Pertanika Journal of

SCIENCE & TECHNOLOGY

JST

VOL. 34 (S3) 2026

A Special Issue devoted to

Technologies and Applications for Sustainable Development

Guest Editors

I. Siva, K. Reddy Madhavi, and Nagaraj Ramrao

PERTANIKA JOURNALS

A scientific journal published by Universiti Putra Malaysia Press

Pertanika Journal of Science & Technology
Pertanika Journal of Science & Technology

PERTANIKA JOURNAL OF SCIENCE & TECHNOLOGY

About the Journal

Overview

Pertanika Journal of Science & Technology is an official journal of Universiti Putra Malaysia. It is an open-access online scientific journal. It publishes original scientific outputs. It neither accepts nor commissions third party content.

Recognised internationally as the leading peer-reviewed interdisciplinary journal devoted to the publication of original papers, it serves as a forum for practical approaches in improve quality on issues pertaining to science and engineering and its related fields.

Pertanika Journal of Science & Technology currently publishes 6 issues a year (*February, April*, *June, August, October, and December*). It is considered for publication of original articles as per its scope. The journal publishes in **English** and it is open for submission by authors from all over the world.

The journal is available world-wide.

Aims and Scope

Pertanika Journal of Science & Technology aims to provide a forum for high quality research related to science and engineering research. Areas relevant to the scope of the journal include: bioinformatics, bioscience, biotechnology and bio-molecular sciences, chemistry, computer science, ecology, engineering, engineering design, environmental control and management, mathematics and statistics, medicine and health sciences, nanotechnology, physics, safety and emergency management, and related fields of study.

History

Pertanika Journal of Science & Technology was founded in 1993 and focusses on research in science and engineering and its related fields.

Vision

To publish a journal of international repute.

Mission

Our goal is to bring the highest quality research to the widest possible audience.

Quality

We aim for excellence, sustained by a responsible and professional approach to journal publishing. Submissions can expect to receive a decision within 90 days. The elapsed time from submission to publication for the articles averages 180 days. We are working towards decreasing the processing time with the help of our editors and the reviewers.

Abstracting and Indexing of Pertanika

Pertanika Journal of Science & Technology is now over 33 years old; this accumulated knowledge and experience has resulted the journal being abstracted and indexed in SCOPUS (Elsevier), Journal Citation Reports (JCR-Clarivate), EBSCO, ASEAN CITATION INDEX, Microsoft Academic, Google Scholar, and MyCite.

Citing Journal Articles

The abbreviation for Pertanika Journal of Science & Technology is *Pertanika J. Sci. & Technol.*

Publication Policy

Pertanika policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. It prohibits as well publication of any manuscript that has already been published either in whole or substantial part elsewhere. It also does not permit publication of manuscript that has been published in full in proceedings.

Code of Ethics

The *Pertanika* journals and Universiti Putra Malaysia take seriously the responsibility of all of its journal publications to reflect the highest in publication ethics. Thus, all journals and journal editors are expected to abide by the journal's codes of ethics. Refer to *Pertanika*'s **Code of Ethics** for full details, or visit the journal's web link at http://www.pertanika.upm.edu.my/code_of_ethics.php

Originality

The author must ensure that when a manuscript is submitted to *Pertanika*, the manuscript must be an original work. The author should check the manuscript for any possible plagiarism using any programme such as Turnitin or any other software before submitting the manuscripts to the Editorial Office, *Pertanika* Journal Publication Section.

All submitted manuscripts must be in the journal's acceptable similarity index range:

≤ 20% – PASS; > 20% – REJECT.

International Standard Serial Number (ISSN)

An ISSN is an 8-digit code used to identify periodicals such as journals of all kinds and on all media–print and electronic.

Pertanika Journal of Science & Technology: e-ISSN 2231-8526 (Online).

Lag Time

A decision on acceptance or rejection of a manuscript is reached in 90 days (average). The elapsed time from submission to publication for the articles averages 180 days.

Authorship

Authors are not permitted to add or remove any names from the authorship provided at the time of initial submission without the consent of the journal's Chief Executive Editor.

Manuscript Preparation

For manuscript preparation, authors may refer to Pertanika's **INSTRUCTION TO AUTHORS**, available on the official website of *Pertanika*.

Editorial Process

Authors who complete any submission are notified with an acknowledgement containing a manuscript ID on receipt of a manuscript, and upon the editorial decision regarding publication.

Pertanika follows a **double-blind peer-review** process. Manuscripts deemed suitable for publication are sent to reviewers. Authors are encouraged to suggest names of at least 3 potential reviewers at the time of submission of their manuscripts to *Pertanika*, but the editors will make the final selection and are not, however, bound by these suggestions.

Notification of the editorial decision is usually provided within 90 days from the receipt of manuscript. Publication of solicited manuscripts is not guaranteed. In most cases, manuscripts are accepted conditionally, pending an author's revision of the material.

The Journal's Peer Review

In the peer-review process, 2 to 3 referees independently evaluate the scientific quality of the submitted manuscripts. At least 2 referee reports are required to help make a decision.

Peer reviewers are experts chosen by journal editors to provide written assessment of the **strengths** and **weaknesses** of the written research, with the aim of improving the reporting of research and identifying the most appropriate and highest quality material for the journal.

Operating and Review Process

What happens to a manuscript once it is submitted to *Pertanika*? Typically, there are 7 steps to the editorial review process:

1. The journal's Chief Executive Editor and Editor-in-Chief examine the paper to determine whether it is relevant to the journal's needs in terms of novelty, impact, design, procedure, language as well as presentation and allow it to proceed to the reviewing process. If it is not appropriate, the manuscript is rejected outright and the author is informed.

2. The Chief Executive Editor sends the article-identifying information having been removed, to 2 to 3 reviewers. They are specialists in the subject matter of the article. The Chief Executive Editor requests that they complete the review within 3 weeks.

   Comments to authors are about the appropriateness and adequacy of the theoretical or conceptual framework, literature review, method, results and discussion, and conclusions. Reviewers often include suggestions to strengthening of the manuscript. Comments to the editor are in the nature of the significance of the work and its potential contribution to the research field.

Pertanika Journal of Science & Technology
Pertanika Journal of Science & Technology

Pertanika Journal of Science & Technology
Pertanika Journal of Science & Technology

3. The Editor-in-Chief examines the review reports and decides whether to accept or reject the manuscript, invite the authors to revise and resubmit the manuscript, or seek additional review reports. In rare instances, the manuscript is accepted with almost no revision. Almost without exception, reviewers' comments (to the authors) are forwarded to the authors. If a revision is indicated, the editor provides guidelines to attend to the reviewers' suggestions and perhaps additional advice on how to revise the manuscript.

4. The authors decide whether and how to address the reviewers' comments and criticisms and the editor's concerns. The authors return a revised version of the paper to the Chief Executive Editor along with specific information describing how they have addressed' the concerns of the reviewers and editor, usually in a tabular form. The authors may also submit a rebuttal if there is a need especially when the authors disagree with certain comments provided by the reviewers.

5. The Chief Executive Editor sends the revised manuscript out for re-review. Typically, at least 1 of the original reviewers will be asked to examine the article.

6. When the reviewers have completed their work, the Editor-in-Chief examines their comments and decides whether the manuscript is ready to be published, needs another round of revisions, or should be rejected. If the decision is to accept, the Chief Executive Editor is notified.

7. The Chief Executive Editor reserves the final right to accept or reject any material for publication, if the processing of a particular manuscript is deemed not to be in compliance with the S.O.P. of *Pertanika*. An acceptance letter is sent to all the authors.

   The editorial office ensures that the manuscript adheres to the correct style (in-text citations, the reference list, and tables are typical areas of concern, clarity, and grammar). The authors are asked to respond to any minor queries by the editorial office. Following these corrections, page proofs are mailed to the corresponding authors for their final approval. At this point, **only essential changes are accepted**. Finally, the manuscript appears in the pages of the journal and is posted online.

Pertanika Journal of

SCIENCE & TECHNOLOGY

Vol. 34 (S3) 2026

A Special Issue devoted to
Technologies and Applications for Sustainable Development

Guest Editors
I. Siva, K. Reddy Madhavi, and Nagaraj Ramrao

A scientific journal published by Universiti Putra Malaysia Press

PJST
Pertanika Journal of Science & Technology
AN INTERNATIONAL PEER-REVIEWED JOURNAL

UNIVERSITY PUBLICATIONS COMMITTEE

CHAIRMAN
Zamberi Sekawi

EDITOR-IN-CHIEF
Mohamed Thariq Hameed Sultan
SEARCA Chair of Agriculture Technology and Innovation

JOURNAL SECTION
Head of Section:
Navaneetha Krishna Chandran

Journal Officers:
Ain Haziqah Ab Razak
Nur Izni Shafiqah Mohd Hamidi
Saundarya Saran

Editorial Assistants:
Zulinaardawati Kamarudin

Pre-press Officers:
Aisyah Umairah Jasmin
Engku Erina Zahirah Engku Mohd Suhaimi
Ku Ida Mastura Ku Baharom

IT Officer:
Kiran Raj Kaneswaran

Journal Section Office
Putra Science Park
1st Floor, IDEA Tower II
UPM-MTDC Technology Centre
Universiti Putra Malaysia
43400 UPM Serdang
Selangor, Malaysia

General Enquiry
Tel. No: +603 9769 1622
E-mail:
pertanika.proceeding@upm.edu.my
URL:
http://www.pertanika.upm.edu.my/

PUBLISHER
UPM Press Centre
Universiti Putra Malaysia
43400 UPM Serdang
Selangor, Malaysia
Tel: +603 9769 3606
E-mail: dir.penerbit@upm.edu.my
URL: http://penerbit.upm.edu.my

ASSOCIATE EDITOR
2024-2026

Miss Laiha Mat Kiah
Security Services Sn: Digital Forensic, Steganography, Network Security, Information Security, Communication Protocols, Security Protocols
Universiti Malaya, Malaysia

Saidur Rahman
Renewable Energy, Nanofluids, Energy Efficiency, Heat Transfer, Energy Policy
Sunway University, Malaysia

EDITORIAL BOARD
2024-2026

Abdul Latif Ahmad
Chemical Engineering
Universiti Sains Malaysia, Malaysia

Hsiu-Po Kuo
Chemical Engineering
National Taiwan University, Taiwan

Mohd. Ali Hassan
Bioprocess Engineering, Environmental Biotechnology
Universiti Putra Malaysia, Malaysia

Ahmad Zaharin Aris
Hydrochemistry, Environmental Chemistry, Environmental Forensics, Heavy Metals
Universiti Putra Malaysia, Malaysia

Ivan D. Rukhlenko
Nonliner Optics, Silicon Photonics, Plasmonics and Nanotechnology
The University of Sydney, Australia

Nor Azah Yusof
Biosensors, Chemical Sensor, Functional Material
Universiti Putra Malaysia, Malaysia

Azlina Harun@ Kamaruddin
Enzyme Technology, Fermentation Technology
Universiti Sains Malaysia, Malaysia

Lee Keat Teong
Energy Environment, Reaction Engineering, Waste Utilisation, Renewable Energy
Universiti Sains Malaysia, Malaysia

Norbahiah Misran
Communication Engineering
Universiti Kebangsaan Malaysia, Malaysia

Bassim H. Hameed
Chemical Engineering: Reaction Engineering, Environmental Catalysis & Adsorption
Qatar University, Qatar

Mohamed Othman
Communication Technology and Network, Scientific Computing
Universiti Putra Malaysia, Malaysia

Roslan Abd-Shukor
Physics & Materials Physics, Superconducting Materials
Universiti Kebangsaan Malaysia, Malaysia

Biswajeet Pradhan
Digital image processing, Geographical Information System (GIS), Remote Sensing
University of Technology Sydney, Australia

Mohd Shukry Abdul Majid
Polymer Composites, Composite Pipes, Natural Fibre Composites, Biodegradable Composites, Bio-Composites
Universiti Malaysia Perlis, Malaysia

Sodeifian Gholamhossein
Supercritical technology, Optimisation, nanoparticles, Polymer nanocomposites
University of Kashan, Iran

Ho Yuh-Shan
Water research, Chemical Engineering and Environmental Studies
Asia University, Taiwan

Mohd Zulkifly Abdullah
Fluid Mechanics, Heat Transfer, Computational Fluid Dynamics (CFD)
Universiti Sains Malaysia, Malaysia

Wing Keong Ng
Aquaculture, Aquatic Animal Nutrition, Aqua Feed Technology
Universiti Sains Malaysia, Malaysia

INTERNATIONAL ADVISORY BOARD
2024-2027

Hiroshi Uyama
Polymer Chemistry, Organic Compounds, Coating, Chemical Engineering
Osaka University, Japan

Mohini Sain
Material Science, Biocomposites,Biomaterials
University of Toronto, Canada

Mohamed Pourkashanian
Mechanical Engineering, Energy, CFD and Combustion Processes
Sheffield University, United Kingdom

Yulong Ding
Particle Science & Thermal Engineering
University of Birmingham, United Kingdom

ABSTRACTING AND INDEXING OF PERTANIKA JOURNALS

Pertanika Journal of Science & Technology is indexed in Journal Citation Reports (JCR-Clarivate), SCOPUS (Elsevier), EBSCO, ASEAN Citation Index, Microsoft Academic, Google Scholar and MyCite.

The publisher of Pertanika will not be responsible for the statements made by the authors in any articles published in the journal. Under no circumstances will the publisher of this publication be liable for any loss or damage caused by your reliance on the advice, opinion or information obtained either explicitly or implied through the contents of this publication.
All rights of reproduction are reserved in respect of all papers, articles, illustrations, etc., published in Pertanika. Pertanika provides free access to the full text of research articles for anyone, web-wide. It does not charge either its authors or author-institution for refereeing/publishing outgoing articles or user-institution for accessing incoming articles.
No material published in Pertanika may be reproduced or stored on microfilm or in electronic, optical or magnetic form without the written authorisation of the Publisher.
Copyright ©2026 Universiti Putra Malaysia Press. All Rights Reserved.

Pertanika Journal of Science & Technology
Vol. 34 (S3) 2026

Contents

Preface xi
I. Siva, K. Reddy Madhavi, and Nagaraj Ramrao

Swin Transfer-Based Auxiliary Classifier Generative Adversarial Network for Colorectal Cancer Analysis 1
G. Kanishka and K. Reddy Madhavi

The Sky Emotional Chatbot Companion: Proactive AI Intervention Fostering Emotional Connectivity to the Mind 27
Baby Vadlana, Karnam Akhil, Tamma Bhuwan Vamsi, and Tejasree Lokireddy

Detecting Fake Reviews on Online Products with Machine Learning 55
P. Dhanalakshmi, Ummadi Janardhan Reddy, Ravikanth Garladinne, S. Ramachandra, E. R. Aruna, and Sharath Chandra Parashara

Deep Learning and Signal Processing Synergy: Multitask Deep Neural Networks (MLT-DNNs) and Second-Order Blind Identification (SOBI) for Cybersickness Detection 71
Bisani Leela Padmavathi and J. Avanija

GenCKD-Net: A Lightweight Multilayer Perceptron Pipeline for Chronic Kidney Disease Classification Using the UCI CKD Dataset 101
Karimulla Sha Shaik, Konda Hari Krishna, and K. K. Baseer

Leveraging Business Analytics for Operational Efficiency in Small-Scale Manufacturing Enterprises in Tamil Nadu 125
Geetha N and U. M. Gopal Krishna

Comparison of Deep Learning Models for Classification of Acute Lymphoblastic Leukaemia 149
Beebi Naseeba, Pranavesh Kumar Talupuri, Nitin Bhimineni, V. Jyothsna, Sharath Chandra Parashara, and Puvi Rajan

An Optimised Deep Learning Framework for Skin Lesion Classification Using ResNet150 and M-GSO 165
Gurram Sunitha, Ahmed Elngar, Pravin R. Kshirsagar, and Subba Rao Polamuri

Enhancing Automotive Safety and Operational Efficiency Through Controller Area Network (CAN) and Internet of Things (IoT) Integration 193
Arava Ramprakash Reddy and Suresh Kallam

SentryNet: Fusion of CNN-LSTM Architecture for Robust Suspicious Activity Detection in Video Surveillance Using Novel Hybrid Optimisation 209
Roopa R and Humera Khanam M

Preface

The 2nd International Conference on Advances in Technologies and Applications for Sustainability (ICATAS 2024), organised by Mohan Babu University, Tirupati, India, on 21–22 December 2024, served as a vibrant platform where scholars, researchers, practitioners, and industry experts came together to deliberate on the role of science, engineering, and technology in shaping a sustainable future.

This conference was envisioned to provide a premier interdisciplinary forum to discuss recent innovations, address pressing challenges, and share practical solutions aligned with the United Nations' Sustainable Development Goals (SDGs). The deliberations reflected the diverse yet interconnected dimensions of sustainability, spanning sustainable engineering, energy, environment, materials science, digital transformation, and societal well-being.

The proceedings of ICATAS 2024, now published by Pertanika, encapsulate the scholarly contributions presented during the conference. The selected works not only highlight theoretical advancements but also emphasise real-world applications, providing pathways to achieve balance between technological growth and environmental responsibility. This special issue thus stands as a testimony to the collective intellectual effort directed towards sustainability, demonstrating how innovations can be harnessed for long-term societal and ecological benefits.

We are deeply grateful to the team members of Universiti Putra Malaysia's Pertanika Journal of Science and Technology for their unwavering support in realising this publication. Our sincere appreciation is also extended to the management of Mohan Babu University, chief patrons, patrons, advisors, reviewers, session chairs, and all committee members whose dedicated efforts ensured the success of ICATAS 2024. Above all, we thank our contributors and participants from across the globe for bringing diverse perspectives and for making this conference a truly enriching experience.

It is our hope that this special issue will inspire researchers, academics, and practitioners to explore novel directions and foster collaborations that contribute to a more sustainable and equitable world.

Guest Editors
I. Siva (Prof. Dr.)
K. Reddy Madhavi (Prof. Dr.)
Nagaraj Ramrao (Prof. Dr.)

Pertanika J. Sci. & Technol. 34 (S3): 1 - 26 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Swin Transfer-Based Auxiliary Classifier Generative Adversarial Network for Colorectal Cancer Analysis

G. Kanishka[1] and K. Reddy Madhavi[2]*

[1]Research Scholar, Department of Computer Science and Engineering, School of Computing, Mohan Babu University, Sree Sainath Nagar, 517102 Tirupati, Andhra Pradesh, India
[2]Department of Data Sciences, School of Computing, Mohan Babu University, Sree Sainath Nagar, 517102 Tirupati, Andhra Pradesh, India

ABSTRACT

Molecular pathways, genetic subtypes, and mutations associated with colorectal cancer (CRC) are important for advancing precision medicine. CRC is an enormous global health issue. Traditional methods of measurement, including gene sequencing, have been shown to be not only costly but also time-intensive, and most deep learning algorithms are inefficient and hard to understand. The present paper proposes a new mechanism to enhance advanced deep learning methods for monitoring biochemical events and predicting significant mutations using cell network information. The goal is to expand the analysis and category of CRC and introduce a distinctive model that integrates an Auxiliary Classifier Generative Adversarial Network (AC-GAN) with Swin Transformer (ST). To address weaknesses commonly encountered in CRC datasets, the AC-GAN component is used to generate high-quality annotated medical images, thereby augmenting the training data and the class distribution. The ST is a high-end vision transformer model that uses these improved datasets to achieve robust organisation by capturing histopathology image-level and global contextual details. The experimental findings from the histopathology CRC dataset show that the AC-GAN-ST combination achieves higher organisation accuracy and sensitivity than a traditional convolutional network. Such a combined approach not only improves the quality of prediction presentation but also enhances the model's overall generalisability, which is why it is a reliable, high-accuracy instrument for detecting CRC and supporting pathologists' clinical decision-making.

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 06 October 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.01

E-mail addresses:
kanishkagrk@gmail.com (G. Kanishka)
kreddymadhavi@gmail.com (K. Reddy Madhavi)
* Corresponding author

Keywords: Auxiliary Classifier GAN (AC-GAN), colorectal cancer, histopathological image analysis, medical image classification, Swin Transformer

INTRODUCTION

Cancer is a group of illnesses characterised by abnormal cells proliferating out of

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

control and forming malignant tumours. Metastasis is the most frequent cause of cancer-related death. It happens when these cells separate, travel, and develop into new tumours elsewhere in the body. Cancer is responsible for one out of every six deaths worldwide, according to the World Health Organisation. The colon, breast, lung, and prostate are among the most affected sites where cancer initially appears. Colorectal cancer (CRC) is the second most deadly and third most common type of cancer (Prezja et al., 2022; Prezja et al., 2023; Skrede et al., 2020). According to data from the American Cancer Society, 56% of patients receive their diagnosis after the cancer has started to spread. Early detection and treatment are crucial. Advances in machine vision, particularly deep neural networks, have improved the automatic classification of diseases, including cancer (Prezja et al., 2022). Despite technological advancements, medical practitioners still need to analyse histologic samples to confirm diagnoses and determine the stage of cancer. Key histological characteristics in these tissues are typically highlighted by hematoxylin and eosin (HE) staining (Khan et al., 2023).

Patients with colorectal cancer are categorised to customise their monitoring and treatment plans. Tumour histopathology, genomic investigations, clinical outcomes, tumour genetics, clinical data, and quantitative biomarkers are among the attributes used for grouping. Molecular and genetic analyses have identified numerous biomarkers. The relevance of the tumour microenvironment to tumour progression has been highlighted by recent studies in tumour immunology. Therefore, identifying new prognostic and predictive biomarkers that accurately characterise cancer traits is crucial (Skrede et al., 2020).

HE-stained whole-slide images have recently enabled the quantification of primary biomarkers using deep learning (DL) methods (Echle et al., 2021; Sirinukunwattana et al., 2021). A measurable indicator generated from medical images, such as histology slides, using deep learning algorithms, is a therapeutically beneficial biomarker produced by convolutional neural networks. This biomarker has the potential to uncover the molecular underpinnings of a disease and predict clinical outcomes such as survival rates and disease progression. To develop a biomarker for colorectal cancer at stages III and IV, Kather and colleagues were among the first teams to use deep learning (Kather et al., 2019). Time and resources may be saved by automatically generating this novel biomarker from images, which performed similarly to the gold standards currently used to forecast the fate of colorectal cancer. It has also been possible to predict known biomarkers using deep learning transformers (Wagner et al., 2023). This strategy reached clinical-level accuracy in colorectal cancer procedures and surpassed existing techniques for detecting microsatellite instability in surgical samples, representing a significant improvement in the area (Ruusuvuori et al., 2023).

In this paper, the AC-GAN collective with the Swin Transformer offers a powerful solution for colorectal cancer analysis by integrating the strengths of generative and

deep learning architectures. The AC-GAN employs a generator and a discriminator: the generator synthesises realistic colorectal tissue images, while the discriminator not only differentiates real from generated images but also predicts class labels, facilitating robust feature extraction. By incorporating Swin Transformers, which excel at capturing long-range dependencies and hierarchical features through shifted windows, the model effectively learns complex spatial patterns in histopathology images, as shown by Khan et al. (2023). The proposed hybrid model leverages the generative nature of the AC-GAN to optimise training data and classification accuracy, and the Swin Transformer's attention model to refine feature representation, enabling accurate detection of cancerous patterns. Subsequently, this architecture overcomes the shortage of annotated data and variability of colorectal cancer images, improves diagnostic accuracy and aids pathologists in the early and accurate detection of cancer.

Motivation of This Research

- The accuracy of colorectal cancer detection should be improved to facilitate faster intervention that would greatly benefit patients.
- Through Swin Transformer and Auxiliary Classifier GANs, we can proceed to train the processing of complex medical imaging data with the state-of-the-art AI methods.
- The use of GANs would aid in producing high-quality synthetic data, which is a solution to the problem of the scarcity of annotated medical datasets.
- By adding Swin Transformers, it is possible to have more flexible modelling of image features, resulting in a better performance in recognising tumour patterns.

The Key Contribution of This Paper

The major contributions of this study are as follows:

- An integrated AC-GAN and Swin Transformer framework is developed for binary classification of colorectal histopathological images.
- AC-GAN is employed to generate class-conditioned synthetic samples to increase training-data diversity in the limited dataset.
- Transfer learning is used to adapt the pre-trained Swin Transformer to colorectal histopathology images.
- The Swin Transformer captures hierarchical and local-to-global visual representations for classification.
- The proposed framework is evaluated under 90:10, 80:20, and 70:30 training/testing configurations using accuracy, sensitivity, and specificity.
- The study provides a contextual comparison with previously reported approaches while acknowledging differences in datasets and experimental results.

LITERATURE SURVEY

Analysis of colorectal cancer has attracted immense attention in recent years because of its emerging prevalence and influence on the health of the population. Early detection and diagnosis have been enhanced due to advances in imaging technologies, machine learning (ML) algorithms, and CT and MRI (Wagner et al., 2023). The research also demonstrates the effectiveness of DL methods (particularly convolutional neural networks (CNNs)) in detecting cancerous lesions and predicting patient outcomes. Moreover, studies indicate that clinical and genomic data should be combined to develop better predictive models (Ruusuvuori et al., 2023). In general, the evolving nature of colorectal cancer analysis requires a multidisciplinary approach to enhance diagnostic and treatment accuracy.

Zhang et al. (2020) proposed a multitask learning model. To address the problems of non-increasing survival likelihood and censored data, we utilise sorting loss and semi-supervised loss approaches. In this approach, a supervised learning model for forecasting the risk of familial polyposis colorectal cancer was constructed with five aspects in mind: semi-supervised logistic regression, clinical practicality, interpretability, and discriminative ability. The comparison of approaches improves our comprehension of the generalisation capacity of various models and serves as a guide for creating clinical prediction models. A prognostic model with valuable performance on multicenter real-world medical data was developed with technical assistance, and the strategy's effectiveness was verified using external data.

Sun et al. (2024) created a special loss function known as Label Distribution and Scale Distribution Aware Loss (LSDA-Loss) to address the problem of colorectal lesion imbalance in white light endoscopy. By accounting for the sample distribution and quantifying the degree of imbalance via Bayesian inference, our innovative loss function addresses category imbalance. Additionally, we evaluate the difficulty of classifying each sample based on the proportional distribution. Tests conducted on three datasets demonstrate that the proposed loss function significantly improves detection accuracy over three conventional FPN models, achieving gains of up to 94.56%. Mulenga et al. (2021) proposed a feature intensification technique that improves a dataset's current features by combining data normalisation techniques. The suggested technique improves the CRC organisation's presentation of a DNN model by integrating feature expansion with data augmentation. The model achieved AUC values of 0.89 and 0.96 on two publicly available microbiome datasets.

Hasan et al. (2024) proposed a technique for identifying key genes associated with colorectal cancer (CC) by combining machine learning-based integrative bioinformatics (ML-IB) and weighted gene co-expression network analysis (WGCNA). ML-IB detected differentially expressed genes (DEGs) and then applied a support vector machine to identify DEDGs, taking the common DEDGs across datasets. Hub genes and hub module genes were identified using MCODE scores and connectivity degrees in protein-protein

interaction networks. ML-IB-based important genes were identified by intersecting hub and hub-module genes. The intersection of WGCNA- and ML-IB-derived important genes resulted in four shared genes. Finally, two genes (AURKA and CCNA2) were identified that strongly correlated with CC patient survival, and their discriminative abilities were confirmed on an independent test dataset using AUC analysis.

To identify and categorise colorectal cancer, Ragab et al. (2023) developed the slime mould algorithm with deep transfer learning (SMADTL-CCDC). The proposed SMADTL-CCDC approach aims to accurately identify colorectal cancer by first enhancing the input image quality for the SMADTL-CCDC model. Additionally, a dense-EfficientNet approach was used to extract feature vectors from pre-processed photos. The SMA was then used in conjunction with the DHNN approach to identify and classify colorectal cancer. The SMA enables more effective parameter selection for the DHNN method.

Sarwinda et al. (2021) propose a DL method for picture organisation to identify colorectal cancer (CRC) using the ResNet architecture. The exceptional performance of DL classifiers has prompted researchers to use them in medical imaging. The model was taught to distinguish between benign and malignant colorectal cancer. Mulenga et al. (2021) proposed a feature intensification strategy that increases the number of features in existing data using data normalisation techniques. This method enhances the accuracy of the DNN architecture's CRC classifier by participating in feature extension through information intensification.

Using H&E-stained WSIs (Yamashita et al., 2021) proposed MSINet, a DL-based method for directly forecasting MSI in CRC. The study found that the model outperformed five pathologists in terms of presentation and performed well on both internal and external datasets. Sasmal et al. (2024) developed a semi-supervised generative adversarial network (GAN) to categorise colorectal polyps using histopathology images. In an adversarial manner, our structure learns discriminative features from large unlabelled datasets and incomplete labelled datasets. As in a conventional GAN, the discriminator of the proposed method is trained to distinguish generated from real images in unsupervised mode and to classify genuine histopathological images in supervised mode (Khalid et al., 2024). A robust, generalised model is produced by the discriminator's capacity to identify and extract minute features from unlabelled images during unsupervised training. Using majority voting techniques at 25% and 50%, our approach achieved classification accuracies of 87.50% and 76.25%, respectively, on the UniToPatho dataset.

Research Gap

Although deep learning has shown promising results in colorectal histopathology analysis, existing studies mainly focus on conventional CNN-based classification, transfer learning, or individual generative approaches. Recent Transformer-based methods can capture

broader spatial relationships in histopathological images, while GAN-based approaches can address limited training data. But their combined use for binary colorectal histopathology classification remains comparatively limited. In particular, the interaction between class-conditioned synthetic-data augmentation and hierarchical Transformer-based feature learning has not been sufficiently explored on small colorectal histopathology datasets. Therefore, this study addresses this gap by integrating AC-GAN with a transfer-learning-based Swin Transformer, in which AC-GAN augments the training data and the Swin Transformer learns hierarchical visual representations for classification. Recent colorectal imaging studies also demonstrate the growing use of Transformer-based architectures in histopathological analysis, supporting the motivation to investigate this integrated framework.

Problem Identification of the Existing System

- The current imaging methods used, including colonoscopy and CT scans, are susceptible to false negatives because they might fail to detect small polyps or early-stage cancers.
- Most existing systems have high false-positive rates, which subject patients to unneeded stress levels and result in invasive follow-up processes. This may also strain the health facilities.
- The current systems are not very effective in incorporating multimodal data, which restricts the ability to analyse the information thoroughly and to plan the treatment individually.
- Traditional analysis methods may require significant time and resources, leading to delays in diagnosis and treatment. Automated systems may not be widely implemented due to high costs or complexity.

PROPOSED SYSTEM

Overall Architecture of Colorectal Cancer Analysis

The proposed framework integrates AC-GAN, transfer learning, and Swin Transformer to address the limited size and visual variability of the colorectal histopathology dataset. AC-GAN is used to generate class-conditioned synthetic training samples, thereby increasing the diversity of the training data. Transfer learning provides the Swin Transformer with pre-trained visual representations (Zidan et al., 2023), which is particularly useful when the target medical dataset is relatively small. The Swin Transformer is then fine-tuned on colorectal histopathology images to learn hierarchical visual features via shifted-window attention, capturing both local tissue patterns and broader spatial relationships. Thus, AC-GAN addresses training-data limitations, transfer learning improves the initialisation of the feature extractor, and Swin Transformer performs the final hierarchical feature learning and classification. The overall architecture is illustrated in Figure 1.

