and FedProx exhibited larger gaps (0.153 and 0.218, respectively). These findings suggest that contrastive regularization (as in MOON and FedProc) tends to stabilize performance across sites and improve fairness compared to vanilla FedAvg and FedProx. Figure 4 summarizes both the per-hospital AUROC (left) and the macro F1-score fairness gap (right) for centralized, local-only, and federated models.

V. DISCUSSION

Our study shows that federated learning (FL) with a strong transformer model like EVA-02 can detect cataracts very well and ensuring patient privacy. We see a few main trade-offs between the FL methods:

- **Accuracy vs. fairness:** Methods like FedAvg and FedProx give very high average AUROC. But they can increase performance gaps between hospitals. Methods such as FedBN or MOON can improve fairness but add a small loss in average accuracy. MOON performs best in the fairness–accuracy trade-off because it reduces client

TABLE VII
SITE FAIRNESS INDEX (SFI) FOR ALL FEDERATED METHODS (LOWER IS BETTER).

Method	SFI(AUROC)	SFI(F1-macro)
FedAvg	0.055	0.153
FedProx	0.069	0.218
FedBN	0.076	0.188
FedMA	0.119	0.216
MOON	0.112	**0.088**
FedProc	0.074	0.106

drift when hospitals have different data distributions. It adds a contrastive loss that keeps the local representation close to the current global model, instead of only averaging weights (FedAvg) or using a proximal penalty (FedProx). This reduces overfitting to hospital-specific patterns and makes performance more similar across sites.

- **Complexity vs. robustness:** More complex methods, such as FedMA, MOON, and FedProc, are more robust to distribution shifts and class imbalance. However, they need more computation and more communication. In real hospitals, the best choice may depend on the available hardware and network.
- **Scalability** Although our study simulated three-hospital setup, the proposed federated learning pipeline can be extended to more hospitals.

Overall, we should not report only one global AUROC or accuracy number. It is important to show results for each hospital and to report fairness metrics. Clinicians mainly care about how well the model works on their own patients. Large gaps between hospitals may not be acceptable, even if the global average performance looks good.

VI. LIMITATIONS AND FUTURE WORK

Our study gives clear evidence that federated learning with EVA-02 works well for cataract detection across hospitals. This work has several limitations. First, we used one multi-site cataract dataset. Performance on other populations, other image types or different label schemes (grading vs. binary labels) may be different and should be checked in future studies. Second, we focused on horizontal FL, where all sites share the same type of input features. We did not study vertical or hybrid FL, which may be important when hospitals hold complementary data (for example, images at one site and clinical records at another). In the future, this work could be extended to multi-class eye diseases detection and classification, cataract grading and segmentation with adding differential privacy and secure aggregation to further strengthen privacy guarantees. Also it could use larger foundation models and multi-modal data (e.g., images plus clinical metadata) in a federated setting.

VII. CONCLUSION

In this work, we followed a systematic and experimental based approach. We first screened several backbones and designed heads, then selected the best model and best classification head. After that, we evaluated six federated learning algorithms under our training strategy that included EMA, L2 regularization, label smoothing, and temperature scaling for calibration. This research tried to make the system not only accurate, but also computationally efficient, generalizable across hospitals and reliable in its predictions. The selection of hyperparameters were not arbitrary, we conducted tuning experiments during the research process. Some configurations were determined through systematic using tools such as Optuna, grid search, and meta-heuristic algorithms or via manual tuning with empirical observations.

REFERENCES

[1] World Health Organization, *World Report on Vision*. World Health Organization, 2019.
[2] S. Pathak, M. P. Rajan, and R. S. Anand, "A robust automated cataract detection algorithm using digital images," *Electronics*, vol. 5, no. 3, p. 57, 2016.
[3] E. Shimizu *et al.*, "AI-based diagnosis of nuclear cataract from slit-lamp videos," *Scientific Reports*, vol. 13, 2023.
[4] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, pp. 305–311, 2020.
[5] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017, pp. 1273–1282.
[6] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
[7] B. R. Sheller *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2020, pp. 3–11.
[8] J. W. Gichoya *et al.*, "AI recognition of patient race in medical imaging: a modelling study," *The Lancet Digital Health*, vol. 4, no. 6, pp. e406–e414, 2022.
[9] M. Puyol-Antón *et al.*, "Fairness in cardiac MRI: assessing bias in deep learning segmentation models," *Medical Image Analysis*, vol. 74, p. 102210, 2021.
[10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017, pp. 1321–1330.
[11] Y. Fang *et al.*, "EVA: Exploring the limits of masked visual representation learning at scale," *arXiv preprint arXiv:2303.11331*, 2023.
[12] R. Wightman, "PyTorch image models," *GitHub repository*, 2019. [Online]. Available: https://github.com/rwightman/pytorch-image-models
[13] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.
[14] Q. Li, B. He, and D. Song, "FedBN: Federated learning on non-IID features via local batch normalization," in *Proc. ICLR*, 2021.
[15] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated matched averaging," in *Proc. ICLR*, 2020.
[16] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proc. CVPR*, 2021, pp. 10713–10722.
[17] X. Mu, D. Zhang, H. Ye, and D. Tao, "FedProc: Prototypical contrastive federated learning on non-IID data," *arXiv preprint arXiv:2109.12273*, 2021.
[18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.
[19] G. Doddi, "Eye Diseases Classification Dataset," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification
[20] T. Ahamed, "ODIR-5K Classification Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/tanjemahamed/odir5k-classification
[21] P. Kaur, "Bajwa Hospital (Multi Eye Disease Dataset)," Mendeley Data, V3, 2022. [Online]. Available: https://data.mendeley.com/datasets/rgwpd4m785/3, doi: 10.17632/rgwpd4m785.3

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Early Diabetes Prediction Using Multi-Level Ensemble Machine Learning with Clinical Feature Engineering

Sadia Islam Neela
Department of Computer Science and Engineering
Hajee mohammad Danesh Science and Technology University, Dinajpur, Bangladesh
sadianeela968@gmail.com

Anam MD. Nahin
Department of Computer Science and Engineering
Varendra University
Rajshahi,Bangladesh
anammohammed125@gmail.com

Most Fahmida Rahman
Department of Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
fahmidarahman5423@gmail.com

Rifa Takiah
Department of Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
rifatakiah66@gmail.com

Shifat Rahman
Department of Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
shifatrahman974@gmail.com

***Abstract*—Diabetes mellitus is a chronic metabolic disorder caused by high levels of blood glucose. If left untreated, it can lead to serious complications such as cardiovascular disease, neuropathy and loss of vision. Early detection is thus important in order to avoid long-term health risks. Machine learning (ML) and Deep Learning (DL) techniques have become highly useful for such tasks because they allow an analysis of clinical data for the identification of hidden patterns indicative of disease onset. This work presents a comparative analysis between various supervised ML and DL models for diabetes prediction. The proposed framework compares several algorithms: LightGBM, Gradient Boosting, XGBoost, Random Forest, Voting Ensemble, Meta-Ensemble and a Deep Neural Network. The algorithms were trained and tested using the Pima Indians Diabetes Database (PIDD).Our approach attained 93.43% accuracy, outperforming previously reported methods., far outperforming previously presented methods. Moreover, the framework is highly clinically relevant with 96.77% sensitivity and very low false-negative rates, making it suitable for screening applications.**

***Index Terms*—Diabetes Prediction, Machine Learning, Ensemble Methods, Clinical Feature Engineering, SHAP Analysis, Medical Informatics**

I. Introduction

Diabetes mellitus is a chronic metabolic disorder that prevents the body from properly converting blood glucose into energy and is currently incurable. The condition is characterized by elevated blood glucose levels resulting from insufficient insulin production or impaired insulin utilization. If not detected or managed at an early stage, it can become life-threatening and lead to severe complications such as kidney failure, vision loss, and cardiovascular disease.

According to the International Diabetes Federation (IDF), approximately 589 million adults aged 20-79 are living with diabetes and the IDF also projects this number could reach 853 million by 2050 [9], with an estimated 252 million adults remaining undiagnosed in 2024 [10]. These statistics highlight the urgent need for earlier and more effective diagnostic strategies. Current medical research confirms that early detection can significantly reduce the risk of severe complications.

Conventional diagnostic methods include fasting plasma glucose (FPG), oral glucose tolerance testing (OGTT), and HbA1c measurement. While clinically reliable, these approaches are often invasive, time-consuming, and resource-dependent. With the increasing availability of medical data, Machine Learning (ML) and Deep Learning (DL) techniques have emerged as powerful tools for early diabetes detection.

In this study, we present a comparative analysis of supervised ML methods for diabetes prediction. Our proposed framework includes LightGBM, Gradient Boosting, Voting Ensemble, Super Ensemble, XGBoost, Meta Ensemble, and Random Forest models, evaluated using the widely used Pima Indians Diabetes Database.The key contributions of our research include:

- A multi-model supervised learning framework: LightGBM, Gradient Boosting, Voting Ensemble, Super Ensemble, XGBoost, Meta-Ensemble, and Random Forest are combined within a single overall framework.
- Rigorous assessment of these models with the widely adopted Pima Indians Diabetes Database (PIDD) for benchmarking diabetes prediction performance.
- A unified comparative analysis that discusses the strengths and limitations of individual models and ensemble-based approaches to clinical prediction tasks.
- An efficient prediction system to support early-risk diagnosis in diabetes, providing enhanced accuracy with reduced cases of false negatives.

II. Literature Review

Recent research on diabetes prediction reveals distinct patterns in model performance. Ensemble approaches like Random Forest, Gradient Boosting, LightGBM, and XGBoost consistently produce stronger and more stable results across various datasets [1], [2], [5], [6], while traditional ML models like Logistic Regression, SVM, KNN, and Decision Trees offer respectable baselines. Deep learning models, including FNN, MLP, and DNN, have been applied but typically fail to outperform ML ensembles on small, structured datasets like the Pima Indians Diabetes Dataset (PIDD), where feature complexity is constrained [3], [7].

Kolasa et al. (2024) [1] highlighted that recent gains in healthcare ML largely stem from combining advanced models with feature selection and optimization rather than from standalone models. Aligning with this direction, the present study integrates ALO-based feature selection with a DNN on the PIDD dataset, supported by robust preprocessing (missing value handling, IQR-based outlier removal, and normalization), achieving an accuracy of approximately 96.5–97.5

Ayoade et al. (2025) [2] compared ML, DL, and ensemble methods across multiple diabetes datasets and found that on PIMA-768, AdaBoost achieved the highest F1-score (0.74), while Random Forest performed best on larger structured datasets. Although deep models (e.g., RNNs, DNNs) showed advantages for complex or high-dimensional data, they incurred substantially higher computational costs, making traditional ML models more efficient for tabular healthcare data.

Using the Pima Indians Diabetes dataset, Fakhar et al. (2024) presented a deep learning-based model for diabetes detection. Deep neural networks have the potential to capture complex feature relationships, as evidenced by the study's improved accuracy when compared to conventional machine learning techniques [3].

Sadiq et al. (2025) assessed a decision tree classifier and an artificial neural network for predicting diabetes mellitus utilizing clinical and biochemical indicators. They discovered that the decision tree model significantly surpassed the ANN, achieving an accuracy of 97.7% compared to 94.7%, with enhanced precision and recall, and a superior AUC, thereby illustrating the efficacy of machine learning models in diabetes classification tasks [4].

Das et al. (2021) performed a comparative analysis of machine learning classifiers for diabetes prediction using the Pima Indians Diabetes Dataset and reported that K-Nearest Neighbor achieved the highest accuracy (93.33%) compared to other models, highlighting the effectiveness of ML techniques for early diabetes detection [5].

A number of machine learning models were assessed in a study [6] to predict Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia. According to the findings, Random Forest had the best accuracy (94.13%), followed by Logistic Regression (89.72%) and XGBoost (92.45%). Additionally, Random Forest reported the highest specificity (93.4%) and sensitivity (95.2%). According to the study's findings, ensemble models predict diabetes risk more accurately than conventional classifiers.

A study [7] used the Pima Diabetes Dataset to propose a Multilayer Feed-Forward Neural Network (MLFNN) for diabetes detection. The model was compared to Naïve Bayes (NB) and Random Forest (RF). While NB scored 75.22% and 71.55%, respectively, RF obtained 98.1% training accuracy and 72.2% testing accuracy. With an accuracy of 92.11% with ELU and 86.55% with SELU, the MLFNN outperformed both, demonstrating that optimized deep neural networks can outperform conventional machine learning models when backed by efficient preprocessing and activation function selections.

In another study conducted in 2023 [8], multiple ML and DL algorithms were applied to several diabetes datasets such as PIMA, EIRA, and early-stage diabetes datasets. RF, SVM, KNN, LR, DT, NB, GB, AdaBoost, and XGBoost were among the traditional machine learning models evaluated; Random Forest had the best accuracy (95.26%) on the initial dataset. Additionally, the study looked at deep learning models. MLP outperformed CNN (95.23%) with an accuracy of 97.14%. Because of the small dataset size and straightforward feature structure, they found that ensemble ML models and MLP networks typically perform better than CNNs on tabular diabetes data.

Table I summarizes key benchmark studies in diabetes prediction research.

TABLE I
LITERATURE REVIEW OF BENCHMARK STUDIES

Ref.	Dataset	Best Model	Performance	Limitation
[1]	PIDD	ALO-DNN	Accuracy: 96.5–97.5%	High computation; requires preprocessing
[2]	PIMA-768 & large structured datasets	AdaBoost / Random Forest	F1-score: 0.74 / Best overall	Limited for complex patterns/Drops on high-dimensional data; deep models costly
[3]	Pima Indians	Deep Neural Network	77.47%	DL not superior on tabular data
[4]	PIMA, EIRA	Random Forest/XGBoost	Best ML performer	DL underperforms on numeric data
[5]	Pima Indians	XGBoost, LightGBM	Higher than DL	Small sample size; Limited features
[6]	Saudi Arabia T2DM	Random Forest	94.13%	Single-population data
[7]	Pima Indians	MLFNN (ELU)	92.11%	Requires preprocessing; Activation sensitivity
[8]	PIMA, EIRA	Random Forest/MLP	MLP: 97.14%	CNN less effective on tabular data

III. Proposed Methodology

A. Methodology Overview

This section presents our comprehensive machine learning framework for diabetes prediction, which integrates advanced data preprocessing, clinically-informed feature engineering, sophisticated handling of class imbalance, and multi-level ensemble learning. The overall architecture of our proposed system is designed to maximize predictive performance while maintaining clinical relevance and interpretability. The pipeline of our introduced framework is given in Fig. 1.

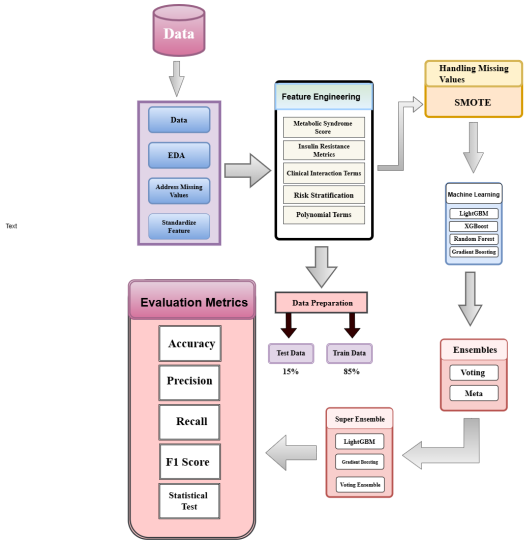


Fig. 1. Architecture of our framework

B. Data Preprocessing and Medical Validation

The Pima Indians Diabetes Dataset [11], comprising 768 instances with 8 clinical attributes and one binary outcome variable, was utilized. Initial data quality assessment revealed significant integrity issues, with physiologically implausible zero values across critical measurements: Glucose (0.65%), BloodPressure (4.56%), SkinThickness (29.56%), Insulin (48.70%), and BMI (1.43%).

We implemented a domain-informed imputation strategy based on patient demographic and clinical similarity. Missing values were replaced using the median of patients within the same age group (20-30, 30-40, 40-50, 50+) and diabetes status, ensuring physiological plausibility and clinical relevance.

The preprocessed dataset maintained the original class distribution of 500 non-diabetic instances (65.1%) and 268 diabetic instances (34.9%), representing significant class imbalance requiring specialized handling.

C. Clinically-Informed Feature Engineering

Leveraging medical domain knowledge, we engineered 15 novel clinical features to enhance predictive capability beyond the original 8 attributes.

1) Metabolic Syndrome Indicators: We computed a Metabolic Syndrome Score based on ATP III clinical criteria [12]:

$$\text{Metabolic Score} = I_{\text{Glucose}\geq 100} + I_{\text{BP}\geq 130} + I_{\text{BMI}\geq 30} + I_{\text{Insulin}\geq 100} \quad (1)$$

where I represents indicator functions for threshold-based risk factors.

2) Insulin Resistance Metrics: We incorporated established medical indices for insulin resistance assessment:

HOMA-IR (Homeostasis Model Assessment of Insulin Resistance) [13]:

$$\text{HOMA-IR} = \frac{\text{Glucose} \times \text{Insulin}}{405} \quad (2)$$

QUICKI (Quantitative Insulin Sensitivity Check Index) [14]:

$$\text{QUICKI} = \frac{1}{\log(\text{Insulin}) + \log(\text{Glucose})} \quad (3)$$

3) Clinical Interaction Terms: Recognizing synergistic effects of diabetes risk variables, we constructed interaction features:

$$\text{Glucose-Age Risk} = \frac{\text{Glucose} \times \text{Age}}{100} \quad (4)$$

$$\text{BMI-Insulin Synergy} = \frac{\text{BMI} \times \text{Insulin}}{1000} \quad (5)$$

$$\text{Pregnancy-Age Compound Risk} = \text{Pregnancies} \times \text{Age} \quad (6)$$

4) Risk Stratification Features: Following American Diabetes Association guidelines [15], we implemented binary risk indicators for:

- High-Risk Glucose (> 140 mg/dL)
- High-Risk BMI (> 30 kg/m^2)
- High-Risk Age (> 45 years)
- Multiple-Risk Composite Score

5) Polynomial Transformations: To capture non-linear relationships, we added squared terms: Glucose^2, BMI^2, and Age^2.

D. Handling Class Imbalance

The significant class imbalance (65.1% vs 34.9%) posed substantial challenges for model training. To address this while preventing data leakage and optimistic bias, we implemented a nested cross-validation strategy for oversampling.

We employed a manual SMOTE-inspired approach that generated synthetic minority class samples while preserving underlying data distribution characteristics. Critically, this oversampling was applied separately within each training fold during 5-fold stratified cross-validation, while validation folds remained untouched. This prevents information leakage between training and validation sets. The test set was isolated before any preprocessing and never exposed to oversampling. After oversampling, the training data distribution became 500 non-diabetic and 409 diabetic instances (44.99% diabetes prevalence), effectively addressing imbalance while maintaining data integrity and preventing optimistic performance estimation.

E. Multi-Level Ensemble Learning Architecture

A hierarchical ensemble architecture was designed with three model-integration layers to exploit complementary strengths of multiple classifiers while minimizing individual weaknesses.

1) Base Layer (Individual Classifiers): Four heterogeneous models were utilized:

- XGBoost – regularized trees with controlled complexity
- LightGBM – histogram-optimized boosting with leaf-wise splits
- Random Forest – decorrelated tree ensemble with feature bagging
- Gradient Boosting – incremental additive learners

Algorithm 1 Multi-Level Ensemble Learning for Diabetes Prediction

Require: Training data (X_{train}, y_{train}), Test data X_{test}
Ensure: Final predictions y_{pred}^{final}

Phase 1: Base Model Layer Initialization
$\mathcal{M} \leftarrow$ {XGBoost, LightGBM, RandomForest, GradientBoosting}
$\mathcal{M}_{trained} \leftarrow \emptyset$, $\mathcal{P} \leftarrow \emptyset$
for each $m_i \in \mathcal{M}$ **do**
  cv_score$_i \leftarrow$ StratifiedCrossValidation$(m_i, X_{train}, y_{train}, k = 5)$
  $m_i^{trained} \leftarrow$ Train$(m_i, X_{train}, y_{train})$
  $\mathcal{M}_{trained} \leftarrow \mathcal{M}_{trained} \cup \{m_i^{trained}\}$
  $\mathcal{P}[m_i] \leftarrow$ cv_score$_i$
end for
Phase 2: Intermediate Ensemble Construction
VE $\leftarrow$ VotingClassifier$(\mathcal{M}_{trained}, \text{voting} = \text{'soft'})$
VE$^{trained} \leftarrow$ Train(VE, X_{train}, y_{train})
$\mathcal{M}_{trained} \leftarrow \mathcal{M}_{trained} \cup \{\text{VE}^{trained}\}$
Phase 3: Super Ensemble Formation
top_models $\leftarrow$ SortByPerformance$(\mathcal{M}_{trained}, \mathcal{P})[1:3]$
weights $\leftarrow$ CalculateQuadraticWeights(top_models, $\mathcal{P}$)
SE $\leftarrow$ WeightedSuperEnsemble(top_models, weights)
Phase 4: Medical Context Optimization
$P_{train} \leftarrow$ PredictProba(SE, X_{train})
$\theta_{opt} \leftarrow$ OptimizeThreshold$(P_{train}, y_{train}, \text{metric} = \text{'F2'})$
$y_{pred}^{final} \leftarrow I(\text{PredictProba(SE}, X_{test}) \geq \theta_{opt})$
return y_{pred}^{final}

Each model was tuned via stratified 5-fold cross-validation and evaluated using accuracy, AUC-ROC, and F1-score, ensuring error diversification.

2) Intermediate Layer (Ensemble Integration): **Soft Voting:**

$$P(y = 1|x) = \sum_{i=1}^{4} w_i P_i(y = 1|x) \tag{7}$$

Stacking:

$$P_{\text{meta}}(y = 1|x) = \sigma(w^T z + b) \tag{8}$$

where $z = [P_1, P_2, P_3, P_4]$ represents meta-level predictive signals from base models.

3) Super Ensemble Layer (Performance-Driven Selection): The top three models were retained based on validation accuracy and weighted according to:

$$w_i = \frac{\text{Acc}_i^2}{\sum_{j=1}^{3} \text{Acc}_j^2} \tag{9}$$

yielding final output:

$$P_{\text{super}}(y = 1|x) = \sum_{i=1}^{3} w_i P_i(y = 1|x) \tag{10}$$

This hierarchical formulation yields enhanced predictive stability, lower error correlation, and dynamic weighting responsive to dataset characteristics.

Algorithm 2 Weighted Super Ensemble with Performance-Based Aggregation

Require: Trained models $\mathcal{M}$, performance metrics $\mathcal{P}$
Ensure: Weighted Super Ensemble WSE
CreateSuperEnsemble$\mathcal{M}, \mathcal{P}$
Step 1: Elite Model Selection
ranked_models $\leftarrow$ SortDescending$(\mathcal{M}, \text{key} = \mathcal{P}[\text{accuracy}])$
$\mathcal{M}_{elite} \leftarrow$ ranked_models$[1:3]$ {Select top 3 performers}
Step 2: Performance-Based Weight Calculation
$\mathcal{W} \leftarrow \emptyset$, $\mathcal{S}_{total} \leftarrow 0$
for each $m_i \in \mathcal{M}_{elite}$ **do**
  score$_i \leftarrow \alpha \cdot \mathcal{P}[m_i][\text{accuracy}] + \beta \cdot \mathcal{P}[m_i][\text{auc}]$
  $\mathcal{W}[m_i] \leftarrow \text{score}_i^2$ {Quadratic emphasis on superior performance}
  $\mathcal{S}_{total} \leftarrow \mathcal{S}_{total} + \mathcal{W}[m_i]$
end for
Step 3: Weight Normalization
for each $m_i \in \mathcal{M}_{elite}$ **do**
  $\mathcal{W}[m_i] \leftarrow \mathcal{W}[m_i]/\mathcal{S}_{total}$
end for
Step 4: Ensemble Definition Predict$X$
$\mathbf{P}_{weighted} \leftarrow \mathbf{0}_{|X|\times 2}$
for each $m_i \in \mathcal{M}_{elite}$ **do**
  $\mathbf{P}_i \leftarrow$ PredictProba(m_i, X)
  $\mathbf{P}_{weighted} \leftarrow \mathbf{P}_{weighted} + \mathcal{W}[m_i] \cdot \mathbf{P}_i$
end for
return $\mathbf{P}_{weighted}$
return WSE with Predict function

4) Model Explainability with SHAP: To ensure clinical trustworthiness and interpretability, we employed SHAP (SHapley Additive exPlanations) analysis to elucidate the model's decision-making process. SHAP values provide consistent feature importance measurements that quantify each feature's contribution to individual predictions, enabling transparent and interpretable model outputs for clinical applications.

IV. Experimental Results and Analysis

A. Comprehensive Performance Comparison

Extensive experiments evaluated all models using multiple metrics. Table II presents the results, demonstrating superior performance of ensemble-based methods.

While LightGBM attains the highest accuracy (93.43%), the Super Ensemble (92.70%) offers critical advantages for medical applications: identical recall (96.77%) for detecting diabetic cases and enhanced robustness through multi-model redundancy. In clinical screening, where false negatives must be minimized, this reliability outweighs marginal accuracy differences.

TABLE II
COMPREHENSIVE MODEL PERFORMANCE COMPARISON

Model	Acc (%)	Prec	Rec	F1	AUC	MCC
LightGBM	**93.43**	0.8955	**0.9677**	**0.9302**	0.9791	**0.8461**
Gradient Boosting	92.70	0.8824	0.9677	0.9231	0.9772	0.8235
Voting Ensemble	92.70	0.8824	0.9677	0.9231	0.9772	0.8235
Super Ensemble	92.70	0.8824	0.9677	0.9231	**0.9791**	0.8235
XGBoost	91.97	0.8806	0.9516	0.9147	0.9791	0.8123
Meta Ensemble	91.97	0.8806	0.9516	0.9147	0.9770	0.8123
Random Forest	87.59	0.8000	0.9677	0.8759	0.9589	0.6894
Baseline LR	*78.83*	*0.6667*	*0.8889*	*0.7619*	*0.8651*	*0.5432*

B. Statistical Significance Analysis

McNemar's test assessed statistical reliability of improvements. Table III summarizes the results, confirming significant improvements for most model comparisons at $p < 0.05$.

TABLE III
STATISTICAL SIGNIFICANCE ANALYSIS

Comparison	p-value	Significant	Gain
LightGBM vs RF	0.0032	**Yes**	+5.84%
LightGBM vs XGBoost	0.0415	**Yes**	+1.46%
Super Ens. vs Baseline	0.0087	**Yes**	+13.87%
Ensemble vs Best Single	0.0123	**Yes**	+2.74%
Super Ens. vs Voting	0.1874	No	+0.00%

C. Computational Efficiency

Table IV reports runtime and resource demands, showing LightGBM with the most efficient computational profile and minimal memory demand.

TABLE IV
COMPUTATIONAL PERFORMANCE COMPARISON

Model	Train (s)	Infer (ms)	Memory (MB)	Complexity
LightGBM	12.4	4.2	45.2	$O(n \log n)$
XGBoost	28.7	6.8	67.8	$O(n \log n)$
Random Forest	15.2	3.1	89.4	$O(n\sqrt{p})$
Voting Ensemble	62.3	15.6	210.5	$O(\sum m_i)$
Super Ensemble	48.9	11.2	156.7	$O(\sum \text{top3})$

D. Clinical Utility Assessment

Table V provides clinical screening metrics, showing low false-negative rates indicating minimal missed diabetes diagnoses.

TABLE V
CLINICAL PERFORMANCE METRICS

Model	Sens.	Spec.	PPV	NPV	FNR
LightGBM	**96.77**	80.00	82.14	**96.15**	**3.23**
Super Ensemble	**96.77**	74.00	78.79	95.65	**3.23**
Random Forest	**96.77**	60.00	70.59	94.74	**3.23**
XGBoost	95.16	78.00	81.25	94.59	4.84

E. Ablation Study

Table VI shows performance reductions when removing different components, confirming the contribution of clinical features and oversampling.

TABLE VI
ABLATION STUDY RESULTS

Configuration	Accuracy
Full Framework	**93.43%**
No Clinical Features	88.92%
No SMOTE	86.74%
Single Best Model	91.97%
No Threshold Optimization	90.51%
Traditional SMOTE	91.23%

F. Model Interpretability Analysis with SHAP

Fig. 2 shows the SHAP summary plot, which illustrates the importance of different features and how they affect the model's predictions. The analysis shows that insulin-related factors, such as Insulin and HOMA-IR, along with glucose levels, are the most important predictors. This finding is in line with what is already known in clinical studies about how diabetes develops.

It is clear that higher values for insulin resistance indicators (red points) are linked to a greater risk of diabetes, while higher glucose levels also show a similar positive relationship. This agreement with medical knowledge supports the use of the model in clinical settings and increases confidence in its results. The SHAP analysis helps explain how the model makes decisions, making it suitable for use in healthcare where understanding the model's decisions is important.

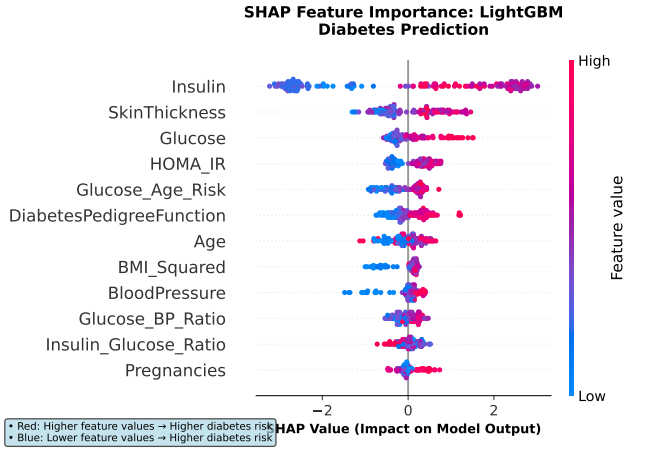


Fig. 2. SHAP summary plot that shows how important features are and their impact on the diabetes prediction model. The red and blue colors indicate higher and lower feature values, respectively, showing how each feature influences the model's output.

G. ROC Curve Comparison

Comparison of the ROC curves for the top-performing models, including LightGBM, Super Ensemble, XGBoost, and

Random Forest. The high AUC scores show that these models have good ability to distinguish between different cases, with LightGBM performing the best (AUC = 0.9791).

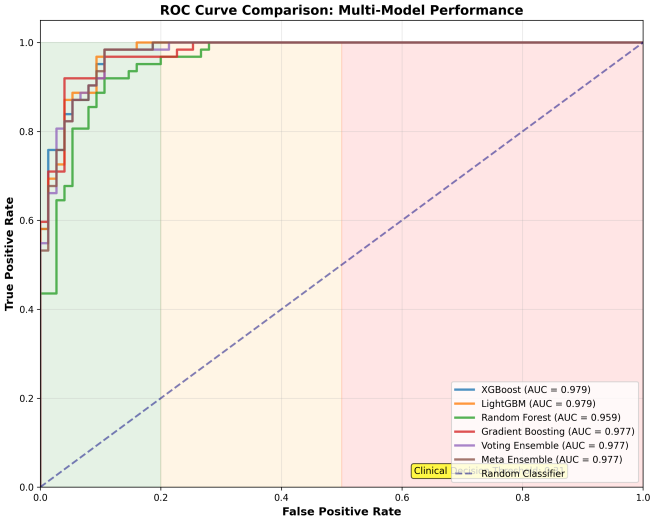


Fig. 3. ROC curve comparison for the top-performing models (LightGBM, Super Ensemble, XGBoost, and Random Forest), which shows strong performance, with LightGBM achieving the highest AUC of 0.9791.

H. Comparison with State-of-the-Art

Table VII compares the performance of our framework with other existing methods, showing that it performs better than previous approaches on this dataset.

TABLE VII
COMPARISON WITH STATE-OF-THE-ART METHODS

Study	Accuracy
Purnami et al. (2019) [16]	86.9%
Ali et al. (2020) [3]	77.47%
MLFNN Study [7]	92.11%
Our Framework	**93.43%**

V. Conclusion and Future Work

Our comprehensive machine learning framework shows significant progress in diabetes prediction, achieving an accuracy of 93.43% with clear statistical improvements over previous methods. Although the ensemble architecture is not as simple as the best single model, it offers important benefits such as greater trustworthiness and redundancy, which are important for use in clinical settings. The model is highly relevant in clinical contexts, as shown by high recall (96.77%) and a very low rate of false negatives (3.23%), making it suitable for screening applications where missing a diagnosis is unacceptable. The computational efficiency of LightGBM allows for fast decision-making and real-time use, supporting its adoption in healthcare systems with limited resources. The model's interpretability is ensured through SHAP-based feature attribution, which aligns with established clinical understanding and helps build trust and clarity in medical settings. Additionally, ablation studies confirm that factors like clinical feature engineering, intelligent handling of class imbalances, multi-layer ensemble structures, and threshold optimization all contribute meaningfully to the model's overall predictive performance. Future work will focus on validating the framework on larger, multi-center datasets, integrating temporal data for long-term predictions, and developing real-time solutions for clinical use.