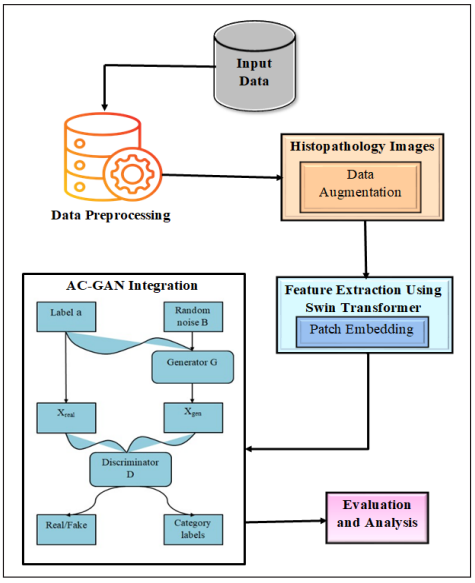


Figure 1. Block diagram of proposed AC-GAN-ST

Dataset

The recently released Warwick-QU dataset, which was the subject of the GlaS@ MICCAI2015 challenge, is used to test our proposed method, AC-GAN-ST, and is publicly available on the contest website (https://www.kaggle.com/code/kerneler/starter-warwick-qu-changed-dataset-d429aecc-a) (Sujatha et al., 2024). It contains 171 images categorised as either cancerous or normal. Figure 2 shows several example images of a colorectal tumour.

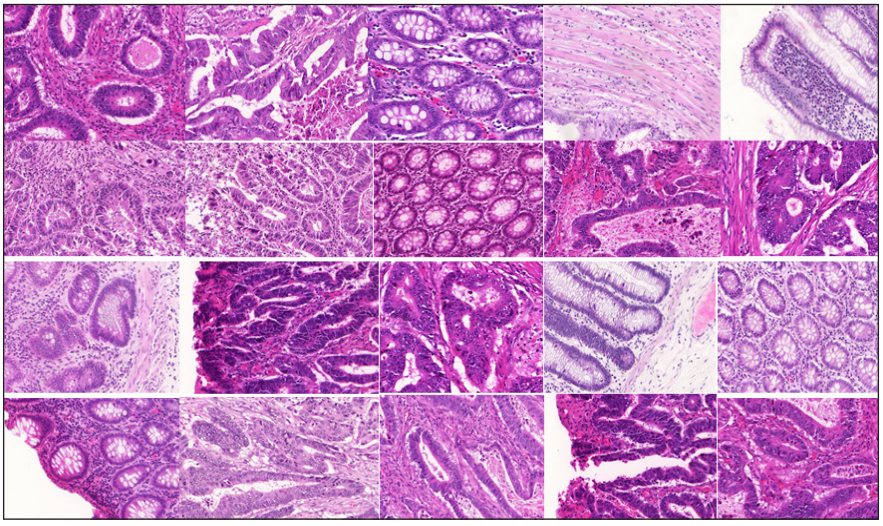


Figure 2. Sample images of colorectal cancer

Data Pre-Processing and Augmentation

Histopathological images may exhibit noise and high visual variability, thereby complicating automatic classification. In the current study, the images have been pre-processed using the proposed Wiener-based Midpoint Weighted Non-Local Means Filter (WMW-NLM) to remove noise while preserving diagnostically significant structures. The geometric augmentation approach, including translation, scaling, and rotation, was used to increase the variability of the training images.

It should be noted that data augmentation was considered a training-stage procedure. Images reserved for testing were left unaltered and were considered evaluation samples. Therefore, the goal of data augmentation in the current study was to expand data diversity during training rather than to enlarge the test set. It helped prevent the artificial improvement in test performance that would otherwise be caused by augmented evaluation samples.

For the AC-GAN part, synthetically generated images conditioned on the class were used to expand the training set. Generated images were used in the training procedure and were not considered additional test cases. As the current study is oriented towards classification accuracy assessment, the synthetically generated images are described as synthetic training samples rather than clinically validated substitutes for real histopathological images. (Khalid et al., 2024).

WMW-NLM Filter for Denoising

The middle weight (MW) in the NLM photo denoising algorithm represents the central pixel's weight. The efficiency of NLM is significantly impacted by the MW. In this work, a Wiener-like estimator is used to determine the MW. The following is the statement model for the proposed WMW-NLM (Equation 1):

$$I_N(mw,n) = I_{NF}(mw,n) + \eta(mw,n) \quad [1]$$

I_N (mw, n) represents a noisy image, I_{NF} (mw, n) signifies a noise-free image, and η (mw, n) specifies Gaussian noise with a variance of σ^2. The spatially adaptive NLM filter transforms the experimental data I_N (mw, n) into U (mw, n) as follows (Equation 2):

$$U(mw,n) = \frac{\sum(mw',n')\in R I_{NF}(mw',n')W(mw,n,mw',n')}{\sum(mw',n')W(mw,n,mw',n')} \quad [2]$$

The weights are represented as W (mw, n, mw′, n′), and R is a search window centred at (mw, n). Two patches, ρ (mw, n), n')) containing pixels at (mw, n) and (mw′, n′), separately, are compared. The midway weight W (mw, n, mw′, n′) is defined in Equation 3:

$$W(mw,n,mw',n') = exp\left(-\frac{||I_N(\rho(mw,n))-||I_N\left(\rho(mw',n')\right)||_{2,\sigma}^2}{f^2}\right) \quad [3]$$

Where f is the filtering level-determining smooth out variable. The weighted Euclidean reserve of image patch ρ (mw, n) is designated by $|\left|I_N(\rho(mw,n)) - -I_N\left(\rho(mw',n')\right)\right||_{2,\sigma}^2$ since patch ρ (mw′, n′) has a size of $2(MW+1)\times 2(MW_\rho+1)$. The following can be shown mathematically using Equation 4:

$$||I_N(\rho(mw,n)) - I_N\left(\rho(mw',n')\right)||_{2,x}^2 = \sum_{i=-MW_\rho}^{MW_\rho}\sum_{j=-M_\rho}^{M_\rho}\zeta_\sigma(i,j)\left(I_N(mw'+i,n'+j) - I_N(mw+i,n+j)\right)^2 \quad [4]$$

In this case, $\zeta_\sigma(i,j)$ stands for the Gaussian Kernel with σ > 0. The (mw, n) whims of the related window R′ are shown here. In accordance with Equation 3, the centre weight is always 1. A Wiener filter is used in this study to build the MWs, thereby achieving high picture-dispensing quality. Let $mw_n = (mw,n), mw_{n'}' = (mw',n'), u_{mw_n}^1 = U(mw,n), I_{mw_n} = I_N(mw,n)$, $I_{mw_n} = I_N(mw,n)$ and $W_{mw_n mw_{n'}'} = W(mw,n,mw',n')\times\lambda_{mw_n}$ is used to represent the MW. Consider R_0 = R − (mw, n) and (2) is simplified as Equation 5:

$$u_{mw_n}^1 - I_{mw_n} = \frac{1}{\lambda_{mw} + \sum_{mw_{n'}',\in R_0} W_{mw_n mw_{n'}'}}\left(\sum_{mw_{n'}',\in R_0} W_{mw_n mw_{n'}'}\left(I_{mw_{n'}'} - I_{mw_n}\right)\right) \quad [5]$$

Equation 5 designates that the noise $u_{mw_n}^1 - I_{mw_n}$ is removed through the NLM filter procedure. The noise that has stood removed is considered by reproducing the two terms: $1/\left(\lambda_{mw_n} + \sum_{mw_{n'}',\in R_0} W_{mw_n mw_{n'}'}\right)$ and $\sum_{mw_{n'}',\in R_0} W_{mw_n mw_{n'}'}\left(I_{mw_{n'}'} - I_{mw_n}\right)$. Here, λ_{m_n} is crucial for the denoising procedure. The quantity of noise removed decreases λ_{m_n}. As a result, excessive smoothness is easily produced. It's conceivable that some visual characteristics will be diminished or eliminated. For instance, NLM will produce an oversmoothed output with a centre weight of 0. The removed noise diminishes as λ_{m_n} enhance. It demonstrates that denoising alone is not sufficient. For example, when the midpoint weight is set to one, NLM produces the desired result. Nevertheless, there is still a sizable amount of residual noise in the denoised image. As a result, the middle weight ought to fluctuate suitably between 0 and 1. This is the foundation for figuring out the intermediate weight. Additionally, smaller midway weights are used to minimise noise while maintaining greater structure in the denoised image. In the frequency domain, higher midway weights

are used to minimise noise without over-smoothing. The calculation of midpoint weights is considerably improved by this rule. The proposed midpoint weight can be expressed using these parameters (Equation 6):

$$\lambda_{mw_n} = \frac{\tau f^2}{V_{mw_n} + \tau f^2} \quad [6]$$

It is comparable to the Wiener filter. In both the gradient and wavelet domains, the proposed V_{m_n}. It is considered better than the previous Wiener filters, as shown in Equation 7.

$$V_{mw_n} = max\left(q_{mw_n} - \tau f^2, 0\right) \quad [7]$$

Here, the image's features are calculated using the feature descriptor q_{m_n}, which is shown in Equation 8:

$$q_{mw_n} = \frac{1}{|R_0|} \sum_{(mw',n') \in R_0} ||I_N\left(\rho(mw,n)\right) - I_N\left(\rho\left(mw',n'\right)\right)||_{2,\sigma}^2 \quad [8]$$

Where the continuous τ is used for tuning the demonising boundary in the local Wiener filter. Equation 6 indicates that λ_{m_n} would be large for the section with reduced V_{m_n}. As a result, Eq. 6 is consistent with the additional rule. Consequently, the WMW-NLM container produces images of higher value than NLM filters with varying midpoint weightiness.

Feature Extraction Using Swin Transfer

The paper "Swin Transformer for Colorectal Cancer Analysis" utilises the Swin Transformer model to identify and analyse colorectal cancer in medical imaging. The Swin (Shifted Window) Transformer is a variation of the Transformer architecture specifically designed for vision tasks. Unlike standard vision transformers, which operate globally, Swin Transformers use attention mechanisms within shifted windows, allowing them to capture both local and global context at a lower computational cost. This makes them ideal for high-resolution medical images, such as those used in colorectal cancer analysis. The Swin Transformer can evaluate colorectal tissue slides by learning intricate spatial relationships within the images. The model identifies key patterns and characteristics that indicate malignant tissue. high-quality diagnostic and analytical accuracy. To enhance the identification of cancer without. This method usually uses transfer learning, since it requires a larger labelled dataset. Here, colorectal cancer images are used to fine-tune a Swin Transformer pre-trained on a large dataset.

- **Self-Attention Mechanism in a Window:** The self-attention within each window is calculated as follows, given a window W with query Q, key K, and charge V conditions as in Equation 9:

$$Attention(Q, K, T) = soft\,max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [9]$$

Where d_k is the dimension of the key vectors.

- **Shifted Window Mechanism:** The Swin Transformer alternates between regular window-based attention and shifted window attention. This shift is achieved by moving the window by a predetermined number of pixels, enabling patches to interact at window boundaries.

- **Patch Embedding:** The image patches are embedded in a high-dimensional space as shown in Equation 10:

$$X = Conv(X_{input}) \quad [10]$$

Conv is a convolution operation that converts input patches into fixed-dimensional vectors.

- **MLP Layer:** The output is passed to a multi-layer perceptron (MLP) for feature transformation, following the attention layers, as given in Equation 11.

$$MLP(X) = \sigma(W_2.Re\,L\,U(W_1.X + b_1) + b_2) \quad [11]$$

Where σ is a motivation function like ReLU, b_1 and b_2 are biases, and W_1 and W_2 are weight matrices.

- **Loss Function:** Colorectal cancer classification can be performed using a basic cross-entropy loss (Equation 12).

$$L_{CE} = -\sum_{i=1}^{N} y_i\,log(\hat{y}_i) \quad [12]$$

Where y_i is the true label and $\hat{y}i$ is the forecast possibility for each class.

This Swin Transformer-based technique enhances the accuracy and robustness of colorectal cancer analysis by utilising sophisticated attention mechanisms that capture both micro-level characteristics within windows and macro-level features across the entire image.

Model the Architecture of AC-GAN

ACGAN is a modified GAN, as illustrated in Figure 3. Unlike GAN, ACGAN can generate samples from specific categories while also recognising and categorising input data. The generator G uses random noise and label b to create new samples G (a,b). After analysing the input samples, discriminator D decides if they are authentic or fake. Through adversarial training, ACGAN's ability to generate and identify samples is continually improved. Ultimately, the model can generate new examples with accurate labels.

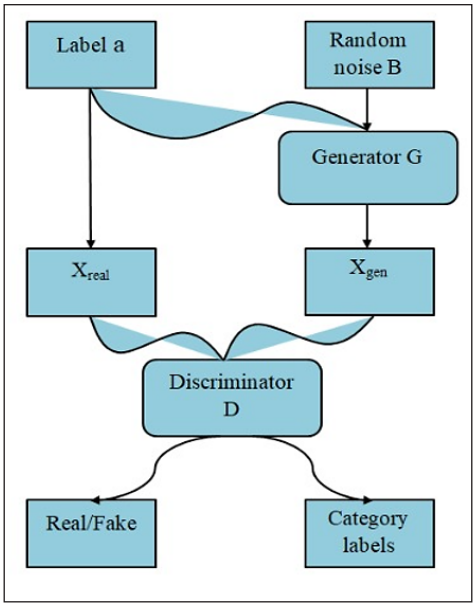


Figure 3. Block diagram of ACGAN

ACGAN's loss function is divided into two halves since it needs to assess the input samples' source and class label information (Equations 13 to 16):

$$L_{source} = E_{X\sim P_{data}}[log\, D\,(x)] + E_{B\sim P_b}\left[log\left(1 - D\big(G(b,a)\big)\right)\right] \quad [13]$$

$$L_{class} = E_{X\sim P_{data}}[log\, P\,(A = a|X_{real})] + E_{B\sim P_b}\left[log\, P\,(A = a|X_{generated})\right] \quad [14]$$

In this case, $P(A = a|X_{real})$ represents the provisional probability distribution of the genuine example, $P(A = a|X_{generated})$ represents the provisional probability distribution of the produced example, and G (a, b) represents the generated example when the generator inputs are random noise and example label b.

Making the most of $L_{source} + L_{class}$ is the impartial function of the D, and maximising $L_{source} - L_{class}$ is the impartial function of the G, as illustrated in the folllowing:

$$L_G = E_{X\sim P_{data}}[log\, D\,(x)] + E_{B\sim P_b}\left[log\left(1 - D(G(b,a))\right)\right] + E_{X\sim P_{data}}[log\, P\,(A = a|X_{real})] + E_{B\sim P_b}[log\, P\,(A = a|X_{generated})] \quad [15]$$

$$L_G = E_{X\sim P_{data}}[log\, D\,(x)] + E_{B\sim P_b}\left[log\left(1 - D(G(b,a))\right)\right] + E_{X\sim P_{data}}[log\, P\,(A = a|X_{real})] - E_{B\sim P_b}[log\, P\,(A = a|X_{generated})] \quad [16]$$

Advantages of the Proposed Model

- The Swin Transformer gathers multi-scale features to improve lesion detection and classification accuracy.
- The Auxiliary Classifier GAN creates high-quality synthetic samples to overcome data scarcity and enhance model resilience.
- The auxiliary classifier in the AC-GAN provides class-conditioned information during generation, allowing synthetic samples to be associated with the corresponding classification categories. In this study, the classification task is binary and should not be interpreted as cancer-stage prediction.

RESULT AND DISCUSSIONS

Experimental Setup

The computer used for the simulation has an Intel Core i7 processor. Additionally, Python 3.8 and PyTorch 1.10 were utilised to develop the proposed AC-GAN-ST model. To investigate the impact of the train-test split on classification performance, three splits were employed: 90:10, 80:20, and 70:30. The validation data used in model development were held separate from the test data. The samples generated by the AC-GAN were utilised only during training to increase the number of training samples. The original test data were not utilised in the image generation process or training process. Selecting and tuning hyperparameters is necessary for optimal model performance. Kather et al. (2019). The suggested network employs a random search methodology, with the main components being batch size, optimisation method, number of epochs, and learning rate. The suggested AC-GAN-ST hyperparameter combinations are listed in Table 1.

Table 1
Hyperparameters setting

Parameter	Value
Epochs	100
Spatial patch sizes	[9 × 9], [11 × 11], [13 × 13], [15 × 15], [17 × 17]
Optimiser	ADAM
Learning rate	0.001
Batch size	64

Sensitivity

Sensitivity, sometimes referred to as recall, is the percentage of true positives a test correctly identifies. It is a significant metric for evaluating the presentation of diagnostic tests, especially in medical contexts, as in Equation 17.

$$Sensitivity = \frac{TP}{TP + FN} \quad [17]$$

Specificity

Specificity is an arithmetical term that estimates the accuracy of a binary diagnostic test. It signifies the fraction of true negatives properly identified among all actual negatives. A height specificity value indicates that the test is effective at recognising negative suitcases (Equation 18).

$$Specificity = \frac{TN}{TN + FP} \quad [18]$$

Accuracy

Accuracy is a statistical indicator that is used to measure the presentation of organisational models. It estimates the fraction of correctly predicted cases among all examples in the dataset. The formula for calculating accuracy is as follows (Equation 19):

$$Accuracy = \frac{Number\ of Correct\ Pr\ e\ dictions}{Total\ Number\ of\ Pr\ e\ dictions} \times 100\% \quad [19]$$

Root Mean Square Error (RMSE)

RMSE is a commonly utilised statistic for calculating the difference between predicted and observed values in regression models. It indicates how far the predictions deviate from the actual data, with lower values indicating better model performance. It can be calculated using Equation 20.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad [20]$$

In the present study, accuracy, sensitivity, and specificity are treated as the primary measures for evaluating the binary classification task. RMSE is retained only as an additional numerical measure reported for consistency with the comparative results included in the manuscript. Since RMSE is more commonly associated with continuous-valued prediction, it is not used as the primary criterion for evaluating classification performance or for establishing superiority over other methods.

Execution Time

Execution time refers to the total duration essential to run a specific procedure or process, from initiation to accomplishment. It is a crucial measure in performance evaluation, typically expressed in seconds or milliseconds, as in Equation 21.

$$T = T_{start} - T_{end} \quad [21]$$

The collection of confusion matrices for the AC-GAN-ST representation across various training/testing (TR/TS) data sizes is shown in Figure 4. Figure 4 presents the confusion matrices obtained under the three training/testing configurations considered in the study. The matrices are shown separately for the 90:10, 80:20, and 70:30 configurations to illustrate the class-wise prediction behaviour of the proposed AC-GAN-ST model. The values in the matrices represent the corresponding classification outcomes for the evaluated samples and should be interpreted together with the quantitative performance measures reported in Tables 2 to 4.

Because the three panels correspond to different training/testing configurations, their sample counts are not expected to be identical. The confusion matrices are therefore intended to provide a visual representation of the classification outcomes for each experimental setting, whereas Tables 2 to 4 provide the summarised accuracy, specificity, and sensitivity values.

Table 2
Result analysis of the AC-GAN-ST technique on TR/TS data of 90:10

	TR/TS (90:10)		
Class Label	**Accuracy**	**Specificity**	**Sensitivity**
Training Phase			
Benign	97.31	92.35	95.47
Malignant	97.31	95.47	92.35
Average	97.31	93.91	93.91

Table 2 (*continued*)

TR/TS (90:10)			
Class Label	**Accuracy**	**Specificity**	**Sensitivity**
Testing Phase			
Benign	94.46	90.00	95.47
Malignant	94.46	95.47	90.00
Average	94.46	92.735	92.735

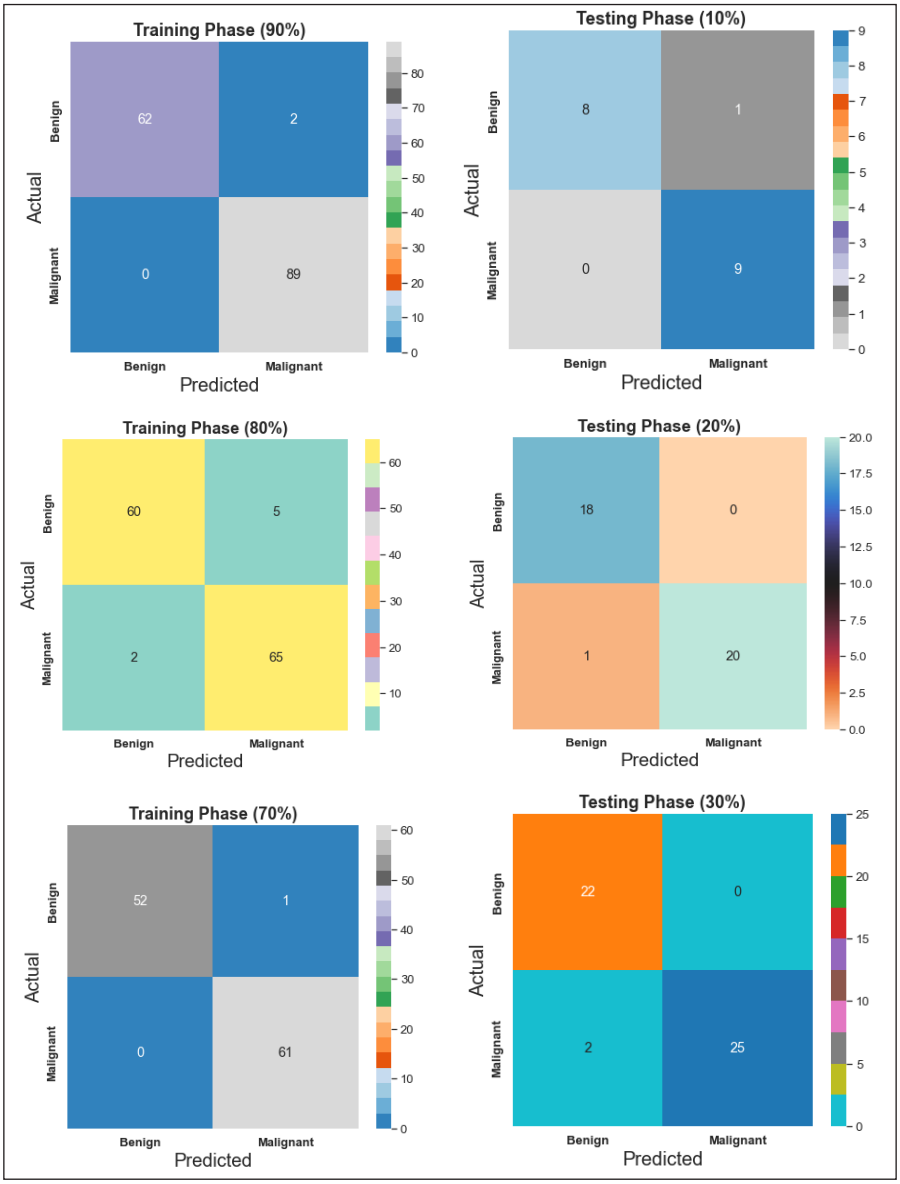


Figure 4. Confusion matrix of the AC-GAN-ST method on distinct sizes of TR/TS data

Table 2 summarises the performance of AC-GAN-ST under the 90:10 training/testing configuration. The model achieved an average training accuracy of 97.31%, with average specificity and sensitivity of 93.91% each. On the testing data, the corresponding average accuracy, specificity, and sensitivity were 94.46%, 92.74%, and 92.74%. These values represent the performance obtained under the 90:10 configurations and are interpreted as a dataset-specific evaluation of the proposed framework. The average metrics for the testing phase were accuracy 94.46%, specificity 92.74%, and sensitivity 92.74%.

Using training/testing (TR/TS) data with a 90:10 ratio, Figure 5 presents the training accuracy (TA), validation accuracy (VA), training loss (TL), and validation loss (VL) for this configuration. The plotted curves show the progression of training and validation metrics across training epochs. The validation and training curves follow a similar overall trend, while the corresponding loss curves decrease during training. These observations provide a qualitative indication of the model's learning behaviour; no statistical significance test was performed on the differences between the curves. The AC-GAN-ST model exhibited the lowest TL and VL values, with VL significantly below TL, as indicated by the experimental data.

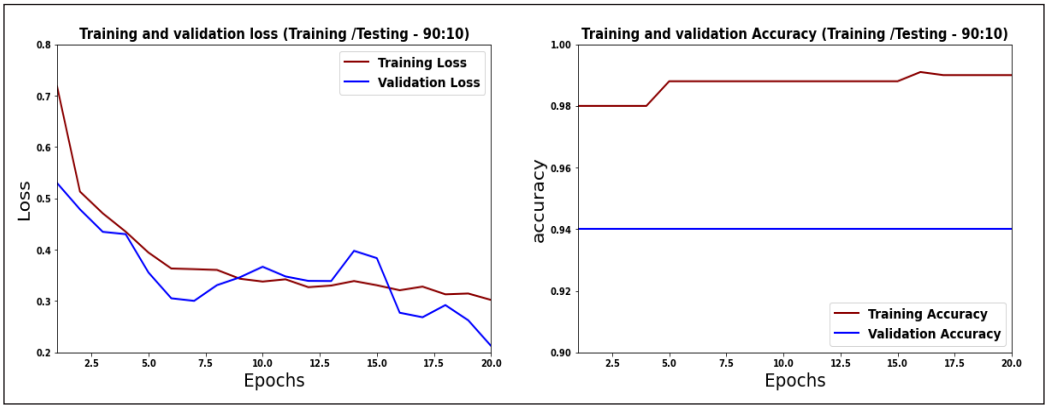


Figure 5. TA/TL and VA/VL analysis of the AC-GAN-ST model on the TR/TS data of 90:10

Table 3 reports the results obtained using the 80:20 training/testing configuration. The model achieved an average training accuracy of 95.23%, with average specificity and sensitivity of 93.05% each. For the test data, the average accuracy was 95.97%, and the average specificity and sensitivity were 92.78% each. The results demonstrate that the reported performance remains relatively consistent as the proportion of testing data increases under this experimental configuration. This resulted in an overall average accuracy of 95.23%, specificity of 93.05%, and sensitivity of 93.05%. In testing, the model maintained high performance, achieving 95.97% accuracy across both classes, with the benign and malignant classes showing specificities of 90.24% and 95.33% and sensitivities

of 95.33% and 90.24%, respectively. The average test results were 95.97% accuracy, 92.78% specificity, and 92.78% sensitivity.

Table 3
Result analysis of the AC-GAN-ST technique on the TR/TS data of 80:20

TR/TS (80:20)			
Class Label	**Accuracy**	**Specificity**	**Sensitivity**
Training Phase			
Benign	95.23	91.73	94.38
Malignant	95.23	94.38	91.73
Average	95.23	93.05	93.05
TR/TS (80:20)			
Class Label	**Accuracy**	**Specificity**	**Sensitivity**
Testing Phase			
Benign	95.97	90.24	95.33
Malignant	95.97	95.33	90.24
Average	95.97	92.78	92.78

Figure 6 displays the TA and VA produced by the AC-GAN-ST method using 80:20 TR/TS data. According to the experimental data, the AC-GAN-ST model obtained the highest TA and VA values, with VA seemingly exceeding TA. Figure 6 also displays the TL and VL derived from 80:20 TR/TS data using the AC-GAN-ST technique. The AC-GAN-ST model exhibited the lowest TL and VL values in the experimental data, with VL appearing to be lower than TL.

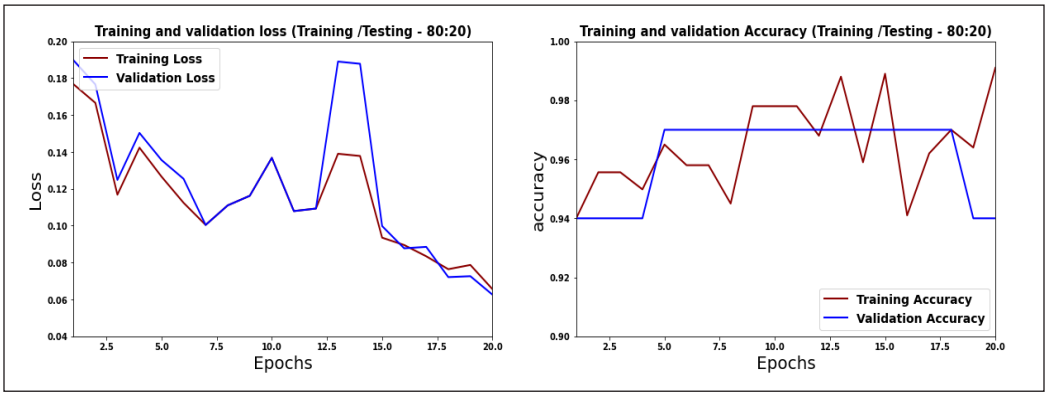


Figure 6. TA/TL and VA/VL analysis of the AC-GAN-ST model on the TR/TS data of 80:20

Table 4 presents an analysis of the AC-GAN-ST method's results using various performance measures on a 70:30 training-to-testing split (TR/TS). In the training stage,

the model achieved 97.13% accuracy across both benign and malignant classes, with specificity and sensitivity averaging 97.19% and 95.73%, respectively. In the testing phase, the model maintained high accuracy at 96.45% for both classes, with an average specificity of 96.38% and sensitivity of 93.45%. These results indicate robust model performance in distinguishing between benign and malignant classes in both training and testing phases.

Table 4
Result analysis of the AC-GAN-ST technique on the TR/TS data of 70:30

TR/TS (70:30)			
Class Label	**Accuracy**	**Specificity**	**Sensitivity**
Training Phase			
Benign	97.13	96.39	96.23
Malignant	97.13	98.00	95.23
Average	97.13	97.19	95.73
TR/TS (70:30)			
Class Label	**Accuracy**	**Specificity**	**Sensitivity**
Testing Phase			
Benign	96.45	95.76	94.18
Malignant	96.45	97.00	93.32
Average	96.45	94.76	93.45

The TA and VA produced by the AC-GAN-ST algorithm utilising 70:30 TR/TS data are displayed in Figure 7. The AC-GAN-ST model produced the highest values for both TA and VA, according to the experimental results, with VA seeming to be higher than TA. Figure 7 also displays the TL and VL derivatives since the AC-GAN-ST model on 70:30 TR/TS data. The AC-GAN-ST method yielded the lowest TL and VL values, with VL apparently lower than TL, consistent with the experimental results.