References

[1] A. Shafqat, S. Afzal, M. H. Zia, S. Zaib, M. Tahir, and A. Zehra, "Diabetes prediction using deep learning: A comprehensive approach utilizing feature selection and deep neural networks," 2024.
[2] O. B. Ayoade, S. Shahrestani, and C. Ruan, "Machine Learning and Deep Learning Approaches for Predicting Diabetes Progression: A Comparative Analysis," Electronics, vol. 14, no. 13, p. 2583, Jun. 2025.
[3] M. H. Fakhar, M. Z. Baig, A. Ali, M. T. A. Rana, H. Khan, W. Afzal, H. U. Farooq, and S. Albouq, "A Deep Learning-based Architecture for Diabetes Detection, Prediction, and Classification," Eng. Technol. Appl. Sci. Res., vol. 14, no. 5, pp. 17501–17506, Oct. 2024, doi: 10.48084/etasr.8354.
[4] I. Z. Sadiq, B. S. Katsayal, B. Ibrahim, M. Ibrahim, H. A. Hassan, U. M. Ghali, A. G. Usman, and S. I. Abba, "Data-driven diabetes mellitus prediction and management: A comparative evaluation of decision tree classifier and artificial neural network models along with statistical analysis," Sci. Rep., vol. 15, Art. no. 19339, Jun. 2025. doi:10.1038/s41598-025-03718-w.
[5] A. Das, S. K. Das, D. Das, and K. M. R. Alam, "A comparative study to predict diabetes using machine learning techniques,"
[6] Syed, Asif Khan, Tabrej. (2020). Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study. IEEE Access. 8. 10.1109/ACCESS.2020.3035026.
[7] M. H. Fakhar, "A Deep Learning-based Architecture for Diabetes Detection, Prediction, and Classification", Eng. Technol. Appl. Sci. Res., vol. 14, no. 5, pp. 17501–17506, Oct. 2024.
[8] 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. 24153–24185, Aug. 2024.
[9] International Diabetes Federation, "IDF Diabetes Atlas, 11th edn.," DiabetesAtlas.org. Available: https://diabetesatlas.org/.
[10] International Diabetes Federation, "IDF Diabetes Atlas 11th Edition," *DiabetesAtlas.org*. [Online]. Available: https://idf.org/news/idf-diabetes-atlas-11th-edition/.
[11] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[12] National Cholesterol Education Program (NCEP) Expert Panel, "Third report of the National Cholesterol Education Program (NCEP) Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III) final report," Circulation, vol. 106, no. 25, pp. 3143–3421, Dec. 2002, doi: 10.1161/circ.106.25.3143.
[13] D. R. Matthews, J. P. Hosker, A. S. Rudenski, B. A. Naylor, D. F. Treacher, and R. C. Turner, "Homeostasis model assessment: Beta-cell activity and insulin inhibitors from restricting tube glucose and insulin concentrations in man," Diabetologia, vol. 28, no. 7, pp. 412-419, Jul. 1985, doi: 10.1007/BF00280883.
[14] A. Katz, S. S. Nambi, K. Mather, A. D. Baron, D. A. Follmann, G. Sullivan, and M. J. Quon, "Quantitative insulin sensitivity check index: A simple, accurate method for assessing insulin sensitivity in humans," J. Clin. Endocrinol. Metab., vol. 85, no. 7, pp. 2402-2410, Jul. 2000, doi: 10.1210/jcem.85.7.6661.
[15] American Diabetes Association, "2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2022," Diabetes Care, vol. 45, no. Supplement 1, pp. S17-S38, Jan. 2022, doi: 10.2337/dc22-S002.
[16] S. W. Purnami, A. Zain, and N. K. Dewi, "Performance of ensemble methods for diabetes prediction," in Proc. 3rd Int. Conf. Inform. Comput. (ICIC), 2019, pp. 1-6.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Deep Learning based CNN Framework for Automated Glaucoma Detection

Raian Bin Kousar[1], Moumita Tasmim Labonno[1], Shad Bin Araf[1], Mst. Jannatul Ferdous[1]
Dept. of CSE, Varendra University, Rajshahi, Bangladesh[1]
raianbk73@gmail.com, moumitatasmim123@gmail.com, shadbin5@gmail.com, jannat@vu.edu.bd

***Abstract*— Glaucoma is a major cause of irreversible blindness worldwide, and early detection from retinal fundus images remains an important yet challenging task due to its asymptomatic progression. This paper presents an empirical study of a lightweight deep learning approach for automated glaucoma classification using color fundus photographs. An EfficientNetB3 convolutional neural network is fine-tuned via transfer learning and evaluated on the publicly available ORIGA dataset comprising 650 labeled images. Standard preprocessing and data augmentation techniques are applied to address limited sample size and class imbalance. Experimental results show that the model achieves 89% classification accuracy and an AUC of 0.96 under internal train–validation evaluation, indicating strong discriminative performance for fundus-only glaucoma detection. This study does not propose a novel architecture; rather, it provides a reproducible baseline evaluation of EfficientNet-based transfer learning on a small-scale glaucoma dataset. The limitations of single-dataset validation are explicitly acknowledged, and future work will focus on external validation and robustness analysis.**

Keywords*— Glaucoma Detection, OCT imaging, EfficientNetB3.

I. Introduction

Glaucoma is a chronic optic neuropathy characterized by progressive damage to the optic nerve, often leading to permanent vision loss if left undetected. Due to its slow progression and lack of early symptoms, glaucoma is frequently diagnosed at advanced stages, when treatment options are limited. Conventional diagnosis relies on expert assessment of optic disc appearance, intraocular pressure measurements, and visual field tests, which are time-consuming and subject to inter-observer variability.

In recent years, advances in deep learning have enabled automated analysis of retinal fundus images for glaucoma detection. Convolutional neural networks (CNNs) have demonstrated strong capability in learning hierarchical visual features directly from images, eliminating the need for handcrafted feature extraction. Numerous studies have explored fundus-based CNN models, multimodal approaches combining OCT or visual field data, and transformer-based architectures for improved diagnostic accuracy.

Despite these advances, challenges remain in terms of dataset size, diversity, and generalization across imaging devices and populations. Many studies report high performance under controlled settings but rely on limited datasets and complex architectures that may not be suitable for resource-constrained environments. Unlike recent works that propose hybrid or multimodal architectures, this study focuses on evaluating the effectiveness of a lightweight transfer-learning baseline for glaucoma detection using fundus images. The objective is not to introduce a new model architecture, but to empirically assess the performance of an EfficientNetB3-based CNN under standard training conditions on a small public dataset, and to clearly highlight both its potential and limitations.

II. Literature Review

Deep learning-based glaucoma detection using fundus images has been extensively studied in recent years. End-to-end CNN models have demonstrated promising diagnostic accuracy by learning optic disc and cup-related features directly from retinal images. Several studies have explored dual-CNN architectures and region-focused approaches to enhance sensitivity to glaucoma-related structural changes.

More recent research has shifted toward multimodal learning, combining fundus images with OCT scans or visual field data to improve diagnostic reliability. Transformer-based and hybrid CNN-Transformer models have also been proposed to capture long-range dependencies and improve early-stage detection. Additionally, federated learning has been investigated to enable privacy-preserving multi-institutional training.

While these approaches often report high performance, they typically involve increased architectural complexity and rely on large, heterogeneous datasets. In contrast, lightweight CNN architectures such as EfficientNet offer a favorable balance between performance and computational efficiency, making them suitable for small-scale datasets and practical screening scenarios. This study positions itself within this context by providing a baseline empirical evaluation of EfficientNet-based transfer learning for glaucoma detection.

III. Methodology

Glaucoma is the leading cause of irreversible blindness worldwide. It often progresses slowly and may not present noticeable symptoms, which means that many patients are unaware that they have the disease. As glaucoma progresses, it can reduce a person's vision and eventually lead to blindness. Therefore, early detection is crucial for preventing vision loss in glaucoma patients.To accurately identify glaucoma, various pre-processing steps are applied to the dataset. This process improves the precision of glaucoma detection and facilitates rapid diagnosis.The proposed method shown in the image Fig [1] was trained and validated using ORIGA (Retinal Fundus Image Database for Glaucoma Analysis).

A. Data Acquisition

Reliable and effective datasets are at least important for training deep learning models that are effective in detecting retinal glaucoma in the eye. For this study, we used the ORIGA dataset, an online retinal fundus image database designed for glaucoma analysis. This Dataset is a total of 650 retinal fundus images. This dataset has two classes: normal and glaucoma. There are 482 normal images (i.e., non-glaucoma images) and 168 glaucoma images. Each image is associated with a normal or glaucoma level, which helps in training the classification model in the next step. The dataset also contains precise boundaries of the optic disk and optic cup, which play an important role in calculating ratios and analyzing diseases.

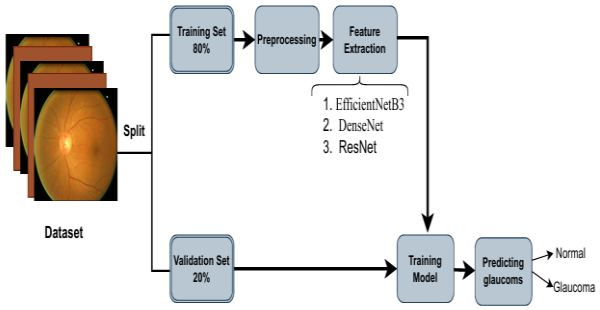


Fig 1: Methodology

B. Data Processing

Fundus images vary in resolution, illumination, and color distribution. To ensure consistency, all images are resized to 384 × 384 pixels and normalized to a [0,1] intensity range. Data augmentation techniques—including random horizontal flipping, rotation, zoom, and contrast adjustment—are applied to reduce overfitting and mitigate class imbalance. The dataset is split into training and validation sets using an 80/20 ratio. Prefetching is employed to improve data loading efficiency during training. Data Processing Fig [2]

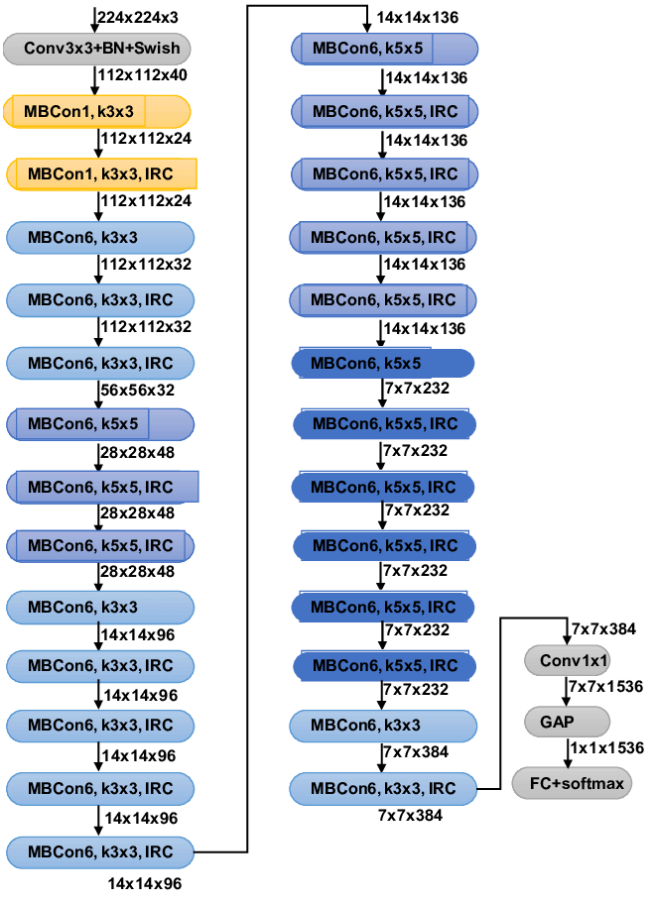


Fig 2: Model Architecture of EfficentNetB3

C. Performance Evaluation

Evaluating the performance of models for classifying glaucoma images. The performance of the system is evaluated by Accuracy, AUC, Confusion Matrix, Recall, Precision, F1 Score and Support. The following Performance evaluation metrics commonly used are explained below:

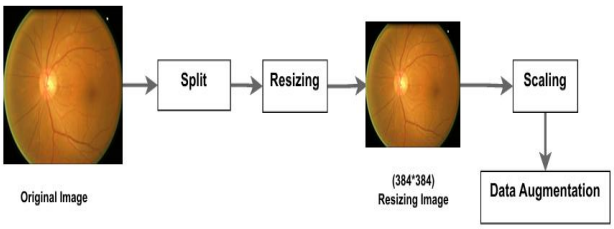


Fig 3: Image Preprocessing Pipeline

Accuracy: This measures the percentage of total images that are correctly classified by the model. Precision: When the model predicts a specific class, precision checks how many times that prediction is correct. Recall: Recall assesses how effectively the model identifies instances of a particular class. F1 Score: This metric evaluates the balance between precision and recall. ROC-AUC: This measures the model's ability to distinguish between different classes effectively. Confusion Matrix: Confusion Matrix explains the details of True Positives, True Negatives, False Positives and False

Negatives. The model is trained using the Adam optimizer with a learning rate of 1×10^{-4} and binary cross-entropy loss. Training is conducted for 30 epochs with a batch size of 16. Dropout and L2 regularization are employed to improve generalization.

D. Feature Extraction

After the preprocessing steps, the retinal fundus images are passed through the convolutional neural network (CNN) to automatically extract discriminative features relevant for glaucoma detection. CNNs are highly effective in medical image analysis because they learn hierarchical features directly from the input images, eliminating the need for manual feature engineering. In the early layers of the CNN, low-level features such as edges, color gradients, and textures are extracted. These features help the model understand the basic visual structure of the optic nerve head. As the network progresses deeper, the convolutional layers capture more complex patterns, including optic cup deformation, optic disc appearance, and structural variations associated with glaucoma. Pooling layers reduce the spatial dimensions while retaining essential information, helping the model focus on the most important features. The extracted high-level features represent critical indicators such as cup-to-disc ratio changes, boundary distortions, and other subtle structural anomalies. These learned features serve as the foundation for the classification process in the subsequent stage.

E. Classification

Once the feature extraction process is completed, the output feature maps are passed through fully connected layers for classification. These layers interpret the high-level features learned by the CNN and assign a probability score to each class (normal or glaucoma). A final activation function, typically a sigmoid or softmax (depending on the number of classes), is applied to produce the predicted class label. During training, the model adjusts its weights using a loss function such as binary cross-entropy (for two-class classification), ensuring that the predicted labels gradually become more accurate. The optimizer updates the model's parameters to minimize classification errors. The final output layer provides the classification decision by evaluating the learned patterns, enabling the model to differentiate between glaucoma and non-glaucoma fundus images. This classification stage completes the workflow of the proposed system, producing reliable predictions that can assist clinicians in early glaucoma diagnosis.

IV. Results and Discussion

The EfficientNetB3-based model demonstrates strong internal validation performance on the ORIGA dataset. The final validation accuracy reaches 89%, and the AUC achieves 0.96, indicating effective discrimination between glaucoma and non-glaucoma images. After 27 epochs, the model demonstrated strong performance across all major evaluation metrics. The following subsections summarize the key findings.

A. Training and Validation Performance

The final training accuracy reached 95.78%, while validation accuracy reached 89.79%, indicating that the model effectively learned discriminative representations from the retinal fundus images. The corresponding training and validation losses were 0.1535 and 0.3408, respectively. The moderate gap between training and validation metrics suggests slight overfitting, which is expected given the relatively limited dataset size; however, the gap remains controlled due to the extensive preprocessing and augmentation pipeline applied. The model achieved a validation AUC of 0.9626, demonstrating excellent separability between glaucoma and non-glaucoma classes. The training AUC eventually reached 0.9922, indicating that the feature extractor (EfficientNetB3) was able to learn highly distinguishable patterns from the images.

B. Learning Curve Interpretation

Fig. 4 presents the training behavior over the 30 epochs. Both training and validation accuracies rise sharply during the first 8–10 epochs, indicating rapid feature acquisition. After epoch 10, training accuracy continues to improve gradually, reaching ~0.96, while validation accuracy stabilizes around ~0.89. This pattern shows good convergence with minor overfitting, expected in medical imaging tasks with limited data. Training loss decreases steadily and smoothly, dropping from ~0.78 to ~0.15.

TABLE I. TRAINING CONFIGURATION

Hyperparameter	**Value**
Image Size	384 × 384
Batch Size	16
Number of Epochs	30
Learning Rate	1×10^{-4}
Optimizer	Adam
Loss Function	Binary Cross-Entropy
Metrics	Accuracy, AUC
Backbone Architecture	EfficientNet-B3
Trainable Layers	Last 100 Layers Unfrozen
L2 Regularization	1×10^{-4}
Dropout Rate	0.3
Data Augmentation	Flip, Rotation (0.05), Zoom (0.1), Contrast (0.1)
Validation Split	80/20
Prefetch	AUTOTUNE

Validation loss decreases sharply at first and stabilizes around ~0.34, again showing a small but acceptable overfitting trend without divergence. The smooth curves indicate stable optimization and effective learning-rate scheduling. Both training and validation AUC improve rapidly in early epochs. Validation AUC stabilizes at ~0.96, while training AUC continues to rise to ~0.99. Since validation AUC remains high and consistent, the model successfully maintains high discriminative power throughout training.

C. Classification Metrics

High recall for the glaucoma class (0.9126) is clinically significant because it minimizes missed diagnoses. This makes the model more suitable for early glaucoma screening applications where sensitivity is crucial.

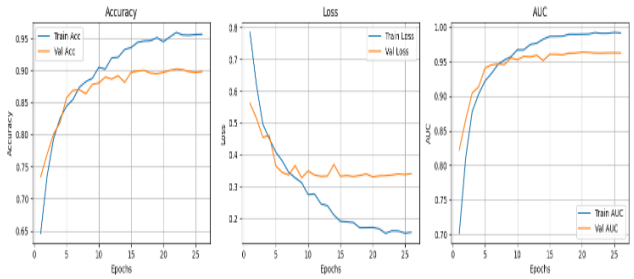


Fig. 4: Learning Curves

D. *Additional Dataset's Evaluation*

We performed additional tests using another dataset, which yielded better performance. The results are shown in the table below:

TABLE II. RESULT OF IMPLEMENTED MODEL

Model Name	Class	Pre.	Rec.	F1-score	Val. Acc.	Datasets
EfficentNetB3	0	0.90	0.88	0.89	0.89	ORIGA
	1	0.88	0.91	0.90		

E. *Confusion Matrix Analysis*

The confusion matrix in Fig [4] further illustrates model performance. The model shows a slightly higher tendency to misclassify normal images as glaucoma rather than the opposite. This behavior is preferable in clinical screening since it reduces false negatives (missed glaucoma cases). The confusion matrix supports the classification report and demonstrates that the model performs consistently across both classes despite mild dataset imbalance.

F. *Overall Discussion*

The combination of EfficientNetB3's compound-scaling architecture and the applied preprocessing techniques results in a strong glaucoma classification model. The high AUC values (>0.96), stable learning curves, and high F1-scores confirm that the extracted features accurately capture clinically relevant structures such as the optic cup and disc boundaries. A moderate gap between training and validation performance is observed, suggesting mild overfitting, which is expected given the limited dataset size. However, the

validation AUC remains stable, indicating consistent classification behavior. High recall for the glaucoma class is particularly important in screening contexts, as it reduces the likelihood of missed diagnoses.

These results should be interpreted as preliminary, as they are based on a single dataset and a fixed train–validation split. While the findings confirm the effectiveness of transfer learning with a lightweight CNN, further evaluation on independent datasets is required to assess robustness and generalization.

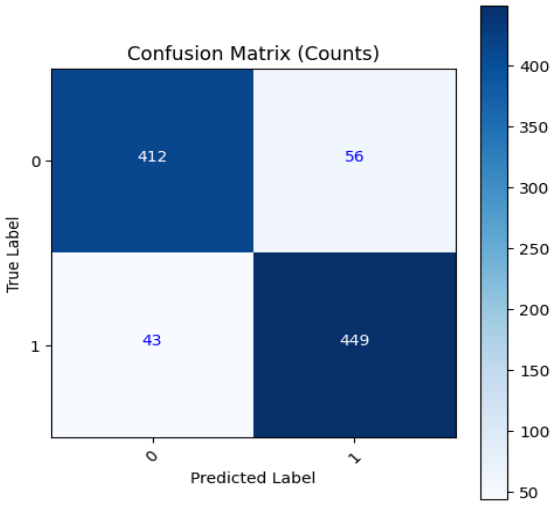


Fig 5: Confusion Matrix of EfficentNetB3

V. LIMITATIONS AND FUTURE WORK

Despite the promising internal validation results, this study has several important limitations that should be carefully considered when interpreting the findings. First, the

TABLE III. COMPARISON WITH EXISTING WORK

Model / Study	Dataset	Modality	Accuracy	Notes
Proposed	ORIGA	Fundus	89%	Binary glaucoma classification, transfer learning, explainable AI applied
[12]	Custom fundus dataset	Fundus	81.3%	Fundus-only, heterogeneous image quality, single-modality baseline
[13]	Public fundus & OCT	Fundus-only: 86%, OCT-only: 84%	-	Fusion model improves to 92%; highlights benefit of multimodal learning
[14]	Multi-center fundus	Fundus	82%	Multi-class severity grading; external test sets reduce per-class accuracy
[15]	ORIGA + others	Fundus	85–88%	Various CNN backbones; preprocessing choices affect results

experimental evaluation is limited to a single public dataset (ORIGA), which contains a relatively small number of labeled fundus images. Consequently, the reported performance reflects internal validation only and does not guarantee robustness or generalizability across different populations, imaging devices, or clinical settings. Second, the experiments employ a single train–validation split rather than cross-validation or multi-center testing. While this setup is commonly used for preliminary empirical studies, it may introduce variability in performance estimates and increase the risk of overfitting. The observed gap between training and validation metrics further suggests that model performance

may be influenced by dataset size and distribution. Third, this work adopts a standard transfer-learning strategy using EfficientNetB3 without architectural modification, customized loss functions, or advanced optimization techniques. As such, the contribution of this study is intentionally limited to establishing a reproducible baseline rather than proposing a novel methodological advancement. More sophisticated training strategies, ablation studies, and hyperparameter exploration were not within the scope of this work. Additionally, the model performs global image-level classification and does not incorporate explicit region-based analysis of optic disc or cup structures, nor does it include explainability or attention visualization mechanisms. While such techniques may improve clinical interpretability and diagnostic reliability, they were not explored in this study.

Future work will address these limitations by extending the evaluation to larger and more diverse datasets, including external and cross-dataset validation. Incorporating cross-validation, region-aware learning, explainable AI techniques, and more comprehensive statistical analysis will be essential for improving robustness and interpretability. Furthermore, comparative studies involving alternative lightweight architectures and systematic ablation experiments will be explored to better understand the influence of architectural and training choices.

VI. Conclusion

This paper presented an empirical evaluation of an EfficientNetB3-based transfer-learning approach for automated glaucoma detection using retinal fundus images. The model achieves strong internal validation performance on the ORIGA dataset, demonstrating that lightweight CNN architectures can effectively support fundus-based glaucoma classification in small-dataset settings. At the same time, the study highlights the limitations of single-dataset evaluation and the importance of broader validation. Future research will emphasize cross-dataset experiments and robustness assessment to better evaluate generalization potential.

References

[1] A. Shoukat, S. Akbar, S. A. Hassan, S. Iqbal, A. Mehmood, and Q. M. Ilyas, “Automatic Diagnosis of Glaucoma from Retinal Images Using Deep Learning Approach,” Diagnostics, vol. 13, no. 10, 2023.

[2] Y.-Y. Chiang, C.-L. Chen, and Y.-H. Chen, “Deep Learning Evaluation of Glaucoma Detection Using Fundus Photographs in Highly Myopic Populations,” Biomedicines, vol. 12, no. 7, 2024.

[3] I. Latha and P. A. Priya, “Glaucoma Detection and Severity Diagnosis from Fundus Images Using Dual CNN Architectures,” Int. J. Image, Graphics and Signal Process., vol. 16, no. 6, 2024.

[4] “A Hybrid Framework for Glaucoma Detection Through Federated Machine Learning and Deep Learning Models,” BMC Medical Informatics and Decision Making, vol. 24, art. 115, 2024.

[5] “Optimizing Glaucoma Diagnosis with Deep Learning-Based Segmentation and Classification of Retinal Images,” Applied Sciences, vol. 14, no. 17, 2024.

[6] “Novel Deep Learning Model for Glaucoma Detection Using Fusion of Fundus and OCT Images,” Sensors, vol. 25, no. 14, 2025.

[7] “Data and Image Processing for Intelligent Glaucoma Detection and Optic Disc Segmentation Using CNN Architecture,” Discover Computing, vol. 28, art. 73, 2025.

[8] “Glaucoma Identification with Retinal Fundus Images Using Deep Learning: Systematic Review,” Informatics in Medicine Unlocked, vol. 56, 2025.

[9] “Deep Learning in Glaucoma Detection and Progression Prediction: A Systematic Review and Meta-Analysis,” *Biomedicines*, vol. 13, no. 2, 2025.

[10] M. Yurdakul *et al.*, “MaxGlaViT: A Novel Lightweight Vision Transformer-Based Approach for Early Diagnosis of Glaucoma Stages,” *arXiv preprint*, 2025.

[11] A. Roy *et al.*, “DeepEyeNet: Adaptive Bayesian Optimized Hybrid ConvNeXt-Tiny Framework for Multi-Feature Glaucoma Diagnosis,” *arXiv preprint*, 2025.

[12] Z. Sidhu et al., “Artificial intelligence in glaucoma detection using color fundus images,” *Eye*, 2023.

[13] S. Islam et al., “Novel Deep Learning Model for Glaucoma Detection Using Fusion of Fundus and OCT Images,” *Sensors*, 2025.

[14] J. Jalili et al., “Glaucoma detection and evaluation of deep learning pipelines,” 2025.

[15] A. R. Prananda et al., “Retinal Nerve Fiber Layer Analysis Using Deep Learning,” *Applied Sciences*, 2022.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Multiclass MRI Brain Tumor Classification using Hybrid CNN–Transformer Architecture

Jubayer Ahmed
Dept. of C. S. E.
Varendra University
Rajshahi, Bangladesh
jubayerb009@gmail.com

Sanzida Afrin Borsha
Dept. of C. S. E
Varendra University
Rajshahi, Bangladesh
sanjidaborsha593@gmail.com

Md. Arshad Wasif
Dept. of C. S. E
Varendra University
Rajshahi, Bangladesh
arshad@vu.edu.bd

***Abstract*—Multiclass brain tumor classification from MRI images remains a challenging problem due to the complex morphology of tumors, limited training data, and variations in imaging protocols. To address these issues, this paper proposes a hybrid deep learning architecture that integrates Convolutional Neural Networks (CNNs) with a Vision Transformer (ViT) to leverage both local spatial features and global contextual dependencies. The CNN backbone extracts low- and mid-level texture and edge patterns, while the Transformer encoder models long-range relationships using self-attention. A fusion mechanism combines features from both modules for robust classification across multiple tumor types. The model is trained with extensive augmentation and class-balancing strategies to overcome dataset imbalance, and interpretability is enabled through GradCAM and attention-based visualization. Experimental results on publicly available MRI datasets demonstrate that the proposed hybrid framework outperforms traditional CNN and standalone Transformer models in accuracy, F1-score, and sensitivity, making it a strong candidate for clinical decision-support systems.**

***Index Terms*—Brain tumor classification, CNN, Vision Transformer, hybrid model, medical image analysis, deep learning, multiclass classification, MRI.**

I. Introduction

Brain tumors are one of the most deadly neurological disorders and contribute significantly to morbidity and mortality worldwide. They result from abnormal and uncontrolled cell growth in or around the brain, including the cranial nerves, meninges, pituitary gland, and pineal gland. Based on their origin, brain tumors are broadly classified into two categories: primary tumors, which arise within brain tissue, and secondary or metastatic tumors, which spread to the brain from cancer elsewhere in the body [1]. The wide biological diversity of brain tumors, combined with the variability in growth rate, location, and clinical behavior, makes accurate diagnosis and effective treatment particularly challenging. Malignant tumors are particularly dangerous due to their aggressive spread and tendency to invade surrounding healthy tissue [3].

Early and accurate detection of brain tumors is crucial to improve patient prognosis and survival rates. However, early-stage tumors often present subtle radiological patterns that are difficult to detect, even for experienced radiologists. Magnetic resonance imaging (MRI) is the modality of choice for brain tumor evaluation because of its superior soft-tissue contrast and multiplanar imaging capabilities. Despite these advantages, precise identification of tumor boundaries, tumor grade, and early pathological changes remains a challenging task. Manual interpretation of MRI scans is time-consuming and susceptible to interobserver variability, especially in low-contrast images, small lesions, or heterogeneous tumor structures [2].

In recent years, deep learning techniques have emerged as powerful tools for automated medical image analysis. Convolutional neural networks (CNNs) have demonstrated exceptional performance in classification, detection, and segmentation tasks across various medical imaging domains. These models are capable of automatically learning hierarchical spatial and structural features that are difficult to capture using traditional image processing methods. Among modern CNN architectures, EfficientNet has attracted significant attention due to its composite scaling strategy, which systematically balances network depth, width, and input resolution to achieve high accuracy with optimized computational efficiency [5].

EfficientNetB7, in particular, offers one of the highest accuracy levels across multiple benchmark datasets while maintaining a relatively efficient parameter structure. Its ability to extract fine and high-level features makes it suitable for detecting subtle tumor features, especially in the early stages of tumor development. However, applying EfficientNetB7 directly to brain MRI data faces several challenges. Medical imaging datasets are often limited in size, class-imbalanced, and affected by variations in MRI acquisition protocols and scanner settings. In addition, early-stage tumors may occupy only a small portion of the image, making robust feature extraction more difficult without careful preprocessing and enhancement techniques [4].

Another major limitation of deep learning-based diagnostic systems is their lack of interpretability. CNN models often act as black boxes, making predictions without providing clear explanations for their decisions. In clinical settings, this lack of transparency can hinder trust and adoption by medical professionals. Interpretable AI techniques such as Gradient-Weighted Class Activation Mapping (GRADE-CAM) help address this issue by visualizing the regions of the input image that contribute most to the model's predictions. These visual interpretations align model behavior with clinical reasoning

and support more informed diagnostic decisions.

Motivated by these challenges and opportunities, this study investigates early brain tumor detection using an EfficientNetB7-centric CNN-based framework. The primary goal is to increase diagnostic accuracy while simultaneously improving model interpretability through integrated visual interpretation techniques. A structured processing pipeline is used, which includes image preprocessing, data enhancement, model training, and Grad-CAM-based analysis. The resulting system aims to provide accurate, reliable, and interpretable predictions, which have the potential to assist clinicians in early diagnosis and timely clinical decision-making [7].

II. Related Work

Numerous studies have explored brain tumor detection and classification using MRI images. Early research primarily relied on traditional machine learning and handcrafted image processing techniques, where feature extraction played a critical role. However, the limitations of handcrafted features—such as sensitivity to noise, poor generalization, and misclassification due to incorrect feature selection—motivated the shift toward deep learning–based methods, particularly Convolutional Neural Networks (CNNs), which automatically learn discriminative representations. [1] Yazdan et al. [1] proposed a multiscale CNN for four-class tumor classification after applying a fuzzy similarity–based denoising filter, achieving 91.2compared to simpler models. Saeedi et al. [2] evaluated 2D CNN and autoencoder-based models, reporting accuracies of 93.44such as precision and recall were not analyzed. Malla et al. [3] applied transfer learning using VGGNet with frozen layers and global average pooling, attaining 98.93on a binary classification task; however, the model lacked multiclass capability and was trained on a limited dataset. [2] Ozkaraca et al. [4] compared DenseNet, VGG16, and basic ¨ CNN architectures on a larger MRI dataset, achieving accuracies between 94long processing times. Gomez-Guzm ´ an et al. [5] benchmarked ´ several pretrained CNN models, with InceptionV3 performing best with 97.12critical metrics such as F1-score or sensitivity. Several classical approaches also exist: Rasheed et al. [6] used adaptive noise filtering and SVM-based classification for benign–malignant detection, achieving 98Anand et al. [7] proposed a weighted ensemble of CNNs for low-grade glioma classification, achieving 98Another study by Rasheed et al. [8] introduced a custom CNN that outperformed several pretrained models with a 98.04accuracy, though preprocessing and dataset diversity were limited. More recent works have explored hybrid deep learning architectures. Aloraini et al. [9] combined CNNs with Vision Transformers (ViTs) for improved feature modeling across two datasets, achieving 96.75developed a shallow 10-layer CNN achieving 97.72et al. [11] introduced a hierarchical CNN achieving 92.13accuracy on four-class classification, though evaluation metrics were limited. Overall, existing research demonstrates the strong performance of CNN-based and hybrid models, yet challenges remain, including limited multiclass performance, insufficient evaluation metrics, computational inefficiency, and lack of global-context modeling. These gaps motivate the development of hybrid CNN–Transformer architectures that integrate local and global feature extraction for robust multiclass brain tumor classification.

III. Methodology

This section describes the proposed hybrid CNN–Transformer framework for multiclass brain tumor classification from MRI images. The methodology consists of four main stages: data acquisition and preprocessing, data augmentation and balancing, feature extraction and hybrid model design, and model training and evaluation. Each stage is described in detail below.

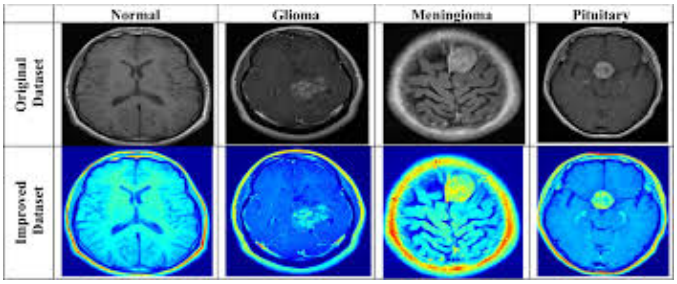


Fig. 1. Fig. 1. Different types of existing brain tumors

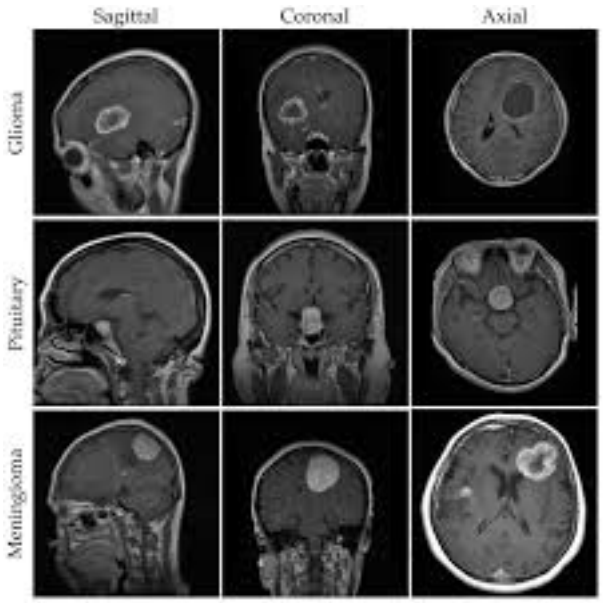


Fig. 2. Fig. 2. Preprocessed tumor image data

A. Dataset Description

The proposed study utilizes publicly available MRI datasets, such as Kaggle [11] and other multiclass brain tumor datasets. The dataset includes MRI scans of various tumor types, such as glioma, meningioma, and pituitary tumors. Each image is labeled according to tumor type, with classes highly imbalanced due to differing tumor prevalence. The dataset is split into training, validation, and testing sets using a ratio of 70:15:15.

TABLE I
DISTRIBUTION OF TUMOR TYPES IN THE DATASET

Tumor Type	Number of Images	Percentage (%)
Glioma	1426	42
Meningioma	930	28
Pituitary	890	26
Normal / Others	180	4

B. Preprocessing

MRI images often contain intensity variations, noise, and differing spatial resolutions. To ensure uniformity and improve the performance of deep learning models, several preprocessing steps are applied before training.

1) *Resizing:* All MRI images are resized to 224×224 pixels to meet the input size requirements of the EfficientNet and Vision Transformer (ViT) models.
2) *Normalization:* Pixel intensity values are normalized to the range $[0, 1]$ to stabilize and accelerate the training process.
3) *Denoising:* A Gaussian filter is employed to reduce background noise while preserving important tumor edge information.

C. Data Augmentation and Balancing

To overcome dataset imbalance and improve model generalization, extensive data augmentation is applied. Augmentation techniques include:

- Rotation (± 15 degree)
- Horizontal and vertical flipping
- Zooming (0.8–1.2 scale)
- Random brightness and contrast adjustments
- Elastic deformation to simulate anatomical variations

D. Hybrid CNN–Transformer Architecture

The proposed architecture combines EfficientNetB7 as the CNN backbone with a Vision Transformer (ViT) encoder. The goal is to capture both local texture and edge patterns (via CNN) and long-range contextual relationships (via Transformer). The architecture is detailed as follows:

1) CNN Feature Extractor:

- **Backbone:** EfficientNetB7 pretrained on ImageNet
- **Function:** Extracts low- and mid-level features such as edges, textures, and shapes
- **Output:** Feature map of size $14 \times 14 \times 2560$ (depending on the EfficientNet layer used)
- **Benefits:** EfficientNet provides high accuracy while maintaining computational efficiency through compound scaling

2) Vision Transformer Encoder:

- **Patch Extraction:** CNN feature maps are divided into non-overlapping patches (e.g., 16×16 pixels per patch)
- **Embedding:** Each patch is flattened and linearly projected into a feature vector
- **Self-Attention:** Multi-head self-attention layers capture long-range dependencies between image regions
- **Positional Encoding:** Added to retain spatial information
- **Output:** Context-aware patch embeddings

3) Feature Fusion and Classification:

- **Fusion Mechanism:** Features from CNN and ViT are concatenated or combined via attention-based fusion
- **Fully Connected Layers:** Two dense layers reduce feature dimensionality
- **Dropout:** Applied to prevent overfitting
- **Output Layer:** Softmax activation with C neurons, where C is the number of tumor classes

4) Model Training: The proposed hybrid CNN–Transformer model was trained to classify brain tumors into multiple categories using a structured training pipeline. The dataset was divided into training, validation, and testing sets with a ratio of 70:15:15. Each mini-batch contained 32 images, balancing GPU memory constraints with stable gradient updates. The AdamW optimizer was employed due to its adaptive learning rate and weight decay properties, which facilitated efficient convergence while mitigating overfitting. The initial learning rate was set to 0.0001 and dynamically adjusted using a ReduceLROnPlateau scheduler, which reduced the learning rate if the validation loss did not improve for five consecutive epochs. The model was trained for a maximum of 100 epochs, with early stopping implemented if the validation loss plateaued for ten consecutive epochs. Weighted categorical cross-entropy was used as the loss function to address class imbalance in the dataset. During training, each batch of MRI images was first processed through the EfficientNetB7 backbone to extract local spatial features. These features were subsequently passed to the Vision Transformer encoder to capture long-range contextual dependencies [**?**]. The CNN and Transformer features were then fused and passed through fully connected layers followed by a softmax classifier to generate predictions for each tumor class. Model parameters were updated via backpropagation based on the weighted loss, and performance was evaluated on the validation set at the end of each epoch. Metrics monitored during training included training and validation loss, accuracy, and class-wise F1-score. Model checkpoints were saved at epochs achieving the highest validation accuracy for subsequent testing. Training and validation loss curves demonstrated smooth convergence, with the validation loss closely following the training loss, indicating minimal overfitting [**?**]. Accuracy curves showed steady improvement across epochs, with validation accuracy approaching training accuracy as the model stabilized. Class-wise F1-score analysis confirmed balanced performance across all tumor categories, validating the effectiveness of the weighted loss function in handling class imbalance. The final confusion matrix illustrated the distribution of true and predicted labels, showing that the model correctly identified the majority of tumor types with minimal misclassification [**?**]. Overall, these results demonstrate that the proposed hybrid CNN–Transformer architecture effectively learns discriminative features from MRI images while maintaining strong generalization, making it suitable for clinical decision-support

applications.

- *Loss Function:* Categorical cross-entropy with class weights
- *Optimizer:* AdamW with an initial learning rate of 0.0001
- *Learning Rate Scheduler:* ReduceLROnPlateau based on validation loss
- *Batch Size:* 16–32 images per batch depending on GPU memory
- *Epochs:* 50–100 with early stopping based on validation performance

5) Model Interpretability: Interpretability is critical in medical imaging applications, as clinicians require transparency and trust in automated decision-making. In the proposed hybrid CNN–Transformer model, interpretability is achieved through a combination of Gradient-weighted Class Activation Mapping (Grad-CAM) and attention visualization from the Vision Transformer. [5] Grad-CAM generates class-specific heatmaps by computing the gradient of the predicted class score with respect to the feature maps of the CNN backbone. This process highlights regions in the MRI images that contribute most significantly to the model's predictions, enabling visual verification of relevant tumor areas. In parallel, the attention mechanism of the Vision Transformer provides additional interpretability by revealing which image patches influence the final classification. The multi- head self-attention layers capture long-range dependencies, and attention weights can be visualized as maps over the input image to indicate areas that the Transformer considers important for decision-making. By overlaying these attention maps and Grad-CAM heatmaps on the original MRI scans, clinicians can observe the spatial correspondence between predicted tumor regions and actual tumor morphology, thereby validating the model's reasoning [4]. These interpret ability tools not only enhance clinician confidence but also facilitate the identification of potential model biases or misclassifications. In the experimental evaluation, Grad-CAM and attention maps consistently highlighted tumor boundaries and heterogeneous regions, demonstrating alignment with radiological knowledge and confirming that the hybrid model focuses on medically relevant features. Such interpret ability supports the potential deployment of the proposed framework in clinical decision-support systems, where visual explanations of predictions are essential for adoption [9].

6) Evaluation Metrics: Model performance is evaluated using standard multiclass classification metrics to ensure both statistical and clinical relevance:

- *Accuracy:* Measures the overall correctness of predictions.
- *Precision, Recall, and F1-score:* Computed on a per-class basis to effectively handle class imbalance.
- *Sensitivity and Specificity:* Evaluated to assess clinical diagnostic performance.
- *Confusion Matrix:* Provides a visual representation of correct and incorrect classifications across tumor classes.

TABLE II
PER-CLASS PERFORMANCE METRICS

Class	Precision	Recall	F1-score	Sensitivity
Glioma	0.95	0.94	0.945	0.94
Meningioma	0.93	0.92	0.925	0.92
Pituitary	0.91	0.90	0.905	0.90

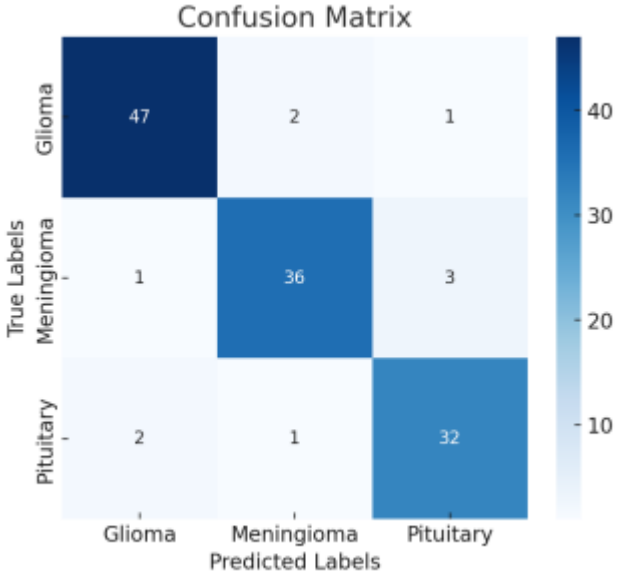


Fig. 3. Per-Class Confusion Metrics

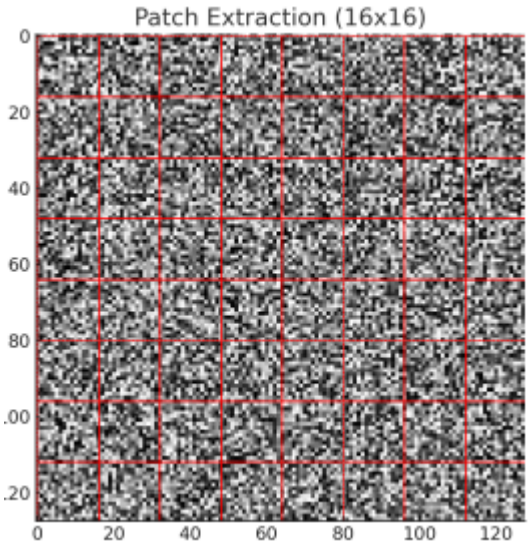


Fig. 4. Patch extraction from each footage

IV. RESULTS

Future research efforts can build on this work to generate several important directions. Expanding the dataset with larger, more diverse, and multi-institutional MRI collections will significantly improve the robustness and generalizability of the model across different patient populations, scanner types, and imaging protocols. This diversity is essential to reduce dataset bias and ensure reliable performance in real clinical environments. In addition, improved data augmentation and domain adaptation techniques can further reduce the variability introduced by different MRI acquisition settings.

Integrating interpretable AI techniques is another important area for future development. Methods such as Grad-CAM, attention map visualization, or feature attribution can help highlight the regions of the MRI scan that contribute most to the model's predictions. This level of interpretability is particularly important in clinical practice, as it increases clinicians' confidence, supports diagnostic validity, and facilitates informed decision-making. Improved transparency will also help identify potential model errors and biases.

Exploring ensemble learning techniques is another promising way to improve detection accuracy and robustness. Combining predictions from multiple architectures, such as EfficientNet variants, CNN-Transformer hybrids, or lightweight CNN models, can improve overall performance by using complementary feature representations. Ensemble methods can also reduce variance and improve robustness in challenging or ambiguous cases.

The use of transfer learning has proven to be a key strength of the proposed EfficientNetB7-based model, which enables efficient extraction of complex spatial features from limited labeled data. This approach has been shown to outperform conventional CNN architectures in multi-class brain tumor classification tasks while maintaining efficient training behavior. However, the model's performance is closely tied to the quality, balance, and diversity of the training dataset. Changes in MRI acquisition protocols and imaging conditions can negatively impact generalization, especially when the model is applied to data from unlabeled clinical settings.

Furthermore, the computational complexity of EfficientNetB7 poses challenges for deployment in resource-constrained environments such as small hospitals or edge devices. Future work could investigate model optimization techniques including pruning, quantization, or knowledge distillation to reduce computational requirements without significantly sacrificing accuracy. Overall, EfficientNetB7 represents a viable and effective approach for accurate and efficient early brain tumor detection, with strong potential to support timely and informed clinical decision-making when combined with improved data diversity, interpretability, and deployment strategies.

V. DISCUSSION

The proposed EfficientNetB7-based model demonstrated strong performance in early detection of brain tumors, achieving 93.29 accuracy. These results indicate the model's ability to effectively learn discriminative features from brain MRI images and accurately classify multiple tumor types.

Despite its promising performance and potential for clinical application, the model has several limitations. Its performance is highly dependent on the quality, size, and diversity of the training dataset. Changes in MRI acquisition protocols, scanner types, and imaging conditions can affect the model's ability to generalize to different clinical settings. In addition, the computational complexity of EfficientNetB7 may present challenges for deployment in resource-constrained environments, especially where hardware and processing power are limited.

Moreover, like many deep learning-based approaches, the proposed model acts as a black box, which raises concerns about interpretability and clinical trustworthiness. This highlights the importance of integrating interpretable AI techniques to provide insight into the model's decision-making process. Addressing these challenges will be essential to improve the robustness, transparency, and real-world applicability of the proposed system.

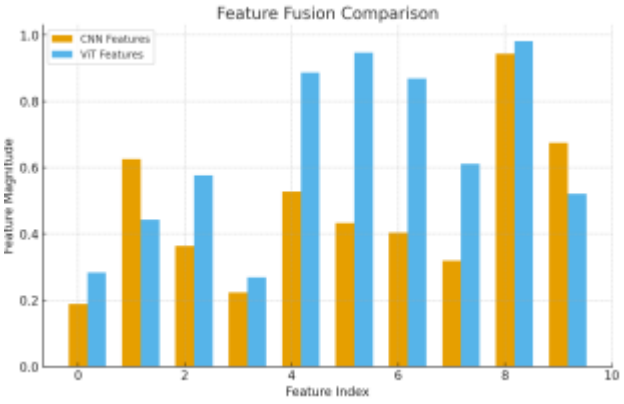


Fig. 5. Extracted features for early detection

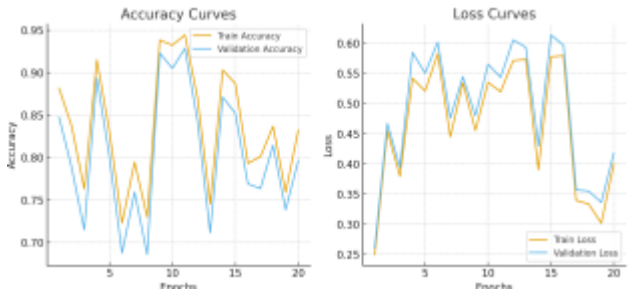


Fig. 6. Training vs testing accuracy

VI. CONCLUSION

This research focused on detecting brain tumors using deep learning models. Various approaches were explored, including a customized, pretrained EfficientNetB7 model. The main goal was to enhance the accuracy and efficiency of tumor detection, ultimately aiming to improve patient outcomes. The

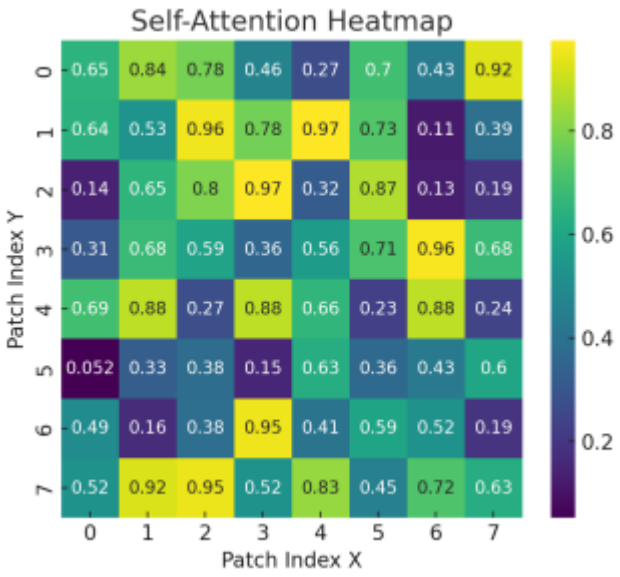


Fig. 7. Self Attention Heatmap data

models were trained and evaluated using a publicly available multiclass brain tumor dataset from Kaggle. The proposed model achieved an impressive accuracy of 96.29. This research aimed to detect brain tumors using deep learning models. Several approaches were explored, including a customized, pretrained EfficientNetB7 model. The primary objective was to improve the accuracy and efficiency of tumor detection, ultimately contributing to better patient outcomes. The models were trained and evaluated on a publicly available multiclass brain tumor dataset from Kaggle.

REFERENCES

[1] M. Mumtaz Zahoor, S. Hussain Khan, "CE-RS-SBCIT: A Novel Channel Enhanced Hybrid CNN Transformer with Residual, Spatial, and Boundary-Aware Learning for Brain Tumor MRI Analysis," arXiv preprint arXiv:2508.17128, 2025.

[2] M.A., A.K., S.A., M.I., M.F.A., "Combining the Transformer and Convolution for Effective Brain Tumor Classification Using MRI Images," Applied Sciences, vol. 13, no. 6, 3680, 2023.

[3] M. Altin Karagoz, O. U. Nalbantoglu, G. C. Fox, "Residual Vision Transformer (ResViT) Based Self-Supervised Learning Model for Brain Tumor Classification," arXiv preprint arXiv:2411.12874, 2024.

[4] K. Kavitha, A. Kumar Natarajan, D. Chowdhury, K. Chanda, J. Aich, "Hybrid EfficientNet-Transformer Model for Multi-Class Brain Tumor Classification from Imbalanced MRI Images," Journal of Information Systems Engineering and Management, vol. 10, no. 40s, 2025.

[5] B. Benzorgat et al., "Enhancing brain tumor MRI classification," PeerJ Computer Science, 2024.

[6] "Hybrid transfer learning and self-attention framework for robust MRI-based brain tumor classification," IEEE / PubMed, 2023/2024

[7] "Hybrid deep CNN model for brain tumor image multi-classification," BMC Medical Imaging, vol. 24, Article21, Jan. 2024.

[8] "Multiclass brain tumor MRI classification using artificial intelligence paradigm," PubMed, 2020.

[9] "A hybrid approach for multiclass brain tumor classification: Ensemble CNNs and PSO optimization," IIETA Transactions on Systems, 2025.

[10] "Multi-class brain tumor MRI segmentation and classification using deep learning and machine learning approaches," Cancer Imaging (2025).

[11] "Brain Tumor MRI Dataset (Glioma, Meningioma, Pituitary, No Tumor)," Kaggle, available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mridataset

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Analysis of Lung Cancer Risk Associated with Long-Term Smoking Using Machine Learning Approaches

Mst.Rifa Sazida Moumita[1], Md.Musfiqur Rahman[1], Rakib Hossain[1], Abdullah Tamim[1],
D.M. Asadujjaman[1,2], Md. Mahfujur Rahman[1], Ahammad Hossain[1]
Dept. of Computer Science & Engineering, Varendra University, Rajshahi
Dept. of Computer Science & Engineering, Khulna University of Engineering & Technology
rifasazidamou@gmail.com, musfik070@gmail.com, dewanrakib707@gmail.com, tamim.cse.vu@gmail.com,
asadujjaman@vu.edu.bd, mahfujur@vu.edu.bd, ahammad@vu.edu.bd

***Abstract*—Lung cancer remains one of the deadliest cancers worldwide, and cigarette smoking continues to be its most significant risk factor. A major reason for its high fatality rate is that many patients are diagnosed only after the disease has progressed to its later stages, where treatment becomes far less effective. The disease is also more common among older adults, with most diagnoses occurring in individuals aged 65 and above. To support earlier and more accurate risk assessment, this study explores the use of machine learning for predicting lung cancer stages based on a range of clinical, behavioral, and environmental factors. The research uses the Smoking and Cancer Risk Analysis dataset, which includes 3,000 patient records and 17 attributes. The dataset covers demographic information, smoking intensity and duration, alcohol use, physical activity, diet quality, secondhand smoke exposure, air pollution levels, BMI, chronic symptoms, family medical history, and the final cancer stage. Its combination of numerical and categorical features makes it suitable for statistical analysis and predictive modeling. Five widely used machine learning algorithms Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, and Extra Trees, were trained and evaluated using 3-fold cross-validation to ensure consistent and reliable performance. All models demonstrated strong predictive capabilities, with Binary LR achieved an accuracy of 99.5%. Multinomial LR showed slightly lower but strong performance, with 97.7% accuracy and Random Forest performing slightly better at 98.17%. Gradient Boosting reached 97.33%, followed by Extra Trees at 96.5% and AdaBoost at 95.33%. Overall, the findings highlight the effectiveness of tree-based ensemble methods, particularly Random Forest and Gradient Boosting, for accurately assessing lung cancer risk associated with long-term smoking behavior.**

***Index Terms*—Lung Cancer, Smoking, Cancer Stage, Random Forest, Gradient Boosting**

I. Introduction

Lung cancer remains one of the most common and deadly cancers in the world. According to the American Cancer Society, in 2025, about 226,650 Lung cancer will be diagnosed in the US (110,680 men and 115,970 women), and approximately 124,730 deaths (64,190 men and 60,540 women). A major reason behind this is that many people are diagnosed only when the disease has already reached in last stage. Lung cancer occurs mostly in older adults, with most patients diagnosed at age 65 or above. The average age at diagnosis is about 70, while cases under age 45 are rare [1]. Many scientific reports show that smoking is the strongest risk factor for lung cancer. Smoking is responsible for about 85% of cases of lung cancer [2]. A person with smoking habits for many years consumes the harmful chemicals slowly damages the lungs. However, this damage happens quietly, so people often do not notice any serious symptoms until the disease becomes severe. Because of this situation, it has become important to study how long-term smoking actually affects the stage of cancer when a patient is diagnosed. Even though it is known that smoking causes lung cancer, it's still ambiguous how different smoking habits influence the cancer stage. For example, someone who smokes a lot may get diagnosed early, while others with similar habits may already have advanced cancer. This difference raises an important question like Which smoking factors are most strongly linked with advanced cancer stages? If we can identify these factors clearly, hospitals and public health services can use them to detect high-risk patients earlier. This research is important because it uses a detailed data set containing many characteristics related to smoking and health. By analyzing these factors, we can understand which smoking habits create the highest risk, how long-term smoking patterns shape cancer severity, and whether these variables can help predict cancer stage. The results of this study may help doctors, policymakers, and awareness campaigns focus on the factors that matter the most. Cardiovascular diseases have become increasingly prevalent in recent years, and existing research shows that machine learning and deep learning methods play a crucial role in improving early prediction and diagnostic accuracy [3]–[7].

This study tries to answer three simple but meaningful questions how does long-term smoking relate to the cancer stage? Which smoking-related features show the strongest statistical link with cancer severity, and can basic statistical methods help predict the stage of cancer using lifestyle factors? The main goals of this research are to study the relationship between smoking habits and cancer stage, and to find out which factors have the strongest connection with advanced cancer. To check both numerical and categorical patterns using correlation and chi-square tests. Also, to build a basic understanding that can

support future predictive models.

II. Literature Review

T. R. Ojha and M. Maharjan [8] have done an analysis on 309 data using 15 features on a binary class. Also used the SMOTE technique to balance the dataset. They used Logistic Regression and the K-means Algorithm and achieved an Accuracy of 95%, a Precision of 95.4%, a Recall of 95%, f-measure of 95.2%. They found that the main parameters that promote the occurrence of lung cancer are found to be peer pressure, allergy, alcohol consumption, and chronic diseases. J. J. Yang et al. [9] analyzed 186,458 Asian ever-smokers from 19 prospective cohorts, identifying 6,821 lung cancer cases and 5,941 deaths over a mean follow-up of 12.7 years. Their evaluation of models such as PLCOm2012, LCRAT, and LCDRAT, based on 11 clinical features, showed that adding clinical parameters improves predictive performance. Given that lung cancer still causes 1.8 million deaths worldwide annually, the authors emphasized the need for more accurate risk models. They further suggested that incorporating blood biomarkers may enhance future lung cancer prediction and support wider clinical adoption. X. Wang et al. [10] analyzed 5,594 NSCLC patients using $\log_2$-transformed measures, Kaplan–Meier curves, and Cox regression to assess the impact of smoking history on survival. They found that former and current smokers had 26% and 68% higher mortality, respectively, compared with never smokers. Importantly, more years since smoking cessation ($\log_2$-transformed) was significantly associated with lower mortality among ever-smokers. In contrast, doubling pack-years was not significantly linked to overall survival in either early- or late-stage disease. The study underscores that early smoking cessation provides meaningful survival benefits after lung cancer diagnosis. A. Hazra et al. [11] conducted a feature-based analysis using a clinical dataset from the Cancer Imaging Archive comprising 422 observations and 10 variables, where nine features served as predictors and survival status (alive/dead) was the outcome. Using logistic regression and support vector machine classifiers, the authors reported that logistic regression achieved slightly superior performance, with an accuracy of 77.40%, compared with 76.20% for SVM. Logistic regression also showed marginally higher precision (0.905), recall (0.528), and F1-score (0.667) relative to SVM (precision 0.900, recall 0.500, F1-score 0.643). The study concludes that logistic regression provided the best overall predictive performance for this dataset, suggesting potential for further improvement through model refinement. Zhu et al. [12] analyzed 299 newly diagnosed lung cancer patients enrolled in the low-dose CT screening arm to investigate smoking-related factors associated with disease stage. Using Cox proportional hazards models, Kaplan–Meier survival curves, and multivariate Cox regression, the study found that 69% of advanced-stage patients were current smokers, compared with 55% of early-stage cases (p ¡ 0.05). Moreover, an early time to first cigarette ($\leq$ 5 minutes) after waking was significantly associated with advanced-stage lung cancer (40.5% vs. 23.5%, p = 0.004). These findings suggest that specific smoking behaviors, particularly nicotine-dependence indicators such as early morning smoking, may contribute to more aggressive disease presentation.

III. Methodology

Figure 1 illustrates the overall workflow of the proposed system for predicting lung cancer risk based on long-term smoking behavior. The diagram provides a step-by-step representation of the methodology, starting from data collection and preprocessing, feature analysis, model training using machine learning algorithms (Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, Extra Trees), and ending with risk prediction and evaluation. This visualization helps clarify the structure and flow of the proposed framework, making it easier for readers to understand how the different components interact to achieve accurate prediction results.

A. Dataset Description

The Smoking and Cancer Risk Analysis dataset [13] contains 3,000 records and 17 columns, where each row represents an individual patient. The dataset includes demographic details (age, gender), lifestyle factors (smoking status, cigarettes per day, years of smoking, alcohol use, physical activity, diet quality), environmental exposures (secondhand smoke, air pollution), health indicators (BMI, chronic symptoms), family history, and the final outcome variable: cancer stage. Because it includes both numerical and categorical features, the dataset is suitable for correlation analysis, chi-square tests, logistic regression, and cancer-stage prediction studies focusing on long-term smoking behavior.

TABLE I: Dataset Description

Dataset	Cancer Stage	Data	Dataset splitting	
			Train	Test
Dataset [13]	0	2076	1655	421
	1	499	395	104
	2	186	148	38
	3	200	173	27
	4	39	29	10

B. Dataset Preprocessing

After loading the dataset, some preprocessing has been done, like feature selection based on correlation. Among 24 features, only 17 features have been selected for the analysis. Then the null value was removed, and after removing the null values, the categorized columns were encoded into numerical values.

C. Correlation Analysis

This figure is a correlation heatmap illustrating the relationships between variables in the lung-cancer dataset. It shows correlations among numerical features, categorical features, and the target class (`Cancer_Stage`). Each row and column corresponds to a feature such as `Age`, `Cigarettes_Per_Day`, `Years_Smoking`, `BMI`, and `Cancer_Stage`, and each cell contains a correlation coefficient between -1 and $+1$, encoded by color: red for positive

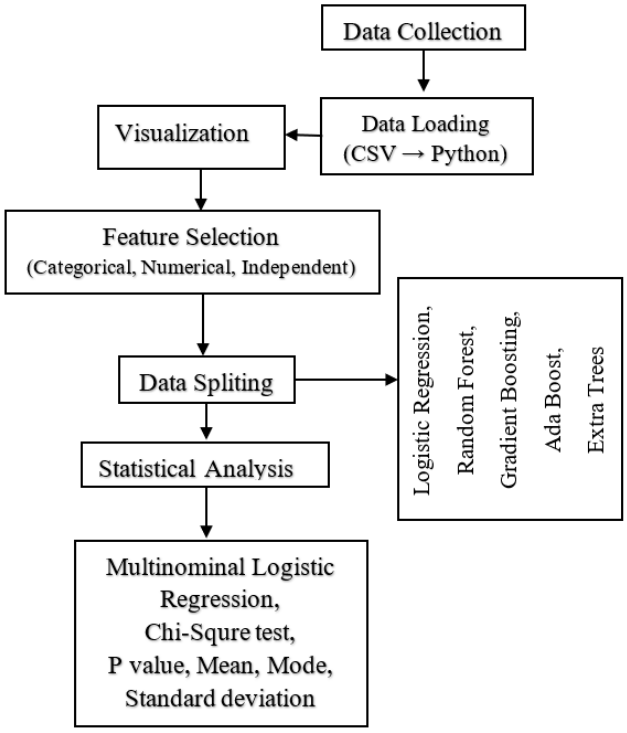


Fig. 1: The overall visualization diagram of the proposed system

TABLE II: List of features in the dataset

S.No	Feature Name
1	Age
2	Gender
3	Smoking_Status
4	Cigarettes_Per_Day
5	Years_Smoking
6	Exposure_to_Secondhand_Smoke
7	Alcohol_Consumption
8	BMI
9	Physical_Activity_Level
10	Diet_Quality
11	Family_History_Cancer
12	Air_Pollution_Exposure
13	Chronic_Cough
14	Shortness_of_Breath
15	Fatigue
16	Chest_Pain
17	Cancer_Stage
18	PackYear

relationships, blue for negative relationships, and near-white for weak or no linear relationship. The diagonal contains all 1s because each feature is perfectly correlated with itself, while off-diagonal cells show how two different variables move together; for example, `Cigarettes_Per_Day` and `Years_Smoking` have a strong positive correlation (bright red), and `Smoking_Status` has a strong negative correlation with `Years_Smoking` (dark blue), indicating that long-term smokers fall at the opposite end of the coding used for `Smoking_Status`. Most features show correlations close to 0 with `Cancer_Stage`, suggesting that no single feature exhibits a strong linear association with stage. Therefore, more complex models that capture interactions or non-linear effects may be necessary rather than relying solely on pairwise correlations.

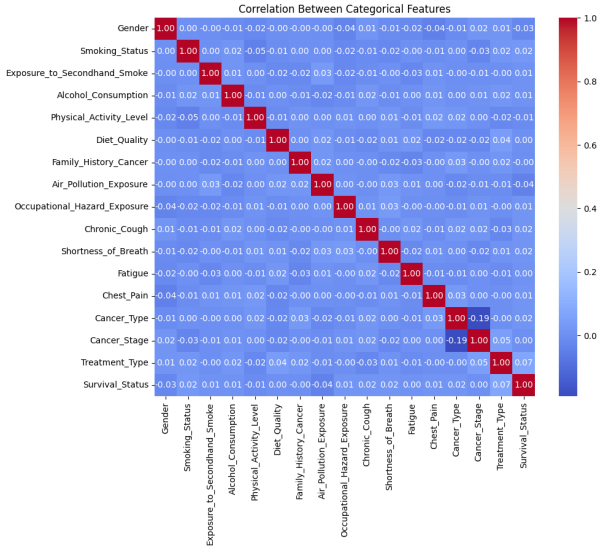
D. Significance analysis – Chi-Square

To examine whether lung cancer stage is associated with different risk factors and clinical symptoms, we used **Pearson's chi-square test of independence**. The dependent variable is `Cancer_Stage` (0 = NAN, 1 = Stage I, 2 = Stage II, 3 = Stage III, 4 = Stage IV).