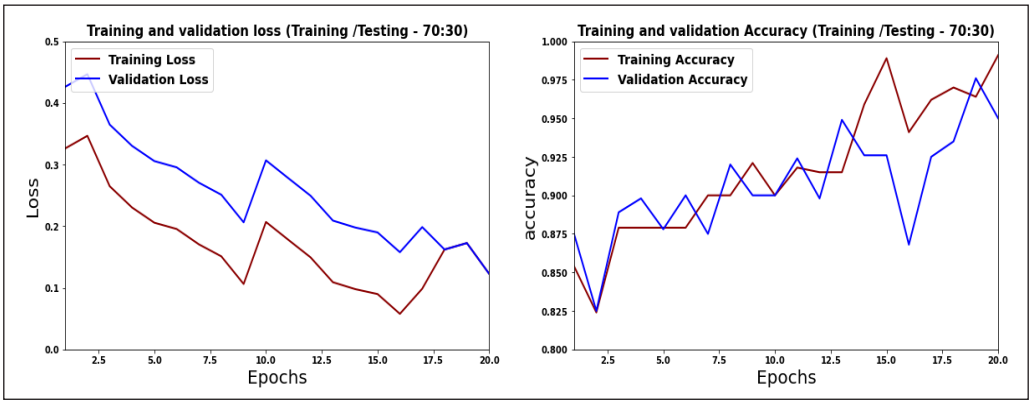


Figure 7. TA/TL and VA/VL analysis of the AC-GAN-ST method on the TR/TS data of 70:30

The presentation was proposed for comparative analysis in Table 5 and Figures 8, 9, 10, 11, and 12. comparisons between metrics of different models, such as the proposed AC-GAN-ST approach and current ones. models from recent research. The table has identified major presentation measures: Sensitivity, Specificity, Accuracy, RMSE, and Time of Execution. The proposed AC-GAN-ST algorithm. has a better presentation of most metrics and has achieved the highest values of Sensitivity (93.45%), Specificity (94.76%), and Accuracy (96.45%). It also shows the lowest RMSE (15.87) and the minimum execution time (1.876 seconds), both of which improve prediction accuracy and the effectiveness of predictive models relative to individuals (Kather et al., 2019; Wang et al., 2021; Yang et al., 2021). This suggests that AC-GAN-ST would be a more effective and valuable solution to the problem under study.

Table 5
Performance measures of the proposed model

Research Work	Sensitivity	Specificity	Accuracy	RMSE	Execution Time
Yang et al. (2021)	76.46	81.23	65.47	31.11	15.224
Wang et al. (2021)	80.44	86.23	80.22	24.55	7.332
Shen et al. (2022)	76.45	80.11	85.13	19.33	8.654
Peng et al. (2019)	74.56	87.56	89.11	17.11	12.117
Kavitha et al. (2022)	79.43	90.45	91.34	22.43	9.223
Neelakandan et al. (2023)	87.99	91.11	93.76	20.55	5.887
Prezja et al. (2023)	92.13	91.45	95.77	21.76	3.876
Proposed AC-GAN-ST	93.45	94.76	96.45	15.87	1.876

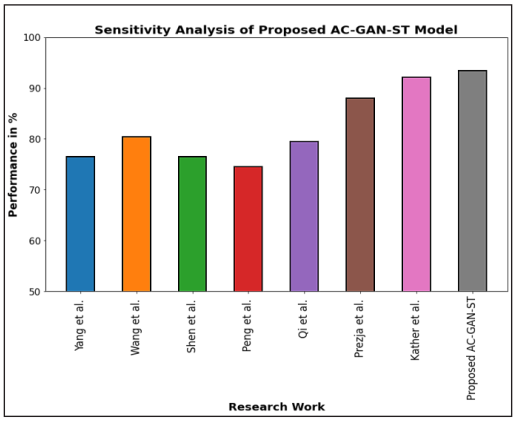


Figure 8. Sensitivity analysis for the proposed model

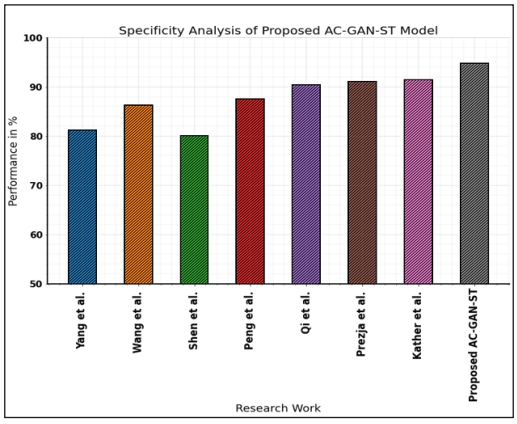


Figure 9. Specificity analysis for the proposed model

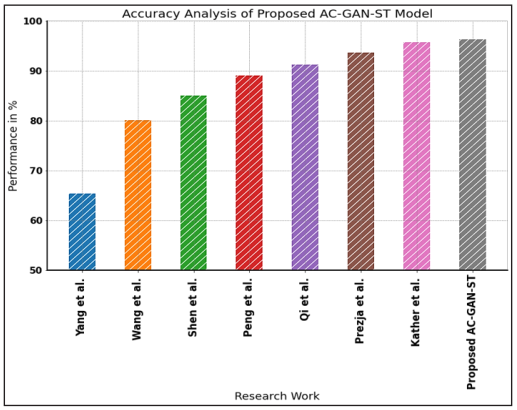


Figure 10. Accuracy analysis for the proposed model

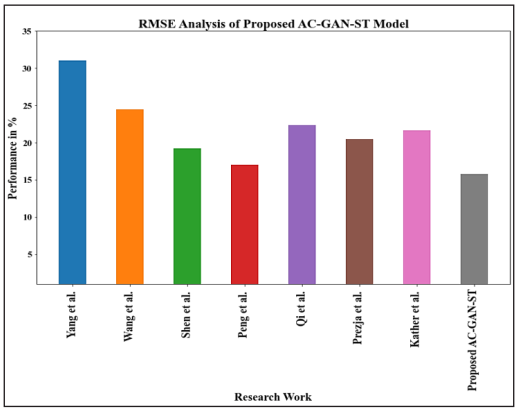


Figure 11. RMSE analysis for the proposed model

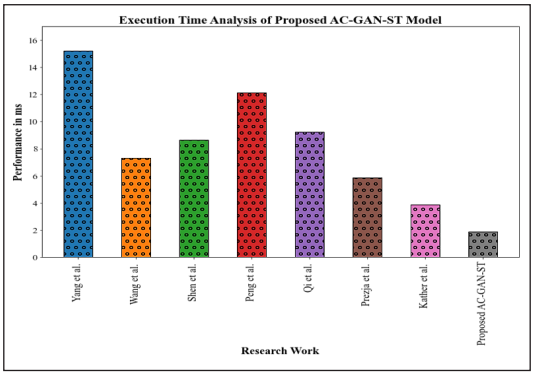


Figure 12. Execution Time analysis for the proposed model

ROC Curve

The ROC curve is an important method of evaluating the efficiency of colorectal cancer detection. models work. The sensitivity-specificity trade-off between the true positive rate and the false positive rate. positive rate) by the ROC curve that relates the true positive rate with the False positive rates of different threshold values as specified in Figure 13. The accuracy of models is measured by the area under the curve. ROC curve (AUC): an AUC of approximately

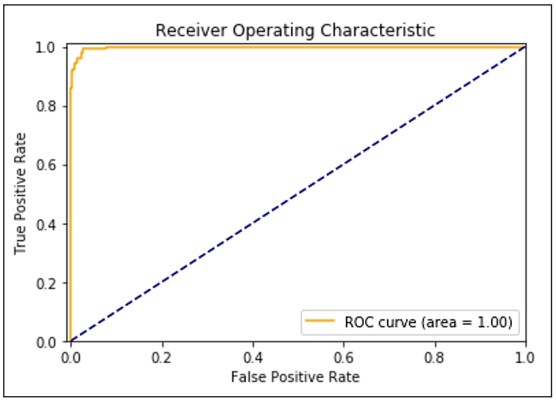


Figure 13. ROC curve analysis for the proposed AC-GAN-ST model

1 indicates that the procedure is highly accurate. In colorectal cancer screening, a high AUC suggests that the diagnostic instrument is useful in differentiating. between malignant and non-cancerous, in favour of early detection and evolution.

Comparative Analysis of the Proposed Framework

Each module in the proposed model serves a significant role. In this section, we present a series of ablation studies on the Warwick-QU colorectal cancer dataset to compare the proposed AC-GAN with other models, including GAN, DNN, ANN-LR, and GRNN, to investigate performance improvements and demonstrate the motivations behind our proposed AC-GAN.

Influence of the GAN

The AC-GAN has shown promising potential for colorectal cancer research by improving diagnostic accuracy and enabling earlier detection. Researchers can create high-quality, realistic images using the generative property of ACGAN with limited information. Simulate various colorectal cancer scenarios to help train and perfect ML models. models used for detection. This approach not only minimises the use of big data, as these are known to be. present in small amounts but similarly increase model strength by producing images that represent different stages and varieties of colorectal cancer, ultimately leading to more precise and timely results. diagnoses. Lastly, the AC-GAN+ST model gave a result. accuracy = 96.45 per cent of our input data. In the evaluation, the current studies by Yang et al. (2021) are utilised. The accuracy performances are 65.44%, 80.26%, 85.17%, 89.17%, 91.34%, 93.76%, 95.77%, and 93.67%, respectively (Peng et al., 2019; Qi et al.; Prezja et al., 2023; Shen et al., 2022; Wang et al., 2021).

Challenges and Limitations

The Combination of Auxiliary Classifier Generative Adversarial Network (AC-GAN) with. Swin Transformers for analysing colorectal cancer have limitations. First, because of the model architecture's sophistication, additional computation can be introduced. It is not so practical in real-time applications because it requires demands. In addition, the performance of the AC-GAN is highly dependent on the quality and quantity of the training dataset, which is limited in clinical settings where annotated images are scarce. There is also a risk of overfitting, especially with small datasets, which can impair the model's generalisability to unseen data. Moreover, it is possible to analyse GAN outputs. be difficult because the generated images might not conform to clinical expectations, raising doubts about their usefulness in diagnostic workflows. Lastly, integrating modern models such as Swin Transformers can be highly education-intensive and time-consuming, limiting their

widespread use in medical institutions. An additional limitation concerns the evaluation of the generated AC-GAN samples. The present experimental protocol focused primarily on downstream classification performance and did not include dedicated synthetic-image quality measures such as Fréchet Inception Distance, Kernel Inception Distance, or an independent expert-based assessment. Consequently, the contribution of the AC-GAN should be interpreted as a training-data augmentation mechanism rather than as evidence of clinically validated realism in synthetic images. Future work should incorporate quantitative and expert-based evaluations of the generated histopathological samples, together with controlled component-wise experiments.

CONCLUSION

Finally, the combination of ACGAN and Swin Transformers for colorectal cancer. The analysis shows great improvement in medical imaging and diagnostics. technology. This hybrid approach will utilise the strengths of both designs to improve the system. The production of high-quality synthetic images and the effective classification of cancerous lesions. The findings suggest enhanced accuracy and strength in detecting and analysing colorectal. cancer, which preconditions more reliable and efficient diagnostic tools. This approach has potential for clinical applications and adds to the growing body of studies on the applications of DL in oncology. There should be further studies on perfecting this in the future. model and discussing its possible integration with the actual workflow in the clinical setting, and the final objective of enhancing patient outcomes by detecting and treating them early.

ACKNOWLEDGEMENT

We are very grateful to the management and senior officials of Mohan Babu University for providing research and infrastructure facilities to carry out this research work.

REFERENCES

Echle, A., Laleh, N. G., Schrammen, P. L., West, N. P., Trautwein, C., Brinker, T. J., Gruber, S. B., Buelow, R. D., Boor, P., Grabsch, H. I., Quirke, P., & Kather, J. N. (2021). Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: A systematic literature review. *ImmunoInformatics, 3-4*, Article 100008. https://doi.org/10.1016/j.immuno.2021.100008

Hasan, M. A. M., Maniruzzaman, M., & Shin, J. (2024). WGCNA and machine learning-based integrative bioinformatics analysis for identifying key genes of colorectal cancer. *IEEE Access, 12*, 144350-144363. https://doi.org/10.1109/ACCESS.2024.3472688

Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N. A., Ferber, D., Jansen, L., Reyes-Aldasoro, C. C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M., & Halama, N. (2019). Predicting survival from colorectal cancer histology slides using

deep learning: A retrospective multicenter study. *PLOS Medicine, 16*(1), Article e1002730. https://doi.org/10.1371/journal.pmed.1002730

Kavitha, T., Mathai, P. P., Karthikeyan, C., Ashok, M., Kohar, R., Avanija, J., & Neelakandan, S. (2022). Deep learning based capsule neural network model for breast cancer diagnosis using mammogram images. *Interdisciplinary Sciences: Computational Life Sciences, 14*(1), 113-129. https://doi.org/10.1007/s12539-021-00467-y

Khalid, M., Deivasigamani, S., V, S., & Rajendran, S. (2024). An efficient colorectal cancer detection network using atrous convolution with coordinate attention transformer and histopathological images. *Scientific Reports, 14*(1), Article 19109. https://doi.org/10.1038/s41598-024-70117-y

Khan, U., Koivukoski, S., Valkonen, M., Latonen, L., & Ruusuvuori, P. (2023). The effect of neural network architecture on virtual H&E staining: Systematic assessment of histological feasibility. *Patterns, 4*(5), Article 100725. https://doi.org/10.1016/j.patter.2023.100725

Mulenga, M., Abdul Kareem, S., Qalid Md Sabri, A., Seera, M., Govind, S., Samudi, C., & Mohamad, S. B. (2021). Feature extension of gut microbiome data for deep neural network-based colorectal cancer classification. *IEEE Access, 9*, 23565-23578. https://doi.org/10.1109/ACCESS.2021.3050838

Neelakandan, S., Rajesh, P., Mahesh, P., Karthik, P., & Mukesh, M. (2023). An efficient WBC cancer cells prediction using dense net algorithm. In *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICAISC58445.2023.10199502

Peng, T., Boxberg, M., Weichert, W., Navab, N., & Marr, C. (2019). Multi-task learning of a deep k-nearest neighbour network for histopathological image classification and retrieval. In D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, & A. Khan (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2019* (Lecture Notes in Computer Science, Vol. 11764, pp. 676-684). Springer. https://doi.org/10.1007/978-3-030-32239-7_75

Prezja, F., Annala, L., Kiiskinen, S., Lahtinen, S., & Ojala, T. (2023). *Adaptive variance thresholding: A novel approach to improve existing deep transfer vision models and advance automatic knee-joint osteoarthritis classification* [Preprint]. arXiv. https://arxiv.org/abs/2311.05799v1

Prezja, F., Äyrämö, S., Pölönen, I., Ojala, T., Lahtinen, S., Ruusuvuori, P., & Kuopio, T. (2023). Improved accuracy in colorectal cancer tissue decomposition through refinement of established deep learning solutions. *Scientific Reports, 13*(1), Article 15879. https://doi.org/10.1038/s41598-023-42357-x

Prezja, F., Paloneva, J., Pölönen, I., Niinimäki, E., & Äyrämö, S. (2022). Deepfake knee osteoarthritis x-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. *Scientific Reports, 12*(1), Article 18573. https://doi.org/10.1038/s41598-022-23081-4

Prezja, F., Pölönen, I., Äyrämö, S., Ruusuvuori, P., & Kuopio, T. (2022). H&E multi-laboratory staining variance exploration with machine learning. *Applied Sciences, 12*(15), Article 7511. https://doi.org/10.3390/app12157511

Ragab, M., Mahmoud, M. M., Asseri, A. H., Choudhry, H., & Yacoub, H. A. (2023). Optimal deep transfer learning based colorectal cancer detection and classification model. *Computers, Materials & Continua, 74*(2), 3279-3295. https://doi.org/10.32604/cmc.2023.031037

Ruusuvuori, P., Valkonen, M., & Latonen, L. (2023). Deep learning transforms colorectal cancer biomarker prediction from histopathology images. *Cancer Cell, 41*(9), 1543-1545. https://doi.org/10.1016/j.ccell.2023.08.006

Sarwinda, D., Paradisa, R. H., Bustamam, A., & Anggia, P. (2021). Deep learning in image classification using residual network (ResNet) variants for detection of colorectal cancer. *Procedia Computer Science, 179*, 423-431. https://doi.org/10.1016/j.procs.2021.01.025

Sasmal, P., Sharma, V., Prakash, A. J., Bhuyan, M.-K., Patro, K. K., Samee, N. A., Alamro, H., Iwahori, Y., Tadeusiewicz, R., Acharya, U. R., & Pławiak, P. (2024). Semi-supervised generative adversarial networks for improved colorectal polyp classification using histopathological images. *Information Sciences, 658*, Article 120033. https://doi.org/10.1016/j.ins.2023.120033

Shen, Y., Luo, Y., Shen, D., & Ke, J. (2022). RandStainNA: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, & S. Li (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2022* (Lecture Notes in Computer Science, Vol. 13432, pp. 212-221). Springer. https://doi.org/10.1007/978-3-031-16434-7_21

Sirinukunwattana, K., Domingo, E., Richman, S. D., Redmond, K. L., Blake, A., Verrill, C., Leedham, S. J., Chatzipli, A., Hardy, C., Whalley, C. M., Wu, C.-H., Beggs, A. D., McDermott, U., Dunne, P. D., Meade, A., Walker, S. M., Murray, G. I., Samuel, L., Seymour, M., … Koelzer, V. H. (2021). Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. *Gut, 70*(3), 544-554. https://doi.org/10.1136/gutjnl-2019-319866

Skrede, O.-J., De Raedt, S., Kleppe, A., Hveem, T. S., Liestøl, K., Maddison, J., Askautrud, H. A., Pradhan, M., Nesheim, J. A., Albregtsen, F., Farstad, I. N., Domingo, E., Church, D. N., Nesbakken, A., Shepherd, N. A., Tomlinson, I., Kerr, R., Novelli, M., Kerr, D. J., & Danielsen, H. E. (2020). Deep learning for prediction of colorectal cancer outcome: A discovery and validation study. *The Lancet, 395*(10221), 350-360. https://doi.org/10.1016/S0140-6736(19)32998-8

Sujatha, R., Mahalakshmi, K., & Yoosuf, M. S. (2024). Colorectal cancer prediction via histopathology segmentation using DC-GAN and VAE-GAN. *EAI Endorsed Transactions on Pervasive Health and Technology, 10*. https://doi.org/10.4108/eetpht.10.5395

Sun, W., Zhao, R., Zhang, K., Gao, J., & Qu, G. (2024). Distribution-aware loss for lesions detection using white-light endoscopy in colorectal region. *IEEE Access, 12*, 46138-46148. https://doi.org/10.1109/ACCESS.2024.3381614

Wagner, S. J., Reisenbüchler, D., West, N. P., Niehues, J. M., Zhu, J., Foersch, S., Veldhuizen, G. P., Quirke, P., Grabsch, H. I., van den Brandt, P. A., Hutchins, G. G.-A., Richman, S. D., Yuan, T., Langer, R., Jenniskens, J. C.-A., Offermans, K., Mueller, W., Gray, R., Gruber, S. B., … Kather, J. N. (2023). Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell, 41*(9), 1650-1661.e4. https://doi.org/10.1016/j.ccell.2023.08.002

Wang, K. S., Yu, G., Xu, C., Meng, X. H., Zhou, J., Zheng, C., Deng, Z., Shang, L., Liu, R., Su, S., Zhou, X., Li, Q., Li, J., Wang, J., Ma, K., Qi, J., Hu, Z., Tang, P., Deng, J., … Deng, H. W. (2021). Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence. *BMC Medicine, 19*(1), Article 76. https://doi.org/10.1186/s12916-021-01942-5

Yamashita, R., Long, J., Longacre, T., Peng, L., Berry, G., Martin, B., Higgins, J., Rubin, D. L., & Shen, J. (2021). Deep learning model for the prediction of microsatellite instability in colorectal cancer: A diagnostic study. *The Lancet Oncology, 22*(1), 132-141. https://doi.org/10.1016/S1470-2045(20)30535-0

Yang, J., Shi, R., & Ni, B. (2021). MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (pp. 191-195). IEEE. https://doi.org/10.1109/ISBI48211.2021.9434062

Zhang, F., Jiang, S., & Li, Y. (2020). Application of massive parallel deep learning algorithm in the prediction of colorectal carcinogenesis of familial polyposis. *IEEE Access, 8*, 129432-129445. https://doi.org/10.1109/ACCESS.2020.3009412

Zidan, U., Gaber, M. M., & Abdelsamea, M. M. (2023). SwinCup: Cascaded Swin transformer for histopathological structures segmentation in colorectal cancer. *Expert Systems with Applications, 216*, Article 119452. https://doi.org/10.1016/j.eswa.2022.119452

Pertanika J. Sci. & Technol. 34 (S3): 27 - 53 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

The Sky Emotional Chatbot Companion: Proactive AI Intervention Fostering Emotional Connectivity to the Mind

Baby Vadlana, Karnam Akhil*, Tamma Bhuwan Vamsi, and Tejasree Lokireddy

Department of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Vignana Jyothi Nagar, Bachupally, 500118 Hyderabad, Telangana, India

ABSTRACT

In recent years, mental health issues, emotional instability and depression have increased exponentially, affecting millions of people of all ages. The increasing percentage of mental illness, integrating with emotional instability and suboptimal perception, has signified the need for innovative and quick solutions to support and enrich emotional well-being. As traditional therapeutic practices become increasingly demanded, with the rise in advancements of technology, specifically with the increased use and evolution of Artificial Intelligence (AI), it paves the way to promising intervention pathways. To address issues of emotional instability, loneliness, and subjectivity, this paper proposes and develops the SKY Emotional Chatbot, designed to support emotional health and promote mental well-being. This proposed framework deploys transformer models and implements the PaLM (Pathways Language Model) framework to enable the chatbot to engage in interactive conversations. The chatbot's conversations will be based on users' previous conversations with it. Training and testing were conducted to ensure the chatbot achieves high accuracy and effectiveness in its conversations. Various transformer models, including DistilBERT, DistilRoBERTa, BERT, RoBERTa, ALBERT, XLNet, DeBERTa, and Electra, were used, achieving 97% precision. The SKY Emotional Chatbot is not merely reactive; it can reference and build on previous conversations, ensuring continuity in dialogue and fostering deeper emotional bonding and resonance with users. By providing a consistent and empathetic virtual companion, the chatbot plays an important role in mitigating negative feelings of isolation and despair, thereby making a proactive contribution to mental health care. The proposed framework focuses on the growing need for creative, innovative mental health solutions, the technological advancements that enable such innovations, and the potential of the SKY Emotional Chatbot to foster emotional well-being and resilience among users.

Keywords: Depression, emotional well-being, SKY emotional chatbot, transformer models

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 06 October 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.02

E-mail addresses:
baby_v@vnrvjiet.in (Baby Vadlana)
akhilresearch18@gmail.com (Karnam Akhil)
bhuwanvamsi@gmail.com (Tamma Bhuwan Vamsi)
tejasreelokireddy@gmail.com (Tejasree Lokireddy)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

The origin of depression means looking at a very complex interplay of biological factors. The main part of this is the disturbance of neurotransmitter systems, notably the monoamine systems that consist of hormones like serotonin, (Agrawal et al., 2025) norepinephrine and dopamine, but also involve the glutamate and GABA systems, according to Colita et al. (2024) and Fouad et al. (2024). Like DNA methylation and histone changes, these can strongly influence the expression of genes associated with depression. Moschny (2020) shed light on stress and external factors. Additionally, prolonged exposure to stress often results in observable morphological changes in brain regions involved in emotional regulation, such as the hippocampus, prefrontal cortex, and amygdala, leading to dendritic regression and alterations in synaptic connectivity (Ren et al., 2025). The neuroimmune system plays an important role because the pathophysiological schema of depression involves immunological dysregulation and inflammatory responses, highlighting the possible neuroimmune origins of disorders (Feng et al., 2025; Tarai et al., 2019). According to Moschny (2020) and Ren et al. (2025), there is a need for a comprehensive approach to treatment and scholarly research because these biological mechanisms are not discrete; rather, they are interconnected and involved in complex interactions that create a multifaceted network that aids both the initiation and continuation of depression.

The availability of emotional and practical support through community aid improves quality of life and increases self-esteem, thereby serving as an important buffer against the impact of social support on depressive symptoms (Alsabi et al., 2022). In the context of postpartum depression, social support, peer support, and community and culturally responsive intervention strategies have been shown to be effective in improving the symptoms, maternal and child health, and health outcomes (Norazman & Lee, 2024). Similarly, among the elderly, the use of technology with social networks has the potential to significantly reduce the social and physical aspects of depression, thereby improving the effectiveness of depression treatment (Lei et al., 2023). The way and quality of social groups are the key. For example, there are correlations between variables such as network size, density, and diversity of support levels and the perception of social support, which is key to people handling depression (Taylor et al., 2022). In clinics, although the effectiveness of social support is a bit lessened, it proves very useful, especially for new parents, because of the stigma and bias which can cause worse depression symptoms (Nakamura et al., 2020). All in all, these observations show why it is important to make and keep social ties and connections and to fight back stigma to prevent and cure depression at any age and in any part of the world.

Regarding the treatment of depressive symptoms in patients with chronic health conditions, Cognitive Behavioural Therapies, or CBTs, have proven to be a benchmark in treatment and have shown effectiveness at various stages. Traditional CBTs such as Beck's

Cognitive Therapy and Behavioural Activation are supported by a substantial body of evidence that demonstrates an individual's capacity to manage their depressive episodes and prevent them, but emerging techniques in the line of Mindfulness-Based Cognitive Therapy are also hinting at promising results, according to Fouad et al. (2024). Systematic review and meta-analysis of research from Mavranezouli et al. (2024) shows that CBTs bring moderate but distinct improvements in both depressive and anxiety levels in people with chronic diseases. What's more, a network meta-analysis, conducted by Li et al. (2024), found that more complex CBT methods, such as ultra-complex CBT, are adequate, albeit with reservations, for individuals with comorbid mental disorders, especially in persons between 18 and 58 years old. This stresses that although basic CBT is efficient, more involved interventions provide more and diverse benefits, and the need to follow up with consistent treatment programmes is paramount when addressing depression in patients with chronic illness.

Moreover, authors of multiple empirical studies also argue for the importance of EI to the analysis and assessment of depression. EI includes self-understanding, emotion regulation, and interpersonal competence in managing depressive states (Fan & Ma, 2024). For instance, a study of a female subject who is suffering from depression showed her self-awareness and mechanisms of coping and behavioural intervention, which would be essential to the successful treatment of her condition, such as social withdrawal and therapy, which are important to managing her disease. (Shen et al., 2024). Also, a study carried out in a medical college examined teachers with severe depressive symptoms and found a varying correlation between EI and depressive states and suggests that heightened EI can reduce depressive symptoms (Zeidner et al., 2012). Further support is found for the claim that individuals with high levels of EI are more adept at dealing with negative emotions such as anxiety and stress, often linked with depression (Volz et al., 2024). For older folks, EI was not equal; women had more feelings of care, and this was linked to sad signs, showing that being more aware can put one at risk of being sad (Bartlett et al., 2024). For teenagers, EI did not have a connection to sadness, but there was a link to being strong, which later helped mental health, and so making EI even better could help mental health get stronger because of stronger strength (Piro et al., 2024). According to the results of these tests, emotional intelligence (EI) is a useful component of new approaches to healing and improved mental health because it plays a significant role in preventing depression in many groups. While there isn't a direct correlation between emotional intelligence (EI) and depression in teenagers, EI plays a significant role in developing strong emotional control, which can improve mental health. Therefore, improving EI is a good way to improve mental health because of its strength (Piro et al., 2024).

For example, Anisha et al. (2024) cited an AI chatbot used in the diagnosis of depression as an evidence-based way to address persistent questions about common treatments and

the need to provide sufficiently skilled mental health professionals worldwide (Anisha et al., 2024; Zhong et al., 2024). AI chatbots provide tailored support and interaction that significantly improve access to mental health care and aid in the early detection and diagnosis of depressive symptoms (Manole et al., 2024). These technologies and digitised innovations are particularly invaluable at the primary care levels, where there is a need to intervene quickly. Chatbots offer round-the-clock assistance, helping users manage mental health concerns more consistently and effectively (Kang & Hong, 2025).

Nevertheless, the interdisciplinary nature of chatbot sentiment and emotion detection raises significant ethical concerns that lie beneath the surface. The biggest problem is the risk that the system might misinterpret or oversimplify users' emotional states due to biases in the training data. These biases can lead to incorrect evaluations and inappropriate system responses. So, in addition to thorough oversight of the work by mental health specialists and adherence to well-defined regulations, it is essential to ensure that the chatbot's outputs are clinically valid, safe, and ethically sound (Denecke & Gabarron, 2024; Guo et al., 2024). Despite these difficulties, the number of empirical studies supporting the idea of AI chatbots as a source of mental and emotional well-being, behavioural change, and therapy supplementation is increasing (Ferrario et al., 2024).

Additionally, even though sentiment and emotion analysis improve chatbot functionality, biased datasets and the misinterpretation of emotional cues continue to raise ethical concerns (Denecke & Gabarron, 2024). Future developments include improving natural language processing models and incorporating AI-enhanced, clinically informed decision-support systems. Improvements in traditional strategies and diagnostic accuracy, as well as strengthening orthodox approaches, can contribute to a more efficient and equitable mental health care system, as noted by Fan (2024).

Problem Statement

Nowadays, there is enormous growth in challenges faced due to depression and emotional instability. Treating individuals with these symptoms of loneliness and emotional inexpressiveness, which are issues of subjectivity, has become a predominant task for better and healthier living. Existing models face many challenges in providing an optimal solution. To achieve this, we deploy an AI-based Chatbot and propose SKY Chatbot as an emotional support companion for continuous conversations.

Research Questions (RQ)

RQ1 : How do we assess the performance of emotions and sentiments associated with subjectivity concerns?

RQ2 : How do you tackle the challenge of Ethical Concerns while dealing with personalised intervention?

RQ3 : How do you deal with the unstructured data in CHATBOTS since the conversation is based on Text analytics?

RQ4 : How can data from previous conversations be mathematically modelled to enhance continuity and contextual awareness in dialogues?

Related Works

Denecke and Gabarron (2024) proposed chatbots for depression that use AI to assess users' emotional states and sentiments, aiming to offer empathetic responses and suitable recommendations. The ethical challenges to this integration of sentiment and emotion analysis are the possibility of misinterpreting emotional expressions among individuals experiencing depression (Yang et al., 2019), the probability of biases associated with training data, and the potential to misunderstand emotions and emotional representations among depressive individuals. The underlying research challenge, highlighted, is whether these chatbots offer the right combination of sentiment and emotion analysis to make a richer understanding of sentiment and emotion. There is ongoing demand for further research into the ethical complications of biased training data and perceptions concerning sentiment analysis, and a lack of understanding of how such issues affect the effectiveness and safety of chatbots and interventions.