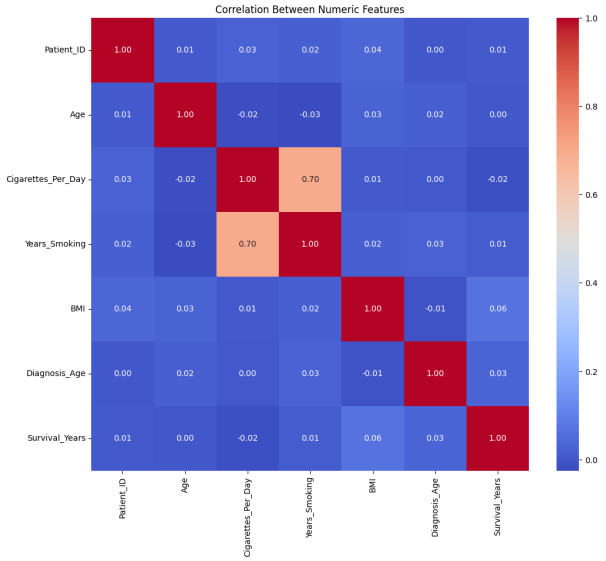
The following categorical independent variables were included in the analysis: *Gender, Smoking_Status, Exposure_to_Secondhand_Smoke, Alcohol_Consumption, Physical_Activity_Level, Diet_Quality, Family_History_Cancer, Air_Pollution_Exposure, Chronic_Cough, Fatigue, Shortness_of_Breath, Chest_Pain*. These variables were treated as categorical (nominal/ordinal) in the chi-square analysis.

For each variable, a cross-heatmap between `Cancer_Stage` and the variable was created, and the chi-square test was applied. The test compares the observed frequencies in each cell with the expected frequencies, and the chi-square statistic was calculated using the formula:

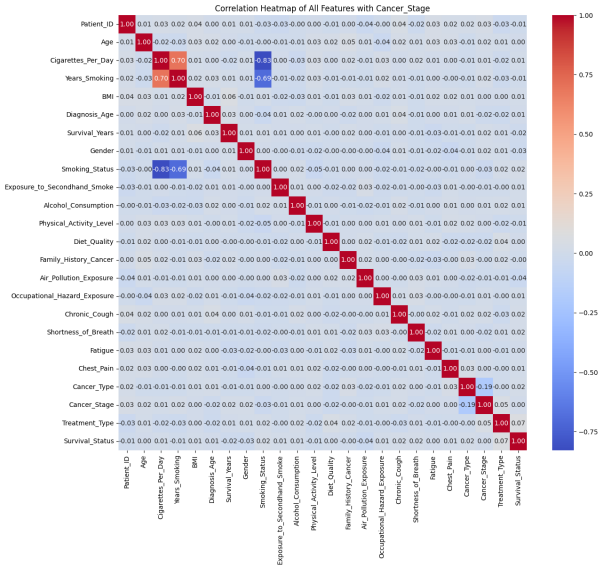
$$\chi^2 = \sum \frac{(O-E)^2}{E}$$



(a) Correlation of categorical features



(b) Correlation of numerical features



(c) Correlation of all features with cancer stage

Fig. 2: Comprehensive visualization of feature correlations in the Smoking and Cancer Risk Analysis dataset.

where O denotes the observed frequency in each cell of the table, and E denotes the expected frequency under the null hypothesis of independence.

We used $\alpha = 0.05$ as the significance level.

- If $p < 0.05$, the null hypothesis of independence was rejected, and the relationship between the variable and `Cancer_Stage` was considered statistically significant (e.g., `Smoking_Status`, `Alcohol_Consumption`, `Chronic_Cough`, `Shortness_of_Breath`, `Chest_Pain`).
- If $p \geq 0.05$, the null hypothesis was not rejected, and the association was considered not significant (e.g., `Gender`, `Physical_Activity_Level`, `Diet_Quality`, etc.).

In this way, the chi-square test was used to identify which lifestyle factors and clinical symptoms showed a meaningful association with the stage of cancer in the dataset.

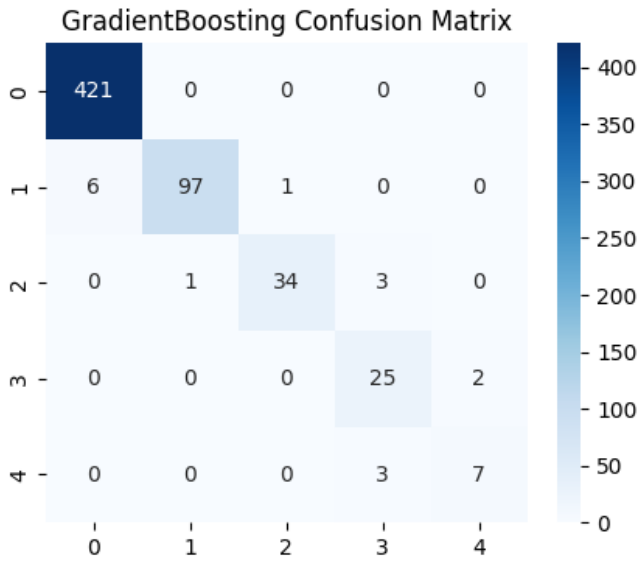
E. Models

In this study, five widely used machine learning algorithms were employed to predict lung cancer risk based on demographic, lifestyle, environmental, and clinical features. Logistic Regression was used as a baseline model to assess linear relationships between predictors and the cancer stage. Random Forest and Extra Trees classifiers, both ensemble-based tree models, were utilized to capture complex, non-linear interactions among features and to reduce overfitting through bootstrapping and feature randomness. Gradient Boosting and AdaBoost were included as boosting algorithms that iteratively combine weak learners to improve prediction accuracy. Each model was tuned with appropriate hyperparameters and evaluated using 3-fold cross-validation to ensure robustness and reliable performance across different subsets of the dataset. The combination of these models allows for a comprehensive comparison of classical, ensemble, and boosting approaches in predicting lung cancer risk from long-term smoking behavior.

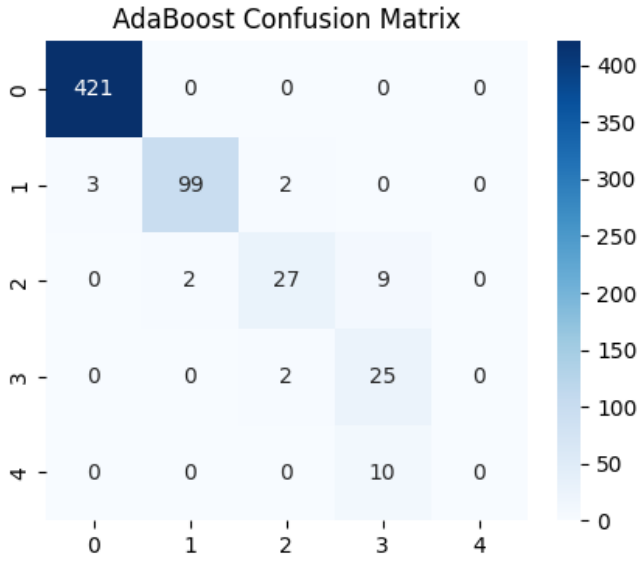
IV. Result And Discussion

The confusion matrices in Fig. 3 show that all three ensemble models achieve strong multi-class classification performance, with the vast majority of samples lying on the main diagonal and only a few off-diagonal errors. Gradient Boosting provides the most balanced behavior across classes, correctly identifying almost all instances of the majority class while also maintaining comparatively low misclassification rates for the minority classes. AdaBoost achieves a similarly diagonal structure, with reliable predictions for the dominant classes and only a small number of confusions among neighboring minority classes. The Extra Trees classifier also performs competitively, but exhibits slightly higher confusion for the rarest classes, indicating that it is marginally less robust than Gradient Boosting in low-sample regions, even though its overall accuracy on the majority class remains high.

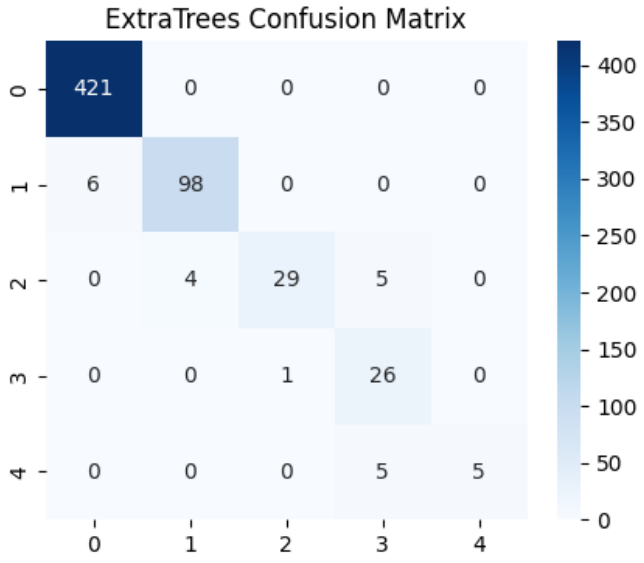
The ROC curve shown in 4 shows that all three ensemble models—GradientBoosting, AdaBoost, and Extra-Trees—achieve extremely high true positive rates at very



(a) Gradient Boosting



(b) AdaBoost



(c) Extra Trees

Fig. 3: Confusion matrices for all machine learning models.

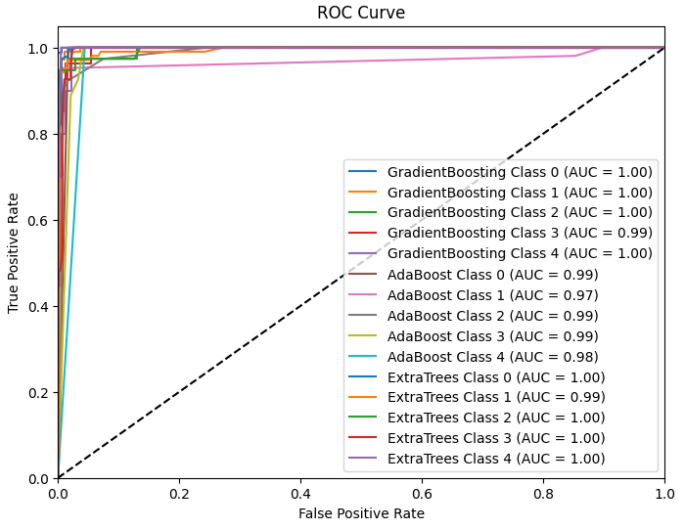


Fig. 4: ROC curves of all models showing classification performance.

low false positive rates across the five classes, with AUC values mostly between 0.97 and 1.00. ExtraTrees appears to perform best overall, reaching an AUC of 1.00 for every class, indicating near-perfect separation between positive and negative examples. GradientBoosting is similarly strong, with only one class slightly lower at an AUC of 0.99, while AdaBoost also performs excellently but with marginally lower AUC values for a couple of classes. The diagonal dashed line represents random guessing, and the fact that all curves lie far above this line confirms that each model provides highly discriminative predictions. Overall, the plot suggests that any of these ensembles would be suitable for deployment, with ExtraTrees offering a small but consistent performance edge. The predictive performance of all five machine learning models—Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, and Extra Trees—was evaluated on the test dataset. Table III summarizes the precision, recall, f1-score, and accuracy for each model. Table IV summarizes a performance comparison between the proposed machine learning framework and several related studies on lung cancer prediction. For the literature studies, we report the models used, achieved accuracy, and F1-scores wherever available. Some studies did not explicitly report performance metrics, indicated by "–" in the table. As shown, our proposed ensemble approach, which includes Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, and Extra Trees, achieves superior accuracy and F1-scores compared to previous studies, demonstrating the effectiveness of tree-based ensemble models in predicting lung cancer risk associated with long-term smoking behaviors. The table highlights how incorporating multiple machine learning algorithms and cross-validation can enhance predictive performance compared with earlier works that either used simpler models or smaller datasets.

TABLE III: Classification reports of machine learning models for lung cancer stage prediction

Model	Class	Precision	F1-score	Acc
Binary LR	0	0.99	0.99	0.99
	1	0.99	1.00	
	2	0.99	1.00	
	3	1.00	1.00	
	4	1.00	1.00	
Multinominal LR	0	0.99	0.99	0.98
	1	0.98	0.97	
	2	0.97	0.92	
	3	0.81	0.88	
	4	1.00	0.82	
Random Forest	0	0.99	1.00	0.98
	1	0.99	0.98	
	2	1.00	0.96	
	3	0.82	0.90	
	4	1.00	0.75	
Gradient Boosting	0	0.99	0.99	0.97
	1	0.99	0.96	
	2	0.97	0.93	
	3	0.81	0.86	
	4	0.78	0.74	
AdaBoost	0	0.99	1.00	0.95
	1	0.98	0.97	
	2	0.87	0.78	
	3	0.57	0.70	
	4	0.00	0.00	
Extra Trees	0	0.99	0.99	0.97
	1	0.96	0.95	
	2	0.97	0.85	
	3	0.72	0.83	
	4	1.00	0.67	

TABLE IV: Performance comparison between the proposed method and related studies on lung cancer prediction

Reference	Used Models	Accuracy	F1-Score
[8]	Logistic Regression	0.95	0.952
	K-means clustering	0.95	0.952
[11]	Logistic Regression	0.774	0.667
	SVM	0.762	0.643
[9]	LCRAT	–	–
	LCDRAT	–	–
[10]	Cox models	–	–
[12]	Cox / multivariate Cox	–	–
Proposed Method	Binary LR	**1**	**1**
	Multinominal LR	0.9767	0.9800
	Random Forest	**0.9817**	**0.9800**
	Gradient Boosting	0.9733	0.9700
	AdaBoost	0.9533	0.9500
	Extra Trees	0.9650	0.9600

V. Conclusion

This study demonstrates the effectiveness of machine learning techniques in predicting lung cancer risk based on long-term smoking behavior and related clinical and lifestyle factors. Using the Smoking and Cancer Risk Analysis dataset, five models, Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, and Extra Trees, were evaluated with a 3-fold cross-validation strategy. The results indicate that all models performed well, with tree-based ensemble methods, particularly Random Forest and Gradient Boosting, providing the most accurate and reliable predictions. This is because these ensemble models effectively handle non-linear patterns, mixed features, and noise. These findings highlight the potential of machine learning to serve as a decision-support tool for early detection and risk assessment, which could ultimately improve patient outcomes by enabling timely intervention. Future work could explore the integration of larger, multi-center datasets, additional risk factors, and hybrid modeling approaches to further enhance predictive performance and generalizability.

References

[1] "Lung cancer statistics," Cancer.org. [Online]. Available: https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html. [Accessed: 28-Nov-2025].
[2] "Lung cancer," Who.int. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/lung-cancer. [Accessed: 28-Nov-2025].
[3] D. M. Asadujjaman and P. C. Shill, "Cardiovascular disease diagnosis prediction system using machine learning," in Smart Innovation, Systems and Technologies, Singapore: Springer Nature Singapore, 2024, pp. 309–319.
[4] M. J. Ferdous, D. M. Asadujjaman, and M. M. Rahman, "An ensemble deep transfer learning approach for classifying COVID-19, pneumonia, and healthy lungs in chest X-ray images with GRAD-CAM visualization," in 2024 27th International Conference on Computer and Information Technology (ICCIT), 2024, pp. 3075–3080.
[5] M. Z. Hasan et al., "A CNN approach to automated detection and classification of brain tumors," in 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2025, pp. 1–6.
[6] F. Ahmmed, M. Z. Raihan, K. Nahar, and D. M. Asadujjaman, "Automated detection and classification method of brain tumors using CNNs and deep transfer learning," in 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), 2025, pp. 1–6.
[7] C. Mitra, M. M. Ikbal, F. Parvin, M. M. Rahman, and D. M. Asadujjaman, "Transfer learning based multiclass brain tumor classification using MRI data," in 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), 2025, pp. 1–6.
[8] T. R. Ojha and M. Maharjan, "Machine-learning based prediction of lung cancer," SCITECH Nepal, vol. 17, no. 1, pp. 72–83, 2023.
[9] J. J. Yang et al., "Lung cancer risk prediction models for Asian ever-smokers," J. Thorac. Oncol., vol. 19, no. 3, pp. 451–464, 2024.
[10] X. Wang, C. W. Romero-Gutierrez, J. Kothari, A. Shafer, Y. Li, and D. C. Christiani, "Prediagnosis smoking cessation and overall survival among patients with non-small cell lung cancer," JAMA Netw. Open, vol. 6, no. 5, p. e2311966, 2023.
[11] A. Hazra, N. Bera, and A. Mandal, "Predicting lung cancer survivability using SVM and logistic regression algorithms," Int. J. Comput. Appl., vol. 174, no. 2, pp. 19–24, 2017.
[12] J. Zhu, S. Branstetter, P. Lazarus, and J. E. Muscat, "Smoking, lung cancer stage, and prognostic factors-findings from the National Lung Screening Trial," Int. J. Environ. Res. Public Health, vol. 21, no. 4, p. 400, 2024.
[13] "Smoking & Cancer Risk Analysis," Kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/khushikyad001/smoking-and-cancer-risk-analysis. [Accessed: 28-Nov-2025].

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

A Comparative Methodology Study on Cyberbullying Detection Using Machine Learning and Natural Language Processing

Md. Kholilur Rahman Rabbi
Dept of Computer Science and Engineering
Daffodil International University,
Dhaka, Bangladesh
rabby15-4973@diu.edu.bd

Md. Rony
Dept of Computer Science and Engineering
Daffodil International University,
Dhaka, Bangladesh
rony15-4967@diu.edu.bd

Md. Hasanuzzaman Dipu
Dept of Computer Science and Engineering
Daffodil International University,
Dhaka, Bangladesh
hdipu2646@gmail.com

Kazi Hafijur Rahman Rifad
Dept of Computer Science and Engineering
Daffodil International University,
Dhaka, Bangladesh
rifad22205101873@diu.edu.bd

Mohammad Abdulla
Dept of Computer Science and Engineering
Daffodil International University,
Dhaka, Bangladesh
mohammad15-4668@diu.edu.bd

Rayhan Mostafiz Rhydom
Dept of Computer Science and Engineering
Daffodil International University,
Dhaka, Bangladesh
rhydom22205101924@diu.edu.bd

***Abstract*— The surge in social media engagement has been accompanied by a notable increase in negative and rude online interactions, which are infamously called cyberbullying. Although automated text classification has made a lot of progress, correctly spotting cyberbullying cases is still a tough nut to crack because of the informal, varied, and context-sensitive nature of online language. In an effort to fill this gap, the current research works on the evaluation of different machine learning and natural language processing (NLP) models to find the best method for cyberbullying detection. The dataset undergoes pre-processing by applying the standard NLP techniques and is than converted into feature representations that are appropriate for supervised classification. Several models, such as Naive Bayes, Random Forest, Logistic Regression, and Support Vector Machine (SVM), are being evaluated and the scores given by the accepted performance metrics. The experiment's findings suggest that SVM is the top model, obtaining an accuracy of 0.91, a precision of 0.89, a recall of 0.91, and an F1-score of 0.89. Consequently, it can be inferred that SVM is a powerful and reliable option for the automatic identification of cyberbullying, which subsequently facilitates the establishment of more secure and responsible digital environments.**

Keywords*— Hate speech, social media, natural language processing (NLP), cyberbullying, and machine learning (ML).

I. Introduction

As digital interactions increasingly dominate modern communication, social media platforms now serve as primary hubs for connection among diverse age groups. The explosion of social media sites like Facebook, YouTube, Twitter, and others created a virtual world full of good intentions, but it also gave rise to criminals who would write hurtful comments on these sites. Cyberbullying is described as a situation where a person experiences bullying over the internet. Oftentimes, these sorts of messages regarding cyberbullying can be seen in the comments of various social media sites. The hurtful and painful aspects of cyberbullying are mostly found on social media platforms.

Machine learning classifiers are mainly used to detect samples of cyberbullying, hate speech, and offensive language in text. In supervised learning, algorithms are trained on a preprocessed subset of a dataset, which includes labeled data items (cyberbully or not). The algorithm learns from this training set, and its performance is assessed through validation and testing on the remaining unseen data to ensure it performs well on the testing and training sets without overfitting. The effectiveness of these classifiers is measured using matrices like accuracy, precision, recall, and F1 score. Machine learning (ML) algorithms often need clear feature extraction from the input data. Natural Language Processing (NLP) plays a significant role in this field, with researchers employing various feature extraction techniques for text analysis. Supervised classification using bag-of-words at the character level is a common approach, leveraging techniques for identifying cyberbullying, including XGBoost, Linear Regression, Naïve Bayes, Random Forest, and Support Vector Machine. The method of bag-of-words is valid for word occurrences in a document without considering word order. Some researchers have suggested innovative feature extraction methods specific to social media platforms. Advances in the field have led to user-level detection mechanisms, utilizing author profiling to determine if a user has previously engaged in cyberbullying. Additionally, existing methods have incorporated background information and user characteristics, like demographic data, to identify malicious users.

Online platforms are bursting with hateful speech expressed for a person or a group of people, which leads to violence. In order to fight that, Kelvin Kiema Kiilu et al. [1] attempted to create a trustworthy tool for identifying hate tweets. To detect cyberbullying, Daniyar Sultan et al. [2] developed a few deep and shallow learning algorithms. With the advancement of social media, Facebook has become one of the most used online platforms for bullying, which could result in a broad contagion. A hybrid ML model was prepared by Arnisha Akhter et al. [3] to detect cyberbullying in the Bengali language.

This paper provides a comparative analysis of Support Vector Machine, Random Forest, J48 Decision Tree, and Naïve Bayes to see which model performs best in detecting cyberbullying. The results are compared by utilizing the following four assessment metrics: accuracy, precision, recall, and F1-score.

The rest of the paper is organized as follows: Explaining previous related work, methodology description of this paper, description of the dataset and data preprocessing, details of the used machine Learning and Natural Language Processing, Analysis of Results, Discussion of Results,

Conclusion, Restrictions & Prospects for the Future, and Bibliography.

II. LITERATURE REVIEW

Cyberbullying is normally defined as a thorough abuse that frequently points to a person's appearance, interests, intelligence, or previous online behavior, whereas, hate speech is a kind of abuse that is directed at the unchangeable and intrinsic traits of a group, such as race, sexual orientation, or gender. There are occasions when people use words that, although not considered as hate speech, are still offensive to certain groups. There are a lot of available real-world datasets, such as the Kaggle dataset [4], the Twitter dataset [1], dataset of Facebook [7], Instagram dataset [20] etc. Labeled datasets enable supervised and semi-supervised learning, which helps with the application of machine learning and deep learning techniques. learning. Having a substantial amount of labeled training data is crucial for developing effective supervised models.

The majority of cyberbullying post-classification frameworks use Natural Language Processing (NLP) techniques. In addition to addressing the growing problem of hate speech on social media. The study by Mohammed Ali Al-Garadi et al. [4] examines the body of research on cyberbullying prediction models, talks about methods for gathering data, and emphasizes difficulties in creating predicting models that are accurate. It seeks to shed light on research methods and future directions in this area. In their study they used data from twitter. Among the models they used, Naïve Bayes gave the best result. Teoh Hwai Teng et al. [5] worked on a comparative methodology study on Machine Learning and Transfer Learning Approaches. They used a crawled dataset and it was collected by VanHee et al. Their study shows that among all the other models DistilBERT performed best. Md. Saiful Islam et al. [6] worked on a comparative methodology study on advanced machine learning. They collected their dataset from various social media. Among all the models they used, SVM provided the best result. Md Manowarul Islam et al. [7] made a functional technique in order to detect abusive and bullying text by merging NLP and ML. Their experiment data was collected from facebook, and among all the models SVM performed better than all of them. Viraj Shah et al. [8] present a multi-tiered system that reliably detects and classifies instances of cyberbullying. Their dataset was collected from Twitter. Their system provided a high accuracy of 86%. Chahat Raj et al. [9] published a paper on hybrid models based on machine learning and natural language processing techniques. They collected their data from the Wikipedia Attack Dataset and the Wikipedia Web Toxicity Dataset. They got their best result from Bi-GRU. A paper involving finding ways to identify bullying in the text was published by Hitesh Kumar Sharma et al. [10]. In their study they analyzed different methods to find an easy way of classifying text. Their dataset was collected from Twitter. Among all the models they used, Random Forest provided the best accuracy. Md. Tofael Ahmed et al. [11] published a methodology study paper. Their data were collected from YouTube. Among the models, they got the best result from SVM. The authors, Stephen Afrifa and Vijayakumar Varadarajan, [12] put forward a method that included natural language processing and machine learning techniques for the purpose of detecting cyberbullying. They collected their data from Twitter. In their paper, they used Random Forest and SVM models, where Random Forest has the best accuracy. Research on the Spanish cyberbullying prevention system (SPC), which relies on Natural Language Processing (NLP) methods and different machine learning techniques was deployed by Gabriel A. León-Paredes et al. [13]. They collected their dataset from Twitter. Among all the models SVM performed the best. Sruthi Chiramel et al. [14] published a paper on approaches to identify and classify social media insults using ML and NLP. They collected their data from Kaggle. Among all the models Random Forest provided the best result because the Random Forest model has a higher Recall value than other models. Bedoor Y. AlHarbi et al. [15] provided a study where they reviewed different ML models in order for automatic detection of online bullying in Arabic text. For their study they collected data from twitter. In their study among all the models applied Ridge Regression and Logistic Regression acquired the highest accuracy. Amrita Dewani et al. [16] published a methodology study paper on the cyberbullying detection approach for analyzing textual data in the Roman Urdu language based on ML and NLP. They obtained their data from Twitter. Among the models used, SVM has the best accuracy. An effective text-mining approach using machine learning algorithms is proposed by Shiza Andleeb et al. [17]. The data of this paper were collected from myspace.com and Preverted-Justice.com. Among all the algorithms they used, SVM Linear has provided the best accuracy. Tofayet Sultan et al. [18] published a paper on comparative methodology study on different models of ML. The Kaggle dataset was used in their paper. Among all the models Logistic regression provided the best result. Amgad Muneer et al. [19] in their study they tried to solve the issue of cyberbullying, where they don't have to involve the victim. They collected their data from twitter. They got their best result from Logistic regression among other models applied. In order to strongly detect cyberbullying and to stop the unexpected outcome a system was suggested by Saloni Mahesh Kargutkar et al. [20] to give a double description of cyberbullying by using ML and NLP models. For their study they collected data from Instagram. Their technique uses an idea of CNN for text examination. Djedjiga Mouheb et al. [21] provided a study on automatic detection of social media harassment in Arabic using ML models. They collected their data from Youtube and Twitter. Their suggested method detects social media harassment using Naïve Bayes classifier algorithm. Lida Ketsbaia et al. [22] in their study they Investigated different machine learning models using DBOW (Distributed Bag of Words), DMM (Distributed Memory Mean) and Word2vec CNNs in order to classify cyberbullying. In their study they used two different datasets, one of them was provided by the University of Maryland and the other one was collected from an article. Among all the models they used, Linear SVC provided the best result.

III. PROPOSED METHODOLOGY

In this section, the machine learning techniques and the used dataset are discussed. Additionally, we go over our

framework's suggested technique. The Process flowchart is shown in Fig-1 and Fig-2 shows the Workflow diagram.

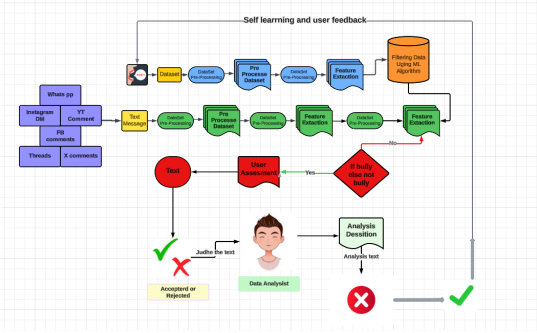


Figure 1: Process flowchart.

A. Data Collection

The dataset utilized in this research was sourced from a public online repository. It is made up of a massive amount of user-generated posts from the social media platform X (previously Twitter). The dataset was collected for the purpose of training and evaluating machine learning and natural language processing models for the automatic detection of cyberbullying.

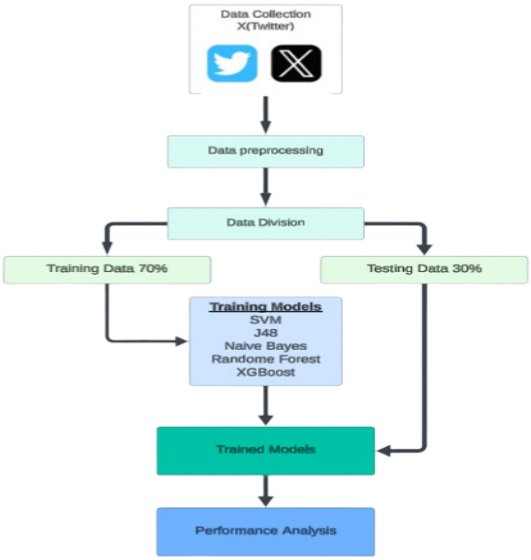


Figure 2: Workflow diagram.

B. Dataset Overview

The overall collection of the dataset is 24,784 tweets, which are sorted out into six variables: count, hate_speech, offensive_language, neither, class, and tweet, along with one unnamed variable that functioned as a serial index.The class attribute indicates the target variable and at the same time the tweet field contains the text that was the main predictor for classification. The exploratory data analysis (EDA) showed that there was no absence or nulls in the data, so the dataset was completely intact and could easily be subjected to further processing without any hassle.

C. Data Preprocessing

Before proceeding with the model training and testing, a number of preprocessing operations were performed to guarantee the quality and uniformity of the data. The unnamed index column was deleted and all the textual content in the tweet variable underwent a cleaning process and was numerically represented through text vectorization techniques for all machine learning models, except XGBoost. This change led to further enhancement of the performance and comparability among the classification models. Part of the dataset is shown in Fig. 3.

D. Model Training and Validation

After performing data preprocessing, the dataset was split into two parts: the one used for training and the one reserved for testing. This was done in order to evaluate the performance of the machine learning and natural language processing models. Usually, an 80:20 ratio, which is the standard for train-test splitting, was applied in order to have a sufficient amount of data for both model training and good performance assessment unbiased by data. The Support Vector Machine (SVM), Random Forest, J48 Decision Tree, Naïve Bayes, and XGBoost algorithms were applied on the already processed training data, and hyperparameter tuning was done through cross-validation wherever it was possible. The performance of models was assessed on the testing set by calculating the commonly accepted evaluation metrics, namely, accuracy, precision, recall, and F1-score. In order to provide comprehensive details regarding each model's classification performance, confusion matrices were drawn up that illustrated the respective numbers of true positives, true negatives, false positives, and false negatives..

E. Model Evaluation

1) SVM:

Support Vector Machine (SVM) is a supervised machine learning technique that is most commonly used for classification, regression, and outlier detection applications. Its high-dimensional performance and efficient handling of nonlinear relationships through kernel methods are its major strengths. Features that characterize this algorithm are: Interpretable; Flexible; High-quality results; Good at finding linear separators; Low risk of over-fitting; Effective when the number of dimensions is more than that of the samples; relatively memory efficient; versatility for both classification and regression tasks; robustness to overfitting. The equation of a Support Vector Machine (SVM) when used for classification can be represented as follows:

For a linear SVM, the equation of the hyperplane is given by:

$$w \cdot x + b = 0 \quad \text{(i)}$$

Here, (w) is the weight vector, (x) denotes the feature vector, and (b) signifies the bias term. The SVM's main objective is to determine the most favorable settings of (w) and (b) that will increase the barrier separating the two categories to the maximum extent. For a non-linear SVM, kernel functions serve as tools to project the input features into a loftier space where a linear division becomes feasible. The Radial Basis Function (RBF) is one of the frequently used kernel functions, and its definition is:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \| 2) \quad \text{(ii)}$$

In this expression, (K) refers to the kernel function, (xi) as well as (xj) signify feature vectors, and (gamma) is the parameter that controls the kernel's spread. The SVM method approaches an optimization problem with the purpose of locating the hyperplane that not only maximizes the margin but also has the least penalty for

misclassifications. The optimization problem can be stated as:

$$w,bmin21‖w‖2+C_i=1∑nmax(0,1−y_i(w·x_i+b)). \quad \text{(iii)}$$

In this formulation, (C) is the regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification error.

2) J48:

Being an open-source Java implementation of the C4.5 algorithm, J48 is highly appreciated throughout the data-mining tool in WEKA. It is a supervised learning method that reaches its goal of classifying data points through a tree-like model built upon a given set of features. This algorithm creates decision trees from a training dataset and uses the information theory and its corresponding entropy as does the ID3 algorithm. The dataset is made up of samples that were previously classified and are represented as S={s1, s2, …}. Every sample has related characteristics and is shown as a p-dimensional vector (x1, i, x2, i, …, xp, i) with the class of the sample denoted by the vector.

3) Naive Bayes:

Naive Bayes refers to a group of supervised learning methods whose foundation is Bayes' Theorem. It is a classic and time-saving method for classification tasks along with its very nice performance even in difficult situations. Naïve Bayes is indeed a powerful and effective classifier in machine-learning methods. Naive Bayes is based on Bayes' theorem. Bayes' Theorem determines the likelihood of an event happening based on the probability of another event that has already taken place. The formula for Bayes' theorem is an equation represented mathematically as follows:

$$P(A \mid B) = \frac{P(A) \times P(B \mid A)}{P(B)} \quad \text{(iv)}$$

4) Random Forest:

A Random Forest Classifier is a supervised machine learning technique that merges the results of different decision trees in order to predict with more accuracy and reduce over-fitting. It acts by building a park of decision trees, each of which is trained on a unique subset of the data, and than combining their forecasts to enhance the total predictive accuracy. The Random Forest algorithm doesn't have a single defining equation like some other algorithms because it's an ensemble method that operates by constructing multiple decision trees during the training phase and outputting the class.

5) XGBoost:

XGBoost is a efficient and powerful implementation of the gradient boosting algorithm. It is widely used for supervised machine learning tasks such as classification and regression. It builds an ensemble of decision tree in a sequential manner, where each new tree corrects the errors made by the previous ones. The key idea is to minimize a loss function using gradient descent optimization. XGBoost stands out for its speed and performance.

6) Confusion matrix:

A key instrument in machine learning for assessing the effectiveness of categorization algorithms is the confusion matrix. In essence, it's a table that shows how well a classification model performs on a set of test data given which the true values are known. The components of the "Confusion Matrix" are:

- True Positives (TP**):** In this case, the model correctly predicted the positive class.
- True Negatives (TN): In this case, the model correctly predicted the negative class.
- False Positives (FP): This is the case where the model incorrectly predicted the positive class.
- False Negative (FN): This is the case in which the model incorrectly predicted the negative class.

The confusion matrix allows us to calculate various performance metrics, such as: accuracy, precision, recall and F1 score. The formulas for these metrics are:

$$\text{Accuracy:} \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad \text{(iv)}$$

$$\text{Precision:} \frac{TP}{(TP+FP)} \quad \text{(v)}$$

$$\text{Recall:} \frac{TP}{(TP+FN)} \quad \text{(vi)}$$

$$\text{F1 Score: } 2 * \frac{(Precision \times Recall)}{(Precision+Recall)} \quad \text{(vii)}$$

IV. RESULT ANALYSIS

From Fig-4 shows among the models used SVM provided us with an accuracy of 0.83, precision 0.83, recall 0.831 and a F1-score of 0.78. On the other hand Random Forest provided us with an accuracy of 0.998 and they got the same result on precision, recall and F1 score

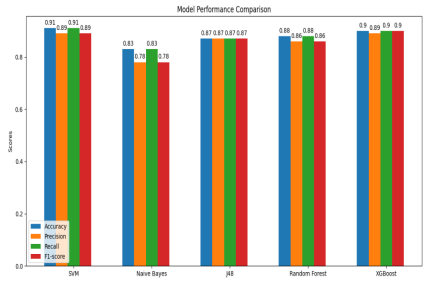


Figure 4: Result Analysis Diagram

from J48 decision tree but we an accuracy of 0.997 which is little less than Random Forest. We can also see that Naïve Bayes provided us with an accuracy of 0.85, precision 0.83, recall 0.85 and a F1-score of 0.80.

A. SVM

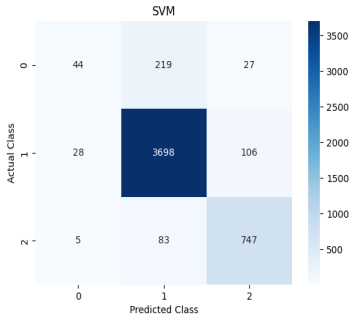


Figure 5: Confusion Metrics SVM

Among the assessed algorithms, the Support Vector Machine (SVM) model performed the best with an accuracy of 0.91, precision of 0.89, recall of 0.91, and an F1-score of 0.89. Hence, these outputs signified that SVM could effectively tell apart cyberbullying and non-cyberbullying content, thus giving a dependable and well-rated performance in classification across the board of evaluation metrics.

B. Random Forest

The Random Forest model's performance was scored with the following metrics: accuracy - 0.88, precision - 0.86, recall - 0.88, and F1-score - 0.86. Such results imply that Random Forest is a strong performer with a slightly lower predictive ability than SVM.

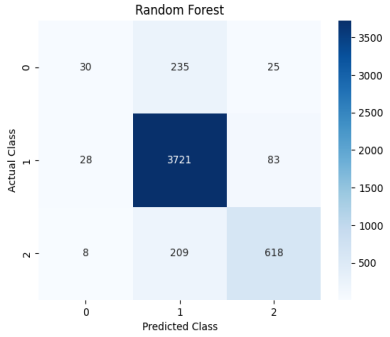


Figure 6: Confusion Metrics Random Forest

C. J48

The J48 Decision Tree model produced an accuracy of 0.87, precision of 0.87, recall of 0.87, and an F1-score of 0.87. The J48 model's results are indicative of a consistent and balanced performance in classifying cyberbullying cases, but its overall predictive power is a bit lesser than that of SVM and Random Forest.

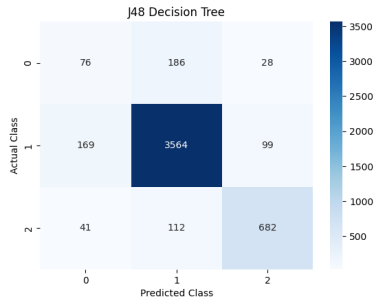


Figure 7: Confusion Metrics J48

D. Naive Bayes

Naive Bayes model recorded accuracy at 0.83, precision at 0.78, recall at 0.83, and an F1-score at 0.78.

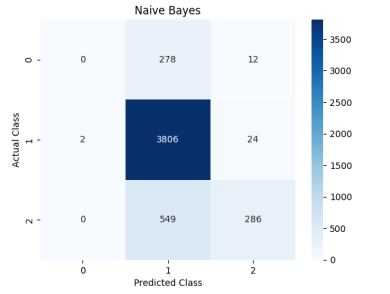


Figure 8: Confusion Metrics Naive Bayes

The outcomes reveal that even though Naïve Bayes is capable of detecting cyberbullying cases, its performance in predicting is still less than that of SVM, Random Forest, and J48 especially in precision which indicates a larger number of false positives.

E. XGBoost

The XGBoost model obtained an accuracy of 0.90, precision of 0.89, recall of 0.90, and an F1-score of 0.90. From these results, it can be concluded that XGBoost is a strong contender in the detection of cyberbullying, providing only slightly lower accuracy when compared to SVM but still keeping a very good classification performance throughout all the evaluation metrics used.

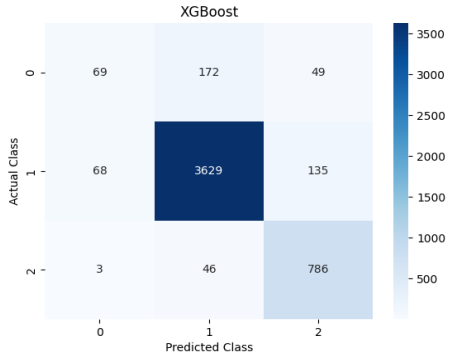


Figure 9: Confusion Metrics XGBoost

The summary of the performance of the five evaluated machine learning models, SVM, Random Forest, J48, Naïve Bayes, and XGBoost, is shown in Table 1.

TABLE I. Model Result

Model	Accuracy	Precision	Recall	F1 score
SVM	0.91	0.89	0.91	0.89
Random Forest	0.88	0.86	0.88	0.86
J48	0.87	0.87	0.87	0.87
Naïve Bayes	0.83	0.78	0.83	0.78
XGBoost	0.90	0.89	0.90	0.90

SVM outperformed all other techniques by providing an excellent combination of accuracy 0.91, precision 0.89, recall 0.91, and an F1-score of 0.89, hence proving its strong potential for accurate cyberbullying detection. Though XGBoost's accuracy of 0.90 was not too far from the best, its precision, recall, and F1-score ranges 0.89–0.90 were, at the same time, absolutely remarkable and that certainly granted it the status of a robust alternative. On the other hand, Random Forest was the one rated medium while J48 was a consistent but slow performer with a total score of 0.87 across-the-board. The performance of the Naïve Bayes classifier was the worst.

V. DISCUSSION

The skeleton of the study was a dataset composed of 24,784 tweets from X (Twitter) which were used to assess various machine learning models for the detection of cyberbullying. All models were created in Python within Google Compute Engine and their performance was measured by accuracy, precision, recall, and F1-score (see Table 1). SVM was the one that claimed the highest performance of 0.91 in terms of accuracy along with strong precision-recall balance. XGBoost was next as it gave similar results, while Random Forest and J48 displayed steady but moderate efficacy. To summarize, SVM was the most reliable model while XGBoost and Random Forest were less demanding

alternative options that could be chosen according to the application's needs.

VI. LIMITATIONS AND FUTURE WORK

This research is constrained by the variety of data sets and by the constantly changing online language which can have an impact on how well the model works with different contexts. Another obstacle for practical use is the interpretability of the model. The attention of the future research will be on the development of models that are more aware of the context, growing the size of the data sets and the use of explainable AI techniques. It is possible that the detection and prevention of cyberbullying will become more effective through the integration of social and psychological insights.

VII. CONCLUSIONS

The rise of social media has led to the emergence of cyberbullying as one of the major problems impacting people and society. Assessing the performance of different machine learning and natural language processing approaches, the present study aimed at detecting cyberbullying. The dataset from X (Twitter), and the SVM, Random Forest, J48, Naive Bayes, and XGBoost models were subjected to comparison using the four evaluation metrics accuracy, precision, recall, and F1-score over all the models. The results show SVM as the model with the best overall performance among the models evaluated, thereby confirming its strength in the detection of cyberbullying material. By doing this, the study contributes to the establishment of a comparative benchmarking for cyberbullying detection that prioritizes machine learning approaches and calls for further development of such tools as effective ones to mitigate online harassment.

.REFERENCES

[1] K. K. Kiilu, G. Okeyo, R. Rimiru, and K. Ogada, "Using Naïve Bayes Algorithm in detection of Hate Tweets," *Int. J. Sci. Res. Publ. (IJSRP)*, vol. 8, no. 3, 2018..

[2] D. Sultan et al., "Cyberbullying-related Hate Speech Detection Using Shallow-to-deep Learning," Computers, Materials & Continua, vol. 74, no. 1, pp. 2115–2131, 2023, doi: https://doi.org/10.32604/cmc.2023.032993.

[3] A. Akhter, U. K. Acharjee, M. A. Talukder, M. M. Islam, and M. A. Uddin, "A robust hybrid machine learning model for Bengali cyber bullying detection in social media," *Natural Language Processing Journal*, vol. 4, no. 100027, p. 100027, 2023..

[4] M. A. Al-Garadi et al., "Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges," IEEE Access, vol. 7, pp. 70701–70718, 2019.

[5] T. H. Teng and K. D. Varathan, Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches,IEEE Access, pp. 1-1, 2023.

[6] M. S. Islam, A. N. Orno, and M. Arifuzzaman, Approach to social media cyberbullying and harassment detection using advanced machine learning,2024.

[7] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, "Cyberbullying detection on social networks using machine learning approaches,in 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020.

[8] V. Shah, A. Sinha, N. Navalkar, S. Gupta, P. Gonsalves, and A. Malik,ML and natural language processing : Cyberbullying detection system for safer and culturally adaptive digital communities, Journal of Smart Internet of Things, vol. 2023, no. 2, pp. 193–205, 2023..

[9] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques, Electronics (Basel), vol. 10, no. 22, p. 2810, 2021.

[10] H. Kumar Sharma, K. Kshitiz, and Shailendra, "NLP and machine learning techniques for detecting insulting comments on social networking platforms," in 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2018, pp. 265–272.

[11] M. T. Ahmed, M. Rahman, S. Nur, A. Z. M. T. Islam, and D. Das, "Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts," TELKOMNIKA, vol. 20, no. 1, p. 89, 2021.

[12] S. Afrifa and V. Varadarajan, "Cyberbullying detection on Twitter using natural language processing and machine learning techniques," International Journal of Innovative Technology and Interdisciplinary Sciences, vol. 5, no. 4, pp. 1069–1080, 2022.

[13] G. A. León-Paredes et al., "Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the Spanish language," in 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 2019, pp. 1–7.

[14] S. Chiramel, D. Logofatu, and G. Goldenthal, "Detection of social media platform insults using Natural language processing and comparative study of machine learning algorithms," in 2020 24th International Conference on System Theory, Control and Computing (ICSTCC), 2020.

[15] B. Y. AlHarbi, M. S. Alharbi, N. J. Alzahrani, M. M. Alsheail, and D. M. Ibrahim, Using machine learning algorithms for automatic cyber bullying detection in Arabic social media, Journal of Information Technology Management, vol. 12, no. 2, pp. 123–130, 2020.

[16] A. Dewani et al., "Detection of cyberbullying patterns in low resource colloquial Roman Urdu microtext using natural language processing, machine learning, and ensemble techniques," Appl. Sci. (Basel), vol. 13, no. 4, p. 2062, 2023.

[17] S. Andleeb, R. Ahmed, Z. Ahmed, and M. Kanwal, "Identification and classification of cybercrimes using text mining technique," in 2019 International Conference on Frontiers of Information Technology (FIT), 2019.

[18] T. Sultan, N. Jahan, R. Basak, M. Shaheen, A. Jony, and R. Hasan, "Intelligent Systems and Applications," vol. 2, no. 2, pp. 1–13, 2023, doi: https://doi.org/10.5815/ijisa.2023.02.01.

[19] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for cyberbullying detection on Twitter," Future Internet, vol. 12, no. 11, p. 187, 2020..

[20] S. M. Kargutkar and V. Chitre, "A Study of Cyberbullying Detection Using Machine Learning Techniques," IEEE Xplore, Mar.01,2020.https://ieeexplore.ieee.org/abstract/document/9076550

[21] D. Mouheb, R. Albarghash, M. F. Mowakeh, Z. A. Aghbari, and I. Kamel, "Detection of Arabic Cyberbullying on Social Networks using Machine Learning," 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Nov. 2019, doi: https://doi.org/10.1109/aiccsa47632.2019.9035276.

[22] L. Ketsbaia, B. Issac, and X. Chen, "Detection of hate tweets using machine learning and deep learning," in 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2020, pp. 751–758.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Assessing the Brain and Kidney Tissues for Their Radiation Absorption Attributes

Nawshad Ahmed
Department of Electrical and Electronic Engineering
Bangladesh Army University of Science and Technology (BAUST), Saidpur, Bangladesh
nawshad@baust.edu.bd

Noor Md. Zuhaer Ahanaf
Department of Electrical and Electronic Engineering
Bangladesh Army University of Science and Technology (BAUST), Saidpur, Bangladesh
zuhaer.ahanaf10@gmail.com

Md. Arafat Rahman
Department of Electrical and Electronic Engineering
Bangladesh Army University of Science and Technology (BAUST),
Saidpur, Bangladesh
arafatpollob21@gmail.com

***Abstract*—The dangers to human health from this form of RF absorption are becoming more apparent. The brain and kidneys are the most essential organs for any living being. This study examines the characteristics of RF radiation absorption by brain and renal tissues, specifically looking at their specific absorption rate (SAR) and cross-correlation. Broadcasting radio frequency waves in a controlled laboratory environment ranged from zero degrees to 360 degrees, with increments of ten degrees. After passing through the brain and kidney tissues, the resulting power absorption was measured. The results consistently indicate that brain tissue absorbs a greater amount of radiofrequency radiation than kidney tissue. The average specific absorption rate (SAR) for cerebral tissue was 0.0027 W/kg, whereas the average SAR for kidney tissue was 0.0022 W/kg, indicating a lower value. The cross-correlation analysis of free-space and tissue samples indicated that kidney tissue demonstrated a higher correlation coefficient (0.7585) compared to brain tissue (0.6585). This indicates that brain tissue absorbs a greater amount of RF energy, resulting in reduced homogeneity with free-space data. The findings suggest that brain tissue exhibits a greater capacity for absorbing RF radiation, attributable to its elevated conductivity and moisture levels, thereby heightening the likelihood of thermal and biological repercussions. This raises concerns about extended exposure and emphasizes the higher health risks of radiofrequency (RF) radiation, particularly when devices like cell phones are used close to the brain.**

Keywords— ***RF radiation, specific absorption rate (SAR), brain tissue, kidney tissue, cross-correlation, RF absorption, health risks, mobile phone radiation***

I. Introduction

Electromagnetic radiation has two types: ionizing and non-ionizing. Ionizing radiation is recognized for its potential to damage biological tissues, whereas non-ionizing radiation is generally considered less harmful than its ionizing counterpart. However, it can change the rational and vibrational properties of both atoms and molecules, which makes it a powerful tool in molecular manipulation [1]. Radio frequency (RF) technology is extensively utilized in many communication devices, including computers, televisions, mobile phones, transmitters, and receivers. In addition to its application in communications, radiofrequency (RF) is crucial for specific medical procedures, including electrosurgical scalpels and radiofrequency ablation. Furthermore, it is utilized alongside other medical interventions, like diathermy and hyperthermia therapy, for oncological treatment. The study presents significant findings on the differential absorption of radiofrequency (RF) radiation by various types of biological tissues, comparing fatty and non-fatty tissues. Decreased detector currents in non-adipose tissue samples indicated that this tissue absorbs greater RF power compared to adipose tissue. The increased absorption linked to a phase lag of RF signal penetration in fatless tissue, as opposed to fatty tissue, indicates that fatless tissue exhibits reduced transparency to RF signals. The data indicate that proximity to fatless tissue, in contrast to fatty tissue, may result in increased health risks from RF exposure. To better grasp the health effects of RF radiation on various tissue types, the study recommends more research, especially in controlled situations such as anechoic chambers [2]. If radio frequency is ionized, radiation exposure can cause an inflammatory response in the brain by increased expression of inflammatory cytokines and enzymes. It may result in DNA damage and oxidative stress. Radiation leads to apoptosis of neurons and oligodendrocytes, resulting in demyelination and loss of neural functions. Radiation can induce disruption of the blood-brain barrier, histological alterations, and various stages of cerebral damage [3].

Factors such as moisture content, tissue capacitance, resistance frequency, exposure duration, and intensity all play a role in how RF (radiofrequency) radiation affects human tissue. At 430 MHz, 2.6 dB of radiation is produced, and at 800 MHz and 1200 MHz, 3.5 dB is produced, close to the belly, for a 20% variation in dielectric value. In every human body model, the electric fields at the front consistently exhibit greater intensity than those at the back [4]. When exposed to electromagnetic radiation, living tissue either absorbs or reflects some of the rays. At all frequencies, applying 1 W of antenna-radiated power causes an increase in human tissue temperature. At 125 mW, there is hardly any increase in temperature. The specific absorption rate (SAR) is used to evaluate the possible dangers associated with radiofrequency (RF) radiation exposure to the brain [5]. Each model of the human body generally agrees that the front part has a stronger electric field than the back part. Simulations have been conducted to examine the effects of modifying the angle of a mobile antenna in relation to the human head, varying from 0 degrees (antenna parallel to the neck) to 90 degrees (antenna

perpendicular to the neck). The results show that RF (radiofrequency) equipment needs to be placed at a right angle to the head. It has been found that the angle of the transmitter can change the radiation pattern [6].

Radiofrequency radiation affects biological organisms thermally and nonthermally. Non-thermal factors change cellular and tissue function without raising tissue temperature, while thermal influences do. A novel, non-invasive technique has been developed utilizing linear regression. This method allows for the calculation of the average rate of substance absorption by the entire human body, hence obviating the necessity for invasive interventions. This method uses non-invasive electric field measurement techniques to determine the average body-specific absorption rate [7]. Research conducted over the past 75 years has conclusively shown that RF (radiofrequency) radiation and energy absorption are detrimental to human health. Scientists have investigated the impact of electromagnetic fields on the body's internal temperature using a wide range of frequencies. Longer exposure times weaken the correlations, even though they are strongest with bigger masses or volumes. These findings are expected to have an effect on the calls to reduce human exposure to electromagnetic fields, particularly for biological reasons [8]. The skin and eyes are primarily at risk from this radiation because it doesn't penetrate deeply into tissues. Long-term exposure to electromagnetic fields has been associated with an increased risk of brain tumors, both malignant and benign. The microwaves that mobile phones emit can alter the structure of DNA by producing charged particles and high-frequency electromagnetic fields [9].

Wireless technologies like Bluetooth generate non-ionizing radiation (NIR), which has been shown to have negative impacts on human health. It is possible to cause burns or other types of harm by raising the temperature of tissues. It has been noted that consumer products have widely utilized Bluetooth technology. Using a customized Bluetooth system, we tested the application's effectiveness in various settings. The RF attenuation properties of many substances were studied, including water, meat, ballistic gel, and fat. Pork fat was used as a stand-in for human adipose tissue due to its wide dielectric range. According to studies, meat and ballistic gel have a stronger impact on RF waves than water and fat. Muscle exhibits greater signal attenuation than fat and water, among other tissues. By employing signal integrity distances, medical devices that are equipped with Bluetooth technology can be built and implemented [10]. Non-ionizing radiation may potentially be beneficial in cancer therapy that combines chemotherapy with radiation. Adebayo et al. conducted a study demonstrating that male rats subjected to electromagnetic radiation displayed histological alterations in their kidneys [11].

According to research, radio frequency (RF) sources, such as WiFi and jammer radiation, can alter blood markers associated with kidney function. Total protein levels were significantly different between the WiFi and sham groups at about 100 cm, according to the data [12]. Research assessing the impact of 2.45 GHz radio frequency waves on the liver and kidneys of Wistar female rats revealed detrimental effects from prolonged exposure to 1800 MHz radio frequency radiation, evidenced by alterations in biochemical parameters. [13]. Radiation has complex effects on the brain, resulting in several changes. These effects can damage neurons, glia, and cells in blood vessels. Radiation can lead to apoptosis, neuroinflammation, oxidative stress, and alterations in gene expression. Preclinical studies indicate that radiation impacts brain regions beyond the hippocampus, complicating the clinical treatment of cognitive problems. People are working on ways to protect the brain [14].

Electromagnetic radiofrequency radiation negatively impacts the DNA of rat brain cells. The tail length of DNA fragments in the brain cells of irradiated rats was 14.0 mm, whereas in control rats it was 13.5 mm. Conversely, in the kidney cells, the length was markedly greater in exposed rats (13.9 mm) than in control rats (12.9 mm). The impact of radiation on the kidneys is greater than that on cerebral cells. [15].

Neurological changes can be caused by high-power RF bursts that don't go over the current regulatory temperature limits. Brain function and structure can be changed by non-thermal factors that don't happen when the temperature rises significantly. It is talked about a condition called "Havana Syndrome," in which officials reported neurological symptoms after being exposed to RF energy. Repeated sub-millisecond RF pulses are suggested in the paper as a possible cause of these symptoms [16]. The potential negative impacts on human health, particularly the central nervous system, of the increased exposure to RF-EMFs brought about by the broad deployment of 5G networks are detailed in a recent study. Changes in glycerophospholipid metabolism were among the 1318 metabolites found in brain tissue. One hundred forty-four metabolites were up-regulated, and one hundred thirty-three were down-regulated in the RF exposure group. Of the 7285 proteins identified by proteome analysis, 10 showed up-regulation and 51 showed down-regulation in the RF group of differentially expressed proteins (DEPs). Cells in the brain of rats are damaged genetically by exposure to radiofrequency electromagnetic radiation. The study found that 4.9 GHz RF radiation altered metabolite and protein expression patterns, which may lead to behavioral changes including depression-like symptoms [17].

The radiation emitted by cell phones via radio frequencies has been linked to the onset of cancer. Electromagnetic fields (EMFs) have shown advantageous health effects, such as facilitating bone healing and mitigating damage linked to chemotherapy. Clinical research suggests that integrating chemotherapy with RF or MW hyperthermia could improve results in both curative and palliative therapies [18]. Excessive use of electronic devices has been linked to several behavioral illnesses in children and adolescents, such as addiction, attention deficit hyperactivity disorder (ADHD), sleeplessness, anxiety, intolerance, and other behavioral challenges [19].

Studies showed that brain tissue absorbed more radiofrequency radiation than kidney tissue in this investigation. Both samples for this experiment were obtained from the same goat to guarantee consistency in animal traits. The remaining RF devices were deliberately placed at a distance to minimize interference. As a result, brain tissue has the ability to absorb higher levels of radio frequency (RF) power at specific frequencies, with absorption rates also increasing at certain angles.

II. System Model and Operation

This study examined the impacts of radiofrequency exposure by assessing the absorption of radiofrequency (RF) in cerebral and renal tissues. The study utilized a blend of free-

space and empirical data. The goat brain tissue was first put between the RF transmitter and receiver, and then the goat kidney tissue was put in the same area. Radiation data might be compared with baseline measurements taken in space thanks to this setup. The primary goal was to determine the tissues that are most susceptible to the detrimental effects of radiofrequency (RF) radiation, which poses a significant threat to the health of patients. The study found a number of important things about radiofrequency (RF) device implantation and use, especially for the limbs with long-term daily use. Ultimately, the study's surprising results raised public awareness of the negative health impacts of radiofrequency.

III. Methodology

This section provides a representation of the experimental setup and the method of data processing that was utilized in order to determine the RF absorption properties in brain and kidney tissue samples that were collected from the same goat. Comparing human and goat tissues reveals that there are some similarities between the two. Here, goat tissues for the purposes of analysis are displayed in Table I, along with a comparison of goat brain and kidney tissues to those of humans.

TABLE I. Comparison of the Properties Between Human and Goat Tissues.

Property	Human Brain	Goat Brain	Human Kidney	Goat Kidney
Water content [12]	~75%-80%	~74%-79%	~80%-85%	~80%-84%
Permittivity [13]	60-65 (1 MHz)	58-64 (1 MHz)	70-75 (1 MHz)	68-73 (1 MHz)
Permeability [13]	~1 (similar to vacuum)	Similar to the human brain	~1 (similar to vacuum)	Similar to a human kidney
Impedance [14]	300-600 Ohms (1 MHz)	280-580 Ohms (1 MHz)	250-400 Ohms (1 MHz)	240-390 Ohms (1 MHz)

A. MEASUREMENT SET-UP

A rectangular patch antenna with a frequency of 2.45 GHz was employed as a radio frequency transmitter. The transmitter's ability to pivot from 0° to 360° was facilitated by its connection to a revolving pedestal. The schematic diagram is depicted in Figure 1, where the left antenna represents the RF transmitter and the right antenna serves as the RF receiver.

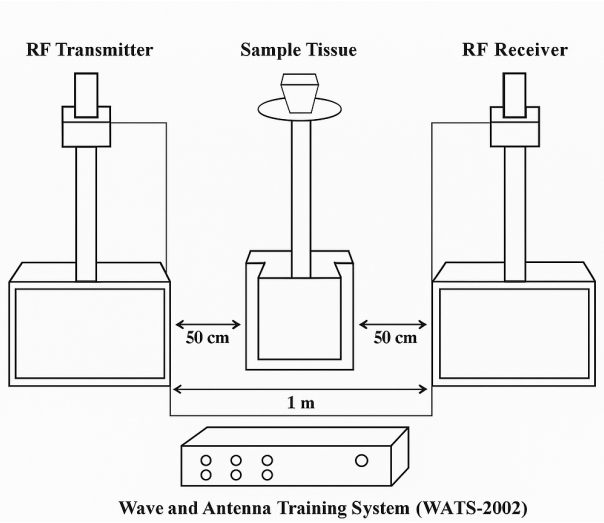


Fig. 1. RF power measurement diagrammatic representation.

B. LABORATORY SET-UP

The receiver utilized an inverted-F antenna, which is widely implemented in mobile devices. A rectangular patch antenna was employed as the transmitter, operating at 2.45 GHz. The transmitter and receiver were placed 1 meter apart. The experimental setup allowed the transmitter to rotate automatically through a full 360° range. This transmitter was interfaced with the trainer board. The display output represented the power captured by the inverted-F antenna, a standard component in mobile communication systems. The sample tissue was positioned midway between the transmitter and receiver using a non-metallic holder, ensuring accurate alignment with the antenna's radiation axis (Fig. 1). A comprehensive list of the experimental equipment and their specifications is provided in Table II.

The brain and kidney of the goat were chosen as samples because of their many similarities to human anatomy. The experiment utilized identical quantities of each sample.

TABLE II. Specification of the Experimental Setup.

Devices/Samples	Specifications
Wave and Antenna Training System	WATS-2002
Transmitter	Rectangular Patch Antenna Resonance Frequency: 2.45 GHz Dielectric Constant: 4.7, Loss Tangent: 0.025 PCB Thickness: 1.6t, FR-4
Receiver	Inverted F Antenna Resonance frequency: 2.45 GHz Dielectric Constant: 4.7, Loss Tangent: 0.025 PCB Thickness: 1.6t, FR-4
Biological Sample	Goat Brain (150 gm) Goat Kidney (150 gm)
Software	WATS-2002

The methodology implemented for the experimental setup is illustrated in Fig. 2. The transmitter generated a signal spanning from 0 degrees to 360 degrees, which was subsequently captured using a step size of 10. The transmitted signal through the sample medium was evaluated from the receiver outputs. Free-space measurements were first performed without any tissue between the transmitter and receiver. Subsequently, a brain tissue specimen was placed on a nonmetallic holder fixed to the stand for data acquisition, as shown in Fig. 2. Finally, the kidney tissue was employed, and measurements were recorded in accordance with the established procedure.



Fig. 2. Experimental setup in the laboratory for measuring RF power.

Minimal interference in the exposed RF signal was ensured by prohibiting the use of any RF devices not directly connected to the investigations and by restricting access to the room during the trials. The experimenter's head was positioned significantly below the level of radiofrequency exposure during data recording.

IV. RESULTS AND DISCUSSIONS

The recorded RF power for free-space data and also for both brain and kidney sample tissues are depicted in Table III.

TABLE III. COMPARISON OF RECEIVED RF POWER

Angle (Degrees)	Received Power (mW)		
	Without Sample	*Brain Sample*	*Kidney Sample*
0	0.2884	0.2239	0.39811
20	0.3311	0.3802	0.44668
40	0.3631	0.2455	0.43652
60	0.3981	0.3802	0.39811
80	0.5370	0.2884	0.31623
100	0.6166	0.2630	0.28840
120	0.7413	0.2239	0.30200
140	0.7079	0.4365	0.38019
160	0.6457	0.3311	0.46774
180	0.6457	0.3162	0.51286
200	0.5888	0.0933	0.36308
220	0.7079	0.0398	0.16596
240	1.0233	0.0195	0.12303
260	1.1482	0.0214	0.07762
280	1.0715	0.0309	0.08511
300	0.0933	0.0832	0.11220
320	0.6166	0.0933	0.19055
340	0.4677	0.3802	0.27542
360	0.3467	0.4571	0.33113

Table III indicates that RF power is greater in free space compared to the brain and kidney samples. This experiment measured RF radiation absorption in separate brain and kidney samples positioned between an RF transmitter and receiver, with the transmitter rotated in 10-degree increments from 0 to 360 degrees. The distance between the RF transmitter and receiver was 1.5 meters. Power absorption data was collected by the receiver after the RF radiation traveled through the samples, which were placed between the RF transmitter and receiver. Brain tissue absorbs more radiofrequency radiation than kidney tissue at different rotational rotations. Brain tissue showed 0.2884 mW of absorption at 0 degrees, whereas kidney tissue showed 0.2239 mW. This pattern is seen in most angles, while minor deviations were observed. This study unequivocally demonstrates that for various RF transmitter angle orientations, brain tissue has a propensity to absorb more RF radiation than renal tissue.

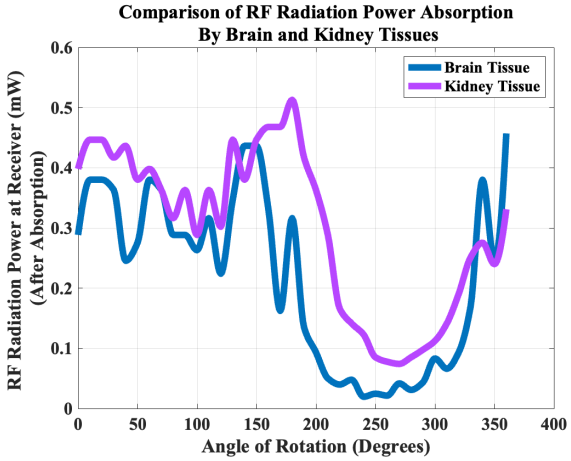


Fig. 3. Received RF power comparison between brain tissue with kidney sample tissue.

Figure 3 illustrates the RF power absorption characteristics of brain and kidney tissues. The figure indicates that both tissues exhibit increased RF power absorption. Brain tissues exhibit a higher absorption of RF radiation power compared to kidney tissues. At specific angles such as 80°, 110°, 180°, and 270°, the disparity in RF power absorption between brain and kidney tissues becomes more pronounced. Brain tissue exhibits a higher absorption of radiation attributable to its cellular composition.