Additionally, prolonged exposure to stress frequently results in observable morphological changes in brain regions involved in emotional regulation, such as the hippocampus, prefrontal cortex, and amygdala, leading to dendritic regression and alterations in synaptic connectivity (Ren et al., 2025). The neuroimmune system plays an important role because the pathophysiological schema of depression involves immunological dysregulation and inflammatory responses, highlighting the possible neuroimmune origins of disorders (Feng et al., 2025; Tarai et al., 2019). According to Moschny (2020) and Ren et al. (2025), there is a need for a comprehensive approach to treatment and scholarly research because these biological mechanisms are not discrete; rather, they are interconnected and involved in complex interactions that create a multifaceted network that aids both the initiation and continuation of depression.

Khan and Bokhari (2024) shared AI driven chatbots that offer 24 hour support, tailored contact, and early signs of disease in general practice however, there are still moral issues to consider evaluate the accuracy of the data privacy and healthcare system integration to succeed, the interest of those involved, the need for extensive training of care workers, and the sensitive ethical issues that could arise if care bots were to identify mental health issues even if the person was unaware of what the bot thought now the author informs us of a research gap that would lead to reliable ways to find out if AI chatbots could cut the rates of depression and protect the data of patients we need a better look at how AI chatbots could join current health systems.

Voon et al. (2026) demonstrated the development of a human-centred chatbot that uses the Depression Anxiety and Stress Scale-21 (DASS-21) to support the prevention of mental health disorders. The chatbot was developed with input from mental health professionals and tested by practitioners and end users. According to the findings, the chatbot is a great tool for preventing mental illness. He et al. (2020) Certain aspects of the application, like the amount of information on designated pages and the application interface, were rated "Average" by a subset of college students. Because there may be knowledge-based issues with the data and the chatbot's general features, this rating indicates future directions for improvement.

Chen et al. (2024) proposed contemporary chatbots that focus on depression. They tend to be general and less personal, merely providing general advice or expressions of emotion. PsyChatbot aims to overcome such limitations by utilising a counselling dialogue framework based on CBT to help users reflect on their negative perceptions through retrieval-based question-and-answer algorithms. Most of the current psychological counselling systems for depressed people support only English speakers, which has left out a large population of depressed Chinese individuals who need improved responsiveness in mental health support. Therefore, PsyChatbot targets the Chinese population.

Heinz et al. (2024) introduced a novel therapy chatbot named Therabot, driven by AI, designed to address mental health issues. In a controlled RCT involving 210 subjects, Therabot showed decreases in symptoms of major depressive disorder (MDD). Users reported engagement and a robust therapeutic rapport, like sessions with human therapists, suggesting that AI chatbots could serve as a viable treatment option for depression. The study primarily examined Therabot's performance in MDD, Generalised Anxiety Disorder (GAD), and Chronic Health-Related Fatigue Disorder (CHR-FED) but did not assess its applicability to other psychological conditions.

Lee and Lee (2024) studied the use of chatbot-based interpretation bias modification (CBM-I) programs to increase cognitive biases associated with depression in early adulthood. Their findings showed that those receiving CBM-I treatment experienced considerably fewer depressive symptoms and better cognitive flexibility and attention control than the control group did. The results of this study, which covers the development of digital interventions for depression and the promotion of mental well-being, are very promising. But there is only evidence that the chatbot-driven CBM-I is working, and it is still not being used to determine the effect this change has on existing conditions.

Dharrao and Gite (2024) created the TherapyBot. TherapyBot is a transformer-based model trained for mental health conversations. TherapyBot offers a safe environment for those who are unwilling to talk about their struggles. TherapyBot has been trained on an open-domain and therapist-client model, enabling it to generate empathetic responses. TherapyBot has been shown to address mental health issues with a loss of 0.29 and a

perplexity of 1.34. The study also offers suggestions for future work. Future work involves developing larger models to achieve better outcomes, applying reinforcement learning for adaptive responses, and implementing generative adversarial networks (GANs) for chatbots. The authors also suggest implementing emotion recognition.

METHODOLOGY

Proposed Architecture

In this research, SKY* Chatbot, an advanced conversational AI system, provided emotional support to individuals facing emotional difficulties. SKY is built on the Pathways Language Model (PaLM) Architecture and can be easily accessed via a web interface. A second feature is the integration of emotion recognition with personalised content recommendation. Based on the emotional state identified from the user's text, the Interests Fetching API can suggest content such as music, videos, and blogs. The intention is not only to generate a conversational reply but also to provide additional content that may help the user remain engaged during the interaction.The chatbot uses eight transformer-based models for training and testing: BERT, RoBERTa, ALBERT, XLNet, DeBERTa, DistilBERT, DistilRoBERTa, and Electra, providing accurate, high-performance comprehension of user input and responses. Thus, SKY's contribution is not based on a single language model but on integrating emotion identification, conversational history, personalised recommendations, and a conversational interface into a single framework.

As Figure 1 shows, SKY has an Interests Fetching Application Programming Interface (API) that gives personalised recommendations for content, including music, videos, and blogs, based on the user's emotional state and interests. The motive of this feature is to uplift and engage the user during their interaction (Reis, 2024). The database stores user credentials, emotional states, previous responses, and preferences, which helps keep the user experience personalised and helps the chatbot understand the user better.

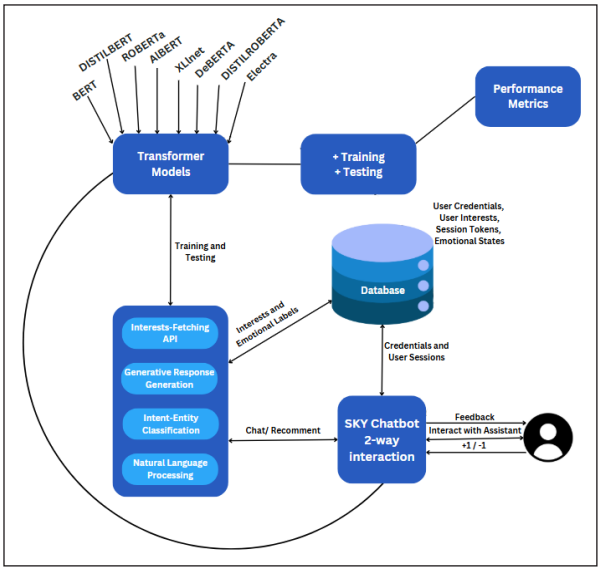


Figure 1. Architecture of the SKY chatbot

Motivation and Hypothesis of SKY Chatbot

The name SKY originates from the authors' hypothesis that the sky, with its beautiful sunrises, magical twilights, and countless stars, projects peace and wonder to anyone who looks at it. The goal is to make users feel the same comfort when interacting with the SKY Chatbot as they do when looking at the night or morning sky.

Scenario 1, as shown in Figure 2, depicts a person who begins with sadness and loneliness, burdened by emotional weight. He resorts to talking to the SKY Chatbot. Throughout, the mood shifts from sadness to happiness as you interact with SKY. He walks outside, standing under a clear, starry sky, where the stars bring hope and a sense of comfort and peace.

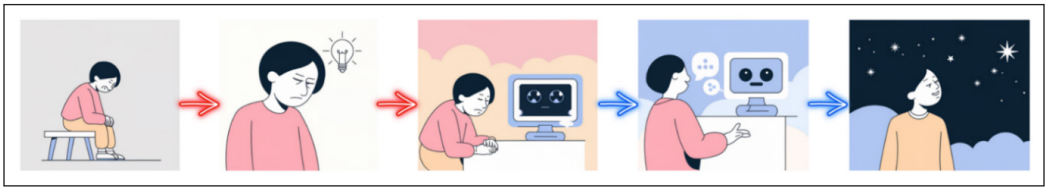


Figure 2. Scenario 1: A person using SKY Chatbot and feeling better

PaLM Model

A single, compelling model applies to several tasks and domains. Creating a new Pathways system to manage accelerator distributed computing was a big step in the right direction. Using the Pathways system, training a single model across several Tensor Processing Unit (TPU) v4 Pods was successful. PaLM is a Transformer model with just dense decoders and 540 billion parameters. Researchers evaluated PaLM using hundreds of language production and comprehension tasks and found that it often outperforms state-of-the-art few-shot performance.

PaLM demonstrates the first large-scale implementation of the Pathways system to expand training to 6144 chips, the most extensive TPU-based system configuration 31 used for training to date. The training is scaled via data parallelism at the Pod level, with traditional data and model parallelism within each Cloud TPU v4 Pod. This knowledge represents a significant increase in scale over most previous LLMs, which were either trained on a single TPU v3 Pod (GLaM, LaMDA, and more), used multiple TPU v3 Pods (Gopher) with a maximum scale of 4096 TPU v3 chips, or scaled to 2240 A100 graphics processing units (GPUs) across GPU clusters using pipeline parallelism (Megatron-Turing NLG). A mix of English and multilingual datasets, including GitHub code, novels, Wikipedia, discussions, and high-quality online articles, was used to train PaLM.

Transformer Models

Transformers are a revolutionary model architecture in NLP that operate independently using the self-attention mechanism introduced by Vaswani et al. (2017). Thus, they can process input data in parallel rather than sequentially (Zandieh et al., 2024). This design allows these architectures to capture long-range dependencies and relationships within the data better than any traditional recurrent neural network (RNN) or convolutional neural network (CNN) ever could. Achieving the highest scores across various tasks, particularly in machine translation, Transformers have loomed large, with reduced training time and computational costs and improved translation quality. To evaluate the SKY Chatbot implemented using the PaLM model, eight transformer models were widely used: BERT, ALBERT, RoBERTa, DeBERTa, DistilBERT, Electra, XLNet, and DistilRoBERTa.

Algorithm for Analysing SKY Chatbot Accuracy: A Mathematical Approach

To evaluate the accuracy of SKY Chatbot, each transformer model must follow the algorithm to obtain the best results:

Step I. Data Collection and Pre-Processing

- Collect conversation data*: D = {(u_i, c_i, t_i)},* where *u_i* is user input, *c_i* is chatbot response, and *t_i* is timestamp.
- Text Cleaning and tokenising data: *D′ = {((u_i)_t, (c_i)_t)}*, where *_t* denotes tokenisation.
- Split data: *D′ = {D_train, D_val, D_test}*.

Step II. Model Training and Evaluation

- Define a loss function *L* (e.g., cross-entropy for classification tasks).
- Initialise model parameters θ.
- For epoch in epochs:
- For *(u, c)* in *D_train*:
- Compute prediction $\hat{y}$ using the model with parameters θ.
- Calculate loss: loss = *L(ŷ, c)*.
- Update parameters: $\theta = \theta - \eta * \nabla_\theta$ loss, where η is learning rate.
- Evaluate on *D_val* and adjust hyperparameters if necessary.

Step III. Integrating Conversation History

a) Feature Engineering:

- Define a function *H* to extract historical features:
- *H(u_i, c_i, t_i) = (u_i-N, ..., u_i-1, c_i-N, ..., c_i-1)*, where *N* is the history window size.
- Augment training data: *D_train_hist = {(H(u_i, c_i, t_i), (u_i)_t, (c_i)_t)}*.

b) Model Adaptation:

- To modify the model for the functioning of the accept historical features as input: $\hat{y} = f(\theta, (u_i)_t, (c_i)_t, H(u_i, c_i, t_i))$.
- Re-train the model on D_train_hist.

Step IV. Iterative Refinement

a) Continuous Monitoring Evaluation:

- Periodically assess the model on various data sources: D_test_new.
- Calculate metrics: $M(D_test_new, \hat{y})$.

b) Model Enhancement:

- If $M(D_test_new, \hat{y})$ falls below a threshold:
- Collect more data: D_new.
- Re-train the model on $D_train \cup D_new$.
- Adjust hyperparameters or model architecture.

Reproducibility and Experimental Procedure

The experimental procedure consists of four main stages: data preparation, model training, conversation-history integration, and performance evaluation.

First, conversational data are collected as user input, chatbot responses, and corresponding interaction information. The text is cleaned and tokenised before being divided into training, validation, and test subsets.

During model training, the selected transformer architecture processes the textual input and generates the predicted class. Cross-entropy loss is used for classification, and the model parameters are updated iteratively during training. Performance on the validation set is monitored to assess the model's behaviour before final evaluation. Conversation history is incorporated by considering a window of previous user inputs and chatbot responses. These historical interactions provide contextual information that is combined with the current interaction. This enables the system to preserve short-term conversational continuity. After training, the models are evaluated using accuracy, precision, recall, and F1-score. Confusion matrices and emotion-wise results are also examined to identify differences in performance among emotion categories.

The reported implementation was executed on a system equipped with an NVIDIA GeForce RTX 4080 GPU and 32 GB RAM. Reporting the preprocessing procedure, data split, model configuration, evaluation measures, and computing environment improves the transparency of the experimental setup and supports the study's reproducibility.

RESULTS AND DISCUSSIONS

The whole SKY Chatbot has been implemented as a system with hardware requirements that include a GeForce RTX 4080 GPU and 32 GB of RAM. The SKY Chatbot integrates with a web interface, making it easily accessible to everyone.

The SKY Chatbot and User Interaction

Figure 3 shows the home page of the SKY Chatbot. Figures 3 and 4 show the interface and user interaction. SKY initially welcomes the users with a greeting message, after which the conversation begins. Users can chat with the chatbot about their emotional weights, and SKY will help them lift them off using its trained ability to understand emotions. Users can chat with the chatbot about their emotional weights, and SKY will help will help them lift them off using its trained ability to understand emotions as shown in Figure 5.

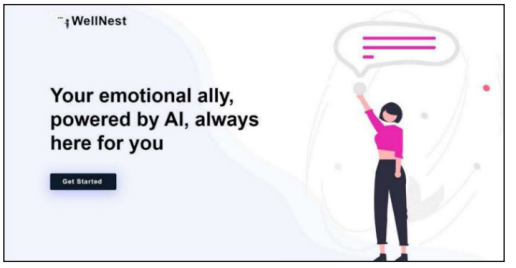


Figure 3. Home page of the SKY Chatbot

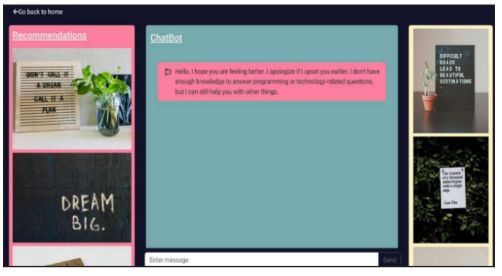


Figure 4. SKY Chatbot page

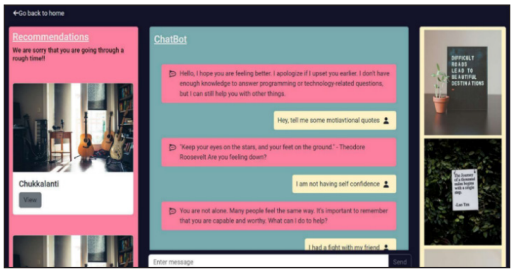


Figure 5. SKY Chatbot responses

Accuracy, F1-Score, Precision, and Recall Evaluation Metrics of All Models

Table 1 represents the metrics: the Accuracies, Precisions, Recalls, and F1-Scores, of all eight Transformer models and the number of epochs each model went through.

DeBERTa clearly outshines all other models, delivering the highest performance across accuracy (0.937), F1-score (0.935), precision (0.9384), and recall (0.937). BERT and ALBERT trail closely, with BERT slightly better at precision (0.9335) but ALBERT showing a more balanced F1-score (0.9323). In contrast, lighter models like DistilBERT and Electra offer efficiency but fall short in performance, with accuracy, F1-scores, precision, and recall all hovering around 0.926 to 0.928. This comparison highlights that while smaller models are resource-efficient, DeBERTa's robust precision and recall make it the clear winner.

Table 1
Summary of statistics of all models

Model	Epochs	Accuracy	Precision	Recall	F1-Score
DistilBERT	6	0.9275	0.926967704	0.9275	0.926177239
DistilRoBERTa	7	0.9305	0.930206809	0.9305	0.928776377
BERT	5	0.933	0.933546594	0.933	0.932790438
RoBERTa	9	0.929	0.929603594	0.929	0.927191908
ALBERT	10	0.9325	0.933316551	0.9325	0.932296927
XLNet	7	0.93	0.932712232	0.93	0.93008278
DeBERTa	8	0.937	0.938377823	0.937	0.934958303
Electra	7	0.9265	0.927955736	0.9265	0.924947401

F1-Score, Precision, and Recall Across Classes

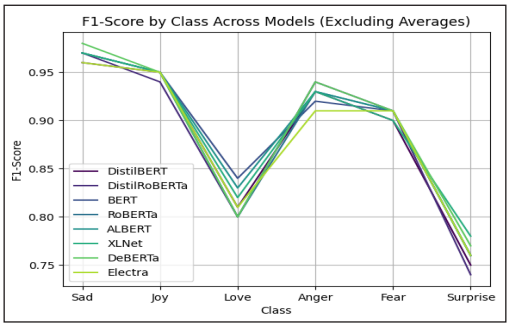


Figure 6. F1-Score per emotion

As shown in Figures 6, 7 and 8, DeBERTa and ALBERT regularly top the list in emotion classification for emotions such as Joy, Fear and Surprise, achieving strong F1-scores, precision and recall. Joy, which is the simplest to detect due to its linguistic signals, achieves the best performance across all models, whereas Love falls short, being the toughest emotion to classify correctly. DeBERTa stands out notably in precision for Fear, minimising positives, and in recall, for Surprise, successfully identifying even subtle occurrences. While DistilBERT and Electra struggle with precision and recall for emotions like Love and Fear, DeBERTa's ability to generalise well, even with imbalanced data, makes it the most reliable model overall.

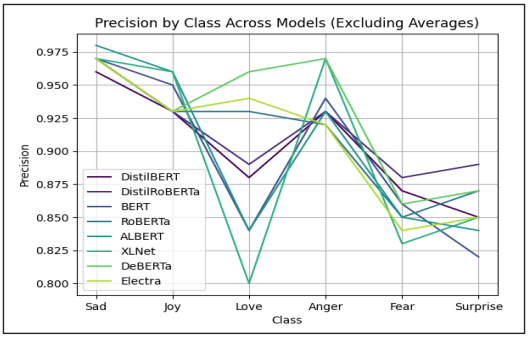


Figure 7. Precision per emotion

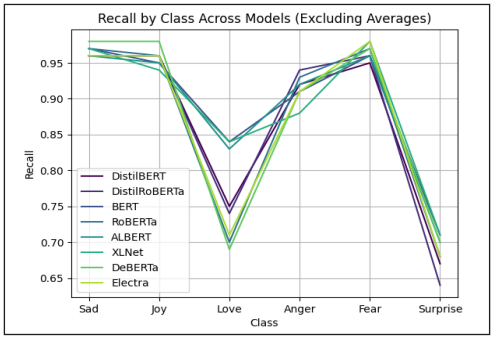


Figure 8. Recall per emotion

Macro and Weighted F1-Score, Precision, and Recall Across All Models

Macro Average F1-Score, Precision, and Recall metrics measure the average performance across all emotional classifications. Each category is treated as equally important, regardless of its frequency in the dataset. Weighted Average F1-Score, Precision, and Recall metrics calculate the mean performance across different classifications while adjusting for the frequency with which instances appear in each category. This method addresses class imbalances by prioritising classes that are more frequently represented.

Figures 9, 10, and 11 show that the transformer model with the best macro scores is XLNet. The transformer model with the best weighted scores is DeBERTa.

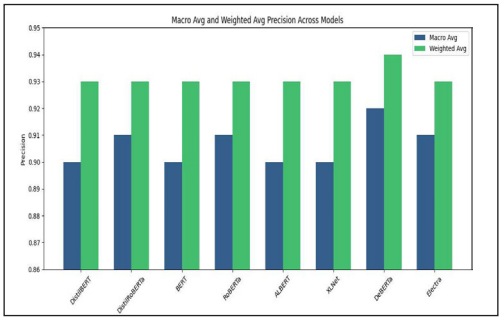


Figure 9. Macro average and weighted average Precision for transformer models

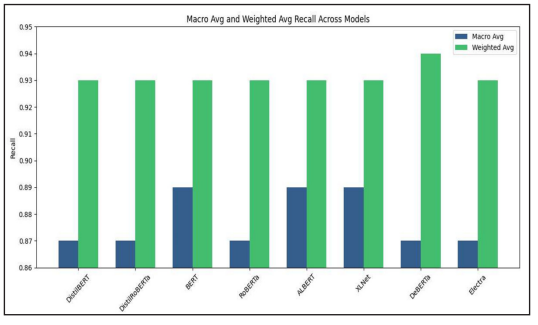


Figure 10. Macro average and weighted average Recall scores for transformer models

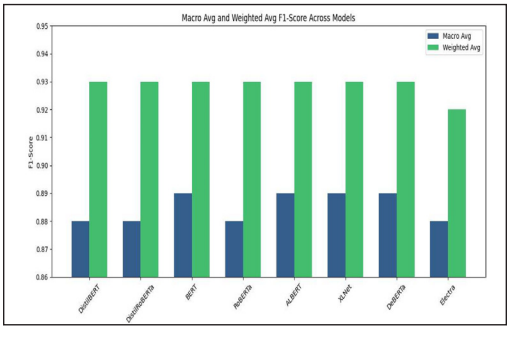


Figure 11. Macro average and weighted average F1-Scores for transformer models

BERT

Based on the information from Devlin et al. (2019), BERT's power is attributed to the bidirectional design that enables it to get the context not only from the left side but also from the right, whereas the traditional models are unidirectional. The incorporation of MLM during its pretraining phase enables it to develop a deep and quite complex understanding of the language, which is the main reason it is very effective at differentiating the emotions of Joy and Sadness. Nevertheless, as a matter of fact, BERT, as a machine learning model in

Figure 12, is less able to identify the emotions of Love and Surprise, which is probably due to the subtle and overlapping nature of these linguistic cues. Here, in Figure 13, it is evident that the entire procedure of its training is heavy on the resources and takes a longer time to converge as compared to the likes of the lightweight model DistilBERT, though the outcome in terms of high accuracy and strong F1-scores in most categories is what really counts. Slight overfitting can be noticed through the validation loss, which is somewhat up and down.

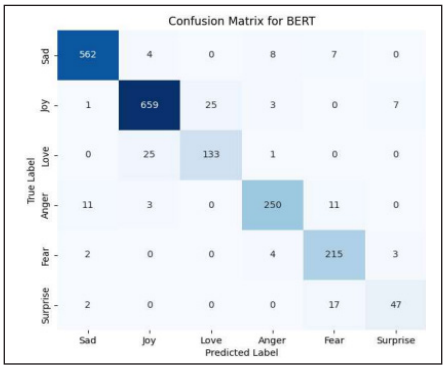


Figure 12. Confusion matrix for BERT

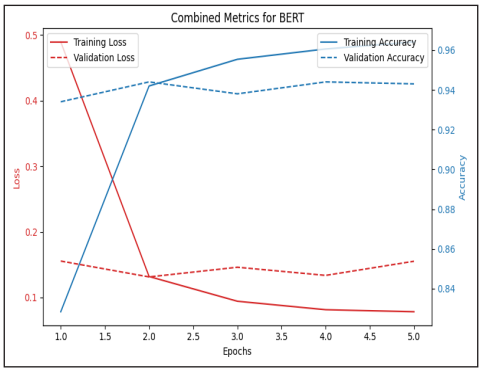


Figure 13. Epochs metrics of BERT

ALBERT

As per Lan et al. (2019), ALBERT stands out for its efficient architecture. As shown in Figure 14, ALBERT excels at classifying emotions such as Joy and Sadness, achieving high precision and recall. Its training process is efficient, with rapid improvements in accuracy and early loss stabilisation, indicating strong generalisation capabilities, as demonstrated in Figure 15. Although ALBERT struggles with classifying more ambiguous emotions like Love, its weighted and macro average F1-scores are among the highest, reflecting consistent performance across emotional categories.

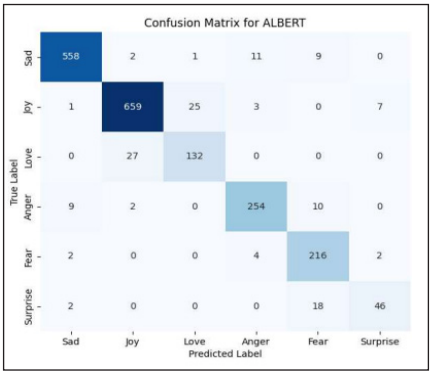


Figure 14. Confusion matrix for ALBERT

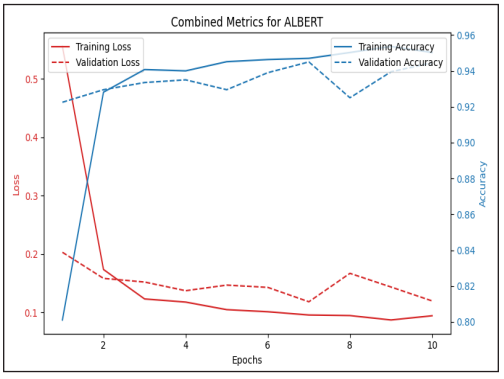


Figure 15. Epochs metrics of ALBERT

RoBERTa

According to Liu et al. (2019), RoBERTa enhances the BERT framework through refining training techniques. Expanding datasets demonstrate that BERT's initial training was suboptimal and could be outperformed by better training approaches. By removing limitations such as Next Sentence Prediction (NSP) and augmenting the training data, RoBERTa attains top-tier performance on benchmarks including GLUE, RACE and SQuAD. Although it performs admirably in classifying emotions such as Joy, Fear, and Surprise, it has difficulty with Love, showing 48 misclassifications, as depicted in Figure 16. RoBERTa demonstrates efficient training by rapidly attaining peak validation accuracy, yet it exhibits overfitting after the fifth epoch, as shown in Figure 17. Its F1-scores remain very competitive, surpassing BERT in categories, especially in accurately identifying complex emotions like Fear, with reduced false positives. Despite occasional inconsistencies in validation loss, RoBERTa demonstrates robust generalisation across diverse emotional categories, effectively balancing accuracy and efficiency.

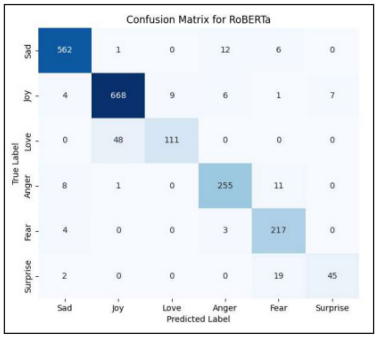


Figure 16. Confusion matrix for RoBERTa

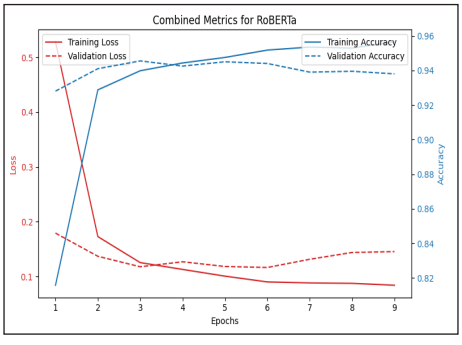


Figure 17. Epochs metrics of RoBERTa

DeBERTa

As He et al. (2020) noted, DeBERTa improves on earlier models such as BERT and RoBERTa with two main innovations: a disentangled attention mechanism and a better masked decoder. By using two separate vectors for each token to distinguish semantic meaning from positional information, DeBERTa enhances attention calculations by accounting for both content and location. Its improved mask decoder includes absolute positions, which helps with masked token predictions during pretraining. With the highest macro and weighted-average F1-Scores, DeBERTa sets a standard for emotion classification, demonstrating impressive precision and recall, particularly for more difficult emotions such as Fear and Surprise. With the highest macro and weighted - average F1-Scores as shown in Figure 18, DeBERTa sets a standard for emotion classification as shown in Figure 19, demonstrating impressive precision and recall, particularly for more difficult emotions such as fear and surprise.

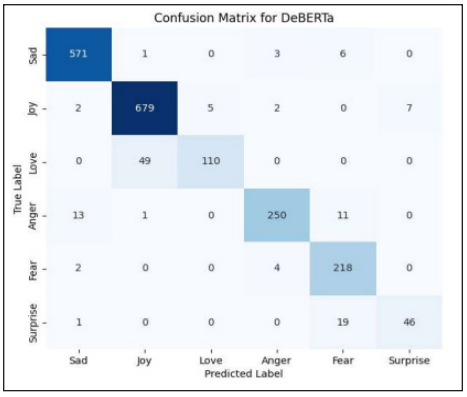


Figure 18. Confusion matrix for DeBERTa

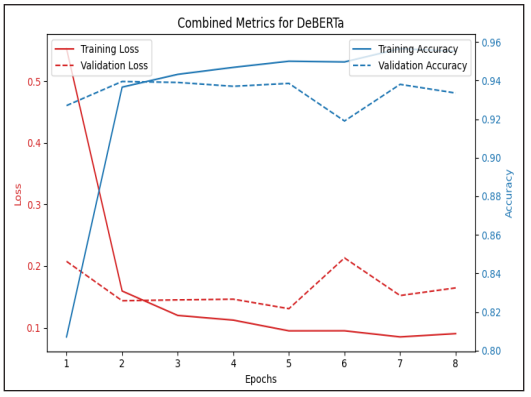


Figure 19. Epochs metrics of DeBERTa

DistilBERT

As Figure 20 illustrates, DistilBERT performs exceptionally well in classifying Joy (668 true positives). It is a very effective option for computationally limited tasks because, as shown in Figure 21, its training process is faster, with a sharp decline in training loss and a quicker plateau in validation accuracy. DistilBERT achieves impressive F1-scores and performs well across most models despite its smaller size. However, it is not as effective as larger models such as DeBERTa, particularly in handling emotional categories like Love and Fear. DistilBERT's recall for uncommon emotions like surprise is lower, underscoring its limitations with underrepresented emotions, but it maintains a relatively high precision, minimising false positives in Joy and Sadness.

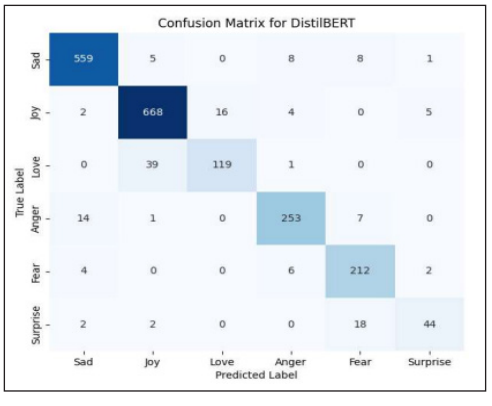


Figure 20. Confusion matrix for DistilBERT

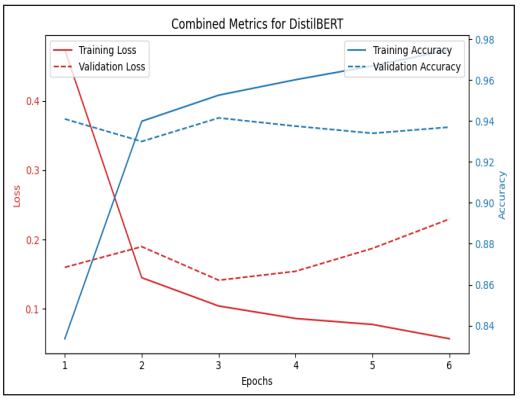


Figure 21. Epochs metrics of DistilBERT

Electra

According to Clark et al. (2020), ELECTRA changes the game by introducing a novel transformer architecture that uses both a generator and a discriminator, thereby forming a dual-transformer model for pretraining. In a masked language modelling (MLM) setting, the generator replaces tokens in the sequence, whereas the discriminator identifies which tokens have been replaced; thus, it performs a replacement-token detection task. The new method is more productive in training than the standard MLM because training proceeds more quickly. As a matter of fact, Figure 22 reveals that ELECTRA is very good at Joy prediction (668 true positives), but it confuses Love (44 misclassifications) and occasionally confuses Surprise (45 false positives). The model learning fast during the initial training epochs is depicted in Figure 23, where high training accuracy was achieved. However, the model exhibits poor generalisation, as evidenced by a high validation loss and an early plateau in validation accuracy. Despite the model's training technique being quite beneficial, ELECTRA's macro-average F1 scores are lower than those of other models. It is strong in Joy and Sadness but weak in rare emotions, and its low precision in misclassifying surprise and love is its major problem. Its recall is good but not as strong as the other models, suggesting it lacks the ability to detect subtle emotional cues.