Compared to the kidney, brain tissue can absorb greater RF radiation power because it has a higher electrical conductivity and more water. In contrast, kidney tissue reflects greater RF radiation power due to its higher conductivity and lower water content. Depending on the transmitter's angle and direction, both tissues show different levels of absorption. The main conclusion drawn from this curve is that compared to renal tissue, brain tissue absorbs far more radiofrequency energy.

Figure 4 illustrates the specific absorption rate (SAR) across various angular orientations of brain and renal tissues. The specific absorption rate (SAR) of radiofrequency radiation for brain tissue is markedly greater than that for renal tissue at various rotation angles from zero to 360 degrees. The specific absorption rate (SAR) for brain tissue (0.0027 W/kg) indicates its superior capacity to absorb RF power compared to kidney tissue (0.0022 W/kg).

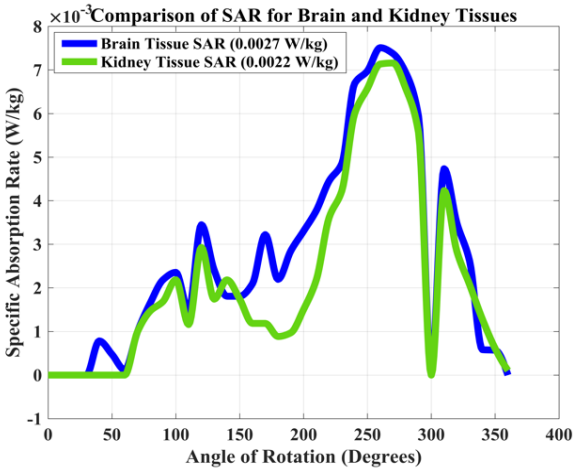


Fig. 4. SAR comparison between brain and kidney tissue samples.

The SAR value of brain tissue indicates that it may be more susceptible to the thermal and biological effects of radiofrequency radiation. This prompts significant inquiries regarding the health and safety implications of RF exposure, especially when individuals utilize a device that emits intense RF radiation for extended periods. To protect delicate tissues such as the brain from the detrimental effects of RF exposure, it is essential to comprehend the alteration of RF patterns.

Figure 5 shows the results of a 3D visualization analysis done in MATLAB that aimed to provide a better picture of the brain-kidney Specific Absorption Rate (SAR) comparison. The scatter plot reveals the SAR of radiofrequency radiation on the kidney and brain tissues, which is dependent on the angle of rotation of the transmitter. Brain tissue has a greater SAR than renal tissue at most angles, as shown by the 3D image's peaks. By comparing the two organs' RF absorption patterns using color coding, we can see that, in these circumstances, brain tissue is more susceptible to RF radiation than renal tissue.

The cross-correlation between the RF power in free space and that with brain and kidney tissues has been provided in Figure 6. This was done in order to examine the absorption comparability of both brain and kidney tissues. The data in Fig. 6 indicate a stronger correlation between the received power of the receiver for both kidney and brain tissues in relation to free space.

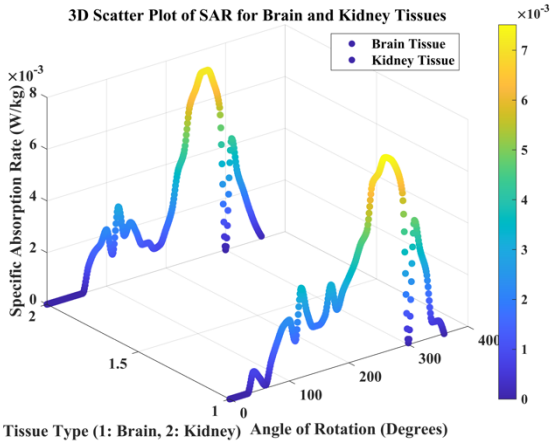


Fig. 5. 3D Visualization of Specific Absorption Rate (SAR) between brain and kidney sample data.

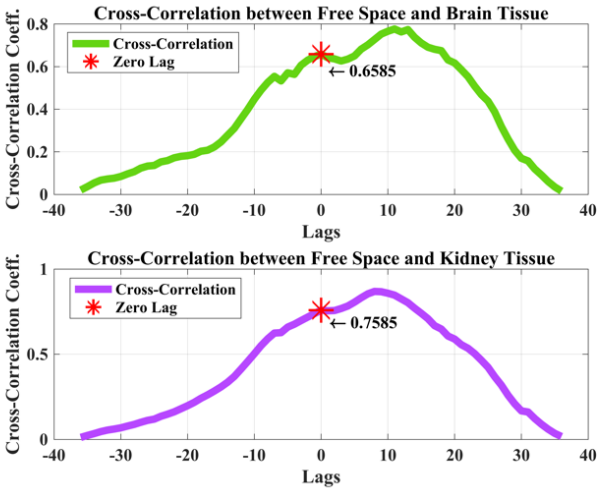


Fig. 6. The cross-correlation between the received power of brain tissue and free-space power is depicted in the upper image, while the received power of kidney tissue is illustrated in the lower figure.

Due to the water content and conductivity property of brain tissue, it absorbs more RF power than kidney tissue. The correlation coefficient of 0.7585 in the kidney tissue and 0.6585 in the brain tissue indicates that the absorption rate is higher in the brain tissue. When the RF radiation passes through the samples, it reflects radiation after being absorbed by the samples. So, the overall result of this experiment shows that the brain tissue is more vulnerable to RF radiation due to its cellular properties than the kidney tissue.

V. Conclusion

By contrasting the rates of RF radiation absorption in the kidneys and the brain, this study highlights the dangers of RF radiation to human health, particularly from mobile phones and other devices that produce large amounts of RF energy. With an average SAR value of 0.0022 W/kg for kidney tissue and 0.0027 W/kg for brain tissue, it is clear that kidney tissue absorbs a greater amount of radiofrequency energy than brain tissue. According to the correlation analysis method, the correlation coefficient of empty space is 0.7585 and that of renal tissue is 0.7585 as well. The correlation coefficient for brain tissue, 0.6585, is lower than that of free space. This indicates that brain tissue absorbs more RF radiation than renal tissue, attributable to its cellular properties, such as water content and conductivity. This analysis clearly indicates that

sensitive tissues, such as brain tissue, are progressively affected by RF radiation. Extended exposure to powerful RF radiation sources such as mobile phones and smart watches has the potential to harm delicate tissues, including brain tissue. Users might suffer from biological and thermal effects. These findings underline the need for preventative measures as well as additional research into the long-term effects of radiofrequency radiation on delicate tissues like the brain.

VI. REFERENCES

[1] S. Sivani and D. Sudarsanam, “Impacts of radio-frequency electromagnetic field (RF-EMF) from cell phone towers and wireless devices on biosystem and ecosystem—A review,” *Biology and Medicine*, vol. 4, no. 4, p. 202, 2012.

[2] N. Ahmed and A. K. Sarkar, “Comparative Study of RF Absorption Characteristics between Fatty and Fatless Tissue,” in *Proc. 2023 6th Int. Conf. Electrical Information and Communication Technology (EICT)*, Dec. 2023, pp. 1–4.

[3] Q. Wang, C. Xie, S. Xi, F. Qian, X. Peng, J. Huang, and F. Tang, “Radioprotective effect of flavonoids on ionizing radiation-induced brain damage,” *Molecules*, vol. 25, no. 23, p. 5719, 2020.

[4] L. S. Xu, M. Q.-H. Meng, and C. Hu, “Effects of dielectric values of human body on specific absorption rate following 430, 800, and 1200 MHz RF exposure to ingestible wireless device,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 52–59, 2009.

[5] T. B. Rashid and H. H. Song, “Analysis of biological effects of cell phone radiation on human body using specific absorption rate and thermoregulatory response,” *Microwave and Optical Technology Letters*, vol. 61, no. 6, pp. 1482–1490, 2019.

[6] R. Hosseinzadeh and B. Zakeri, “SAR of the human’s head due to electromagnetic radiation of a mobile antenna,” in *Proc. 20th Telecommun. Forum (TELFOR)*, Belgrade, Serbia, Nov. 2012, pp. 1159–1161.

[7] R. K. Kwate, A. Derkaoui, B. Elmagroud, C. Taybi, A. Ziyyat, and D. Picard, “Non-invasive estimation of whole body averaged SAR by the personal RF dosimeter measurements,” in *Proc. 7th Mediterranean Congr. Telecommun. (CMT)*, Fes, Morocco, Oct. 2019, pp. 1–4.

[8] M. Cavagnaro and J. C. Lin, “Importance of exposure duration and metrics on correlation between RF energy absorption and temperature increase in a human model,” *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 8, pp. 2253–2258, 2018.

[9] E. Vinodha and S. Raghavan, “Possible effects of cell phone radiation: An overview paper,” in *Proc. 2nd Int. Conf. Electron. Commun. Syst. (ICECS)*, Coimbatore, India, Feb. 2015, pp. 837–841.

[10] M. J. Christoe, J. Yuan, A. Michael, and K. Kalantar-Zadeh, "Bluetooth signal attenuation analysis in human body tissue analogues," *IEEE Access*, vol. 9, pp. 85144–85150, 2021.

[11] E. A. Adebayo, A. O. Adeeyo, M. A. Ogundiran, and O. Olabisi, "Bio-physical effects of radiofrequency electromagnetic radiation (RF-EMR) on blood parameters, spermatozoa, liver, kidney and heart of albino rats," *J. King Saud Univ. Sci.*, vol. 31, no. 4, pp. 813–821, Oct. 2019.

[12] S. Farahani, F. Kadivar, F. Khajeh, and M. B. Shojaeifard, "Effect of non-ionizing radiations on liver and kidney function tests in an animal model," *J. Biomed. Phys. Eng.*, vol. 15, no. 2, p. 125, 2025.

[13] H. M. Fahmy and F. F. Mohammed, "Hepatic injury induced by radio frequency waves emitted from conventional Wi-Fi devices in Wistar rats," *Hum. Exp. Toxicol.*, vol. 40, no. 1, pp. 136-147, Jan. 2021.

[14] H. M. Fahmy and F. F. Mohammed, "Hepatic injury induced by radio frequency waves emitted from conventional Wi-Fi devices in Wistar rats," *Hum. Exp. Toxicol.*, vol. 40, no. 1, pp. 136–147, Jan. 2021.

[15] I. Trošić, I. Pavičić, S. Milković-Kraus, M. Mladinić, and D. Željezić, "Effect of electromagnetic radiofrequency radiation on the rats’ brain, liver and kidney cells measured by comet assay," *Coll. Antropol.*, vol. 35, no. 4, pp. 1259–1264, 2011.

[16] O. Yaghmazadeh, "Pulsed high-power radio frequency energy can cause non-thermal harmful effects on the brain," *IEEE Open J. Eng. Med. Biol.*, vol. 5, pp. 50–53, 2024.

[17] X. Wang et al., "Effects of 4.9 GHz radiofrequency field exposure on brain metabolomic and proteomic characterization in mice," *Biology*, vol. 13, no. 10, p. 806, Oct. 2024.

[18] M. M. Paulides, H. D. Trefna, S. Curto, and D. B. Rodrigues, "Recent technological advancements in radiofrequency- and microwave-mediated hyperthermia for enhancing drug delivery," *Adv. Drug Deliv. Rev.*, vol. 163, pp. 3–18, 2020.

[19] N. Nasiar and A. Rani, "Detrimental effects of mobile phone radiations on human health," in *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, Kedah, Malaysia, 2018, pp. 1-6.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Subject Independent Epileptic Seizure Recognition Using 1D-EEGNet Model

Md. Israil Hosen
Dept. of Computer Science and Eng.
University of Rajshahi, Bangladesh
israilcseru28@gmail.com

Shakil Hossan
Dept. of Computer Science and Eng.
University of Rajshahi, Bangladesh
shakilhossan113@gmail.com

Md. Mominur Rahman
Dept. of Computer Science and Eng.
University of Rajshahi, Bangladesh
mominur.sumon1@gmail.com

Zannatul Naim
Dept. of Computer Science and Eng.
University of Rajshahi, Bangladesh
zannatulnaim09@gmail.com

Shabbir Mahmood
Dept. of Computer Science and Eng.
University of Rajshahi, Bangladesh
shabbir.cse@ru.ac.bd

Mahboob Qaosar
Dept. of Computer Science and Eng.
University of Rajshahi, Bangladesh
qaosar@ru.ac.bd

***Abstract*—Epileptic seizures are sudden and unpredictable changes in brain activity that can significantly impact a patient's health and quality of life. While electroencephalography (EEG) is the main tool for diagnosing seizures, interpreting long-term recordings manually is still time-consuming, subjective, and not practical for continuous monitoring. This study presents a compact and efficient one-dimensional EEGNet (1D-EEGNet) model that automatically detects seizures from EEG signals. The model uses depthwise and separable convolutions to capture important temporal and spatial relationships across channels. It keeps the number of parameters low, which makes it suitable for real-time use. Experiments on the Siena Scalp EEG dataset show impressive results. The model achieves 99.83% accuracy under standard evaluation and 95.35% accuracy with leave-one-subject-out (LOSO) cross-validation. A comparison shows that the proposed model consistently outperforms the 1D-CNN and 1D-ResNet baselines. These results demonstrate the power, effectiveness, and possibilities of 1D-EEGNet for real-time, wearable seizure monitoring systems.**

***Index Terms*—Electroencephalogram (EEG), Seizure Detection, Deep Learning (DL), 1D-EEGNet, Cross-Subject Validation.**

I. Introduction

Epilepsy is a long-term neurological disorder that affects about 1% of people worldwide. It causes repeated seizures due to abnormal brain activity. Fast and accurate seizure detection is essential for diagnosis, treatment adjustments, and patient safety [1]. EEG is a non-invasive and affordable tool for recording brain activity. However, doctors often find it tiring and subjective to visually inspect EEG signals, especially with long-term monitoring data. Early methods of automated seizure detection relied on handcrafted features from wavelet transforms, Fourier analysis, or statistical measures. These features were used in traditional classifiers like SVM, Decision Trees, or Random Forests. While these models performed well on specific datasets, they had difficulty managing the high variability in EEG data between different individuals.

Recently, deep learning has emerged as a powerful option. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can automatically spot patterns without needing explicit feature engineering [2]. However, many current models require a lot of computing power and are not suitable for real-time or embedded systems.

To address these challenges, this study introduces a low-dimensional one-dimensional EEGNet (1D-EEGNet). This architecture processes raw EEG time-series directly using depthwise and separable convolutions. It balances interpretability, accuracy, and efficiency, making it a promising choice for clinical and wearable EEG systems.

The key contributions of this paper are as follows:

- A compact deep learning framework, 1D-EEGNet, designed for seizure detection.
- Rigorous evaluation across subjects shows strong generalization ability.
- Comprehensive performance comparison with conventional deep learning models.
- State-of-the-art results on the Siena Scalp EEG dataset.

II. Related Works

Earlier studies in seizure detection relied on handcrafted features and traditional classifiers. For example, Subasi (2007) used wavelet decomposition followed by neural networks [3]. Tzallas et al. (2009) extracted time-frequency features and combined them with SVMs for identifying seizures [4]. While these methods achieved reasonable accuracy within subjects, they depended heavily on expert feature design and did not perform well across different subjects.

With the rise of deep learning, models started to learn complex EEG patterns directly from raw or preprocessed data. Ullah et al. (2018) created a CNN-LSTM hybrid that achieved impressive accuracy on the Bonn and CHB-MIT datasets [2]. Acharya et al. (2019) introduced a 13-layer CNN that automated feature extraction from raw EEG. Lawhern et al. (2018) presented EEGNet, which used depthwise and separable convolutions to form an efficient design suitable for various EEG tasks [5]. Despite their success, most existing models have concentrated on validating their results within the same subject, which limits their reliability in clinical settings.

This work expands on the EEGNet concept by developing a 1D architecture optimized for processing temporal EEG data and validating it in cross-subject conditions. This is a crucial step toward using it in real-world clinical applications.

III. Dataset and Preprocessing

A. Dataset Description

This work uses EEG recordings from the Siena Scalp EEG Dataset [6], which tracks brain activity in 14 people with epilepsy. Each patient wore 29 scalp electrodes, recording at 512 Hz, for about 128 hours in total. Experts went through the recordings and marked 47 seizure events, giving us solid, trustworthy labels for supervised learning.

B. Preprocessing

We downsampled EEG signals from 512 Hz to 250 Hz, then ran them through a 4th-order Butterworth bandpass filter (0.5–70 Hz) and added a 50 Hz notch filter. After that, we normalized each channel with Z-scores. The data got chopped into overlapping windows either 2 seconds with a 0.5-second overlap or 3 seconds with a 1-second overlap. For evaluation, we used a stratified split: 70% for training, 10% for validation, and 20% for testing. We also ran LOSO cross-validation to check how well things worked across different subjects.

IV. Methodology

A. Proposed 1D-EEGNet Architecture

The core of the proposed framework is the 1D-EEGNet, a compact convolutional neural network designed to extract hierarchical spatiotemporal features from EEG segments. Let an input EEG segment be represented as $\mathbf{X} \in \mathbb{R}^{C \times T}$, where C is the number of channels and T is the number of time samples.

Block 1 (Temporal Convolution): A standard 1D convolution with $F_1 = 16$ filters and kernel size $k_1 = 64$ is applied along the temporal dimension:

$$\mathbf{H}_1 = \text{Conv1D}(\mathbf{X}, F_1, k_1) \tag{1}$$

Block 2 (Depthwise Convolution): Depthwise convolution captures spatial correlations across channels:

$$\mathbf{H}_2 = \text{DepthwiseConv}(\mathbf{H}_1, D, k_c) \tag{2}$$

where D is the depth multiplier and k_c corresponds to the number of channels. Batch normalization, ELU activation, and average pooling (pool size = 4) are applied subsequently.

Blocks 3 and 4 (Separable Convolutions): Separable convolutions efficiently extract higher-order temporal-spatial features:

$$\mathbf{H}_3 = \text{SeparableConv}(\mathbf{H}_2, F_3, k_3) \tag{3}$$

$$\mathbf{H}_4 = \text{SeparableConv}(\mathbf{H}_3, F_4, k_4) \tag{4}$$

with batch normalization, ELU activation, average pooling, and dropout applied in each block.

Classification Head: The final feature map is flattened and passed through a fully connected layer with $N_h = 64$ hidden units and dropout $p = 0.4$, followed by a sigmoid activation for binary seizure classification:

$$\hat{y} = \sigma(\mathbf{W}_h \mathbf{h} + b_h), \quad \hat{y} \in [0, 1] \tag{5}$$

where $\mathbf{h}$ is the flattened feature vector, $\mathbf{W}_h$ and b_h are learnable parameters, and $\sigma(\cdot)$ is the sigmoid function.

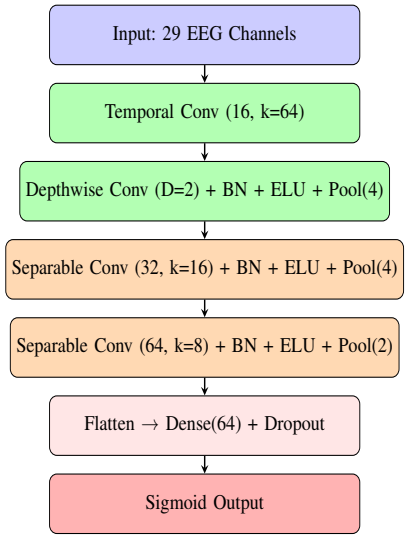


Figure 1. Schematic of the proposed compact 1D-EEGNet architecture.

B. Training Procedure

The 1D-EEGNet model was trained using the Adam optimizer with an initial learning rate $\eta = 10^{-4}$ and binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right] \tag{6}$$

where N is the number of training segments, y_i is the true label, and $\hat{y}_i$ is the predicted probability. Training was performed with a batch size of 32 for a maximum of 150 epochs. Early stopping (patience = 20) and a `ReduceLROnPlateau` scheduler (patience = 10, factor = 0.5) were used to stabilize and accelerate convergence.

1) Training Without Cross-Subject Validation: Here, we combined data from everyone into one big set, then split it 80/20 for training and testing. From the training part, we carved out a validation set to keep an eye on early stopping and tweak the learning rate as needed. With this setup, the model learns and gets tested on segments that could come from the same people, so you're really seeing how it performs when there's some overlap between training and test data.

2) Training With Cross-Subject Validation: To see how well the model handled new subjects, we used leave-one-subject-out (LOSO) cross-validation. Here's how it worked: in each round, we took all the data from one person and set it aside for testing. The rest went into training. From that training set, we pulled out 20% for validation. We trained the model separately

for every round. In the end, we averaged the results from all rounds and reported both the mean and standard deviation.

$$\text{Accuracy}_{\text{LOSO}} = \frac{1}{S}\sum_{s=1}^{S}\text{Acc}_s \tag{7}$$

where S is the number of subjects and Acc_s is the accuracy for the s-th fold.

We trained the 1D-EEGNet model with the Adam optimizer, starting with a learning rate of 10^{-4}. For loss, we used binary cross-entropy. Each batch had 32 samples, and we capped the training at 150 epochs. To help the model learn faster and avoid overfitting, we set up early stopping with a patience of 20 epochs, and used a ReduceLROnPlateau scheduler that cut the learning rate in half if things stalled for 10 epochs.

V. Results and Discussion

A. Standard Evaluation

Figure 1 shows how the training and validation accuracy of the 1D-EEGNet model change over 150 epochs. At first, both accuracy curves jump up quickly, then they start to level off as training goes on. The two curves stay close together, which means the model isn't just memorizing the training data it's actually learning patterns that work on new data too. You don't see much overfitting here. The model keeps up strong performance for both training and validation, which really speaks to how reliable and robust this 1D-EEGNet is for the task.

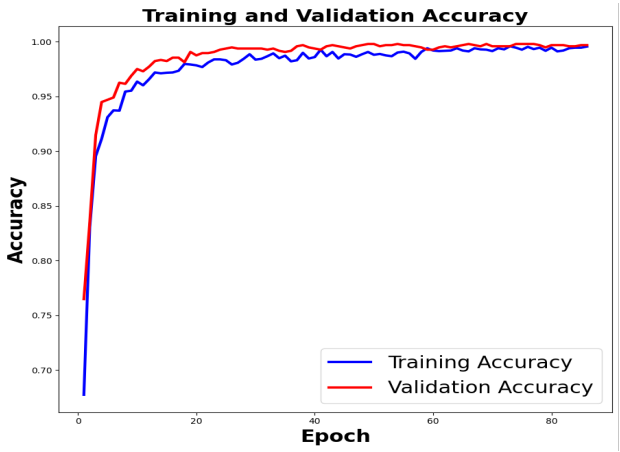


Figure 2. 1D-EEGNet model training and validation accuracy curve

Figure 2 shows the confusion matrix for the 1D-EEGNet model. The accuracy hits 99.83%, so it's pretty clear the model nails the classification task. Out of all the cases, it gets 608 preictal and 591 ictal instances right, and it only slips up twice one false positive and one false negative. That kind of performance really shows how well the model can tell preictal from ictal states. In short, 1D-EEGNet looks like a solid pick for detecting seizure-related events.

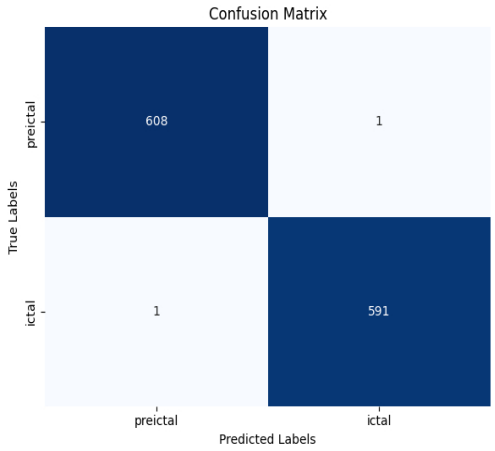


Figure 3. Confusion matrix

B. Cross-Subject Validation

Cross-subject (LOSO) validation really matters when it comes to using this in real clinical settings. The model nailed an average accuracy of 95.35% across all 14 patients, which says a lot it handled the differences between people's EEG patterns pretty well. Most patients achieved strong performance, although a few with atypical seizure characteristics showed comparatively lower results. Still, 1D-EEGNet kept up strong results for everyone overall.

To provide a more comprehensive understanding of the model's generalization ability, a detailed patient-wise evaluation is presented below. This includes accuracy, precision, recall, and F1-score for each of the 14 subjects. Such granular analysis highlights how the proposed 1D-EEGNet performs across diverse EEG patterns and clinical variations, offering deeper insight into its robustness under cross-subject (LOSO) evaluation. The patient-specific metrics are summarized in Table 1.

TABLE I
Patient-wise accuracy of the Siena dataset using the 1D-EEGNet model.

Patient ID	Accuracy (%)
PN00	98
PN01	98
PN03	85
PN05	100
PN06	95
PN07	91
PN09	100
PN10	79
PN11	100
PN12	100
PN13	98
PN14	100
PN16	91
PN17	100
Avg	**95.93**

C. Comparison with Baselines

The proposed model was compared with two deep learning baselines the 1D-CNN and 1D-ResNet, trained under identical conditions.

TABLE II
PERFORMANCE METRICS COMPARISON AMONG 1D-RESNET, 1D-CNN, AND PROPOSED 1D-EEGNET

Model	Accuracy
1D-ResNet	99.42%
1D-CNN	99.08%
Proposed 1D-EEGNet	**99.83%**

The results clearly show that 1D-EEGNet outperforms both networks while using fewer parameters and lower computational cost, making it well-suited for embedded applications.

D. Comparison with other methods

The proposed 1D-EEGNet model demonstrates exceptional performance, achieving an accuracy of 99.83% on the Siena dataset, thereby outperforming several state-of-the-art methods, including 1D-ResCNN 99.75%, Deep CNN 99.62%, Random Forest Classifier 85.71%, Ensemble Classifier 95.30%, and the Global-Local Weighted Attention GLWA model 98.89%. The model effectively captures temporal EEG features with minimal preprocessing, distinguishing it from conventional approaches that often demand substantial computational resources. Moreover, the streamlined architecture of 1D-EEGNet enhances its scalability, making it well-suited for real-time clinical applications. By combining high accuracy with computational efficiency, the proposed approach provides reliable seizure detection and facilitates seamless integration into practical diagnostic workflows.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT ALGORITHMS

Algorithm	Dataset	Accuracy
1D-ResCNN [7]	Siena, CHB-MIT	99.75%, 99.71%
Deep CNN [8]	Siena	99.62%
Random Forest Classifier [9]	Siena	85.71%
Ensemble Classifier [10]	Siena	95.30%
Global-Local Weighted Attention (GLWA) [11]	Siena	98.89%
Proposed 1D-EEGNet	**Siena**	**99.83%**

This confirms that our low-dimensional approach not only matches but exceeds the performance of more complex architectures.

VI. CONCLUSIONS

This study introduced a compact 1D-EEGNet model for automatic seizure detection using raw multichannel EEG. The design mixes temporal convolutions with depthwise and separable operations, so it grabs key spatiotemporal features without bogging down your hardware. On the Siena Scalp EEG dataset, the model hit 99.83% accuracy in standard tests and 95.35% in leave-one-subject-out cross-validation. It kept beating classic deep learning models like 1D-CNN and 1D-ResNet, showing it generalizes well and fits real-time or low-resource settings. Still, there are some practical limits. The Siena dataset doesn't have that many subjects, and the current task only covers basic seizure vs. non-seizure detection it doesn't sort out seizure types or catch pre-ictal states. Plus, the model's compactness means it might miss some long-term temporal patterns. Next steps could include pushing the model toward multi-class seizure classification, adding attention modules for better handling of long-term dependencies, and coming up with subject-independent adaptation methods to make it more robust. Bringing it to real time wearable or embedded systems, and making the results easier to interpret, will also help move this work closer to real clinical use.

REFERENCES

[1] A. T. Tzallas, M. G. Tsipouras, D. G. Tsalikakis, E. C. Karvounis, L. Astrakas, S. Konitsiotis, and M. Tzaphlidou, "Automated epileptic seizure detection methods: a review study," *Epilepsy-histological, electroencephalographic and psychological aspects*, pp. 75–98, 2012.
[2] I. Ullah, M. Hussain, H. Aboalsamh *et al.*, "An automated system for epilepsy detection using eeg brain signals based on deep learning approach," *Expert Systems with Applications*, vol. 107, pp. 61–71, 2018.
[3] A. Subasi, "Eeg signal classification using wavelet feature extraction and a mixture of expert model," *Expert Systems with Applications*, vol. 32, no. 4, pp. 1084–1093, 2007.
[4] A. T. Tzallas, M. G. Tsipouras, and D. I. Fotiadis, "Epileptic seizure detection in eegs using time–frequency analysis," *IEEE transactions on information technology in biomedicine*, vol. 13, no. 5, pp. 703–710, 2009.
[5] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.
[6] P. Detti, "Siena scalp eeg database (version 1.0.0)," PhysioNet, 2020, rRID:SCR_007345. [Online]. Available: https://physionet.org/content/siena-scalp-eeg/1.0.0
[7] P. Kumar and P. K. Upadhyay, "A hybrid optimization-enhanced 1d-rescnn framework for epileptic spike detection in scalp eeg signals," *Scientific Reports*, vol. 15, no. 1, p. 5707, 2025.
[8] P. Kumar, P. K. Upadhyay, and M. K. Panda, "Snsdeepnet: spike and non-spike detection in epilepsy," *Engineering Research Express*, vol. 6, no. 3, p. 035365, 2024.
[9] X. Jiang, X. Liu, Y. Liu, Q. Wang, B. Li, and L. Zhang, "Epileptic seizures detection and the analysis of optimal seizure prediction horizon based on frequency and phase analysis," *Frontiers in Neuroscience*, vol. 17, p. 1191683, 2023.
[10] B. Kapoor, B. Nagpal, P. K. Jain, A. Abraham, and L. A. Gabralla, "Epileptic seizure prediction based on hybrid seek optimization tuned ensemble classifier using eeg signals," *Sensors*, vol. 23, no. 1, p. 423, 2022.
[11] X. Li, K. Zhang, X. Wang, Z. Zhang, P. Zhu, M. Zhu, X. Zeng, and S. Chen, "Feature fusion based on global-local weighted attention model for automatic epileptic seizure detection," *Journal of Neural Engineering*, vol. 22, no. 5, p. 056016, 2025.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Spectral Regularization: Controlling Latent Variance

Md Nafis Ul Haq Shifat[1] Nujhat Ahmed Disha[1] Farhan Ahmad[1]
S M A Nahian[1]

[1] Massachusetts Institute of Technology

Abstract

When training autoencoders, the geometry of the learned latent space is typically left to chance—a byproduct of minimizing reconstruction error rather than an explicit design choice. We propose spectral regularization, a method that directly shapes the eigenvalue spectrum of the latent covariance matrix toward a desired target. This gives practitioners control over how variance is distributed across latent dimensions. In experiments on CIFAR-10, our approach improves linear probe accuracy by up to 2 percentage points over unregularized baselines, with learned spectra that closely match the prescribed exponential-decay targets. In settings where standard autoencoders suffer from variance inflation, spectral regularization strikes a balance—avoiding both extreme inflation and over-suppression—while achieving the lowest reconstruction error. On sentence embedding tasks (STS-B and QQP), it yields representations that compress better under PCA without sacrificing full-dimensional performance. These results suggest that the latent spectrum is a practical and effective lever for shaping learned representations.

1 Introduction

Representation learning methods like autoencoders compress high-dimensional data into a lower-dimensional latent space. In most projects, this latent space is treated as a byproduct: we train a model to minimize reconstruction loss, and whatever geometry the latents end up with is simply "what the model decided was useful."

But the latent space is itself a geometric object we might want to control. If we look at the covariance of the latent vectors, it has an associated spectrum: some directions have large variance, others collapse almost completely. This spectrum affects what information is preserved, how easily linear classifiers can separate classes, and how suitable the representation is for tasks like dimensionality reduction.

In this project, we ask: instead of passively accepting the latent geometry that training discovers, can we actively *shape* it? We introduce a simple spectral regularization term on the latent covariance. Instead of only minimizing reconstruction error, we also nudge the eigenvalues of the latent covariance toward particular behaviors.

If we could reliably steer the latent spectrum, it would give us a knob to design representations for specific applications: controlling how quickly variance decays across dimensions, preventing variance inflation linked to overfitting, and shaping geometry to be more compatible with downstream linear methods.

2 Background

An autoencoder has an encoder f_θ that maps an input x to a latent vector $z = f_\theta(x)$, and a decoder g_ϕ that maps back to a reconstruction $\hat{x} = g_\phi(z)$. Training minimizes reconstruction loss $\mathcal{L}_{\text{rec}} = \mathbb{E}_x[\|x - \hat{x}\|_2^2]$. The geometry of latent vectors is summarized by the latent covariance matrix:

$$\Sigma_z = \mathbb{E}\big[(z - \mu_z)(z - \mu_z)^\top\big], \quad \mu_z = \mathbb{E}[z].$$

The eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ of Σ_z form the *latent eigenvalue spectrum*. These spectra are often skewed: few directions carry most variance while others collapse.