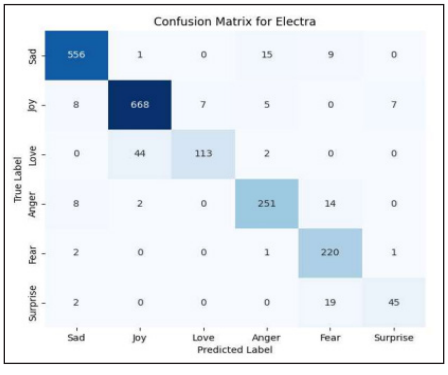


Figure 22. Confusion matrix for Electra

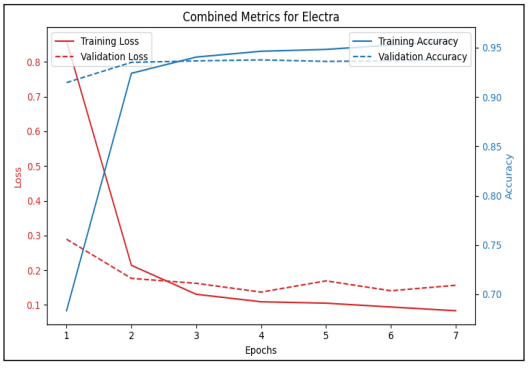


Figure 23. Epochs metrics of Electra

DistilRoBERTa

As discussed by Sanh et al. (2019) DistilRoBERTa, a compressed variant of the RoBERTa model, achieves a smaller footprint and lower complexity via knowledge distillation while preserving robust performance across several natural language understanding tasks, particularly in real-time scenarios such as text classification, sentiment detection, and content moderation. Its confusion matrix in Figure 24 highlights accuracy for Joy (669 true positives) and a few errors, for Sadness (6 misclassifications) as shown in Figure 25.

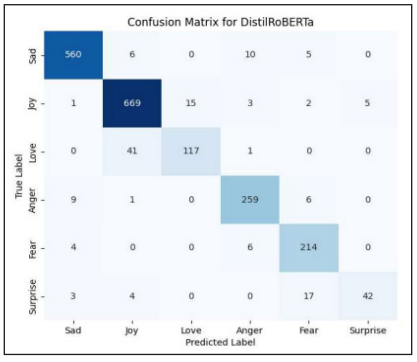


Figure 24. Confusion matrix for DistilRoBERTa

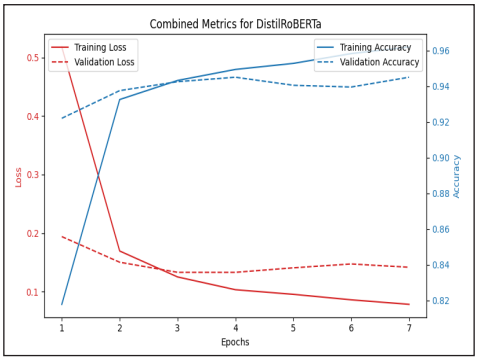


Figure 25. Epochs metrics of DistilRoBERTa

XLNet

By adding a permutation-based autoregressive pretraining technique, XLNet overcomes some of BERT's shortcomings by learning bidirectional context and improving its modelling of token dependencies. As seen in Figure 26, XLNet struggles to correctly classify Love and occasionally misclassifies Fear and Surprise, but it is excellent at classifying Sadness and Joy. Figure 27 shows that XLNet is efficient during training, but there is some instability in validation loss and accuracy, indicating potential difficulties in generalising to new data. Despite these variations, it performs better than BERT and ALBERT. Even though XLNet outperforms its predecessors by a small margin, further tweaking may be necessary to maximise its performance, especially for less common emotional categories such as Surprise and Love.

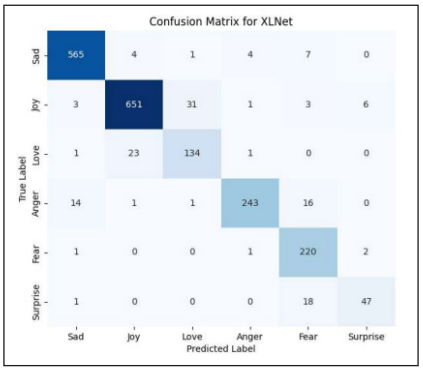


Figure 26. Confusion matrix for XLNet

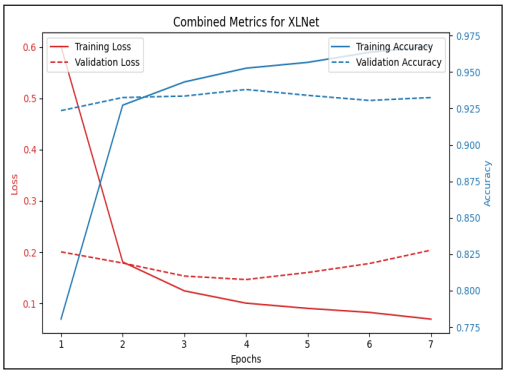


Figure 27. Epochs metrics of XLNet

Comparison of Metrics Across Each Class

Figures 28, 29, 30, 31, 32 and 33 compare Precision, Recall, and F1-Scores for all eight transformer models of the emotional categories Anger, Fear, Joy, Love, Surprise, and Sadness, respectively.

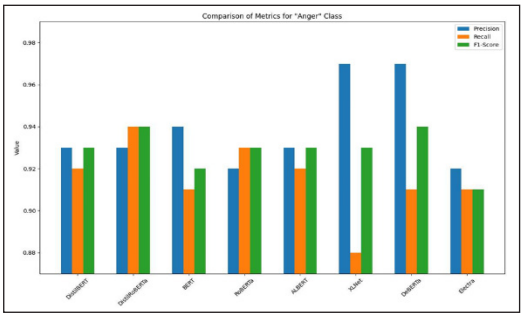


Figure 28. Scores for Anger emotion

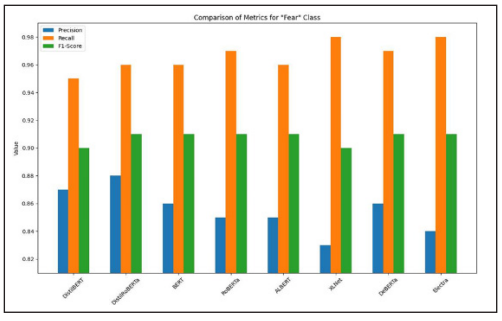


Figure 29. Scores for Fear emotion

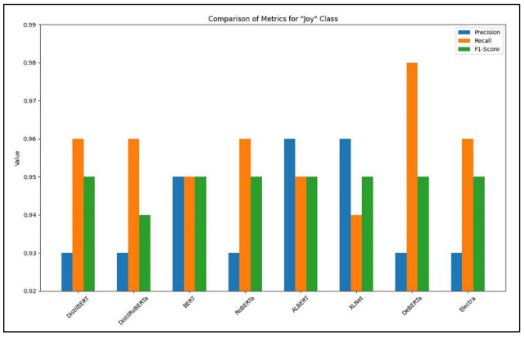


Figure 30. Scores for Joy emotion

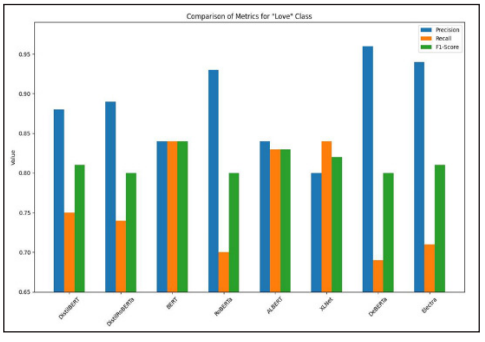


Figure 31. Scores for Love emotion

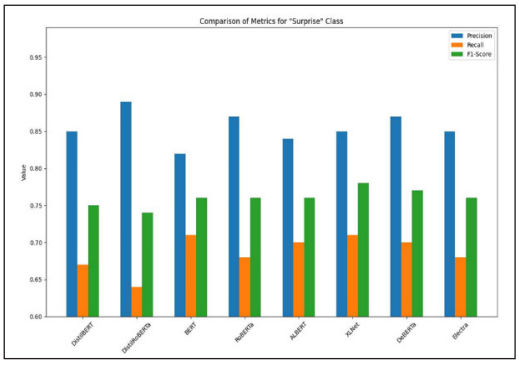


Figure 32. Scores for Surprise emotion

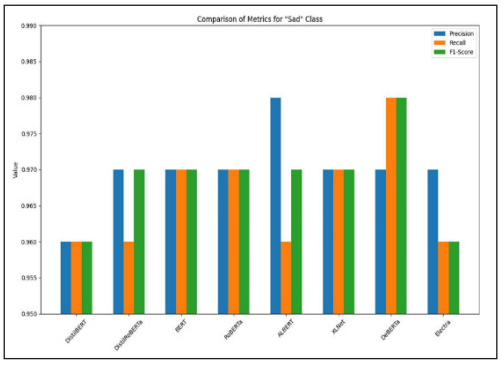


Figure 33. Scores for Sad emotion

DistilBERT and DistilRoBERTa both performed well, especially in recognising Sad and Fear emotions, with DistilRoBERTa being a bit more consistent overall. BERT did great with Joy and Surprise, while RoBERTa was slightly better at Joy but struggled with Sad and Fear. ALBERT was decent across the board but had trouble with Anger and Surprise. XLNet really stood out for Anger, Fear, and Surprise, outperforming the others in those areas. DeBERTa was the most reliable overall, especially strong in Love and Fear, with top precision for Love and Surprise—though it missed some instances in those categories.

Electra had the weakest performance, particularly with Sad and Anger, but did okay with Fear and Surprise. Overall, DeBERTa and XLNet were the most consistent and effective models, while Electra fell behind.

DISCUSSION

Formally, depression and mental issues connected to the mind are discussed. The emergence of chatbots in the current scenario is illustrated. The SKY Chatbot was developed using the PALM model to enable continuous, chat-based interactions that build on previous conversations. To evaluate the model (chatbot), various AI-integrated Deep Learning transformers were used for performance analysis.

The performance metrics, including Accuracy, F1-score, Precision, and Recall, of multiple models are illustrated in the paper. DeBERTa demonstrates the model's supremacy on text classification tasks. Accuracy was achieved by models such as ALBERT and BERT. Nevertheless, DeBERTa consistently outperformed every other model on every metric. This finding suggests that DeBERTa might be the ideal model in circumstances where obtaining a higher percentage of true positives is crucial. Even though they perform relatively poorly, lightweight models like DistilBERT and Electra can be used in situations where the main goal is efficient resource utilisation and minor trade-offs on accuracy and other metrics are acceptable. It becomes clear how model size, training time, and performance are balanced. While lightweight models are better suited for tasks where resource utilisation is the top priority, more comprehensive models, such as ALBERT and DeBERTa, are better at generalisation.

The models under consideration in this study mostly show good results, with Joy being the easiest emotion to predict by all the platforms. Sadness, similarly, shows good results; for example, DeBERTa, BERT, and ALBERT correctly identify most Sadness instances. On the other hand, Love is the most difficult category for the models to identify, as many models mislabel a high percentage of actual Love instances as other emotions, e.g., joy or Anger. It is worth mentioning that, in general, Fear and Surprise have fewer classification issues across models, suggesting that these models are better at differentiating these emotions from others. While all models perform well, DeBERTa stands out a bit more in terms of its ability to classify all emotional categories with a relatively lower number of errors, especially in the more complex category of Surprise.

In conclusion, DeBERTa and ALBERT are recommended for large-scale emotion classification tasks, given their ability to achieve high cross-performance across all metrics. In the future, where computing resources are limited, DistilBERT and DistilRoBERTa can strike a helpful trade-off between computational speed and accuracy. The results indicate that DeBERTa achieved the best overall performance among all models, especially in the Love and Fear categories. At the same time, XLNet ranked relatively high in distinct

dimensions, particularly in terms of Anger. The best models in the Surprise and Joy categories were BERT and DistilBERT. However, the ones that worked are more heavily covered in categories such as Electra and have been quite limited, thereby hinting that higher-performing models, such as DeBERTa and XLNet, might have a penchant for the finer classification in sentiment tasks.

DeBERTa and ALBERT are recommended for large-scale emotion classification tasks, given their ability to achieve high cross-performance across all metrics. In the future, where computing resources are limited, DistilBERT and DistilRoBERTa can strike a helpful trade-off between computational speed and accuracy.

The following is the list of solutions for research Questions (RQ) addressed in the paper.

RQ1 : The mention of using various transformer models suggests that the SKY Emotional Chatbot is equipped to analyse emotions and sentiments effectively. The high precision (97%) indicates that the chatbot can accurately assess emotional states, addressing this research question.

RQ2 : Ethical concerns can be addressed within personalised interventions. This requires more information on privacy, data usage, and the ethical guidelines used in the chatbot's design. It is a challenging task.

RQ3 : Using transformer models like BERT and its variants shows a strong ability to process unstructured text data. These models are built to handle linguistic nuances, enabling the chatbot to effectively manage unstructured conversational data.

RQ4 : The proposed chatbot can refer to previous conversations, indicating that methods are in place to improve continuity and contextual awareness. This suggests a mathematical or algorithmic approach to modelling past interactions for coherent dialogue.

FUTURE SCOPE

Several interesting research sources on sentiment and emotion classification are currently being investigated using transformer architectures. Given some of these transformer models fine-tuned on specialised datasets, an interesting direction of research is to examine the effects of fine-tuning. A larger extent of how these models can generalise across a wider spectrum of emotional states is to make the model appropriate to use on individual emotion datasets or to include additional emotions. Such fine-tuning may improve performance in specific emotion categories that are harder to determine and would provide much more insight into the model’s capabilities. This could help improve performance in contexts such as customer feedback, social media interactions, or customer service experiences, where contextual details and subtle tone are important.

Text, audio, and video are all included in multimodal data integration, which opens new possibilities for chatbots to develop emotional intelligence. To gain a deeper understanding of the user's emotional state, the chatbot can employ multimodal deep learning models and feature fusion. For example, the system can identify subtle emotional cues from text, facial expressions, and vocal tone to provide more accurate and sympathetic responses. These enhanced perceptions will not only make it easier for the chatbot to recognise emotions but also enable it to adapt in complex scenarios where a single input might be ambiguous or insufficient in context.

Incorporating novel techniques in mathematical modelling, specifically aimed at tensor-based representations, could significantly improve model performance. Certain tensor methods are effective at capturing interactions among multiple representations, while the modal attention mechanism can focus on the most pertinent information for context-based emotional recognition. This technological advancement could enable chatbots to respond to users consistently and accurately. In addition, dynamic attention to different modalities will enable SKY to function in various environments as an articulate emotional support system.

The ethical problems with the subject's applications become increasingly important as systems improve at identifying emotions. Better, more accurate models may make it easier for people to exploit weaknesses or manipulate users, which calls for close supervision. Maintaining trust and avoiding overly intrusive emotional interventions present ethical challenges. The systems should incorporate precise consent mechanisms to inform users that their emotional data is being used, given these ethical concerns. To maintain the emotional interventions' non-intrusiveness and respect, oversight mechanisms should ensure transparency in their execution.

Finally, exploring ways to minimise problems such as overfitting and underfitting in low-data environments is an important area for improvement. Through ways such as knowledge distillation, adaptive loss functions, and pre-training strategies, we can better generalise emotional categories. This allows smaller models, such as DistilBERT, to significantly improve efficiency and accuracy, creating more responsive and robust emotional recognition models for real-time applications.

Therefore, future research should focus on fine-tuning these transformer-based models. They offer great opportunities to integrate multimodal data, adopt hybrid approaches, and apply specific methods to address various discrepancies. As emotion recognition continues to develop, these efforts could lead to more reliable, scalable, and practical research applications.

CONCLUSION

These findings demonstrate a thorough examination of transformer-based models using a range of performance metrics. With consistent scores in Accuracy, F1-Score, Precision, and

Recall, these results demonstrate that DeBERTa is the most adaptable model. However, models like DistilBERT or Electra can be considered depending on task requirements (e.g., resource constraints or the need to achieve better results).

DeBERTa demonstrated superiority across all emotions and confusion-matrix metrics, even though all models performed well. However, Love remains a major challenge for all models, and future work should focus on methods to improve classification accuracy for these emotionally complex categories.

Emotion classification can be done well by transformer-based models such as DeBERTa and ALBERT, which are able to show very high performance in complicated categories like Fear and Surprise. Nevertheless, issues remain, particularly in the classification of Love, a category in which all models have struggled to achieve higher F1-Score, Precision, and Recall scores. Although complex models like DeBERTa and ALBERT have generally yielded better results, lighter models such as DistilBERT and Electra have incurred efficiency-accuracy trade-offs, making them suitable for situations where considerable resource consumption is required.

Studying just about each of the eight different transformer-based models for emotion classification, this research has essentially figured out their comparison criteria to be F1-Score, Precision, and Recall. The heatmap representations of these statistics are quite informative, as they show how the different models align with the various emotion classes in terms of their strengths and weaknesses. DeBERTa and ALBERT, as usual, are the top two performers, whereas DistilBERT and DistilRoBERTa can be considered good choices for the task of efficient performance to some extent.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to VNR Vignana Jyothi Institute of Engineering and Technology for providing a supportive academic environment. The authors also extend their appreciation to everyone who contributed guidance, encouragement, and support throughout the development and completion of this research.

REFERENCES

Agrawal, R., Monteiro, K., Narasimhamurti, N. S., Sharma, S., Suryawanshi, A., Bariya, A., Narvekar, S., Bagnoli, L., Saxena, M., Magoun, L., Parsekar, S. S., Pozuelo, J. R., Lesh, N., Sood, M., Sharma, T., Yadav, H., Bhan, A., Nadkarni, A., & Patel, V. (2025). A conversational agent (PracticePal) to support the delivery of a brief behavioral activation treatment for depression in rural India: Development and pilot-testing study. *JMIR Formative Research, 9*, Article e73563. https://doi.org/10.2196/73563

Alsabi, R. N. S., Zaimi, A. F., Sivalingam, T., Ishak, N. N., Alimuddin, A. S., Dasrilsyah, R. A., Basri, N. I., & Jamil, A. A. M. (2022). Improving knowledge, attitudes and beliefs: A cross-sectional study of postpartum depression awareness among social support networks during COVID-19 pandemic in Malaysia. *BMC*

Women's Health, 22(1), Article 221. https://doi.org/10.1186/s12905-022-01795-x

Anisha, S. A., Sen, A., & Bain, C. (2024). Evaluating the potential and pitfalls of AI-powered conversational agents as humanlike virtual health carers in the remote management of noncommunicable diseases: Scoping review. *Journal of Medical Internet Research, 26*, Article e56114. https://doi.org/10.2196/56114

Bartlett, C., Littlewood, A., McCool, R., Mak, C., & Chen, S. Y. (2024). EE546 postpartum depression: Systematic literature reviews of economic evaluations of pharmacological and non-pharmacological treatments, utilities, costs, and healthcare resource use [Conference abstract]. *Value in Health, 27*(12), S163. https://doi.org/10.1016/j.jval.2024.10.826

Chen, T., Shen, Y., Chen, X., & Zhang, L. (2024). PsyChatbot: A psychological counseling agent towards depressed Chinese population based on cognitive behavioural therapy. *ACM Transactions on Asian and Low-Resource Language Information Processing*. Advance online publication. https://doi.org/10.1145/3676962

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). *ELECTRA: Pre-training text encoders as discriminators rather than generators* [Preprint]. arXiv. https://arxiv.org/abs/2003.10555v1

Colita, C.-I., Hermann, D. M., Filfan, M., Colita, D., Doepnner, T. R., Tica, O., Glavan, D., & Popa-Wagner, A. (2024). Optimizing chronotherapy in psychiatric care: The impact of circadian rhythms on medication timing and efficacy. *Clocks & Sleep, 6*(4), 635-655. https://doi.org/10.3390/clockssleep6040043

Denecke, K., & Gabarron, E. (2024). The ethical aspects of integrating sentiment and emotion analysis in chatbots for depression intervention. *Frontiers in Psychiatry, 15*, Article 1462083. https://doi.org/10.3389/fpsyt.2024.1462083

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dharrao, D., & Gite, S. (2024). TherapyBot: A chatbot for mental well-being using transformers. *International Journal of Advances in Applied Sciences, 13*(1), 1-12. https://doi.org/10.11591/ijaas.v13.i1.pp1-12

Fan, H., & Ma, S. (2024). Editorial: Effects of exercise intervention on emotional health and emotional processing: Evidence related to psychology and physiology. *Frontiers in Psychology, 15*, Article 1474722. https://doi.org/10.3389/fpsyg.2024.1474722

Fan, J. (2024). *AI psychotherapists and positive psychology: A systematic literature review of current literature and future directions* [Master's thesis, Lappeenranta–Lahti University of Technology LUT]. LUTPub. https://lutpub.lut.fi/handle/10024/168817

Feng, X., Ma, X., Li, J., Zhou, Q., Liu, Y., Song, J., Liu, J., Situ, Q., Wang, L., Zhang, J., & Lin, F. (2025). Inflammatory pathogenesis of post-stroke depression. *Aging and Disease, 16*(1), 209-238. https://doi.org/10.14336/ad.2024.0203

Ferrario, A., Sedlakova, J., & Trachsel, M. (2024). The role of humanization and robustness of large language models in conversational artificial intelligence for individuals with depression: A critical analysis. *JMIR*

Mental Health, 11, Article e56569. https://doi.org/10.2196/56569

Fouad, M. A., Tadros, M. G., & Michel, H. E. (2024). A comprehensive review of the pathophysiology of depression. *Archives of Pharmaceutical Sciences Ain Shams University, 8*(1), 122-132. https://doi.org/10.21608/aps.2024.271710.1160

Guo, Z., Lai, A., Deng, Z., & Li, K. (2024). Evaluating the feasibility and acceptability of a GPT-Based chatbot for depression screening: A mixed-methods study. In X. Xie, G. Powathil, I. Styles, & M. Ceccarelli (Eds.), *Artificial intelligence in healthcare* (Lecture Notes in Computer Science, Vol. 14975, pp. 249-263). Springer. https://doi.org/10.1007/978-3-031-67278-1_20

He, P., Liu, X., Gao, J., & Chen, W. (2020). *DeBERTa: Decoding-enhanced BERT with disentangled attention* [Preprint]. arXiv. https://arxiv.org/abs/2006.03654v1

Heinz, M. V., Mackin, D., Trudeau, B., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A., Griffin, T., & Jacobson, N. C. (2024). *Evaluating Therabot: A randomized control trial investigating the feasibility and effectiveness of a generative AI therapy chatbot for depression, anxiety, and eating disorder symptom treatment* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/pjqmr

Kang, B., & Hong, M. (2025). Development and evaluation of a mental health chatbot using ChatGPT 4.0: Mixed methods user experience study with Korean users. *JMIR Medical Informatics, 13*, Article e63538. https://doi.org/10.2196/63538

Khan, H., & Bokhari, S. F. H. (2024). Integrating artificial intelligence (AI) chatbots for depression management: A new frontier in primary care. *Cureus, 16*(8), Article e66857. https://doi.org/10.7759/cureus.66857

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *ALBERT: A lite BERT for self-supervised learning of language representations* [Preprint]. arXiv. https://arxiv.org/abs/1909.11942v1

Lee, J., & Lee, D. (2024). The effects of a chatbot-based interpretation bias modification on early adulthood depression [Conference abstract]. *European Psychiatry, 67*(S1), S532. https://doi.org/10.1192/j.eurpsy.2024.1105

Lei, X., Wu, H., Deng, Z., & Ye, Q. (2023). Self-disclosure, social support and postpartum depressive mood in online social networks: A social penetration theory perspective. *Information Technology & People, 36*(1), 433-453. https://doi.org/10.1108/itp-12-2020-0825

Li, J., Li, Y., Mei, X., He, H., Yu, L., Ho, G. W. K., & Chan, E. A. (2024). Effectiveness of cognitive-based interventions on psychological distress in adolescents with physical disabilities: A systematic review and meta-analysis. *International Journal of Mental Health Nursing, 33*(6), 1729-1742. https://doi.org/10.1111/inm.13391

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach* [Preprint]. arXiv. https://arxiv.org/abs/1907.11692v1

Manole, A., Cârciumaru, R., Brînzaș, R., & Manole, F. (2024). Harnessing AI in anxiety management: A chatbot-based intervention for personalized mental health support. *Information, 15*(12), Article 768. https://doi.org/10.3390/info15120768

Mavranezouli, I., Megnin-Viggars, O., Pedder, H., Welton, N. J., Dias, S., Watkins, E., Nixon, N., Daly, C.

H., Keeney, E., Eadon, H., Caldwell, D. M., O'Donoghue, K. J.-M., Stockton, S., Arnold, S., Thomas, J., Kapur, N., & Pilling, S. (2024). A systematic review and network meta-analysis of psychological, psychosocial, pharmacological, physical and combined treatments for adults with a new episode of depression. *eClinicalMedicine, 75*, Article 102780. https://doi.org/10.1016/j.eclinm.2024.102780

Moschny, N. K. (2020). *Epigenetics in psychiatry: Characterization of the immune system and its methylome in depressed patients receiving electroconvulsive therapy* [Doctoral dissertation, University of Veterinary Medicine Hannover]. https://nbn-resolving.org/urn:nbn:de:gbv:95-114539

Nakamura, A., Sutter-Dallay, A.-L., El-Khoury Lesueur, F., Thierry, X., Gressier, F., Melchior, M., & van der Waerden, J. (2020). Informal and formal social support during pregnancy and joint maternal and paternal postnatal depression: Data from the French representative ELFE cohort study. *International Journal of Social Psychiatry, 66*(5), 431-441. https://doi.org/10.1177/0020764020911409

Norazman, C. W., & Lee, L. K. (2024). The influence of social support in the prevention and treatment of postpartum depression: An intervention-based narrative review. *Women's Health, 20*, Article 17455057241275587. https://doi.org/10.1177/17455057241275587

Piro, A., Curci, A., & Mangiulli, I. (2024). Detection of cognitive-emotional feedback in interviewer-interviewee interaction [Conference abstract]. *Preliminary Reports and Negative Results in Life Science and Humanities, 1*(2, Suppl.), 170. https://cab.unime.it/journals/index.php/PRNR/article/viewFile/4892/pdf

Reis, P. M. (2024). *Systematic review on psychotherapy in the context of psychedelic therapy* [Master's thesis, University of Porto]. https://doi.org/10.34626/bj3a-rw68

Ren, S., Guo, Z., Zhang, J., He, Y., Sun, Z., & Yang, J. (2025). Transcriptomic alterations in the hippocampus and prefrontal cortex of rats with chronic unpredictable stress induced by low-intensity pulsed ultrasound. *Molecular Neurobiology, 62*(5), 5917-5929. https://doi.org/10.1007/s12035-024-04656-w

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* [Preprint]. arXiv. https://arxiv.org/abs/1910.01108v1

Shen, Y., Zou, X., & Hu, G. (2024). Editorial: Digital technology in the management and prevention of diabetes. *Frontiers in Endocrinology, 15*, Article 1511236. https://doi.org/10.3389/fendo.2024.1511236

Tarai, S., Mukherjee, R., Gupta, S., Rizvanov, A. A., Palotás, A., Chandrasekhar Pammi, V. S., & Bit, A. (2019). Influence of pharmacological and epigenetic factors to suppress neurotrophic factors and enhance neural plasticity in stress and mood disorders. *Cognitive Neurodynamics, 13*(3), 219-237. https://doi.org/10.1007/s11571-019-09522-3

Taylor, B. L., Nath, S., Sokolova, A. Y., Lewis, G., Howard, L. M., Johnson, S., & Sweeney, A. (2022). The relationship between social support in pregnancy and postnatal depression. *Social Psychiatry and Psychiatric Epidemiology, 57*(7), 1435-1444. https://doi.org/10.1007/s00127-022-02269-z

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 5998-6008). Curran Associates. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Volz, F., Wolf, K., Fung, C., Carroll, I., Lahmann, C., Lützen, N., Urbach, H., Klingler, J.-H., Beck, J., & El Rahal, A. (2024). Impact of spinal CSF leaks on quality of life and mental health and long-term reversal by surgical closure. *Neurology: Clinical Practice, 14*(2), Article e200272. https://doi.org/10.1212/cpj.0000000000200272

Voon, V. L. W., Tan, P. P., & Sim, C. W. (2026). Towards human-centric mental health chatbot design: Leveraging on DASS-21. *Journal of Advanced Research in Applied Sciences and Engineering Technology, 63*(1), 37-51. https://doi.org/10.37934/araset.63.1.3751

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized autoregressive pretraining for language understanding* [Preprint]. arXiv. https://arxiv.org/abs/1906.08237v1

Zandieh, S., Abdollahzadeh, S. M., Sadeghirad, B., Wang, L., McCabe, R. E., Yao, L., Inness, B. E., Pathak, A., Couban, R. J., Crandon, H., Torabiardakani, K., Bieling, P., & Busse, J. W. (2024). Therapist-guided remote versus in-person cognitive behavioural therapy: A systematic review and meta-analysis of randomized controlled trials. *Canadian Medical Association Journal, 196*(10), E327-E340. https://doi.org/10.1503/cmaj.230274

Zeidner, M., Matthews, G., & Roberts, R. D. (2012). The emotional intelligence, health, and well-being nexus: What have we learned and what have we missed? *Applied Psychology: Health and Well-Being, 4*(1), 1-30. https://doi.org/10.1111/j.1758-0854.2011.01062.x

Zhong, W., Luo, J., & Zhang, H. (2024). The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. *Journal of Affective Disorders, 356*, 459-469. https://doi.org/10.1016/j.jad.2024.04.057

Pertanika J. Sci. & Technol. 34 (S3): 55 - 69 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Detecting Fake Reviews on Online Products with Machine Learning

P. Dhanalakshmi[1]*, Ummadi Janardhan Reddy[2], Ravikanth Garladinne[3] , S. Ramachandra[4], E. R. Aruna[5], and Sharath Chandra Parashara[6]

[1]Department of Artificial Intelligence and Machine Learning, School of Computing, Mohan Babu University, Sree Sainath Nagar, 517102 Tirupati, Andhra Pradesh, India
[2]Department of Computer Science and Engineering, Faculty of Engineering and Technology, Alliance University, Chikkahagade Cross, Chandapura–Anekal Main Road, Anekal, 562106 Bengaluru, Karnataka, India
[3]Department of Computer Science and Engineering, KLEF College of Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, 522302 Guntur, Andhra Pradesh, India
[4]Department of Mechanical Engineering, Mother Theresa Institute of Engineering and Technology, Melumoi Post, 517408 Palamaner, Chittoor, Andhra Pradesh, India
[5]Department of Computer Science and Engineering, Vardhaman College of Engineering, Kacharam, Shamshabad, 501218 Hyderabad, Telangana, India
[6]Fenix Commerce Inc., 5434 Castle Manor Drive, San Jose, 95129 California, United States

ABSTRACT

Reviews play a crucial role in potential buyers' decision-making and enable companies to showcase the quality of their products. A growing number of deceitful reviews, published either intentionally or automatically as computer-generated, pose a significant threat to online review platforms. For consumers and companies, it is thus crucial to distinguish between reliable and deceptive online reviews. For this reason, the automatic classification of online reviews as being fake or genuine is a highly relevant text classification task. In this work, we present a framework that relies on the feature extraction from textual data, using Term Frequency–Inverse Document Frequency (TF-IDF) in combination with the vectorisation techniques Count Vectoriser and Hashing Vectoriser. The extracted feature vectors are then fed into three classification models: Support Vector Machine (SVM), Random Forest, and Logistic Regression. The Amazon review dataset was used for the experiments. The performance of all classifiers was measured using accuracy, precision, recall, and F1-score. Our results show that SVM performs best among the tested approaches, with TF-IDF in combination with the Count Vectoriser yielding strong results. By combining vectorisation with standard machine learning, our approach offers an efficient, easy-to-understand framework that is less complex

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 06 October 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.03

E-mail addresses:
mallidhana5@gmail.com (P. Dhanalakshmi)
ummadi.janardhan@gmail.com (Ummadi Janardhan Reddy)
garladinne.ravikanth@gmail.com (Ravikanth Garladinne)
mail2sramachandra@gmail.com (S. Ramachandra)
er.aruna@vardhaman.org (E. R. Aruna)
hellosharathchandra@gmail.com (Sharath Chandra Parashara)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

than Deep Learning or Transformer-based approaches. Thus, classical machine learning methods can serve as a good starting point for automatically detecting and avoiding the display of fake online reviews across various applications, while also efficiently processing text with limited computational resources.