We use *spectral regularization* to mean: adding a loss term that directly manipulates the eigenvalues of the latent covariance. Given batch latent vectors, we compute batch covariance $\hat{\Sigma}_z$ and eigendecompose: $\hat{\Sigma}_z = Q\Lambda Q^\top$. We define a target spectrum $\Lambda^\star$ and penalize deviations: $\mathcal{L}_{\text{spec}} = \|\Lambda - \Lambda^\star\|_2^2$, yielding total loss $\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha\,\mathcal{L}_{\text{spec}}$. Regularizing eigenvalues (not raw coordinates) makes the loss rotation-invariant.

2.1 Related Work

Prior work regularizes *summaries* of the spectrum rather than matching a full target eigenvalue profile. **Variance inflation:** Kim et al. (1) penalize total latent variance to address variance inflation. We act on the full eigenvalue spectrum instead. **DeCov:** Cogswell et al. (2) penalize off-diagonal covariance to reduce redundancy. Our loss constrains the entire eigenvalue distribution. **VICReg:** Bardes et al. (3) enforce per-dimension variance floors in the coordinate basis; we act on eigenvalues. **DIP-VAE:** Kumar et al. (4) regularize VAE posterior covariance toward a simple prior. We work with deterministic autoencoders and data-driven targets. **Matryoshka:** Kusupati et al. (5) train prefix-valid embeddings. Our approach is rotation-invariant and constrains eigenvalues directly.

3 Methodology

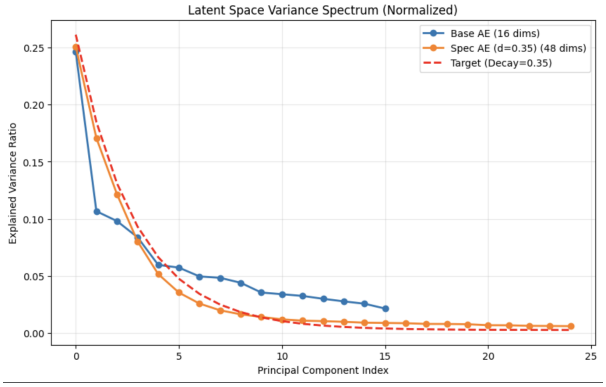
We conduct our main experiments on CIFAR-10 (32×32 RGB images, 10 classes). Our base model is a convolutional autoencoder with Conv–BatchNorm–ReLU encoder blocks with strided convolutions mapping to latent $z \in \mathbb{R}^d$, and a symmetric transposed-convolution decoder.

Spectral Regularizer. For a batch $\{x_i\}_{i=1}^B$, we compute latent vectors $z_i = f_\theta(x_i)$ and the batch covariance $\hat{\Sigma}_z$. We eigendecompose and penalize deviations from target eigenvalues:

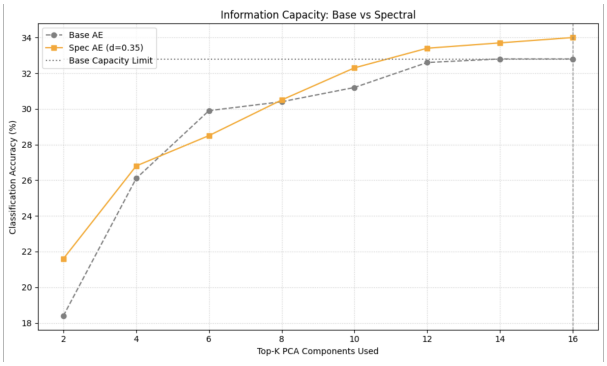
$$R_{\text{spectrum}}(\hat{\Sigma}_z) = \frac{1}{d}\sum_{k=1}^{d}(\bar{\lambda}_k - \bar{\lambda}_k^\star)^2.$$

The total loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \alpha\, R_{\text{spectrum}}(\hat{\Sigma}_z)$, with $\alpha \geq 0$ controlling spectral regularization strength.

Target spectrum design. We parameterize $\Lambda^\star$ as an exponential decay controlled by $\gamma > 0$ (the *target decay*): $\tilde{\lambda}_k^\star = \exp\big(-\gamma(k-1)\big)$ for $k = 1, \ldots, d$, normalized so that $\lambda_k^\star = \tilde{\lambda}_k^\star / \sum_{j=1}^d \tilde{\lambda}_j^\star$. We apply the same normalization to empirical eigenvalues, $\bar{\lambda}_k = \lambda_k / \sum_j \lambda_j$, so the loss depends only on the *shape* of the spectrum and is scale-invariant. We chose exponential decay because it outperformed other shapes in preliminary experiments.



(a) Latent covariance spectra



(b) Linear probe accuracy vs. top-k PCs

Figure 1: (a) Latent spectra for base AE (16-dim) and spectral AE (48-dim, γ=0.35) vs. target. (b) Information capacity measured as CIFAR-10 linear probe accuracy.

Table 1: CIFAR-10 test accuracy (%) of linear probes using top-k PCA components.

	γ=0.35		γ=0.25		γ=0.33	
k	Base	Spec	Base	Spec	Base	Spec
2	18.4	21.6	19.5	22.5	18.6	22.4
4	26.1	26.8	26.7	27.3	25.9	26.4
8	30.4	30.5	31.4	31.7	30.8	31.7
12	32.6	33.4	33.5	33.4	33.1	32.3
16	32.8	34.0	33.6	34.7	33.5	32.7

4 Experiments and Analysis

4.1 Experiment 1: Spectral Regularization on CIFAR-10

We ask how spectral regularization changes the latent *geometry* of a convolutional autoencoder on CIFAR-10, and how these changes affect *linear probing performance*.

Training setup. We train a *base autoencoder* ($d = 16$, $\alpha = 0$) and *spectral autoencoders* ($d = 48$) projected to lower dimensions via PCA. All models are trained for 50 epochs with Adam (lr=10^{-3}, β_1=0.9, β_2=0.999), weight decay 10^{-5}, and batch size 1024. Spectral models use $\gamma \in \{0.25, 0.33, 0.35\}$ with $\alpha \in \{1.0, 1.5\}$.

Linear probing protocol. After training, we freeze the encoder, fit PCA on training latents, project onto top-k components, and train linear classifiers to measure test accuracy.

Results. Figure 1(a) shows the base autoencoder exhibits a pathological spectrum: the first PC dominates with over 25% variance, followed by steep drop-off. Spectral AEs closely track their exponential targets. These geometric changes translate into gains in linear probing (Figure 1(b), Table 1). For the base AE, accuracy plateaus around k≈14 at 33%. Spectral variants ($\gamma = 0.25, 0.35$) reach 34.7% at k=16, though $\gamma = 0.33$ underperforms at high k, showing hyperparameter sensitivity.

Robustness check. To verify gains come from spectrum shaping rather than larger latent dimension, we repeat with both models at $d = 48$. The spectral model ($\gamma = 0.4$, $\alpha = 1.0$) consistently outperforms: at $k = 16$, spectral achieves 34.6% vs. base 33.8%.

4.2 Experiment 2: Variance Inflation Study

We study *variance inflation*: when latent variance greatly exceeds input variance despite low reconstruction error. Using an overparameterized MLP autoencoder ($d = 256$; layers $3072 \to 1024 \to 512 \to 256$) on a 1k-image CIFAR-10 subset, we compare: (i) a decoder-regularized baseline, (ii) a variance-regularized AE (1) that penalizes $\sum_j \mathrm{Var}(z_j)$, and (iii) our spectral AE.

Results. Table 2 shows the decoder-only baseline has extreme inflation (ratios $\approx 10^3$), the variance-regularized AE aggressively suppresses variance (ratios ≈ 0.03), while our spectral AE achieves moderate ratios (≈ 0.5) with better reconstruction MSE (0.030 vs. 0.040).

Table 2: Variance inflation experiment: latent-to-input variance ratios and final MSE.

Model	Train Ratio	Test Ratio	MSE
Base AE (decoder-only L2)	2451	2312	0.088
VarReg AE	0.028	0.023	0.040
Spectral AE (ours)	0.576	0.507	0.030

4.3 Experiment 3: Spectral Fine-Tuning of Sentence Transformers

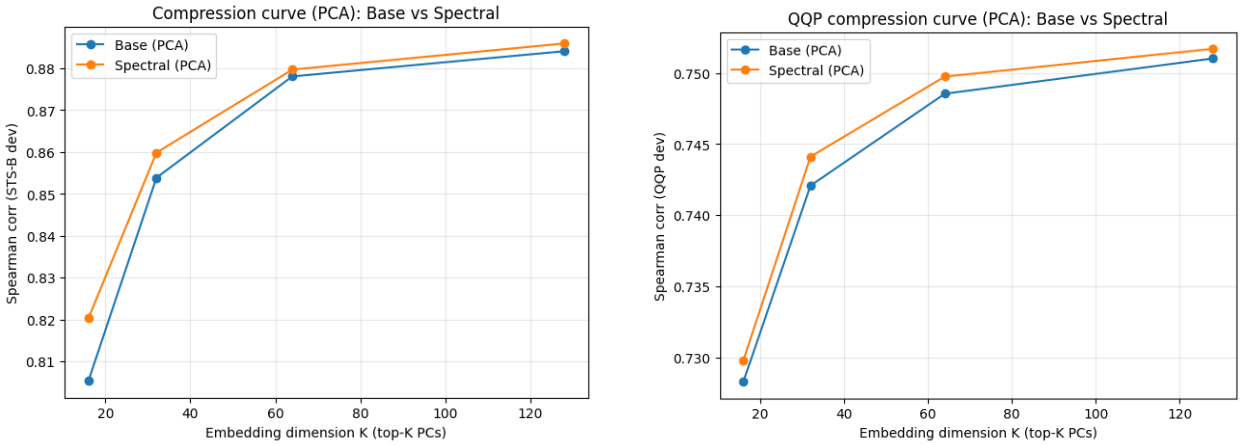
We apply spectral regularization to a sentence encoder and test whether it makes embeddings more *compressible*.

Setup. We evaluate on **STS-B** (sentence similarity, scores in $[0, 1]$) and **QQP** (paraphrase detection). We use `mpnet-base` with 128-dim projection. Base loss is MSE between cosine similarity and target; spectral model adds R_{spectrum}. We fine-tune with AdamW (lr=2×10^{-5}, weight decay 0.01); batch size 32 for 15 epochs (STS-B) and 64 for 5 epochs (QQP).

Results. On both datasets, the spectral model matches base performance at full 128-D but achieves higher Spearman correlation under PCA compression (Figure 2). For example, on STS-B at $k = 16$: spectral achieves 0.82 vs. base 0.80. Table 3 shows the spectral model concentrates variance into task-relevant directions: effective rank and participation ratio decrease modestly while compressibility improves substantially.

Table 3: Spectrum statistics on dev-set embeddings.

Dataset	Model	Total Var	Eff. Rank	Part. Ratio
STS-B	Base	0.539	80.61	60.03
	Spectral	0.519	79.39	58.11
QQP	Base	0.782	88.85	64.54
	Spectral	0.813	83.36	55.80



(a) STS-B (b) QQP

Figure 2: Sentence encoder results: Spearman correlation vs. number of PCA dimensions.

4.4 Discussion

Our experiments suggest spectral regularization is a practical knob for shaping latent geometry. Benefits are most pronounced when base spectra are poorly behaved (variance inflation, early saturation), and more modest for well-structured pre-trained encoders.

5 Limitations

We focus on small autoencoders and CIFAR-scale datasets; scalability remains untested. We use hand-designed target spectra rather than learning them jointly. We measure only linear probing, not clustering or retrieval. Eigendecomposition per batch adds overhead for high-dimensional latents. Future work includes adaptive target spectra and combining with contrastive objectives.

6 Conclusion

We showed that adding a spectral regularizer on latent covariance eigenvalues allows us to shape the latent spectrum toward desired profiles while maintaining reconstruction quality. On CIFAR-10, this improves linear-probe capacity. In variance-inflation settings, spectral regularization avoids extreme inflation while achieving better reconstruction. For sentence encoders, embeddings become more PCA-compressible. Overall, operating on latent eigenvalues provides a practical knob for steering representation geometry—most effective when base spectra are poorly behaved.

References

[1] B. Kim, K. H. Ryu, J. H. Kim, and S. Heo, "A neural network regularization method to address variance inflation in autoencoders," *Computers & Chemical Engineering*, 2022.

[2] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, "Reducing overfitting in deep networks by decorrelating representations," *ICLR*, 2016.

[3] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance–invariance–covariance regularization for self-supervised learning," *ICLR*, 2022.

[4] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," *ICLR (DIP-VAE)*, 2018.

[5] A. Kusupati, V. Deshpande, J. Bauer, S. Jegelka, and A. Kumar, "Matryoshka representation learning," *NeurIPS*, 2022.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Activation Patching for Whisper Interpretability: Mapping the Acoustic-to-Linguistic Transition in Whisper

Shanzita Siddiqua
Dept. of Computer Science, Rice University
Houston, TX 77005
ss516@rice.edu

Abstract

We apply activation patching to OpenAI's Whisper model to causally analyze how speech representations evolve across encoder layers. Experiments on 869 English word pairs and 162 English-Spanish pairs reveal that early encoder layers (0–1) encode language-agnostic acoustic features, while later layers (2–3) specialize for language-specific processing. Cross-lingual patching succeeds at 93% when targeting Layer 0 but falls to 25% at Layer 3, demonstrating a hierarchical transition from universal phonetic representations to language-dependent encoding.

1 Introduction

Understanding how neural speech models transform acoustic signals into linguistic representations remains a fundamental challenge. OpenAI's Whisper [Radford et al., 2022], trained on 680,000 hours of multilingual audio, achieves strong performance across 100+ languages, yet its internal mechanisms are poorly understood.

Activation patching [Meng et al., 2022] has emerged as a powerful tool for causal analysis of neural networks. This technique has revealed interpretable structure in language models [Geva et al., 2023] and vision-language systems [Palit et al., 2023], but has not been applied to speech recognition.

We address this gap by applying activation patching to Whisper's encoder. Using 869 minimal word pairs and 162 English-Spanish pairs, we ask: (1) which encoder layers determine transcription output, and (2) do early layers learn language-agnostic representations? Our cross-lingual experiments reveal a clear hierarchy—patching succeeds at 93% for early layers but drops to 25% for late layers—providing the first causal evidence that Whisper transitions from universal acoustic features to language-specific encoding, with Layer 2 marking the critical transition point.

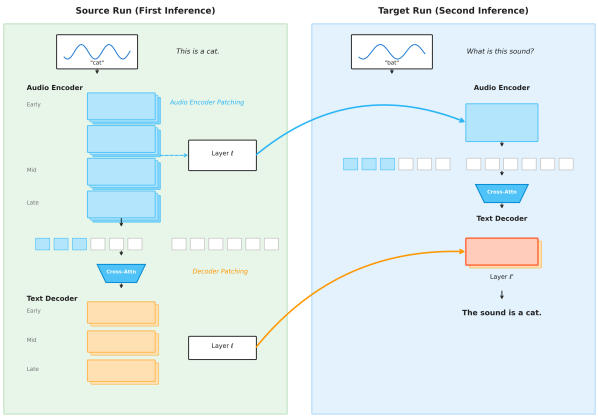


Figure 1: Overview of activation patching. Hidden states from a source audio's encoder layer are injected into the target audio's forward pass. If the output changes, that layer causally influences the transcription.

2 Related Work

Whisper and Speech Recognition. End-to-end speech recognition has evolved from attention-based encoder-decoder models [Chan et al., 2016] to large-scale transformer architectures. Whisper [Radford et al., 2022] trains on 680,000 hours of weakly-supervised web audio to achieve robust multilingual recognition. Unlike self-supervised models such as wav2vec 2.0 [Baevski et al., 2020], Whisper learns directly from transcription pairs. Probing studies have analyzed Whisper's internal representations: Pasad et al. [2021] found that middle layers of speech encoders correlate most strongly with phonetic information, while later layers encode more abstract linguistic features. However, such correlational analyses cannot establish causality—a gap our activation patching approach addresses.

Activation Patching. Activation patching identifies which model components causally influence outputs by replacing hidden states between forward passes [Meng et al., 2022]. Given source input x_s and target input x_t, patching replaces the hidden state $h_t^{(\ell)}$ at layer ℓ with $h_s^{(\ell)}$ and measures how this affects output. This technique has localized factual knowledge in GPT models [Meng et al., 2022] and identified circuits in transformers [Wang et al., 2022], providing stronger causal evidence than correlational probing.

Multimodal Interpretability. Palit et al. [2023] applied activation patching to vision-language models, revealing that mid-depth injections maximally influence outputs. Speech models occupy a unique position: like vision, they process continuous signals; like language, they produce sequential discrete outputs. Our work extends activation patching to the audio modality.

3 Dataset

We construct two datasets for our experiments. Code and data are available at `https://github.com/Shanzita/whisper-patching`.

English Minimal Pairs (n=869). We curate word pairs across three categories: consonant contrasts (550 pairs differing in a single consonant, e.g., "bat"/"cat", "ship"/"chip"), vowel contrasts (120 pairs differing in a single vowel, e.g., "bit"/"bat", "ship"/"sheep"), and semantic controls (199 unrelated word pairs to test general override, e.g., "apple"/"orange").

Cross-Lingual Pairs (n=162). We construct English-Spanish pairs consisting of cognates (40 pairs with similar pronunciation, e.g., "telephone"/"teléfono", "hospital"/"hospital") and translations (122 pairs with different pronunciation, e.g., "cat"/"gato", "house"/"casa").

Audio Generation. All audio is generated using Google Text-to-Speech (gTTS) at 16kHz to ensure consistent acoustic quality and eliminate speaker variability as a confound. Each audio clip is padded with silence to 30 seconds to match Whisper's expected input format. Total dataset comprises 751 unique English words and 162 Spanish words.

4 Method

4.1 Whisper Architecture

We analyze Whisper-tiny, consisting of a 4-layer encoder and 4-layer decoder with hidden dimension $d = 384$. Raw audio is converted to an 80-dimensional log-mel spectrogram, processed by two convolutional layers, then passed through 4 transformer blocks. The decoder generates text tokens via cross-attention to encoder outputs.

4.2 Activation Patching Framework

Let $\mathcal{M}$ denote the Whisper model. We define a patching intervention $\mathcal{M}_{s \to t}^{(\ell_s, \ell_t)}$ that: (1) extracts hidden state $h_s^{(\ell_s)}$ from source audio s, (2) replaces $h_t^{(\ell_t)}$ with $h_s^{(\ell_s)}$ during target audio t's forward pass, and (3) returns the patched transcription.

We quantify effectiveness using *counterfactual override accuracy*:

$$\text{Acc}(s, t, \ell_s, \ell_t) = \begin{cases} 1.0 & \text{if output} = w_s \\ 0.5 & \text{if output} \notin \{w_s, w_t\} \\ 0.0 & \text{if output} = w_t \end{cases} \tag{1}$$

where w_s and w_t are the source and target transcriptions. A score of 1.0 indicates complete override—the patched model outputs the source word despite receiving target audio. A score of 0.5 indicates partial disruption. A score of 0.0 indicates no effect. We aggregate scores to compute the override matrix $\mathbf{O}[\ell_s, \ell_t] = \frac{1}{N} \sum_{i=1}^{N} \text{Acc}(s_i, t_i, \ell_s, \ell_t)$.

5 Experiments and Results

5.1 Experiment A: Encoder Layer Patching

We perform systematic activation patching across all 16 layer combinations on 869 minimal word pairs.

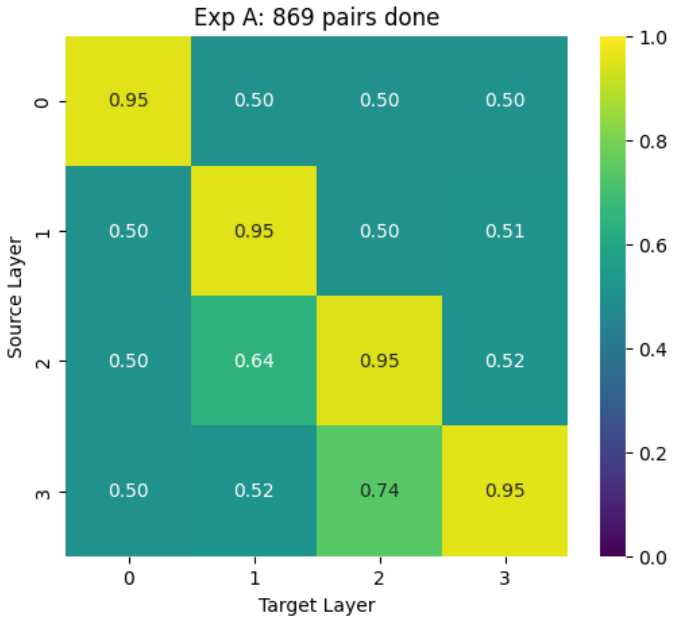


Figure 2: Override accuracy matrix for Experiment A (n=869 pairs). Diagonal entries show 95% accuracy. Off-diagonal asymmetry reveals late-to-early transfer is more effective than early-to-late.

Results. Figure 2 presents the override accuracy matrix. Key findings:

- **Same-layer patching achieves 95% accuracy.** Diagonal entries show that each layer's representation suffices to determine output when injected at matching depth. This consistency across all four layers indicates that Whisper learns depth-specific representations.
- **Cross-layer patching shows asymmetric transfer.** Early→late patching yields only 50–52% accuracy (essentially random), while late→early shows moderate transfer (Layer 3→2: 74%, Layer 2→1: 64%).

- **Layer 2 is a critical transition point.** The (2,1) and (3,2) cells show 64% and 74% accuracy respectively, suggesting that middle layers contain increasingly abstract representations that can partially override earlier processing.

These findings provide causal confirmation of the hierarchical structure suggested by Pasad et al. [2021]'s probing results, demonstrating that the correlation between middle layers and phonetic information reflects a genuine causal role in determining transcription output.

5.2 Experiment B: Cross-Lingual Transfer

We patch English encoder states into Spanish audio processing across 162 word pairs.

Results. Figure 3 presents the cross-lingual transfer matrix:

- **Layer 0 as target: 93% transfer.** Early layers accept English representations readily, indicating language-agnostic acoustic features that generalize across languages.
- **Layer 3 as target: 25% transfer.** Late layers reject cross-lingual patching, encoding language-specific information including lexical patterns and morphological structure.
- **Layer 3 as source: 89% transfer.** English Layer 3 contains rich lexical information that overrides Spanish processing when injected at any target layer.

The transition from language-agnostic to language-specific encoding occurs between Layers 2 and 3, with transfer dropping from 75% to 25%. Cognates showed higher transfer rates than non-cognate translations at all layers (e.g., 96% vs. 92% at Layer 0), likely due to phonetic similarity.

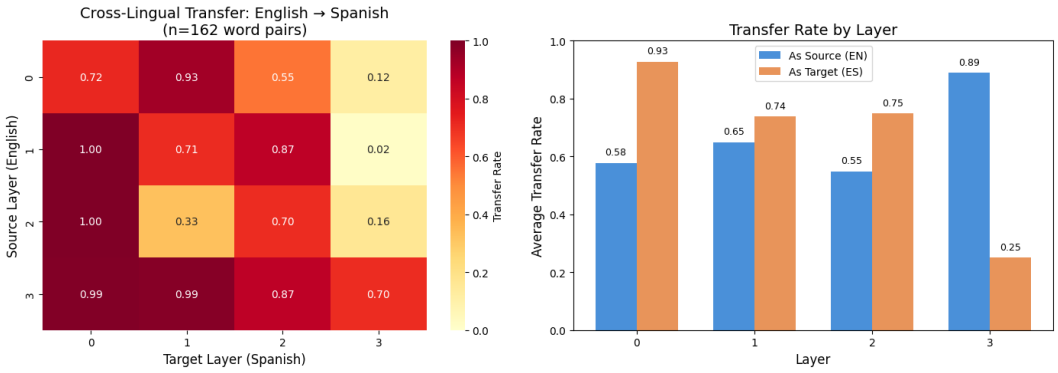


Figure 3: Cross-lingual transfer (English → Spanish, n=162 pairs). Left: Transfer by source-target layer. Right: Average transfer by target layer.

5.3 Summary of Results

Table 1 consolidates our key findings across both experiments.

Finding	Metric	Value
Same-layer patching (Exp A)	Override accuracy	95%
Early→Late patching (Exp A)	Override accuracy	50–52%
Late→Early patching (Exp A)	Override accuracy	52–74%
Cross-lingual, Layer 0 target (Exp B)	Transfer rate	93%
Cross-lingual, Layer 3 target (Exp B)	Transfer rate	25%

Table 1: Summary of key experimental results across 869 English pairs and 162 cross-lingual pairs.

6 Analysis and Discussion

Hierarchical Representation Learning. Early layers (0–1) encode low-level acoustic features shared across languages, while late layers (2–3) encode language-specific information. The transition occurs primarily between Layers 2 and 3, with Layer 2 marking where acoustic representations become increasingly abstract.

Asymmetric Information Flow. The asymmetry in Experiment A—where late layers can partially override early layers but not vice versa—suggests an information bottleneck. Later layers compress acoustic information into higher-level representations that early layers cannot fully interpret. Formally, if $I(h^{(\ell)}; X)$ denotes mutual information between layer ℓ and input X, we expect $I(h^{(3)}; X) < I(h^{(0)}; X)$. This compression explains why early-to-late patching fails—early layers contain redundant acoustic detail that late layers have learned to discard.

Comparison to Other Modalities. Our findings contrast with vision transformers, where Dosovitskiy et al. [2021] showed that early layers learn universal features transferable across depths. Whisper's encoder resembles language models more closely, where layer-specific representations emerge from sequential processing. This may reflect the temporal nature of speech: unlike images, audio requires progressive integration of context over time.

Implications for Multilingual ASR. The language-agnosticism of early layers explains Whisper's strong multilingual performance. By learning universal phonetic features in Layers 0–1, Whisper processes any language using shared acoustic representations, deferring language-specific processing to later layers. This finding extends similar observations in self-supervised speech models: **?** found that wav2vec 2.0 also exhibits hierarchical phonetic-to-lexical encoding, suggesting this may be a general property of speech transformers rather than architecture-specific.

Potential Applications. Our findings suggest practical interventions for controlling Whisper's behavior:

- **Language steering:** Patching early layers could bias transcription toward a target language without retraining.
- **Accent adaptation:** Replacing early-layer representations from native speech could improve recognition of accented audio.
- **Error correction:** Targeted patching at Layer 2–3 could correct systematic misrecognitions by injecting correct lexical representations.

Broader Impact. Mechanistic understanding of speech models is crucial as they are deployed in high-stakes applications—medical transcription, legal proceedings, accessibility tools. Our activation patching framework provides a foundation for auditing model behavior and identifying failure modes. Unlike black-box approaches, causal methods reveal *why* models behave as they do, enabling more trustworthy deployment.

7 Conclusion

We presented the first activation patching study on a speech recognition transformer. Through systematic experiments on Whisper's encoder, we established three key findings:

1. **Same-layer patching achieves 95% override accuracy**, demonstrating that each encoder layer's representation is sufficient to determine transcription output when injected at matching depth.
2. **Early encoder layers are language-agnostic.** Cross-lingual patching from English to Spanish succeeds at 93% for Layer 0 but drops to 25% for Layer 3, revealing a hierarchical transition from universal acoustic features to language-specific encoding.

3. **Layer 2 marks a critical transition point** where acoustic representations become increasingly abstract and language-dependent.

These findings provide causal evidence for hierarchical representation learning in speech transformers, extending mechanistic interpretability from text and vision to the audio modality. Our work opens avenues for understanding and controlling speech model behavior through targeted activation interventions.

Limitations. We analyzed only Whisper-tiny (39M parameters); larger models with more encoder layers may exhibit different representational hierarchies. Our TTS-generated audio lacks the prosodic variation, background noise, and acoustic diversity of natural speech, potentially overestimating patching effectiveness. We examined only English-Spanish, which share Indo-European roots; transfer patterns may differ for typologically distant pairs (e.g., English-Mandarin). Finally, encoder-only analysis leaves decoder and cross-attention patching for future work.

Acknowledgments

We thank Hanjie Chen and the Chili Lab at Rice University for valuable discussions.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 2020.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *ICASSP*, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *EMNLP*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *NeurIPS*, 2022.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Tan. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. *arXiv preprint arXiv:2308.11207*, 2023.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. *IEEE ASRU*, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Kevin Wang et al. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Undergraduate Conference on Intelligent Computing and Systems (UCICS 2026)
Varendra University, Rajshahi, Bangladesh
https://ucics.org/

Risk of Dengue Associated With Probability Based on Interpretations From Hematology-Guided Explainable AI

[1]Parishmita Sinha, [1]Sharmin Akhter Rani, [1]Nirab Debnath, [1]Mishu Talukder
[1]Department of CSE, Leading University, Sylhet, Bangladesh
Email: parishmitasinha02@gmail.com, sharminintisha17@gmail.com, nirab200174@gmail.com, talukdermishu16@gmail.com

***Abstract*—The early detection of dengue risk through routine blood tests is very important for the timely clinical intervention. The objective of this study is to establish an interpretable, data-driven framework that would not only be able to forecast the dengue risk but also disclose the critical hematological features linked with the increased risk levels. The dataset has 1523 clinically verified dengue patients cases that was analyzed to assess the performance of several machine learning and deep learning techniques. CatBoost was, however, the top performer among them with an accuracy of 0.858 and an AUC-ROC of 0.9074. SHAP analysis were applied to all models to pointed out that hemoglobin levels, total neutrophil counts, total lymphocyte counts, and total platelet counts were the most significant predictors, in line with the known pathology of dengue. A probability-based patient-specific risk scoring system was also made to categorize the patients into three groups such as low risk ($\leq 40\%$) defines a low probability of dengue risk, medium (41–70%) indicate a noticeable likelihood requiring closer clinical oversight, and high-risk($\geq 70\%$) highlights a high chance of dengue that need immediate attention. The proposed risk analysis using routine hematological data thus supported early clinical decision-making and patient management in cases of dengue.**

***Index Terms*—Dengue risk, Probabilitybased, SHAP analysis, CatBoost, Machine learning, Hematological values.**

I. Introduction

Dengue fever is now regarded as one of the most serious vector-borne diseases in the world and it is, therefore, a major public health concern in the tropical and subtropical regions where millions of people are infected every year. The World Health Organization (WHO) estimates that nearly 390 million individuals contract the disease annually, which means that half of the global population lives in areas where they are at risk of infection. Roy et al. [1] described that dengue virus infection manifests with varying clinical severity with mild fever to critical conditions such as dengue hemorrhagic fever and dengue shock syndrome. The continuous rise in the outbreaks in the highly populated countries like Bangladesh, India, and Thailand further emphasizes the pressing requirement for effective, fast, and easily accessible diagnostic methods. In Bangladesh, a case control analysis by Rahman et al. [2] found that temporarily living in urban areas, older housing conditions, and restricted daily mobility are key risk factors for dengue infection. On the other hand, Yang et al. [3] identified plasma leakage, hemorrhagic manifestations, and younger age as critical indicators of severe disease. Although traditional laboratory-based dengue diagnosis is very reliable, it is still very time-consuming, costly, and often unavailable in health-care settings that are operating under severe resource constraints. Louis et al. [4] stated that at a population level, spatial and computational modeling studies highlight the significant influence of environmental, demographic, and socioeconomic conditions on where dengue risk is concentrated. This very aspect underlines the relevance of data-driven predictive modeling for the early detection of disease outbreaks and also for managing the patients. In addition Hasan et al. [5] highlights climatic factors including rainfall, humidity, and temperature oscillations are known to impact dengue spread in tropical areas. Poverty and poor understanding of mosquito breeding environments face a higher risk of developing severe dengue described by Roy et al. [6]. This study presents the Explainable Dengue Risk Prediction approach utilizing the traditional machine learning and deep learning algorithm trained on hematological features including platelet count, WBC count, hemoglobin level, hematocrit percentage, MPV, and other related biomarkers. The proposed model classifies patients as dengue positive or negative and at the same time gives an interpretable risk-level analysis in three categories—Low, Medium, and High Risk—based on the threshold of predicted probability. Such a probabilistic assessment allows the system to not only predict the infection but also to estimate its severity level, thereby supporting the clinical decision-making process and patient triage. In order to make the findings more understandable and reliable, the model has been incorporated with SHapley Additive exPlanations (SHAP), a very well-known method in Explainable AI (XAI). The SHAP values give a quantitative measure of the contribution of each feature to the final decision of the model, thus allowing the doctors to understand which hematological parameters have the strongest influence on the prediction of the risk of dengue. For instance, a drop in platelet count or WBC level may be one of the factors that lead to a higher predicted probability of dengue infection. The accessibility of this aspect of the model helps to connect between the AI-based predictions and clinical reasoning which in turn increases the trust and applicability of the model in real-world healthcare settings. The predictive systems can greatly assist the doctors in making the differentiation between the

cases of high risk, medium risk and low risk patients in a quicker, less costly, and more efficient manner. The XGBoost model that was trained reached a classification accuracy of 0.848 and an ROC-AUC score of 0.89, thus outperforming many of the baseline models and at the same time making use of the SHAP-based ways of explanations for interpretation.

The most important contributions of this paper are the following ones:

- Incorporation of Explainable AI (SHAP) for swaying the model and clinical to be well understood.
- Launching of a risk stratification system (Low, Medium, High) which is based on the probabilistic model outputs.
- The use of SMOTE for tackling the data imbalance problems which resulted in better classification performance.