Keywords: Count Vectoriser, fake review detection, Hashing Vectoriser, logistic regression, machine learning, natural language processing, random forest, support vector machine, TF-IDF

INTRODUCTION

Online shopping has dramatically changed the nature of information gathering for products and services, and buyers now rely heavily on previous buyers' experiences with a product or service. Hence, online reviews contain substantial information regarding product characteristics, perceived quality, usability, and customer satisfaction. Review systems assist customers when purchasing products online.

However, in most cases, customers rely on authentic reviews. Furthermore, online reviews can be deceptive, as they may be intentionally written to promote a product positively, make a product more popular, or even be fake. There are even programs that automatically generate reviews, which makes manual checking of online reviews very time-consuming and nearly impossible when many reviews are uploaded to an online platform. Deceptive content in online reviews influences both buyers and sellers.

Reviewing buyers may base their purchasing decisions on false information. In addition, sellers may be adversely affected by competitors' fake and fraudulent reviews. Therefore, methods to detect false online reviews have become a highly relevant research topic. We treat fake-review detection as a binary classification problem between deceptive and legitimate reviews. In a binary classification problem, a machine learning model decides between two categories, e.g., whether a review is deceptive or legitimate.

A variety of aspects need to be addressed when building a classifier, such as sufficient labelled data, proper data preprocessing, suitable feature representation, a proper classification model, and an appropriate way to measure performance. Reviews are typically text data that cannot be used directly by most conventional algorithms. Thus, the first step is converting the reviews into numerical feature vectors.

This will result in possibly very different feature spaces depending on the method used to transform the text. There are two main methods: Count Vectoriser (CV) and Hashing Vectoriser (HV). To improve feature extraction using CV, we can use TF-IDF weighting to assign greater importance to the few informative terms in a review. We employ three standard classifiers (SVM, Random Forest, and Logistic Regression) and analyse their behaviour in conjunction with the methods for representing text as feature vectors.

All experiments have been done using the same dataset of labelled Amazon reviews and the same evaluation protocol. The purpose of this research was to investigate the capability of relatively simple machine learning methods to correctly distinguish between fake and genuine online reviews. The research made use of a number of methods to transform the reviewed text into numerical vectors (features) that can be used by machine learning methods for classification, and compared the performance of Support Vector Machine, Random Forest and Logistic Regression classifiers when used in conjunction with Term Frequency-Inverse Document Frequency (TF-IDF) transformed reviews, generated using the Count Vectoriser and Hashing Vectoriser features.

The study, therefore, provides a controlled experimental analysis of six classifier-Vectoriser configurations. Accuracy, precision, recall, F1-score, and confusion matrices are considered to obtain a broader view of classification performance. Rather than attempting to establish superiority over modern transformer or large-language-model systems, the work focuses on understanding the behaviour of conventional machine-learning techniques under a common experimental setup.

LITERATURE SURVEY

Research on deceptive online reviews has grown dramatically with the advent of machine learning, deep learning, and natural language processing. Most studies investigated different types of information, including text characteristics, sentiment, user behaviour, relationships between users and products, and multimodal information.

Mutemi and Bação (2024) presented and evaluated machine-learning approaches for e-commerce fraud detection. We expand on this work and discuss the growing arsenal of automated detection methods for the vast array of online fraud emerging on the e-commerce scene.

Supervised learning approaches for detecting fake reviews have also been extensively studied. Mir et al. (2023) examined the application of supervised learning with BERT to identify fake reviews. They used supervised learning to classify deceptive reviews, using contextual language representations rather than simple frequency features. These contextual language features are created by a transformer-based language model such as BERT.

Related to the earlier discussion of misleading online content, Akhtar et al. (2023) have developed an approach using Artificial Intelligence (AI) to identify other types of on-line deception. In this paper, they presented an overview and comparison of current state-of-the-art approaches to detecting and preventing fake news and disinformation, using AI and machine learning techniques to analyse text-based content.

Hajek et al. (2023) investigated the use of fake reviews for aspect-based sentiment analysis. This is an important line of work, as it can reveal more about the relationship between a reviewer's expressed opinions and the relevant aspects of the product under review.

In related work, Zhang et al. (2023) explore the use of sentiment intensity for effective fake-review detection, employing Positive-Unlabeled learning to address the challenge of obtaining reliable labels for fraudulent reviews in practical settings.

Using such methods for learning with limited or even no positive examples for fraudulent reviews is an interesting topic for future research. Deceptive review classification using deep learning approaches has also been investigated. Mohawesh et al. (2023) introduced an explainable multi-view deep-learning ensemble for detecting fake reviews. The paper offers in-depth insight into the approach, highlighting its ability to combine multiple representations and to provide explanations for its predictions.

Additional work has examined the discrepancy between review scores and corresponding text. Work here has identified instances of deceptive reviews in which review scores and corresponding text are "inconsistent" (Sharifpour & Lahmiri, 2023).

We can gain insight not found by examining reviews one by chance. Zabeen et al. (2023) presented an uncertainty-aware framework that utilised LSTM and BERT to develop sequential and contextual representations for training and assessing a deceptive-review classification model.

Several recent studies have taken advantage of the wealth of additional information beyond the review text. Multimodal approaches, as well as behavioural studies, can leverage this data, which is often not available in plain text reviews. Li and Chen (2023), for example, propose a deep neural network that incorporates review text with other multimodal features and performs feature fusion to merge them into a single set of inputs.

Recently, many hybrid approaches have been introduced to exploit both the sequential information in reviews and the contextual information in words. For this purpose, Mohawesh et al. (2024) enhanced an LSTM network by integrating it with the powerful pre-trained language model RoBERTa. Reviews are also associated with reviewers' behaviours.

We can therefore combine review text with information from reviewers' behaviour in a detection system. Liu et al. (2024) use the RoBERTa model to generate text representations and, on top of this, add several types of reviewer behaviours to their framework. Often, suspicious activity leaves a 'fingerprint' with the behaviour of reviewers on a website rather than in the linguistic content of individual reviews.

To model the relationships among reviewers, products, and corresponding reviews, graph-based methods have been extended in a strong manner so far. Specifically, Yao et al. (2024) presented a graph contrastive learning framework that incorporates hierarchical relationships and label consistency for review detection.

Recently, transformer-based models have become popular for detecting deceptive content. In this line of work, Hyder et al. (2024) proposed a method, DeceptiveBERT, to identify deceptive reviews using BERT representations. Since these models are equipped with contextual language models, they can capture inter-word relationships, which are challenging for frequency-based methods to capture.

More recent work has used large language models and advanced deep learning techniques. In this category of work, Fan et al. (2025) introduced a deep learning method that uses characteristics derived from chaos theory to classify sophisticated deceptive online reviews. Thilini Jayasinghe and Dassanayaka (2025) compared the performance of various deep neural networks in detecting fraudulent Amazon reviews.

The area of fake-review detection has moved on from simple text classification using features to more complex deep learning models, including transformers, behavioural models such as graph-based methods, and, more recently, multimodal and large language models. All the studies reviewed here are influenced by both the content representation of the review and its classification. Importantly, no single source can solve the problem of fake reviews for all scenarios. The useful signals are hidden in the review textual patterns as well as in the reviewer's sentiment, in the reviewer's ratings, in the reviewer's behaviour, as well as in the entities' inter-relations.

RESEARCH GAP

The state of the art for the detection of fake reviews is presented in increasing order of technological complexity, i.e., from conventional machine-learning methods to deep learning methods, to transformer-based methods, to behavioural models, to graph models, to multimodal methods, and finally to large language models (Fan et al., 2025; Hyder et al., 2024; Liu et al., 2024; Mir et al., 2023; Mohawesh et al., 2023; Mohawesh et al., 2024; Thilini Jayasinghe & Dassanayaka, 2025; Yao et al., 2024). Even if the presented methods can successfully detect fake reviews, most are very complex and require substantial computational power.

Most studies also have several limitations. The performance of existing models is often reported across different data sets, with different types of labelling, class distributions, and preprocessing steps, feature representations, and evaluation measures. Thus, even if a model is reported to outperform other models on a particular benchmark, this does not necessarily mean it will also outperform them on other, similar benchmarks. In order to provide a simple and understandable framework for the comparison of various text representations and machine learning approaches within a fixed setup of experiments, the use of two simple text representations (Count Vectoriser and Hashing Vectoriser) in combination with TF-IDF, and three standard classifiers (Support Vector Machine, Random Forest, and Logistic Regression) for the detection of fake reviews appears to be promising. These methods are easy to evaluate and to interpret.

Position of the Present Study

This study investigates the capability of TF-IDF-based representations, built with the Count Vectoriser and the Hashing Vectoriser, to evaluate the performance of SVM, Random

Forest, and Logistic Regression classifiers in detecting fake reviews. The same experimental framework is then used to analyse the impact of vectorisation and classification on performance. The goal of this study is to provide a simple, interpretable tool for detecting fake reviews and to determine the best combination of text representation and a traditional classifier on a given dataset of Amazon reviews. The results will not be compared against the state of the art, which comprises more recent transformer- and LL_miss-based approaches, but they will provide a concrete, and thus practical, evaluation of fake-review detection using lightweight methods.

EXISTING METHODOLOGY

Existing GPT-2-Based Approach

Previous research has investigated the use of generative language models to create and identify deceptive online reviews. GPT-2 can be used to generate synthetic review content by learning linguistic patterns from existing review data. Such generated reviews can be used as an experimental source for studying the characteristics of artificially produced review text. In the existing approach considered in this study, GPT-2 was used to generate synthetic reviews with different review lengths and rating characteristics.

However, the GPT-2-based synthetic review generation approach is considered an existing methodology and is not used as the dataset for the proposed machine-learning experiments presented in Section 4. The proposed experiments are conducted independently on the Amazon review dataset, which contains labelled computer-generated and human-written reviews. This distinction is made to clearly separate the previously investigated synthetic review generation approach from the dataset used to train and evaluate the proposed classification models.

The existing methodology provides a basis for understanding how artificially generated review content can be produced and how such content may differ from human-written reviews. However, the present study focuses on the subsequent classification task of distinguishing fake reviews from genuine reviews using conventional machine-learning algorithms and alternative text-vectorisation techniques.

Proposed Methodology

The proposed methodology focuses on detecting fake and genuine online product reviews using text-based feature representations and supervised machine learning classification as shown in Figure 1. The approach consists of data preparation, text preprocessing, feature extraction, TF-IDF-based feature weighting, classification, and performance evaluation. Count Vectoriser and Hashing Vectoriser are considered alternative text vectorisation techniques, while Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) are used as classification algorithms.

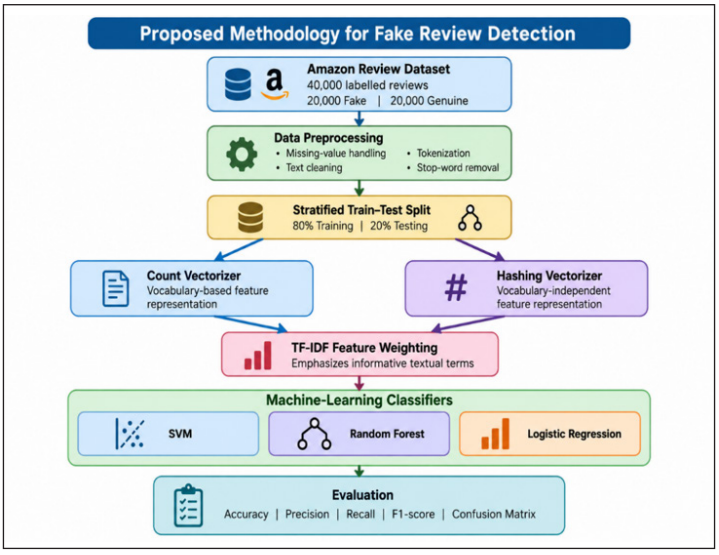


Figure 1. Proposed methodology for fake review detection using text vectorisation and machine learning

The proposed methodology is designed to provide a consistent comparison of different combinations of text representations and machine-learning classifiers. The overall process begins with the labelled Amazon review dataset and proceeds through preprocessing and numerical feature representation before classification. The trained models are finally evaluated using standard classification metrics to assess their effectiveness at distinguishing fake reviews from genuine ones.

Dataset Description

The proposed model uses an Amazon product review dataset containing 40,000 labelled reviews. The dataset consists of 20,000 reviews labelled as computer-generated (fake) and 20,000 labelled as human-written (genuine), resulting in an equal class distribution of 50% fake and 50% genuine. The balanced class distribution allows the classification models to be evaluated without introducing a class-frequency bias toward either category, as shown in Table 1.

Table 1
Distribution of fake and genuine reviews in the Amazon review dataset

Review Class	Number of Reviews	Percentage
Fake / Computer-generated	20,000	50%
Genuine / Human-written	20,000	50%
Total	40,000	100%

In contrast to the GPT-2-based synthetic review generation approach discussed in Section 3, the Amazon review dataset described in this section is the dataset used for the proposed machine-learning experiments. Each review is assigned a predefined class label, enabling supervised machine-learning algorithms to learn the textual characteristics of fake and genuine reviews.

Data Preprocessing

The review text is first pre-processed to remove unnecessary information and prepare the data for feature extraction. The preprocessing procedure includes removing irrelevant characters, handling missing or incomplete text, converting text into a consistent format, and preparing the reviews for numerical representation. These preprocessing steps reduce noise in the review text and improve the data's suitability for subsequent feature extraction and classification.

Data Partitioning and Validation

The labelled Amazon review dataset is divided into training and testing subsets using an 80:20 stratified split. The training subset is used for model development, while the testing subset is kept separate and is used only for final performance evaluation. Stratified partitioning preserves the proportion of fake and genuine reviews in both subsets.

The classification models are trained using the training data and subsequently evaluated using the unseen test data. Accuracy, precision, recall, F1-score, and confusion matrices are considered to assess the classification performance. The test subset is not used during model training, thereby providing an independent evaluation of the trained models.

Feature Representation

After preprocessing, the textual reviews are converted into numerical representations because machine-learning classifiers cannot directly process raw text. In this study, the Count Vectoriser and the Hashing Vectoriser are investigated as alternative approaches to converting review text into numerical feature representations. TF-IDF weighting is subsequently used to represent the importance of textual terms.

Tokenisation

Tokenisation divides the review text into individual words or tokens. During preprocessing, unnecessary stop words and irrelevant textual elements are removed to obtain a cleaner representation of the review content. The resulting tokens are then used to represent numerical features.

Count Vectoriser

Count Vectoriser converts the pre-processed review text into a numerical document-term matrix. Each row represents a review, and each column represents a term in the vocabulary, while the corresponding matrix value represents the occurrence of that term in the review.

In this study, Count Vectoriser provides an explicit vocabulary-based representation of the review text before TF-IDF weighting is applied. The purpose of using Count Vectoriser is to preserve word-frequency information, allowing the classifiers to identify differences in word usage between fake and genuine reviews. The resulting representation is subsequently combined with TF-IDF weighting to emphasise informative textual features.

Hashing Vectoriser

Hashing Vectoriser provides an alternative text representation to Count Vectoriser. Instead of explicitly constructing and storing a vocabulary, it applies a hashing function to map terms directly into a predefined feature space. By not creating an explicit vocabulary, such an approach can avoid consuming a lot of memory during feature generation.

We compare the performance of Hashing Vectoriser to that of the Count Vectoriser in the experiments proposed in this work, in order to verify whether or not a vocabulary-independent representation can deliver results as good as those obtained by using a vocabulary-dependent representation when combined with the aforementioned classification methods (i.e., SVM, Random Forest and Logistic Regression).

Count Vectoriser and Hashing Vectoriser are feature-representation techniques. They transform text into numerical features which can then be used with classification algorithms such as SVM, Random Forest or Logistic Regression.

TF-IDF Feature Weighting

The Term Frequency-Inverse Document Frequency (TF-IDF) score is used to weight the importance of features in a review. The feature's Term Frequency (TF) within a review is combined with the feature's Inverse Document Frequency (IDF) across the entire review collection. The Inverse Document Frequency score decreases for features that occur in a large proportion of documents.

High weight will be given to terms that are very frequent in a single review but rare elsewhere in the dataset. To enhance the numerical representations derived from the review text, TF-IDF is used in this study. Two alternative Vectorisers, Count Vectoriser and Hashing Vectoriser, are also investigated. The use of TF-IDF weighting is furthermore compared to the weighting of no features at all.

Model Training and Classification

Three supervised learning methods in machine learning, SVM, RF, and LR, are implemented for the binary classification between fake and genuine reviews in this study.

The classifiers are trained individually using the feature vectors from the Count Vectoriser and Hashing Vectoriser with TF-IDF weighting. In the training phase, the classifiers are developed to recognise two review types, and in the testing phase, the individual classifiers are used to predict whether reviews are fake or real on previously unseen test data.

The same data processing and feature representation are applied to all classifiers to allow for a fair comparison of their classification performance.

Model Evaluation

We evaluate the performance of our proposed classification approaches by using standard measures such as accuracy, precision, recall, F1-score and a confusion matrix. Accuracy is the proportion of correctly classified reviews, as shown in Table 2. Precision measures the proportion of identified positive samples that are positive. Recall measures the ability to identify samples of interest. The F1-score is the harmonic mean of both precision and recall.

A confusion matrix has been used to look at the number of correctly/incorrectly classified instances of fake and genuine reviews. Using multiple measures to assess a classification model's performance yields more accurate results than relying on accuracy alone.

Table 2
Experimental configurations used for fake-review classification

Vectoriser	Feature Weighting	Classifier	Accuracy
Count Vectoriser	TF-IDF	Support Vector Machine (SVM)	91%
Count Vectoriser	TF-IDF	Random Forest (RF)	87%
Count Vectoriser	TF-IDF	Logistic Regression (LR)	86%
Hashing Vectoriser	TF-IDF	Support Vector Machine (SVM)	88%
Hashing Vectoriser	TF-IDF	Random Forest (RF)	85%
Hashing Vectoriser	TF-IDF	Logistic Regression (LR)	82%

RESULTS AND DISCUSSIONS

We conduct an experimental investigation of 6 combinations of 2 text vectorisation methods and 3 supervised classification methods, all using the same set of labelled reviews. The goal of the comparison is to test whether the text representation of reviews influences classification performance. Therefore, Count Vectoriser and Hashing Vectoriser are tested separately from the TF-IDF feature-weighting tool.

Comparative Performance

Experimental results revealed that the choice of Vectoriser greatly affected the performance of the machine learning methods, as shown in Figure 2. Using a combination of TF-IDF and Count Vectoriser, the SVM method achieved the highest accuracy of 91%, while the Random Forest method achieved 87% and the Logistic Regression method achieved 86%.

For TF-IDF with Hashing Vectoriser, the SVM achieved an accuracy of 88%, the Random Forest an accuracy of 85%, and the Logistic Regression an accuracy of 82%, as shown in Figure 3.

Across all tested setups, the highest accuracy was achieved by SVM combined with TF-IDF and Count Vectoriser, at 91%. The results indicate that the vocabulary-based Count Vectoriser representation provides better classification performance than the Hashing Vectoriser representation for all three classifiers under the experimental conditions considered in this study shown in Table 3.

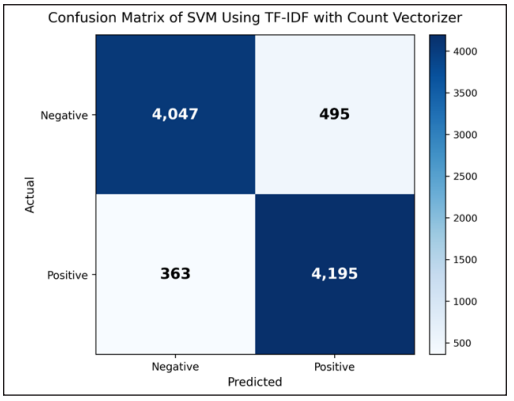


Figure 2. Confusion matrix of SVM using TF-IDF with Count Vectoriser

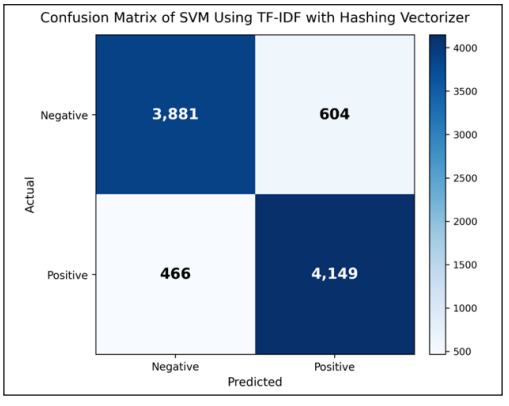


Figure 3. Confusion matrix of SVM using TF-IDF with Hashing Vectoriser

Table 3
Comparative accuracy of the evaluated machine-learning models

Vectoriser	Feature Weighting	Classifier	Accuracy
Count Vectoriser	TF-IDF	Support Vector Machine (SVM)	91%
Count Vectoriser	TF-IDF	Random Forest (RF)	87%
Count Vectoriser	TF-IDF	Logistic Regression (LR)	86%
Hashing Vectoriser	TF-IDF	Support Vector Machine (SVM)	88%
Hashing Vectoriser	TF-IDF	Random Forest (RF)	85%
Hashing Vectoriser	TF-IDF	Logistic Regression (LR)	82%

DISCUSSION

The results demonstrate that both the feature representation and the classification algorithm influence the performance of fake-review detection shown in Figure 4. Count Vectoriser combined with TF-IDF provides higher accuracy than Hashing Vectoriser combined with TF-IDF for SVM, Random Forest, and Logistic Regression.

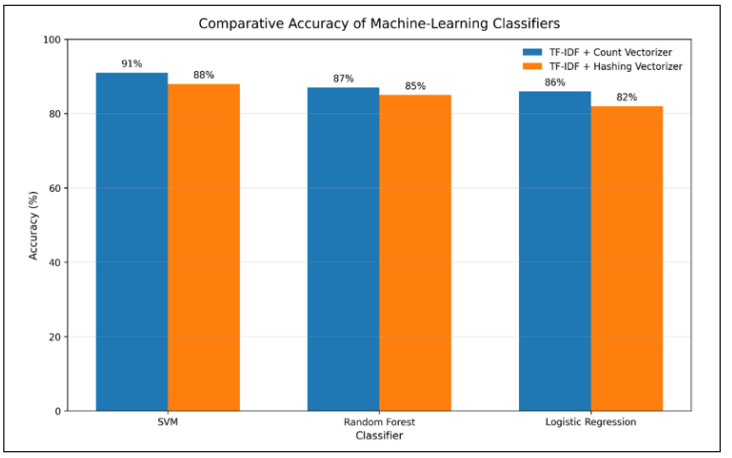


Figure 4. Comparative accuracy of machine-learning classifiers using Count and Hashing Vectorisers

SVM provides the best performance among the evaluated classifiers, achieving 91% accuracy with Count Vectoriser and 88% accuracy with Hashing Vectoriser. Random Forest achieves 87% and 85%, respectively, while Logistic Regression achieves 86% and 82%.

The comparatively higher performance of SVM indicates that it is effective in distinguishing the textual patterns associated with fake and genuine reviews in the selected dataset. However, the results are specific to the dataset and experimental configuration used in this study and should not be interpreted as evidence that SVM will outperform other approaches on all fake-review datasets.

Overall, the experiments demonstrate that conventional machine-learning algorithms, combined with appropriate textual feature representations, can achieve effective performance in fake-review classification. The best-performing configuration in the present study is SVM with TF-IDF and Count Vectoriser, which achieves an accuracy of 91%.

CONCLUSION

The purpose of this research is to investigate online product review classification (i.e. deceptive vs genuine) by using conventional machine learning approaches and text-based feature representations. We conducted experiments on a large dataset of 40,000 labelled Amazon reviews, comprising 20,000 computer-generated and 20,000 human-written reviews.

The text was transformed into numeric vectors by two different methods. We used both the Count Vectoriser and the Hashing Vectoriser. Both methods were combined with TF-IDF weighting and then used in three different supervised learning classifiers: Support Vector Machine, Random Forest and Logistic Regression. Our experiments achieved different accuracy values on the six different configurations. The best results were achieved by SVM (91%) with Count Vectoriser and TF-IDF, followed by Random Forest (87%) and Logistic Regression (86%). Using the Hashing Vectoriser with TF-IDF yielded accuracy values of 88%, 85% and 82% for SVM, Random Forest and Logistic Regression, respectively.

Our results show that, when classifying fake reviews with conventional machine-learning classifiers, the quality of the feature representation is crucial for achieving high accuracy. In our experiments, the best results were achieved using the Count Vectoriser and its default parameters across all methods. Our lightweight setup enables experimental investigations into deceptive-review classification without relying on large, compute-intensive transformer or large language model architectures.

It can also serve as a simple baseline in experimental setups or even in production with very limited computational resources. The findings reported here are based on a single dataset with a specific set of experimental parameters. A well-balanced, large dataset of labelled reviews for e-commerce systems, which may include many more features (linguistic, behavioural, temporal, etc.) than those used in this study, will likely need to be evaluated with other models and experimental settings. Thus, the results obtained here should not be used to assess the performance of the given models on other datasets.

Future work can explore additional features, such as reviewer-level behavioural features, temporal features, product-level features, and more linguistic features. We also plan to explore the use of Transformer-based architectures and large language models to investigate the added value of the contextual information encoded in these representations compared to the results obtained with the conventional methods used in this work.

ACKNOWLEDGEMENT

The authors would like to thank our universities for providing the facilities and support needed to carry out this research. We also extend our gratitude to our colleagues for their helpful insights and feedback during the preparation of this manuscript.