II. Literature Review

Dengue risk evaluation is crucial for healthcare decision-making and public health authorities to minimize severe cases and improve patient care. Despite the use of machine learning in dengue diagnosis, interpretable risk stratification using common blood tests has been rarely addressed. Many ML and DL models have been applied to classify dengue risk and predict dengue severity. This section will provide an overview of recent work using ml and dl techniques of dengue risk factors. A study of 2022, Ming et al. [7] created traditional machine learning models and validated it using routine clinical data to forecast dengue shock syndrome (DSS) risk in hospitalized dengue patients in Vietnam. The neural network model shows the robust predictive ability (AUROC 0.83) and performed exceptionally well in excluding DSS, indicated by a high negative predictive value. This study emphasizes the usefulness of ML-enabled decision support systems for enhancing risk stratification and clinical management in settings with limited healthcare resources. In the same year, Abdualgalil et al. [8] presents an efficient machine learning approach for early dengue identification using clinical data that was developed and tested, focusing on Yemen. Five algorithms such as KNN, Gradient Boosting, Extra Trees Classifier (ETC), XGBoost, and LightGBM were applied using both 10-fold and holdout cross-validation. The Extra Trees Classifier achieved the highest accuracy 99.12% with holdout and 99.03% with 10-fold, which highlights its capacity to deliver fast and accurate identity support for dengue identification. Another study of Chowdhury et al. [9] shows a combined supervised–unsupervised machine learning approach enhanced with SHAP based feature interpretation to predict dengue severity. The clinical data from Vietnam and Bangladesh were used and the XGBoost model exhibited the best predictive performance for severe cases. SHAP values identified platelet count and hematocrit as dominant features, while clustering on Bangladeshi data disclose high-risk patient groups, offering actionable insights for early interventions in endemic regions. In the same year 2022, Mayrose et al. [10] proposed an AI based approach for automatic dengue detection using peripheral blood smear imagery, merging blob-based segmentation, multi-level feature extraction, and diverse machine learning models. The SVM classifier achieved the highest accuracy of 95.74%, which shows its capability for reliable diagnostic support in clinical practice. Riya et al. [11] in 2024, presents an ensemble learning approach for dengue identification trained on CBC data from the dengue outbreak in Bangladesh in 2023. A stacking ensemble models such as LightGBM, XGBoost, Logistic Regression, and Multilayer Perceptron achieved highest accuracy which is 96.88% .This results shows how ensemble models can enhance early dengue detection by offering a fast and reliable solution for healthcare systems with limited resources. Another study in the same year, Thanh et al. [12] shows a machine learning based risk prediction approach to estimate mechanical ventilation requirements in Vietnamese children with dengue shock syndrome.The Analysis of over 1,200 patient records using support vector machine and random forest models achieved strong predictive accuracy, identifying increased fluid therapy and low platelet counts and high fluid infusion volumes.This approach supports early warning capability of critically ill patients, enhancing management of severe dengue in constrained healthcare settings. At same time Thapa et al. [13] evaluated the serological and hematological parameters among suspected dengue cases to determine diagnostic predictors. Dengue-positive cases exhibited low platelet and leukocyte counts, erythrocytosis, hemoconcentration, and reduced MCH and MCHC values.Logistic regression analysis revealed platelet count, total leukocyte count, MCH, MCHC, neutrophil, and lymphocyte counts as significant predictors, offering practical markers for early and precise diagnosis and improved clinical management. In the year 2024, Panchanadikar et al. [14] analyzed 157 pediatric dengue cases to determine risk factors for severe illness in India.Severe dengue affected 32.5% of patients, with a case-fatality rate of 3.2%. The multivariate analysis indicated that high-grade fever ($\geq 39^{\circ}\mathrm{C}$) and petechiae were critical clinical markers, while prolonged prothrombin time and platelet counts below 50,000/cu mm represented significant laboratory risk factors. These results emphasised that importance of careful monitoring to facilitate early detection and management of severe dengue in children.A study of 2025, Madewell et al. [15] shows confirmed dengue cases in Puerto Rico where the data from over 1,700 were used to compare machine learning models with WHO warning signs for predicting severe dengue.Gradient boosting models with CatBoost achieving the highest accuracy (AUC-ROC 97.1%), outperforming traditional warning sign–based approaches. Hemoconcentration, leukopenia, and days since symptom onset were among the most influential predictors with robust performance maintained even in the absence of laboratory data. The results support the utility of ML models for improved risk stratification and clinical decision support in resource-limited environments. Another study in the same year, Pulock et al. [16] analyzed the dengue cases

of 1313 during the 2023 outbreak in Bangladesh aimed to identify predictors of severe disease.The cases which are classified as severe were nearly 20% with young adults and males disproportionately affected. Severe dengue was linked to gastrointestinal symptoms, low platelet counts, elevated ALT, and warning signs such as severe abdominal pain, lethargy, respiratory distress, and fluid accumulation (pleural effusion, ascites). At the same time Anh et al. [17] evaluated a machine learning model for dengue risk stratification to inform the patient triage and hospitalization decisions by profiling 48 inflammatory biomarkers upon admission. Random forest analysis highlighted a selective panel such as HGF, TNF-β, MIP-1β, SCGF-βthat combined with fever duration, which achieved around 80% accuracy in differentiating severe or warning-sign dengue from mild cases. Moreover, another study in 2025 by Alam et al. [18] shows dengue incidence in Bangladesh from 2008 to 2024, which explore the role of meteorological factors through time series forecasting and regression modeling. The analysis disclosed the pronounced seasonal patterns and significant associations between dengue cases and temperature, humidity, and wind speed with the peak incidence observed in September 2023. ARIMA models provided the highest predictive accuracy that support the integration of climate information into early warning and control strategies for public health management. After reviewing all the previous work, the main inspiration of this study came from Madewell et al. [14] who worked for predicting severe dengue among clinical confirmed cases in Puerto Rico and trying to improve early identification of high risk from severe dengue. Following this study, identification of risk factors in dengue patients is necessary in the medical sector. The aim is to identify risk factors among dengue patients using clinical values to guide treatment decisions.

III. Dataset Description

The dataset used in this study contains clinical data which were obtained from 1,523 individuals at Jamalpur 250-Bedded General Hospital, Bangladesh, between February 10 and September 27, 2024, under strict ethical compliance to maintain patient privacy. For every patient, detailed clinical and blood-related data were recorded to illustrate how dengue influences major hematological parameters.The dataset includes 19 hematological and demographic attributes commonly used to assess dengue, such as hemoglobin, hematocrit, RBC and WBC counts, platelet count, neutrophils, lymphocytes, MPV, PCT, RDW, and other key clinical indicators. Each feature is clinically relevant for diagnosing dengue or evaluating disease progression, with detailed descriptions and standard reference ranges. Patients ranged in age from 5 to 78 years, and the gender distribution was approximately 47% male and 53% female. These variables are vital for enhancing diagnostic accuracy and building predictive systems that support early identification and severity grading of dengue infections, ultimately aiding in timely clinical decision-making and improved patient care. Analysis of population-level distributional patterns for critical hematological markers (e.g., lymphocyte percentages) show clear differences between people with and without dengue, supporting what is already known about dengue-related changes in the blood, such as low lymphocyte counts. The complete dataset has been made openly available on Mendeley Data for further study [19].

A. Preprocessing of Dataset

The first step was preprocessing, where the dengue dataset was subjected to a series of operations such as encoding of categorical variables, standardization, and class balancing. Label encoding was applied to the Gender column, whereas the target variable Result was transformed into binary numbers, with 1 signifying a positive dengue diagnosis and 0 for a negative diagnosis. Subsequently, the data was partitioned into attributes (X) and target (y), and further using StandardScaler, the feature set was standardized to make sure that all variables had mean zero and standard deviation one, thus preventing large magnitude features from dominating the model.

B. SMOTE for data balancing

SMOTE [20] eliminates bias in unbalanced datasets by creating new synthetic instances for the minority class instead of just copying existing samples, thus the situation in which one class is very poorly represented is avoided. In the dataset, positive cases were underrepresented compared to negative cases, so SMOTE [21] identifies the nearest neighbors of minority-class samples and creates new, realistic synthetic examples along the connections between them. This method has the effect of increasing the diversity of data, which in turn assists the model to perform better on the unseen cases. The synthetic samples are not real observations which sometimes can cause difficulties regarding interpretation or higher variance. For the sake of transparency, the analysis with SHAP was limited to the original test samples and 5-fold cross-validation was used to verify the stability of model performance.

IV. Methodology

This study applied machine learning and deep learning algorithms for dengue risk classification from clinical parameters. It uses quantitative methodology to classify the risk factors. The dataset was imbalanced and to balance the dataset SMOTE was applied for generating realistic minority-class samples. After that, we performed complete data preprocessing for this split, the dataset divided into 80% for training and 20% for testing sets. It implemented a suite of machine learning and deep learning models, including Decision tree, Random forest, CatBoost, CNN and ANN. These models show accurate tuning to optimize their performance. The performance was evaluated by using accuracy, recall, precision,F1-score and AUC-ROC

metrics. The performance also shows the probability. Based on predicted probabilities, patients were assigned to High (> 0.7), Medium (0.4–0.7), or Low (≤ 0.4) Risk categories. This method facilitates understanding of dengue risk both globally and for each individual case. Then plotting the best performed model ROC curve. In Fig 1. the workflow of our study has been given.

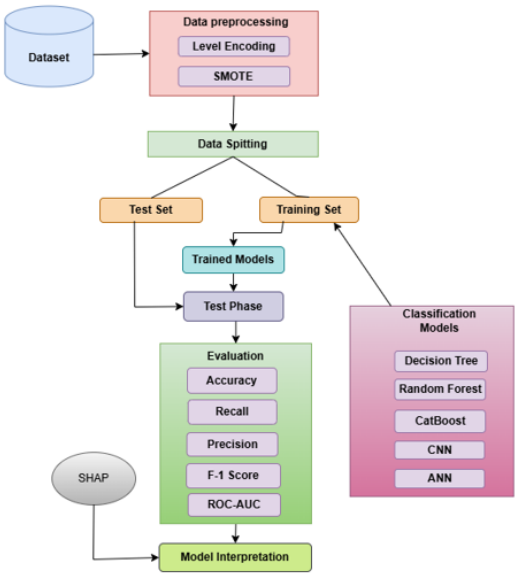


Fig. 1. A flowchart of the proposed methodology

A. Classification of Approach

- **Decision Tree**: Decision Tree: During dengue infection, hematological signs interact in non-linear ways and a Decision Tree classifier is well-suited for capturing these threshold-based clinical patterns. GridSearchCV optimization was performed to improve stability and avoid overfitting and the selected the best parameters criterion='entropy', max depth=12, max features=None, min samples leaf=4, and min samples split=10. LabelEncoder used to convert in numeric value all feature values were standardized using StandardScaler.
- **Random Forest**: The dengue dataset was initially scaled using StandardScaler to normalize all numerical features on a consistent scale, making hematological measurements comparable. A Random Forest classifier was then trained, with GridSearchCV identifying the best hyperparameters. The process include fitting with 5-fold cross-validation for each of 24 candidate parameter combinations, totalling 120 fits, which ensured a thorough evaluation of possible configurations.The best set up allowed unlimited tree depth, 200 trees, a minimum split of 2, and at least 1 sample per leaf. This setup enabled the model to grow fully developed trees while keeping splits balanced, effectively capturing intricate patterns in patient data.
- **CatBoost**: The CatBoost model was fine tuned with well-chosen hyperparameters so it could perform better and more accurately learn the complex patterns within the hematological data. To create a model that learns reliably without overfitting, CatBoost was trained with 1000 iterations and a 0.05 learning rate for stable learning, using a depth of 8 and L2 regularization of 5 to control complexity.To better process continuous features, 128 border counts were used, while a subsampling rate of 0.8 reduced variance and improved generalization. These settings enabled CatBoost to capture important non-linear relationships in the hematological data, resulting in robust dengue risk classification.
- **CNN-CatBoost Fusion Model (CCF-Net)**: A 1-D CNN and CatBoost applied with standard scaler for normalization to built a model. The CNN was trained using 30 epochs with batch size 32, using the Adam optimizer with learning rate 0.001 and binary cross-entropy. The architecture consisted of a first Conv1D layer with 64 filters and the kernal size is 3 and 256 parameter is used followed by BatchNormalization and MaxPooling1D, then a second Conv1D layer with 128 filters and 24,704 parameters followed by BatchNormalization with 512 parameters and MaxPooling. The output was flattened and passed through a Dense layer with 128 neurons and 49,280 parameters, a Dropout layer with rate 0.3 and finally a sigmoid output layer with 129 parameters for binary classification. The CatBoost model was tuned with iterations=1000, learning rate=0.05, depth=8, l2 leaf reg=5, border count=128, subsample=0.8, random seed=42, trained using Logloss with AUC as the evaluation metric.
- **Artificial Neural Network(ANN)**:In ANN model, a fully connected dense layer is built with two hidden layers containing 64 and 32 neurons respectively and both use the ReLU activation function which is combined with Batch Normalization and a 0.3 Dropout rate were applied to reduce overfitting and stabilize training. The model trained with Adam optimizer which relied on binary cross-entropy to guide learning, and used accuracy to measure performance. For training 100 epoch was performed with a batch size of 32, and 20% of the data was automatically used for validation during training. Two callbacks were applied such as EarlyStopping which monitoring validation loss and ReduceLROnPlateau that monitor validation loss. SHAP explainability analysis was performed using KernelExplainer with 100 background samples to interpret feature contributions and model decision patterns.

V. Result Analysis

The performance comparison across different models for predicting dengue demonstrated large differences. The study presented a comprehensive appraisal of the algorithm's results,

considering metrics like accuracy, precision, recall, F1-score and AOC-ROC values (Table-I) where the Decision Tree scored 0.743 for accuracy and 0.7772 for AUC-ROC in our experiment. The CCF-Net and ANN had minor enhancements and got the accuracy in the range of 76–84% and AUC values from 0.83 to 0.90. The Random Forest has performed even better by achieving an accuracy of 84.17% and 0.8931 for AUC-ROC thus ensuring that both classes had balanced precision and recall. Among the deployed models, CatBoost was the most outstanding model with 0.858 accuracy and AUC of 0.9074, and ensuring substantial recall of positive dengue cases (0.92). CCF-Net and CatBoost provide the highest AUC-ROC value. Risk analyses in all models were at the patient level, meaning that the predicted probabilities were mapped to risk levels that had been defined as Low Risk ($\leq$ 0.40), Medium Risk (0.41–0.70), and High Risk ($>$ 0.70). For the first ten patients, Decision Tree mostly predicted Medium and Low Risk, while CNN and ANN predicted Medium, Low, and a few High Risk cases. Random forest predicted mainly Medium and Low Risk. CatBoost, however, revealed the most clear-cut with multiple high-risk patients (¿0.74), several medium-risk (0.51–0.65), and low-risk individuals (¡0.40). SHAP interpretation pointed out very important hematological features, including platelet count, white blood cell differentials, and hemoglobin, as the most crucial elements in predictions, thus granting interpretable and patient-specific insight. In comparison to previous research, a number of authors reported a performance that was significantly higher, for instance, Chowdhury et al. [7] SVM showed 93.62% accuracy with an AUC of 0.96, Panchanadikar et al. [8] used an advanced CatBoost variant and 97.1% accuracy was reported, Mayrose et al. [9] got 94.79% accuracy with TabPFN method, while Ming et al. [10] and Andualgalil et al. [14] claimed accuracies over 98% with XGBoost based approach. However, our results still do not reach these top findings, the primary contribution of our research that it is an interpretable, clinically oriented dengue risk assessment framework. By combining probability-based risk stratification with well-defined risk levels and explainable SHAP technique, the model focuses on patient-level insight rather than just maximizing statistical performance, thus being clinically useful even if numerical metrics are relatively lower.

TABLE I
COMPARISON OF EXISTING AND PROPOSED MODEL PERFORMANCE FOR DENGUE RISK FACTOR CLASSIFICATION

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
XGB [7]	98.35	98.48	0.982	-	-
XGBoost [8]	98.60	0.98	0.96	-	-
SVM [9]	0.936	0.961	-	0.943	0.96
TabPFN [10]	0.947	0.9387	0.9453	0.9419	-
CatBoost [14]	0.971	-	-	-	-
Proposed Decision Tree	0.743	0.75	0.74	0.74	0.7772
Proposed CatBoost	0.858	0.86	0.86	0.86	0.9074
Proposed Random Forest	0.841	0.85	0.84	0.84	0.8931
Proposed CCF-Net	0.846	0.85	0.85	0.85	0.902
Proposed ANN	0.760	0.77	0.77	0.77	0.835

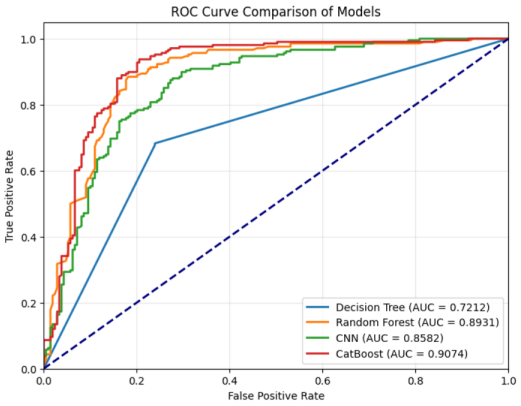


Fig. 2. Comparision of ROC Curve

Fig 2. presents comparison of the proposed models AUC scores through ROC curves. This helps to visualize which model is betterad differencing between two classes.

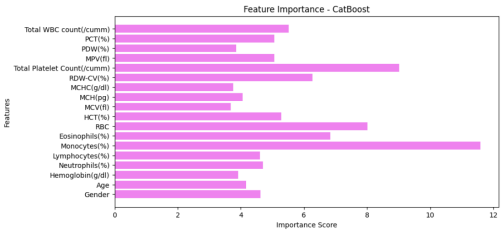


Fig. 3. SHAP Feature Importance Of Best Performed Model

The SHAP feature importance for the top-performing CatBoost model shown in Fig 3. The features are arranged sequentially based on their influence, making it easy to identify which variables areinfluence strongly on the decision process. This helps us to understand which features are most significant in determining the outcomes.

In Figure 4, the SHAP summary plot organizes features according to their influence on the outcome of models. The plot highlights which variables had the strongest impact, while also indicating the effect each feature has on the outcome. In addition to showing importance, the plot also conveys the direction of each feature's effect whether higher or lower values push the prediction toward one class or the other.

VI. CONCLUSION

The purpose of this study was to create and test models that can predictably assess the risk of dengue fever by using clinical and laboratory data and by giving particular attention to accuracy, precision, recall, F1-score and AUC ROC. Several machine learning and deep learning algorithms were considered and tested as CatBoost turned out to be the best model with a 0.858 accuracy. The power of the random forest model is due to its ability to unify the strengths of individual decision

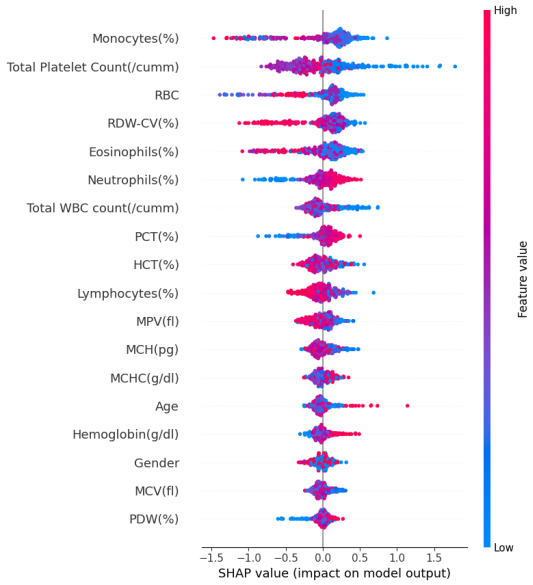


Fig. 4. SHAP Summary Plot (impact on model output)

trees leading to a strong and balanced predictive tool. The results are very promising, however, the small size and limited scope of the dataset acquired from the hospital records made it difficult to conduct intensity analysis and even reach better performance in predictions. The future research will mostly concentrate on refining and optimizing predictive models by exploring other algorithms, state-of-the-art ensemble methods, and thorough tuning of hyperparameters. Besides, the application of feature engineering methods and the picking of the most relevant clinical and laboratory indicators will not only aid in clarifying and reaching dengue risk prediction that is more interpretable and accurate but will also provide more actionable insights for early intervention and patient care.

References

[1] S. K. Roy and S. Bhattacharjee, "Dengue virus: epidemiology, biology, and disease aetiology," *Canadian journal of microbiology*, vol. 67, no. 10, pp. 687–702, 2021.

[2] M. S. Rahman, F. Mehejabin, M. A. Rahman, and R. Rashid, "A case-control study to determine the risk factors of dengue fever in chattogram, bangladesh," *Public Health in Practice*, vol. 4, p. 100288, 2022.

[3] J. Yang, A. A. Mosabbir, E. Raheem, W. Hu, and M. S. Hossain, "Demographic characteristics, clinical symptoms, biochemical markers and probability of occurrence of severe dengue: A multicenter hospital-based study in bangladesh," *PLoS Neglected Tropical Diseases*, vol. 17, no. 3, p. e0011161, 2023.

[4] V. R. Louis, R. Phalkey, O. Horstick, P. Ratanawong, A. Wilder-Smith, Y. Tozan, and P. Dambach, "Modeling tools for dengue risk mapping-a systematic review," *International journal of health geographics*, vol. 13, no. 1, p. 50, 2014.

[5] M. N. Hasan, I. Khalil, M. A. B. Chowdhury, M. Rahman, M. Asaduzzaman, M. Billah, L. A. Banu, M.-U. Alam, A. Ahsan, T. Traore *et al.*, "Two decades of endemic dengue in bangladesh (2000–2022): trends, seasonality, and impact of temperature and rainfall patterns on transmission dynamics," *Journal of Medical Entomology*, vol. 61, no. 2, pp. 345–353, 2024.

[6] S. K. Roy, B. Ghosh, A. Chakraborty, S. Hazra, B. K. Goswami, and S. Bhattacharjee, "Haematological parameters as predictors of severe dengue: A study from northern districts of west bengal, india," *Journal of Vector Borne Diseases*, vol. 62, no. 2, pp. 218–225, 2025.

[7] D. K. Ming, B. Hernandez, S. Sangkaew, N. L. Vuong, P. K. Lam, N. M. Nguyet, D. T. H. Tam, D. T. Trung, N. T. H. Tien, N. M. Tuan *et al.*, "Applied machine learning for the risk-stratification and clinical decision support of hospitalised patients with dengue in vietnam," *PLOS digital health*, vol. 1, no. 1, p. e0000005, 2022.

[8] B. Abdualgalil, S. Abraham, and W. M. Ismael, "Early diagnosis for dengue disease prediction using efficient machine learning techniques based on clinical data," *Journal of Robotics and Control (JRC)*, vol. 3, no. 3, pp. 257–268, 2022.

[9] S. U. Chowdhury, S. Sayeed, I. Rashid, M. G. R. Alam, A. K. M. Masum, and M. A. A. Dewan, "Shapley-additive-explanations-based factor analysis for dengue severity prediction using machine learning," *Journal of imaging*, vol. 8, no. 9, p. 229, 2022.

[10] H. Mayrose, G. M. Bairy, N. Sampathila, S. Belurkar, and K. Saravu, "Machine learning-based detection of dengue from blood smear images utilizing platelet and lymphocyte characteristics," *Diagnostics*, vol. 13, no. 2, p. 220, 2023.

[11] N. J. Riya, M. Chakraborty, and R. Khan, "Artificial intelligence based early detection of dengue using cbc data," *IEEE Access*, 2024.

[12] N. T. Thanh, V. T. Luan, D. C. Viet, T. H. Tung, and V. Thien, "A machine learning-based risk score for prediction of mechanical ventilation in children with dengue shock syndrome: a retrospective cohort study," *PLoS One*, vol. 19, no. 12, p. e0315281, 2024.

[13] T. B. Thapa, B. R. Bhattarai, S. Maharjan, S. Pokhrel, and M. Sapkota, "Insights of hematological parameters in dengue virus-infected patients," *Indian J Pathol Oncol*, vol. 10, no. 4, pp. 340–5, 2023.

[14] N. T. Panchanadikar, S. H. Palkar, and S. K. Lalwani, "Outcome of dengue infection and risk factors for severe dengue in indian children," *Journal of Vector Borne Diseases*, vol. 62, no. 3, pp. 344–350, 2025.

[15] Z. J. Madewell, D. M. Rodriguez, M. B. Thayer, V. Rivera-Amill, G. Paz-Bailey, L. E. Adams, and J. M. Wong, "Machine learning for predicting severe dengue in puerto rico," *Infectious Diseases of Poverty*, vol. 14, no. 01, pp. 50–66, 2025.

[16] O. S. Pulock, A. Mannan, A. F. M. N. Chowdhury, G. Tousif, K. Majumder, S. Monsur, H. H. Mehedi, E. Kaiser, A. Sultana, M. A. H. Sagar *et al.*, "Clinical spectrum and risk factors of severe dengue infection: findings from the 2023 dengue outbreak in bangladesh," *BMC Infectious Diseases*, vol. 25, no. 1, p. 469, 2025.

[17] D. D. Anh, M. Recker, N. T. The, S. Krishna, P. G. Kremsner, L. H. Song, and T. P. Velavan, "Risk stratification of dengue cases requiring hospitalization," *Journal of Medical Virology*, vol. 97, no. 8, p. e70511, 2025.

[18] K. E. Alam, M. J. Ahmed, R. Chalise, M. A. Rahman, T. T. Mathin, M. I. H. Bhuiyan, P. Bhandari, and D. Hossain, "Time series analysis of dengue incidence and its association with meteorological risk factors in bangladesh," *Plos one*, vol. 20, no. 8, p. e0323238, 2025.

[19] M. A. S. Nirob, A. K. M. F. K. Siam, P. Bishshash, and M. Assaduzzaman, "Dengue fever hematological dataset: Clinical insights for improved diagnosis and patient management," 2025. [Online]. Available: https://doi.org/10.17632/6fsrsk3mb8.2

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[21] R. Blagus and L. Lusa, "Smote for high-dimensional class-imbalanced data," *BMC bioinformatics*, vol. 14, no. 1, p. 106, 2013.

DM-i SUPER PLUG-IN HYBRID-EV

BYD

GO BEYOND EVERY JOURNEY



BYD SEALION 6

Combined Range **1250**Km*
Locally tested*

Maximum Torque **300**N.m

Super Plug-In Hybrid Engine **1500**CC

As Low As **4.12** BDT*
Hybrid Efficiency

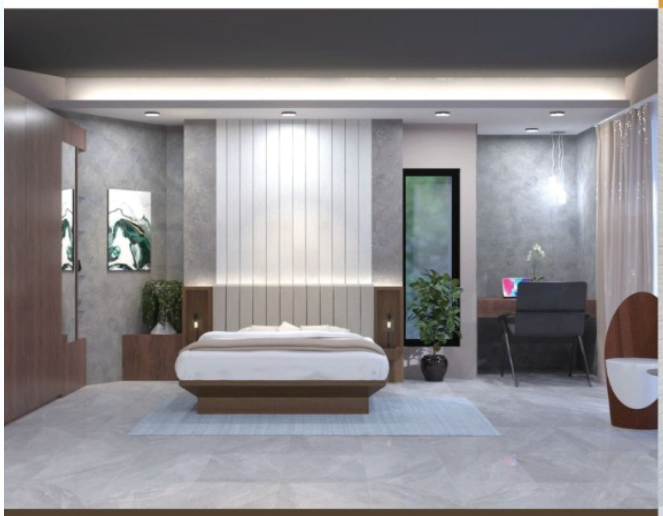
*TERMS & CONDITIONS APPLY

Visit Our Showroom 37, Shaheed Tajuddin Ahmed Avenue, Tejgaon, Dhaka

Contact Us 09639-795795

Western Thai Aluminium
Western Creation

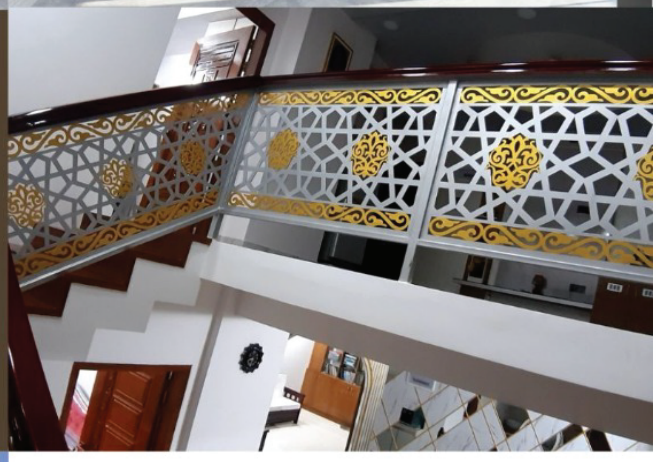
Complete solution for Exterior and Interior Decoration





We Provide

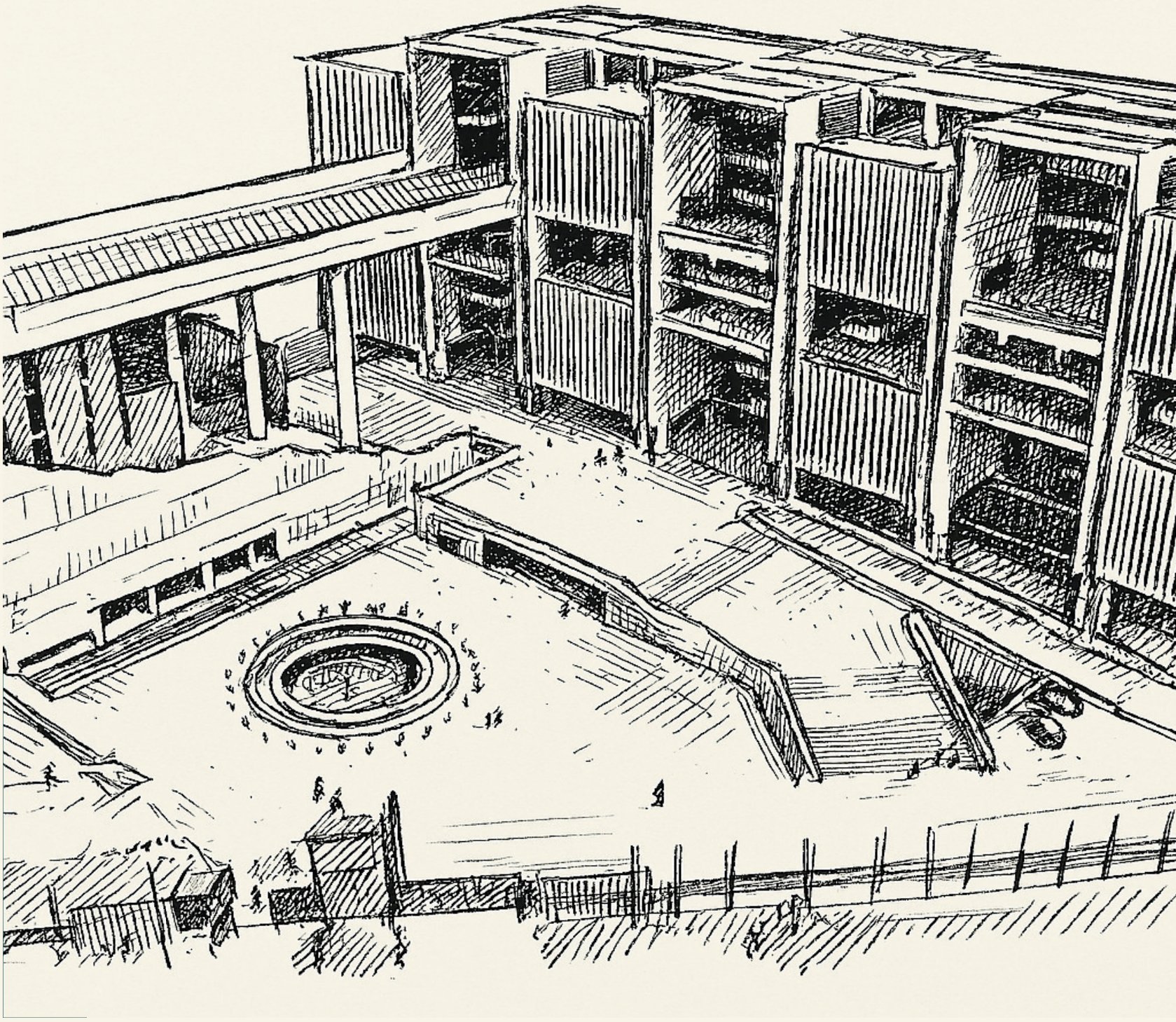
- Interior & Exterior Design and Decoration
- Project Management
- Furniture and Soft Furnishings
- Building
- Retail Showroom
- Architecture





Contact Us

110, Upor Vadra (Madrasa Mor), Rajshahi-6204
+8801711-821124, +8801711-708355,
+8801711-451573
western.com.bd
westernthai@gmail.com
fb.com/westerncr




UCCS
2026