REFERENCES

Akhtar, P., Ghouri, A. M., Khan, H. U. R., Amin ul Haq, M., Awan, U., Zahoor, N., Khan, Z., & Ashraf, A. (2023). Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions. *Annals of Operations Research, 327*(2), 633-657. https://doi.org/10.1007/s10479-022-05015-5

Fan, Y., Fan, H., Zhang, F., & Wang, Z. (2025). Large language model meets chaos: A new deep learning model for fake review detection. *Electronic Commerce Research and Applications, 73*, Article 101521. https://doi.org/10.1016/j.elerap.2025.101521

Hajek, P., Hikkerova, L., & Sahut, J.-M. (2023). Fake review detection in e-commerce platforms using aspect-based sentiment analysis. *Journal of Business Research, 167*, Article 114143. https://doi.org/10.1016/j.jbusres.2023.114143

Hyder, S. B., Tariq, N., Moqurrab, S. A., Ashraf, M., Yoo, J., & Srivastava, G. (2024). BERT-based deceptive review detection in social media: Introducing DeceptiveBERT. *IEEE Transactions on Computational Social Systems, 11*(6), 7234-7243. https://doi.org/10.1109/tcss.2024.3403937

Li, X., & Chen, L. (2023). Fake review detection using deep neural networks with multimodal feature fusion method. In *2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS)* (pp. 2869-2872). IEEE. https://doi.org/10.1109/icpads60453.2023.00411

Liu, J., Quan, P., & Zhang, W. (2024). A study on fake review detection based on RoBERTa and behavioral features. *Procedia Computer Science, 242*, 1323-1330. https://doi.org/10.1016/j.procs.2024.08.131

Mir, A. Q., Khan, F. Y., & Chishti, M. A. (2023). *Online fake review detection using supervised machine learning and BERT model* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2301.03225

Mohawesh, R., Salameh, H. B., Jararweh, Y., Alkhalaileh, M., & Maqsood, S. (2024). Fake review detection using transformer-based enhanced LSTM and RoBERTa. *International Journal of Cognitive Computing in Engineering, 5*, 250-258. https://doi.org/10.1016/j.ijcce.2024.06.001

Mohawesh, R., Xu, S., Springer, M., Jararweh, Y., Al-Hawawreh, M., & Maqsood, S. (2023). An explainable ensemble of multi-view deep learning model for fake review detection. *Journal of King Saud University - Computer and Information Sciences, 35*(8), Article 101644. https://doi.org/10.1016/j.jksuci.2023.101644

Mutemi, A., & Bação, F. (2024). E-commerce fraud detection based on machine learning techniques: Systematic literature review. *Big Data Mining and Analytics, 7*(2), 419-444. https://doi.org/10.26599/bdma.2023.9020023

Sharifpour, K., & Lahmiri, S. (2023). Fake review detection using rating-sentiment inconsistency. In *2023 International Conference on Machine Learning and Applications (ICMLA)* (pp. 926-931). IEEE. https://doi.org/10.1109/icmla58977.2023.00137

Thilini Jayasinghe, J. M., & Dassanayaka, S. (2025). Detecting deception: Employing deep neural networks for fraudulent review detection on Amazon. *Neural Computing and Applications, 37*(26), 21715-21742. https://doi.org/10.1007/s00521-025-11485-y

Yao, J., Jiang, L., Shi, C., & Yan, S. (2024). Fake review detection with label-consistent and hierarchical-relation-aware graph contrastive learning. *Knowledge-Based Systems, 302*, Article 112385. https://doi.org/10.1016/j.knosys.2024.112385

Zabeen, S., Hasan, A., Islam, M. F., Hossain, M. S., & Rasel, A. A. (2023). Robust fake review detection using uncertainty-aware LSTM and BERT. In *2023 IEEE 15th International Conference on Computational*

Intelligence and Communication Networks (CICN) (pp. 786-791). IEEE. https://doi.org/10.1109/cicn59264.2023.10402342

Zhang, S., Zhu, A., Zhu, G., Wei, Z., & Li, K.-C. (2023). Building fake review detection model based on sentiment intensity and PU learning. *IEEE Transactions on Neural Networks and Learning Systems, 34*(10), 6926-6939. https://doi.org/10.1109/tnnls.2023.3234427

Pertanika J. Sci. & Technol. 34 (S3): 71 - 100 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Deep Learning and Signal Processing Synergy: Multitask Deep Neural Networks (MLT-DNNs) and Second-Order Blind Identification (SOBI) for Cybersickness Detection

Bisani Leela Padmavathi[1]* and J. Avanija[2]

[1]Department of Computer Science and Engineering, School of Computing, Mohan Babu University, Sree Sainath Nagar, 517102 Tirupati, Andhra Pradesh, India
[2]Department of Artificial Intelligence and Machine Learning, School of Computing, Mohan Babu University, Sree Sainath Nagar, 517102 Tirupati, Andhra Pradesh, India

ABSTRACT

Cybersickness in virtual reality (VR) environments presents an important challenge for user experience. Accurate evaluation approaches are essential for mitigating or preventing cybersickness among users. Compared to traditional subjective user studies, such as forms, electroencephalography (EEG) offers the advantage of high temporal resolution, enabling more precise real-time evaluation. EEG signals offer a promising avenue for objective detection of cybersickness, as they capture real-time neural responses. This study introduces a novel framework that combines Multitask Deep Neural Networks (MLT-DNNs) with Second-Order Blind Identification (SOBI) to enhance the accuracy and robustness of cybersickness detection. The MLT-DNN architecture leverages multitask learning to simultaneously predict cybersickness severity and related physiological states, such as stress and cognitive load, thereby enabling a comprehensive understanding of user experiences. SOBI is employed to pre-process EEG signals by separating and denoising independent source components, effectively reducing artefacts and noise. This integration of SOBI with MLT-DNN ensures high-quality input features, improving classification performance. The proposed framework is validated using a benchmark EEG dataset collected from VR users exposed to varying cybersickness-inducing scenarios. Experimental results demonstrate superior performance over baseline models, with significantly higher detection accuracy and reduced computational overhead. By combining advanced signal processing and deep learning methodologies, this work provides a scalable solution for real-time cybersickness monitoring

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 29 September 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.04

E-mail addresses:
leelaashwin.b@gmail.com (Bisani Leela Padmavathi)
avans75@yahoo.co.in (J. Avanija)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

in VR environments, contributing to the development of user-centric VR systems. Furthermore, this approach underscores the potential of multitask learning in addressing complex neurophysiological phenomena in human-computer interaction.

Keywords: Cybersickness, EEG, machine learning, multitask deep neural network, second-order blind identification, signal processing, virtual reality

INTRODUCTION

A wide range of businesses, including education, healthcare, and entertainment, have embraced VR. However, a major obstacle to its safe adoption is cybersickness, which is often associated with VR use. When people use virtual reality technology to immerse themselves in a virtual environment, they may experience cybersickness, a form of motion sickness (Chang et al., 2023). This condition results from perceived visual motion and the amount of detected vestibular motion; it is characterised by nausea, oculomotor pain, and disorientation. Some effects of using VR include dizziness, loss of balance, and impaired coordination. Cybersickness affects 50-80% of viewers of virtual reality content. Eradication of cybersickness is crucial to the growth of virtual reality. Thus, it is necessary to determine the extent of cybersickness associated with VR content to properly control it during its creation (Caserman et al., 2021).

The most frequently used assessment methods in the past were questionnaires and other subjective measures. As a word of caution, questionnaire-based methods rely on participants' perceptions and therefore can be sensitive to variation among participants (Martin et al., 2020). Furthermore, because these evaluations are usually conducted post-trial, data collection relies solely on retrospective measurements, which makes continuous monitoring of cybersickness challenging. The temporal lobe is involved in vestibular function, and the occipital lobe is involved in vision. The human brain is the primary organ of the central nervous system (Krokos & Varshney, 2022). When studying the physiological correlates of cybersickness, brain data are closely linked to the phenomenon. Previous studies have found a high level of dependency of participants' cybersickness in immersive virtual environments and EEG frequency characteristics. Moreover, because EEG is continuous and offers high temporal resolution, cybersickness can be evaluated automatically and continuously using EEG-based assessment methods. Consequently, cybersickness can be quantified and assessed with high reliability and practicability using EEG-based approaches (Heo & Yoon, 2020).

There are many VR devices available for creating an interesting experience. Among consumer VR systems, VR head-mounted displays (HMDs) are the most widely used. HMDs feature two screens positioned in front of the user's eyes. To represent each eye's perspective, digital images are presented in the appropriate perspective on each screen (Ahir et al., 2020).

A shell housing the screens and sensors continuously monitors the user's head movements and adjusts the virtual environment accordingly. Information obtained through the senses forms the basis of awareness of one's immediate surroundings. VR HMDs must replace real sensory experiences with computer-generated ones so that users can actively perceive them using their typical sensorimotor dependencies. The term "being" is used to describe the illusion of existing within a virtual environment. VR HMDs enable users to experience virtual surroundings as real by providing a wide field of view, stereoscopic vision, and real-time sensory updates (Porcino et al., 2020). This sense of presence leads people to behave realistically within virtual reality. With their compact size, lightweight design, and expanded field of view, HMDs enhance immersion in virtual environments by offering a variety of interactive experiences. The newest generation of HMDs enables users to "look around," "turn around," and engage realistically in a virtual environment by simulating changes in visual display as the head and torso are tilted. The latest generation of HMDs can generally be described as highly immersive and subjective (Teixeira & Palmisano, 2021).

Research on Our Work

The present research concerns the integration of deep learning with signal processing techniques to improve the recognition of cybersickness. Therefore, in this paper, we propose using a signal processing technique called SOBI in conjunction with a newly introduced multitask DNN to process physiological signals, including EEGs. This integration enhances the identification of cybersickness by incorporating the classification features of the multitask DNNs with the feature extraction features of the SOBI. The results of this work show how deep learning and modern signal processing techniques can be applied to address several issues in the analysis of physiological signals.

Motivation of This Research

This work is inspired by the growing application of virtual and augmented reality (VR/AR) technologies, which, although they hold transformative promise, can cause cybersickness that hinders their adoption. Existing approaches to detecting such behaviour are inaccurate and not applicable in real time. This study will seek to fill this gap using deep learning and signal processing – multitask DNNs and Second-Order Blind Identification (SOBI) for enhancing cybersickness detection. In this way, the study aims to create a new, effective, and practical system by integrating these approaches to improve VR/AR usability and detect cybersickness.

Main Contribution of Our Work

- To enhance the precision and resilience of cybersickness detection, this study introduces a novel framework that integrates SOBI with MLT-DNNs.

- The MLT-DNN architecture uses multitask learning to estimate cybersickness intensity and other correlated physiological conditions like stress and cognitive workload, which will enhance the understanding of cybersickness.
- Experimental results highlight superior performance compared to baseline models, demonstrating a significant increase in detection accuracy and reduced computational overhead.
- This work proposes a real-time cybersickness detection framework that integrates signal processing and deep learning to create user-oriented VR systems.

Paper Organisation

Figure 1 displays the organisation of the paper.

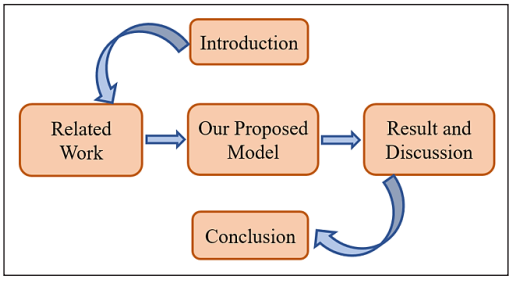


Figure 1. Organisation of the research

Related Work

Cybersickness, a condition characterised by discomfort, nausea, dizziness, and other physiological symptoms, often arises during interactions with virtual or augmented reality environments. As these technologies become more integrated into our daily lives, ranging from gaming and education to healthcare and professional training, understanding and addressing cybersickness is essential to ensure user comfort and widespread adoption. This survey aims to explore perceptions, experiences, and potential indicators of cybersickness to enhance detection and mitigation strategies.

Kourtesis et al. (2024) examined the background of cybersickness and its relationship to cognitive and motor performance. The CSQ-VR and the MSSQ were completed by 30 participants, aged between 20 and 45 years, during the VR roller coaster. In addition to the cognitive and psychomotor activities during the VR session, the participants were administered the CSQ-VR before and after the ride. It was found that the most important factors influencing adult cybersickness were susceptibility to motion sickness and pupil size.

Papaefthymiou et al. (2024) studied the effects of immersion, task execution, and subject characteristics on cybersickness symptoms across various stages of VR exposure. To this end, 47 subjects aged between 18 years and 45 years completed the Deary–Liewald Reaction Time (DLRT) task and the Cybersickness in Virtual Reality Questionnaire (CSQ-VR). Symptoms of cybersickness were assessed at four stages: participants' responses to questionnaires completed before and after the 12-minute rollercoaster ride (programmed to induce cybersickness) and before and after VR immersion. The results showed that

subjects' motion sickness symptoms worsened after the rollercoaster ride but were partially alleviated during the post-ride tasks.

Martirosov et al. (2022) conducted an experiment to examine the correlation between immersion and cyber illness among participants. Ninety participants, aged 19-36, were divided into two groups, with participants in the first group having higher levels of virtual reality exposure than those in the second group. The low-immersive group used monocular, screen-based PCs; the semi-immersive group used a CAVE with stereoscopic projectors; the highly immersive group used VR head-mounted displays; and the control group received no immersion.

According to Zaidi et al. (2023), to predict outcomes in computer science, the following ML and AI methods should be employed: SVM, DT, and KNN. Our findings revealed that DT and SVM achieve higher test accuracy than KNN, and DT achieves higher training accuracy than both SVM and KNN. When using machine learning to develop better solutions to computer science challenges, developers can anticipate potential problems when designing VR projects.

The issue of indirectly identifying and quantifying cybersickness in virtual reality based on physiological signals is discussed by Sameri et al. (2024). As a result, they propose a multimodal strategy that combines self-reported measurements with physiological data. Our experimental paradigm involves a roller coaster simulation and employs a mixed-methods approach integrating quantitative and qualitative analyses. By utilising statistical analysis and machine learning techniques, the study examines physiological data collected from 22 participants who were exposed to a virtual-reality roller coaster experience.

The systematic review by Yang et al. (2022) focuses on 26 selected studies on the automatic detection of cybersickness using machine learning applied to biometric and neurophysiological signals collected via wearable technology. The review examines techniques, data processing methods, machine learning architectures, and recommendations for further research on cybersickness prediction and detection. Various machine learning architectures, features, participant activities, and immersion environments were identified. Despite advancements in cybersickness detection methods, there is currently no model in the literature capable of predicting first-instance incidents.

The authors (Qu et al., 2022) used physiological sensors to monitor cybersickness in real time, collect data, and apply a deep learning model to quantify the elements contributing to cybersickness. Additionally, they developed a passive navigation task and a VR experimental platform designed to induce cybersickness in users. In the experiment, to train the LSTM Attention neural network model, the position and rotation of the user's virtual figure's bone, as well as physiological signals such as electrodermal activity (EDA) and electrocardiogram (ECG), were captured in real time.

Research Gap

While there has been significant progress in VR technology, the precise and timely identification of cybersickness remains a major challenge. Current approaches can be based on self-estimations, which are notoriously unreliable, or rely on physiological data, which are not suitable for real-time monitoring and diffusion. Furthermore, the interaction between content-specific subcomponents, such as motion patterns and latency, and user-specific factors, such as age, gender, and prior VR exposure, remains poorly understood, leading to variable detection performance. This highlights the importance of robust, accurate, and versatile detection systems that integrate data from multiple sensors with state-of-the-art machine learning techniques to improve predictive accuracy and practicality across various VR applications.

Problem Identification of the Existing System

- Existing systems often fail to provide precise, real-time detection of cybersickness symptoms, limiting their usability in dynamic virtual environments.
- Many systems rely on generalised models that do not account for individual variability in susceptibility to cybersickness, leading to inconsistent results.
- Current detection methods primarily depend on a single source of data, such as physiological signals or user behaviour, without integrating multiple modalities for a more comprehensive analysis.
- Numerous solutions require users to wear bulky or intrusive sensors, which can exacerbate discomfort and reduce user acceptance.

Proposed Model

The study presents a new approach of combining SOBI and MLT-DNNs to improve the performance and reliability of cybersickness detection. The MLT-DNN architecture uses multitask learning to estimate cybersickness intensity and related physiological states, including stress and cognitive load, thereby providing a holistic view of the user experience. In this paper, SOBI is used to preprocess EEG signals through independent source separation and filtering out artefacts and noise. The integration of SOBI and MLT-DNN guarantees high-quality input features, thereby enhancing classification. The applicability of the proposed framework is demonstrated using a benchmark EEG dataset acquired from VR users experiencing different cybersickness-eliciting conditions. The model diagram is illustrated in Figure 2.

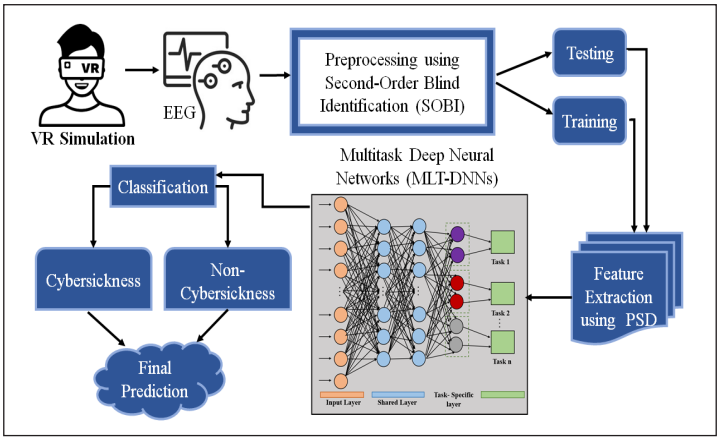


Figure 2. Block diagram of the proposed model

Dataset

The dataset is a collection of user feedback obtained in VR settings, specifically user choices, mood, and physiological characteristics such as heart rate variability and skin conductance (Joshi, 2023). The main goal of this study is to contribute to the development of VR technologies by focusing on users' experiences, with particular attention to enhancing VR design, user protection, and personalisation. By analysing the physical and emotional reactions to different VR situations, the dataset helps developers design appropriate and engaging VR experiences for people. The main variables in the dataset are Age (the participant's age during the VR session), Gender (the participant's gender), and User ID (the participant's identification number). Other variables, such as the VR Headset Type, the duration the respondent spent with the VR, and the Motion Sickness Rating, also give information about the VR equipment and the respondent. The dependent variable in this study is Immersion Level, which measures the extent to which users felt immersed while using the mobile application, on a 1-5 Likert scale. This dataset will be useful to developers for improving VR systems, increasing user satisfaction and immersion, and designing VR for various clients.

Signal Pre-Processing

This makes signal preprocessing important for improving data quality and for extracting useful features for identifying cybersickness. This process involves three key steps: noise removal, normalisation, and Second Order Blind Identification (SOBI) (Liu et al., 2019). These steps help ensure that the signals are clean and free of interference and can be used to decompose independent components associated with cybersickness.

Noise Removal

Signal filtering is used to eliminate many unwanted noise and artefacts present in the signal. To remove baseline drifts at low frequencies and high-frequency noise, a band-pass filter is employed. The filtered signal $y(t)$ is computed using the convolution of the raw signal $x(t)$ with the filter impulse response $h(t)$ as in Equation 1:

$$y(t) = \int_{-\infty}^{\infty} x(\tau) \quad h(t-\tau)d\tau \qquad [1]$$

For physiological signals such as EEG, the normal frequency is between 0.5Hz and 50Hz as in Equation 2.

$$h(t) = A[\sin(2\pi f_{high}t) - \sin(2\pi f_{low}\, t)] \qquad [2]$$

Where f_{low} and f_{high} signify the lower and upper cutoff frequencies, correspondingly.

Additionally, baseline drift can be removed using a moving average filter, which subtracts the local average over a window N as in Equation 3:

$$y(t) = x(t) - \frac{1}{N}\sum_{i=t-N/2}^{t+N/2} x(i) \qquad [3]$$

This helps in making the signals free from long-term trends and high-frequency noise.

Standardisation

To ensure consistency across participants and sessions, the signals are normalised. Each signal value x is transformed into a standardised value z using Equation 4:

$$z = \frac{x-\mu}{\sigma} \qquad [4]$$

Here, μ is the mean of the signal, and σ is the standard deviation. This makes it possible to eliminate individual variability and hence improve the accuracy of the analysis by reducing it to a standard form with a mean of zero and variance of one.

Second-Order Blind Identification (SOBI)

SOBI is an enhanced signal processing method used to extract independent components from observed signals. It starts by computing the covariance matrix of the observed signal vector $x(t)$ at different time lags τ as in Equation 5:

$$C_x(\tau) = \mathbb{E}[x(t)x^T(t+\tau)] \qquad [5]$$

Here, $C_x(\tau)$ captures the relationships between signal channels at a time lag τ. SOBI performs joint diagonalisation of these covariance matrices to estimate a transformation matrix W that minimises off-diagonal elements as mentioned in Equation 6:

$$WC_x(\tau)W^T\Lambda(\tau) \quad [6]$$

Where $\Lambda(\tau)$ is a diagonal matrix representing independent signal components.

The independent components $s(t)$ are then computed as specified in Equation 7:

$$s(t) = W_x(t) \quad [7]$$

These components are then compared to determine which are most associated with cybersickness symptoms, such as EEG patterns associated with motion sickness or cognitive load.

Feature Extraction Using Power Spectral Density (PSD)

In all the feature extraction methodologies for physiological signals, especially the EEG, PSD is among the most efficient (Alam et al., 2021). PSD is a quantitative measure of how a signal's power is distributed across its frequencies and is used to identify frequency-specific patterns of cybersickness. PSD can determine the physiological and cognitive conditions that lead to cybersickness symptomatology by employing the delta, theta, alpha, and beta bands.

To compute the PSD, the signal's autocorrelation function $R_{xx}(\tau)$ is analysed in the frequency domain using the Fourier Transform as mentioned in Equation 8:

$$PSD(f) = \int_{-\infty}^{\infty} R_{xx}(\tau)e^{-j2\pi fr}\,d\tau \quad [8]$$

Here, $R_{xx}(\tau) = \mathbb{E}\lfloor x(t)x(t+\tau)\rfloor$ is the autocorrelation function of the signal $x(t)$, which represents how the signal correlates with itself at different time lags τ, and f is the frequency. For discrete signals, PSD can be efficiently estimated using the Fast Fourier Transform (FFT):

Where N is the length of the signal, $X(f_k)$ is the Fourier Transform of the signal at frequency f_k.

The power within specific frequency bands, such as delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz), can be computed using the PSD. These bands provide meaningful insights, such as elevated theta power reflecting cognitive stress or altered alpha power indicating discomfort.

To standardise the characteristics for model input, the relative power of each frequency band is computed as in Equation 9:

$$P_{relative} = \frac{\int_{f_{low}}^{f_{high}} PSD(f)df}{\int_{f_{min}}^{f_{max}} PSD(f)df} \quad [9]$$

Here f_{low} and f_{high} is the boundaries of the specific frequency band, f_{min} and f_{max} is the full frequency range of the signal.

By converting raw physiological signals into PSD features, the model can extract more frequency-domain information, which helps it identify patterns associated with cybersickness. These features are very useful for the severity levels and the early signs of the disease.

Multitask Deep Neural Networks (MLT-DNNs)

Advanced machine learning architectures, known as MLT-DNN, are designed to handle multiple related tasks simultaneously by leveraging shared representations to enhance performance across all tasks (Chang et al., 2024). These models operate on the principle of joint learning, wherein shared parameters W are trained to optimise multiple loss functions L_t corresponding to each task t. The objective function of an MLT-DNN can be expressed as in Equation 10:

$$\min_{W,\{W_t\}} \sum_{t=1}^{T} \lambda_t L_t(W, W_t) \quad [10]$$

Where T is the number of tasks, W_t are task-specific parameters, and λ_t The weights control the importance of each task. The shared parameters W enable the model to capture common features across tasks, improving generalisation and reducing overfitting compared to training separate models in Figure 3 shows the MLT-DNN model.

A standard MLT-DNN consists of shared layers followed by task-specific layers. Let h_s represent the output of the shared

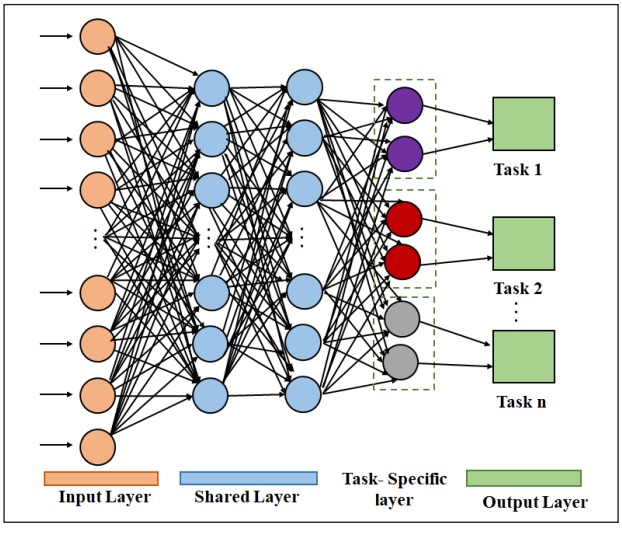


Figure 3. The architecture of the MLT-DNN model

layers and $h_{t,i}$ the output for the i-th input of the task t. This can be expressed as in Equations 11 and 12:

$$h_s = f_s(x; W_s) \quad [11]$$

$$h_{t,i} = f_t(h_s; W_t) \quad [12]$$

Here, W_s and W_t are the weights of the shared and task-specific layers, respectively. During backpropagation, gradients from all tasks contribute to updating the shared parameters W_s as in Equation 13:

$$\frac{\partial L}{\partial W_s} = \sum_{t=1}^{T} \lambda_t \frac{\partial L_t}{\partial W_s} \quad [13]$$

Another advantage of the MLT-DNNs is that they enable the sharing of data distributions across tasks to overcome data scarcity in individual tasks. For instance, in a multitasking environment with tasks A and B, the model can leverage task dependencies. If $g(x)$ represents a shared latent feature, then predictions for tasks can be given by Equations 14 and 15:

$$yA = g(x)W_A \quad [14]$$

$$yA = g(x)W_B \quad [15]$$

Regularisation is crucial in MLT-DNNs to prevent the overfitting of shared parameters to a single task at the expense of others. Common approaches include l_2-regularisation and dropout as in Equation 16:

$$R(W_s) = \frac{\lambda}{2}||W_s||^2 \quad [16]$$

The design of loss functions and weighting factors λ_t significantly impacts multitasking performance. A dynamic weighting scheme can be employed to balance losses as mentioned in Equation 17:

$$\lambda_t = \frac{exp\,(-L_t)}{\sum_{k=1}^{T} exp\,(-L_k)} \quad [17]$$

The cross-stitch units are also a new concept that allows certain features to be selectively shared between tasks. If z^A and z^B represent task-specific features, cross-stitching is defined as specified in Equations 18 and 19:

$$z^{A'} = \alpha z^A + \beta z^B \quad [18]$$

$$z^{B'} = \gamma z^A + \delta z^B \quad [19]$$

In MLT-DNNs, $\alpha, \beta, \gamma, \delta$, these are learned parameters used in cross-stitch units to control the degree of feature sharing between tasks. These parameters control the relative weighting of task-specific and general features to accommodate task dependencies while preserving task-specific differentiation.

Some of the fields where MLT-DNNs are applied include NLP, which involves tasks such as POS tagging and syntactic parsing that require similar representations. For instance, in NLP, word embeddings (E) used across different tasks can be trained together as mentioned in Equation 20:

$$E' = E - \eta \frac{\partial L}{\partial E} \quad [20]$$

Finally, the optimisations in MLT-DNNs include gradient clipping and adaptive learning rates, with optimisers such as Adam or RMSProp as specified in Equation 21.

$$\Delta W = \frac{\eta}{\sqrt{v_t + \epsilon}} m_t \quad [21]$$

In the context of optimisation algorithms like Adam, m_t and v_t represent the first and second moment estimates of the gradients, respectively. m_t is the exponentially weighted moving average of the gradients, capturing the mean, while v_t tracks the variance by computing the moving average of squared gradients. These estimates help control the learning rate and enable stable convergence. The integration of shared representation, dynamic task weighting, and regularisation strategies makes MLT-DNNs a reliable model for effectively solving complex multitask problems.

Advantages of the Proposed Model

- Combines the strengths of multitask DNN and SOBI to enhance detection performance.
- Combines deep learning's ability to extract nonlinear characteristics with SOBI's skill in isolating and processing signal parts.

- Simultaneously addresses multiple related tasks, such as classification and regression, for a comprehensive cybersickness assessment.
- Efficiently processes complex, multi-channel signal data, ensuring adaptability to diverse real-world scenarios.

RESULT AND DISCUSSIONS

Experimental Setup

To train the network, we applied the Adam optimiser with a learning rate of 0.001. The experiment was performed on a server system with an Intel i9-10900X CPU, 128 GB of RAM and Nvidia RTX 3090 GPU.

Performance Metrics

Root Mean Square Error (RMSE)

RMSE, which calculates the difference between expected values $(\hat{y}_i)$ and actual values (y_i). It is a widely used statistic for assessing a model's accuracy. Better model performance is indicated by a lower RMSE, which measures the average magnitude of prediction errors as in Equation 22:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad [22]$$

Pearson's Linear Correlation Coefficient (PLCC)

It measures the linear relationship between two variables and assesses both its direction and strength. A complete negative correlation is represented by a value of -1, a perfect positive correlation by a value of 1, and no correlation by a value of 0. In regression tasks, PLCC is frequently used to evaluate the consistency between expected and actual data, as in Equation 23:

$$PLCC = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad [23]$$

Spearman Rank Order Correlation Coefficient (SROCC)

A monotonic relationship between two ranked variables is measured by the SROCC, which also indicates the direction of the relationship. This test evaluates whether a monotonic

function can adequately characterise the relationship between two datasets and is non-parametric. To calculate the SROCC, use Equation 24:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad [24]$$

Accuracy

One popular metric for assessing a classification model's performance is accuracy. This is the percentage of total instances that are correctly classified. The formula for accuracy is specified in Equation 25:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad [25]$$

F1-Score

The F1-Score is a performance measure that incorporates both precision and recall in problems of classification. It is particularly useful in handling an imbalanced dataset. The F1-Score is an average measure defined as the harmonic mean of precision (P) and recall (R) as in Equation 26.

$$F1 - Score = 2 \times \frac{Pr\ ecision\ \times Re\ call}{Pr\ ecision\ + Re\ call} \quad [26]$$

Processing Time

The total time required to complete a particular computation or process. It is used to assess the efficiency of algorithms in different areas, such as artificial neural networks and signal processing. Processing time can be expressed as in Equation 27:

$$T_{Proces\ sin\ g} = T_{Comp} + T_{IO} \quad [27]$$

Recall

In classification problems, recall measures the percentage of actual positive samples the model recognises. It indicates the extent to which the model can identify positive cases as specified in Equation 28:

$$Re\ c\ all = \frac{TP}{TP+FN} \quad [28]$$

Precision

Precision measures the number of correctly classified positives among all positives predicted by the model in classification problems as mentioned in Equation 29. In this

case, it is especially beneficial to measure the precision of positive predictions when the cost of false positives is high.

$$Pr\,e\,cision = \frac{TP}{TP + FP} \quad [29]$$

Comparative Methods

- **3D-CNN (Three-Dimensional (3D) - Convolutional Neural Network (CNN)) (Lee et al., 2019):** This is accomplished by using disparity, saliency, and optical flow maps from an input video as the inputs for the 3D CNN, representing depth, velocity, and eye movement, respectively.
- **SVM+TDA (Support Vector Machine + Topological Data Analysis) (Hadadi et al., 2022):** When compared to various feature descriptor combinations for physiological time series, the TDA + SVM combination performs statistically similarly to or better than, or matches, more complex and harder-to-interpret approaches.
- **LSTM-Attention (Long Short-Term Memory-Attention) (Qu et al., 2022):** We collected real-time physiological data from participants throughout the experiment, including their skin electrical activity (EDA) and electrocardiogram (ECG), as well as location and bone-rotation data for their virtual avatar, to train the LSTM Attention neural network model.

Table 1 and Figure 4 provide a comprehensive comparison of the F1-score and accuracy across different models: 3D-CNN, SVM+TDA, LSTM-Attention, and the Proposed Framework, evaluated on multiple subjects. The Proposed Framework consistently outperforms the other models, achieving the highest average F1-score (88.03%) and accuracy (97.16%). This marked improvement underscores the success of combining multitask learning with preprocessing techniques for identifying cybersickness. These results show that the Proposed Framework is reliable and extensible in practice, which establishes a new standard for identifying cybersickness.

Table 1
Comparison of F1-score and accuracy analysis with existing systems

Subject	3D-CNN		SVM+TDA		LSTM-Attention		Proposed	
	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
S1	65.35	65.91	70.91	68.89	62.67	76.56	89.34	94.35
S2	61.22	68.54	79.44	76.45	65.98	70.77	81.44	95.98
S3	76.34	76.34	67.23	71.99	75.98	66.71	90.23	97.77
S4	79.91	76.76	61.55	75.34	76.18	60.91	88.18	98.81
S5	77.44	77.34	69.98	72.26	79.91	79.99	90.99	98.91
Average	72.05	72.97	69.82	72.98	72.14	70.98	88.03	97.16

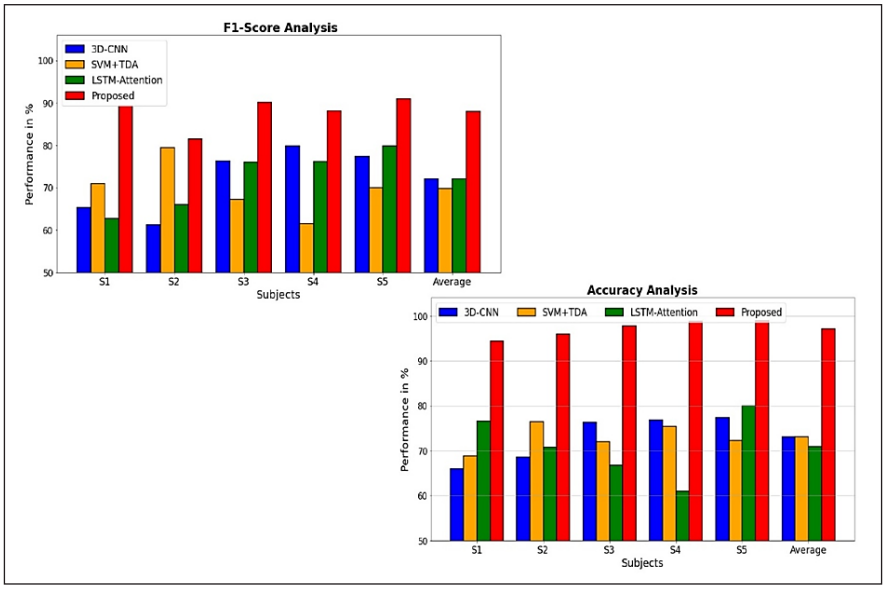


Figure 4. Comparison of F1-score and accuracy analysis with existing systems

Table 2 and Figure 5 highlight the precision and recall performance of various models, including 3D-CNN, SVM+TDA, LSTM-Attention, and the Proposed Framework, across multiple subjects. The Proposed Framework maintains the highest overall average precision (95.97%) and recall (95.92%), while the other approaches are inferior. Such superior performance indicates the high reliability of the Proposed Framework in detecting instances of cybersickness while minimising false negatives. The findings confirm the efficiency of the proposed multitask learning approach and the effectiveness of the more sophisticated preprocessing methods, thereby strengthening its practical applicability to real-world VR scenarios.

Table 2
Comparison of precision and recall analysis with existing systems

Subject	3D-CNN		SVM+TDA		LSTM-Attention		Proposed	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
S1	88.45	78.87	76.54	76.91	75.55	75.18	94.65	93.25
S2	80.22	66.19	79.44	78.87	74.14	76.34	95.99	95.55
S3	81.66	60.34	85.44	65.23	76.67	72.18	93.66	97.44
S4	77.65	76.16	80.41	69.91	70.98	76.66	97.66	96.61
S5	70.33	79.44	81.43	69.99	81.33	78.91	97.91	96.78
Average	79.62	72.21	80.65	72.18	75.73	75.85	95.97	95.92

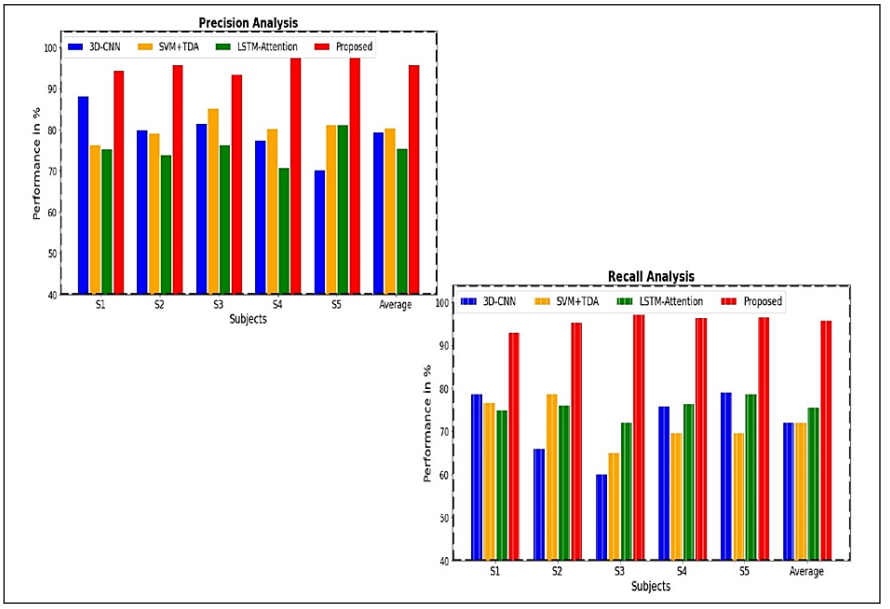


Figure 5. Comparison of precision and recall analysis with existing systems

The PLCC and SROCC are compared across several models, including 3D-CNN, SVM+TDA, LSTM-Attention, and the Proposed Framework, as presented in Table 3 and Figure 6. The Proposed Framework performs best across all experiments, achieving the highest average PLCC of 0.947 and SROCC of 0.952. These results demonstrate its enhanced capacity to sustain high levels of correlation with true cybersickness measures, which in turn suggests reliability and resilience in predictive performance. This supports the Proposed Framework for accurate, scalable assessment of cybersickness in virtual reality settings.

Table 3
Comparison of PLCC and SROCC analysis with existing systems

Subject	3D-CNN		SVM+TDA		LSTM-Attention		Proposed	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
S1	0.543	0.678	0.765	0.891	0.823	0.611	0.934	0.967
S2	0.765	0.761	0.667	0.875	0.733	0.865	0.913	0.913
S3	0.897	0.789	0.611	0.743	0.776	0.734	0.987	0.954
S4	0.881	0.666	0.788	0.711	0.678	0.771	0.911	0.966
S5	0.880	0.732	0.843	0.764	0.651	0.798	0.904	0.989
Average	0.793	0.725	0.734	0.796	0.730	0.755	0.947	0.952

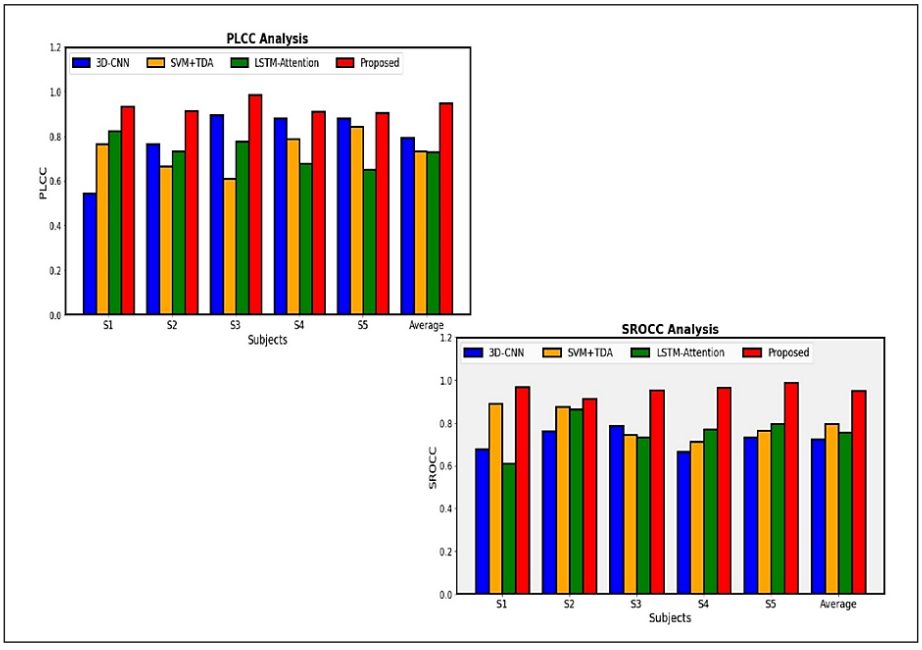


Figure 6. Comparison of PLCC and SROCC analysis with existing systems

Table 4 and Figure 7 show the RMSEs of the 3D-CNN, SVM+TDA, LSTM-Attention, and the Proposed Framework across different subjects. The Proposed Framework also has the lowest RMSE across all subjects, but its overall RMSE of 13.15 is higher than that of the other models. On the other hand, the average RMSE values for the other algorithms, namely SVM+TDA, LSTM-Attention, and 3D-CNN, are 25.85, 34.49, and 37.06, respectively. These results show that the Proposed Framework achieves higher prediction accuracy than the Baseline Model and can minimise prediction errors for cybersickness. This decrease in RMSE indicates the effectiveness of the proposed framework in accurately assessing cybersickness in a real-life VR environment.

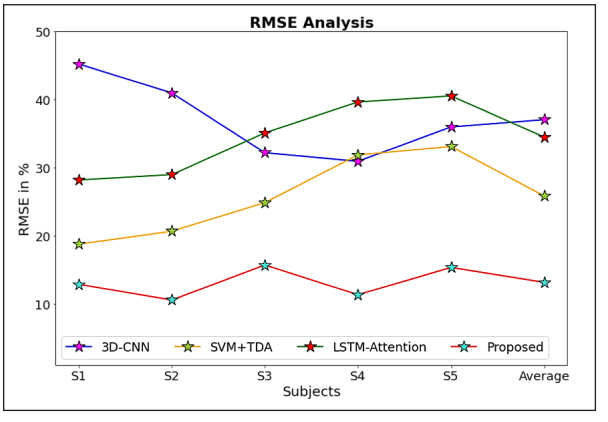


Figure 7. RMSE Analysis for proposed models with existing systems

Table 4
RMSE Analysis for proposed models with existing systems

Subject	3D-CNN	SVM+TDA	LSTM-Attention	Proposed
	RMSE			
S1	45.23	18.76	28.18	12.86
S2	40.98	20.66	28.98	10.55
S3	32.19	24.87	35.11	15.69
S4	30.93	31.87	39.66	11.33
S5	35.98	33.11	40.55	15.33
Average	37.06	25.85	34.49	13.15

Table 5 and Figure 8 present the processing times of the models 3D-CNN, SVM+TDA, LSTM-Attention, and the Proposed Framework for various subjects. The Proposed Framework takes the least time, 4.50ms on average, compared to the other models, which take 17.19ms, 16.26ms, and 10.17ms for 3D-CNN, SVM+TDA, and LSTM-Attention, respectively. This remarkable reduction in processing time underscores the effectiveness of the Proposed Framework, which can be implemented for real-time cybersickness detection in VR environments.

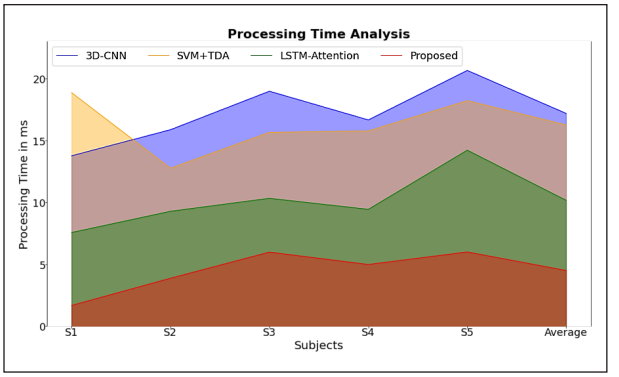


Figure 8. Processing time analysis with existing systems

Table 5
Processing time analysis with existing systems

Subject	3D-CNN	SVM+TDA	LSTM-Attention	Proposed
	Processing Time			
S1	13.765	18.876	7.567	1.675
S2	15.876	12.765	9.286	3.876
S3	18.987	15.664	10.334	5.987
S4	16.667	15.776	9.443	4.987
S5	20.667	18.222	14.223	5.999
Average	17.19	16.26	10.17	4.50

Statistical Analysis

In the next section, the provided dataset provides a strong basis for exploring the application of deep learning and signal processing to counteract cybersickness in VR. It collected 1000 entries reflecting user characteristics, the type of VR headset used, the duration of each session, and users' self-reported levels of cybersickness and immersion, which are beneficial for determining the causes of cybersickness. Factors such as the degree of immersion, the severity of motion sickness, and session duration can indicate critical points that help maintain the necessary level of immersion and reduce discomfort. Moreover, differences in motion sickness across devices indicate where hardware or software improvements can reduce discomfort and enhance VR experiences. The proposed dataset is particularly well-suited to MLT-DNNs and SOBI. While MLT-DNNs can estimate motion sickness intensity and immersion simultaneously from dataset features, SOBI enables signal filtering to extract user-specific factors, such as head movements, that contribute to discomfort. This combination enables more accurate identification of cybersickness, laying the foundation for developing VR systems that adjust their content based on the user's feedback and physiological indicators. The results of this study help to solve not only the problem of cybersickness but also the problem of making VR technologies comfortable for everyone.

Age Distribution Analysis

Table 6 and Figure 9 present users' ages, providing information on the demographic characteristics of VR participants. The data is aggregated by age, with frequencies given across different age brackets. For instance, the dataset shows that the largest user group is between 20 and 30 years old, with a frequency of about 200 users, which means that young people are the main consumers of VR technology. On the other hand, users in the 40 to 50-year age group are relatively few, with approximately 80 users. There are also 50 users under 20, indicating that teenagers are interested in VR technology. However, participation declines among users over 60, with less than 5% of the dataset in this category. Such an age-related trend might be explained by such factors as familiarity with technology, its availability, and a focus on the immersive experience among the younger population. The distribution of the overall data by age shows that

Table 6
The frequency of users across different age groups

S. No.	Age	Frequency
1	18	16
2	19	21
3	20	23
4	21	14
5	22	11
.	.	.
.	.	.
.	.	.
42	26	17
43	27	30

VR is not exclusive to any age group but is more prevalent among the younger generation. These are crucial findings for VR developers to understand their target market, create the right products, and market them accordingly.

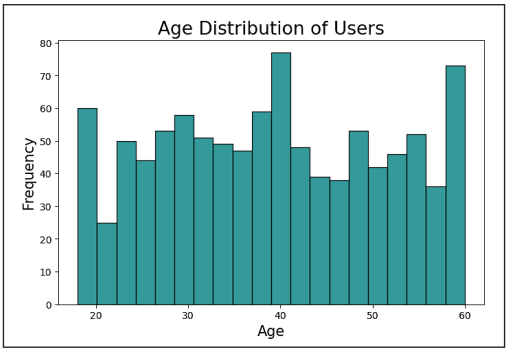


Figure 9. The frequency of users across different age groups

VR Headset Popularity Analysis

Table 7 and Figure 10 present an analysis of VR headset popularity, highlighting user preferences across three primary devices: HTC Vive, Oculus Rift, and PlayStation VR. The distribution of these headsets in the dataset is relatively equal, with Oculus Rift slightly dominating, accounting for 345 respondents. The HTC Vive is second with 336 users, while the PlayStation VR has 325 users. This means there is no single headset with a large user base in virtual reality hardware. The slight edge that Oculus Rift has over its competitors can be attributed to its availability and to certain features that might be more appealing to the masses. The HTC Vive, known for its tracking performance, remains a formidable contender for those seeking a high level of immersion and precision. PlayStation VR is also slightly less popular, but it still has a considerable market share, which can be attributed to the compatibility with PlayStation consoles and relatively low price for the gaming audience. The near-even distribution of these headsets suggests that users have different preferences, which may depend on price, available content, and compatibility. Such findings could help manufacturers and content producers to better understand how to market their products to a broader audience or satisfy specific preferences. More insights into this would be gained if additional information about the users of these headsets were available.

Table 7
The number of users for each VR headset

S. No.	Number of Users	VR Headset
1	Oculus Rift	345
2	HTC Vive	336
3	PlayStation VR	325

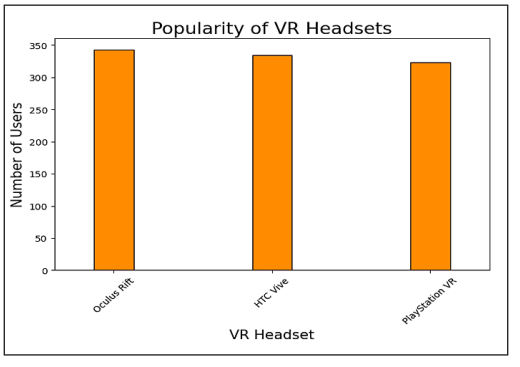


Figure 10. The number of users for each VR headset

Gender Distribution Analysis

Table 8 and Figure 11 present the dataset, including information on the gender of users who engaged with VR. Of the total users, the majority are males, constituting about 65% of the sample, while females are about 35%. This means that the gender distribution of users in VR is imbalanced, with male users more dominant than female users by a margin of 1.8:1. Such a disparity could be attributed to several factors, including differences in interests in technology, games, or other immersive experiences. Knowledge of this distribution is essential for VR developers and marketers who want to design content that is appealing to minorities and targets them effectively. For instance, developing VR experiences that target specific niches might attract more female users and help reduce the gender gap. Furthermore, this gender perspective might influence headset design, interface interactivity, and content selection, making VR technology more appealing to a wider public. In sum, the results presented shed light on the potential to improve gender representation in VR and make it more gender-balanced.

Table 8
The count of users by gender

S. No.	Number of Users	Gender
1	348	Others
2	330	Male
3	326	Female

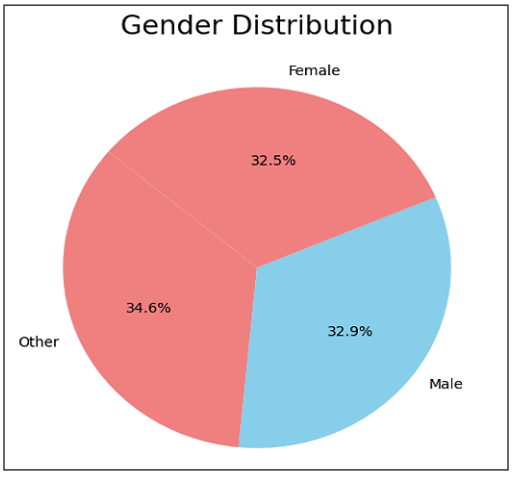


Figure 11. The count of users by gender

Table 9
Classification session lengths into ranges and their individual frequencies

S. No.	Duration Range	Frequency
1	(27.987, 31.456]	119
2	(48.978, 54.476]	112
3	(11.516, 17.003]	106
4	(38.927, 43.502]	101
5	(31.498, 38.893]	103
6	(21.501, 25.936]	98
7	(5.927, 11.423]	97
8	(44.493, 48.738]	94
9	(55.593, 59.963]	90
10	(16.002, 22.522]	87

Session Duration Distribution

Table 9 and Figure 12 present the session durations of VR experiences, indicating that the distribution of user time in VR differs significantly. The data was split into 10 equal-duration categories. A significant number of users, especially in the short-session category (less than 10 minutes), indicated that their use of the site was brief, exploratory, or casual. The most frequent range, however, is between 20 and 30 minutes, suggesting a sweet spot where users feel comfortable and

engaged before fatigue or other factors set in. Longer sessions, extending beyond 50 minutes, were relatively rare, representing only a small percentage of users. This could be attributed to motion sickness, hardware discomfort, or a natural tendency among users to limit their VR exposure in a single sitting. Notably, mid-range durations (30–40 minutes) still saw significant activity, reflecting deeper engagement with content such as games or simulations. These insights suggest that VR developers might consider optimising experiences for the 20–30-minute range to align with common user behaviour. Additionally, improving comfort and content for longer durations could help capture the smaller but valuable segment of extended-session users.

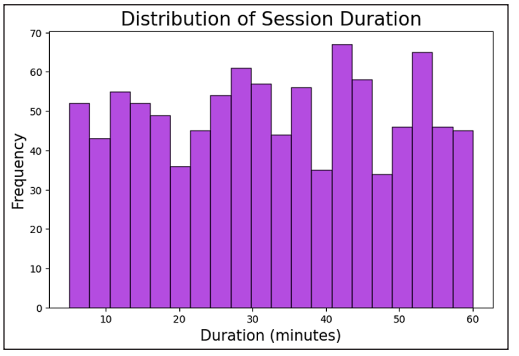


Figure 12. Classification session lengths into ranges

Session Duration by VR Headset Analysis

Table 10 and Figure 13 show that session durations vary significantly across VR headsets. Users of the HTC Vive had an average session duration of 34.7 minutes, with a standard deviation of 16.4 minutes, indicating a moderately variable range of session lengths. The shortest session observed was 6.0 minutes, while the longest reached 68.3 minutes. The median session duration was 35.6 minutes, suggesting that typical users spend approximately this amount of time in VR. For the Oculus Rift, the average session duration was slightly lower, at 34.1 minutes, with a standard deviation of 16.9 minutes, showing a similar range of session diversity. Sessions ranged from 6.5 minutes to 65.8 minutes, with a median duration of 31.0 minutes. Lastly, users of the PlayStation VR had the shortest average session duration, at 29.7 minutes, with a standard deviation of 16.8 minutes. Sessions ranged from 6.1 minutes to 60.1 minutes, with a median duration of 29.7 minutes. This comparison shows that HTC Vive users tend to engage in longer VR sessions, while PlayStation VR users tend to have shorter sessions on average.

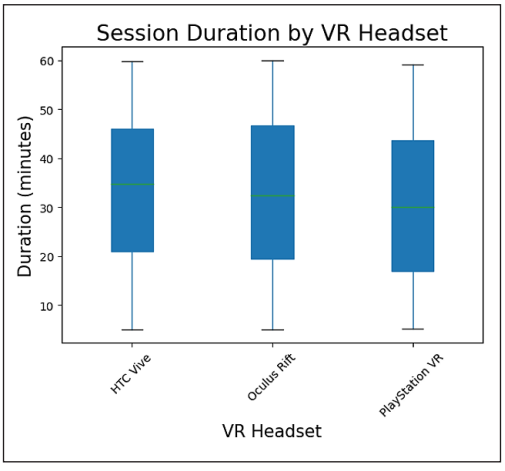


Figure 13. Durations for each VR headset

Table 10
Statistics (mean, median, quartiles) of session durations for each VR headset

S. No.	VR Headset	count	mean	std	min	25%	50%	75%	max
1	HTC Vive	335.0	34.78623	16.42787	6.003567	22.07534	35.65768	47.35789	60.88746
			97724418	66789976	68789890	56963435	62334577	90227889	67878990
			54	89	6	67	89	35	91
2	Oculus Rift	343.0	34.12249	16.96542	6.509987	17.49992	33.48966	47.34787	60.75664
			75398546	56892684	52457886	87567799	46587882	89646721	53367899
			89	57	7	78	56	94	92
3	PlayStation VR	323.0	29.78755	16.84567	6.157468	15.78843	29.78422	44.73678	60.18442
			34013789	99900365	88900787	29075799	45768907	78934328	46678999
			55	90	8	53	65	76	01

Motion Sickness vs Immersion Level Correlation Analysis

Table 11 and Figure 14 illustrate the correlation between motion sickness and immersion level, providing insights into their relationship. The correlation coefficient between motion sickness and immersion level is calculated as -0.12, indicating a weak negative correlation. This suggests that as motion sickness slightly increases, immersion levels may tend to decrease, though the effect is minimal. Further analysis of the correlation matrix reveals a self-correlation of 1.0 for both motion sickness and immersion level, as expected, since they are perfectly correlated with themselves. However, the cross-correlation value indicates that the relationship between these two variables is not strongly linear, suggesting that other factors may influence immersion levels in VR experiences. This weak negative correlation provides a basis for exploring whether reducing motion sickness symptoms could enhance user immersion, thereby informing future design improvements in VR applications and hardware.

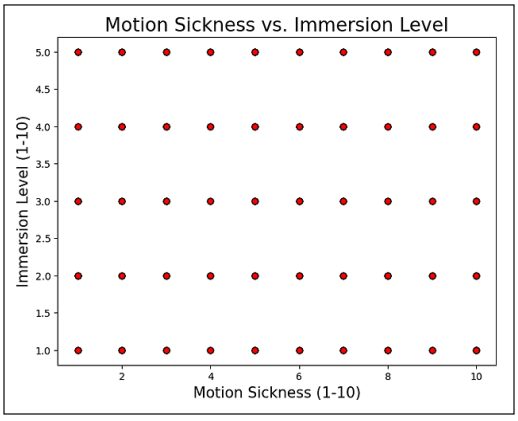


Figure 14. Correlations between motion sickness and immersion levels

Table 11
Correlations between motion sickness and immersion levels

S. No.	Metric 1	Metric 2	Correlation
1	Motion Sickness	Motion Sickness	1.0000000
2	Motion Sickness	Immersion Level	-0.038766778099755
3	Immersion Level	Motion Sickness	-0.038766778099755
4	Immersion Level	Immersion Level	1.00000000

Motion Sickness by VR Headset Analysis

Table 12 and Figure 15 show that the analysis of motion sickness levels across different VR headsets reveals interesting insights. On average, PlayStation VR users reported the highest motion sickness score of 5.73, with a standard deviation of 2.75. This suggests slightly greater variability in user experiences compared to other headsets. In contrast, HTC Vive users reported an average motion sickness score of 5.48, while Oculus Rift users had the lowest average score of 5.38, both with comparable variability (standard deviations of 2.96 and 2.88, respectively). The minimum motion sickness level was 1 across all headsets, indicating that some users experienced no symptoms. The maximum score of 10 was also consistent across headsets, showing that all devices caused severe motion sickness for a subset of users. Median scores indicate that the typical experience involved moderate symptoms, with medians ranging from 5 to 6 across devices. This data highlights that motion sickness is a significant consideration for all VR devices, though variations exist between headsets. PlayStation VR may warrant further investigation due to its slightly higher average motion sickness.

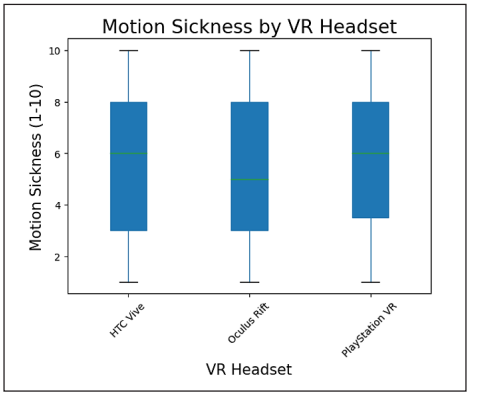


Figure 15. Descriptive statistics of motion sickness levels for each VR headset

Table 12
Descriptive statistics of motion sickness levels for each VR headset

S. No.	VR Headset	count	mean	std	min	25%	50%	75%	max
1	HTC Vive	335.0	5.476899 0456578 27	2.945678 7985434 77	1.0	3.0	6.0	8.0	10.0

Table 12 (*continued*)

S. No.	VR Headset	count	mean	std	min	25%	50%	75%	max
2	Oculus Rift	334.0	5.378752 1986645 78	2.856758 9464657 89	1.0	3.0	5.0	8.0	10.0
3	PlayStation VR	324.0	5.772389 7664323 79	2.746932 6789975 34	1.0	3.0	6.0	8.0	10.0

Ablation Study

The proposed model's modules each perform an important role. In this section, we conduct a series of ablation studies on cybersickness detection to compare the proposed MLT-DNN with current models, including 3D-CNN, SVM+TDA, and LSTM-Attention. To study performance development and validate our proposed MLT-DNN, new features derived from the model are gradually incorporated.

Table 13 and Figure 16 present an ablation study on the accuracy of various model configurations for cybersickness detection. The baseline model achieves 87.56% accuracy, and using only SOBI achieves 80.55%. The MTL-DNN configuration improves accuracy to 94.55%. Nevertheless, the accuracy of the Proposed Framework, which combines SOBI and MTL-DNN, is the highest at 97.16%, demonstrating that integrating such advanced signal pre-processing with multitask learning is beneficial. These outcomes indicate that the suggested framework increases the performance of cybersickness detection.

Table 13
Ablation study for accuracy analysis

Model Configuration	Accuracy
Baseline	87.56
SOBI Only	80.55
MTL-DNN Only	94.55
Proposed Framework	97.16

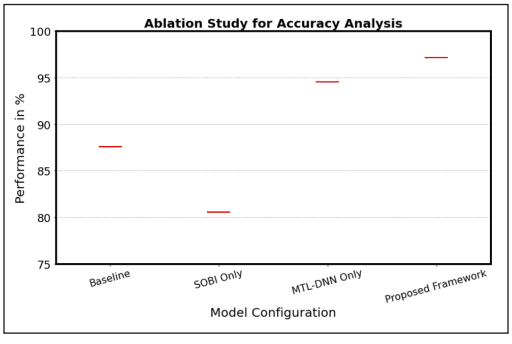


Figure 16. Ablation study for accuracy analysis

Table 14 and Figure 17 compare the accuracy of the proposed method with other state-of-the-art methods for physiological data-based assessment. The authors of Garcia-Agundez et al. (2019) achieved an accuracy of 82% using the SVM, KNN, and NN algorithms on ECG, EOG, EDA, and RESP signals. Pane et al. (2018) achieved 88.9% accuracy using EEG

data with SVM and CNN. Kim et al. (2019) improved the accuracy to 90.4% using a CNN-RNN architecture with video and EEG signals. Yalcin et al. (2024) achieved 91.7% accuracy using a Bidirectional LSTM and a CNN with multimodal physiological signals, including ECG, ACC, EDA, BVP, and TEMP. Such progression is attributed to the enhanced efficiency of the complex neural network architecture and multi-modal data fusion in enhancing accuracy.

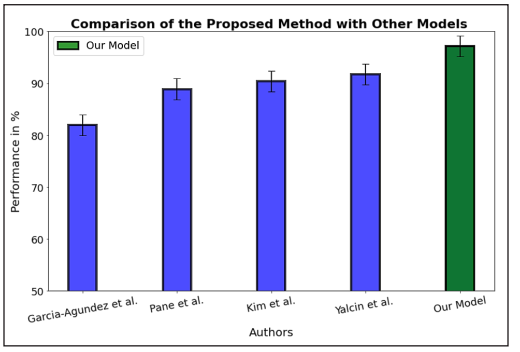


Figure 17. Comparison of the proposed method with the state-of-the-art methods for accuracy analysis

Table 14
Comparison of the proposed method with the state-of-the-art methods for accuracy analysis

Author's	Setup	Physio.Data	Methods	Accuracy
Garcia-Agundez et al., 2019	HMD, Electrode-based	ECG, EOG, EDA, RESP	SVM, KNN, NN	82
Pane et al., 2018	EEGSetup	EEG	SVM, CNN	88.9
Kim et al., 2019	EEGSetup	Video, EEG	CNN-RNN	90.4
Yalcin et al., 2024	HMD,PolarH10, Empatica	ECG, ACC, EDA, BVP, TEMP	Bid-LSTM, CNN	91.7
Our Model	128 GB memory, Intel i9-10900X CPU, and Nvidia RTX 3090 GPU	-	MLT-DNN	97.16

Limitations & Challenges

The application of deep learning and signal processing methods, including MTL-DNN and SOBI, in cybersickness detection has some important limitations and challenges. However, there are some drawbacks to the proposed MTL-DNN model: its high computational complexity may limit its real-time application on low-power devices. It also suggests that the assumption of signal independence, on which the method is based, may not be as advantageous for SOBI when noise levels are high and correlated or overlapping. Furthermore, for deep learning models, acquiring a large, high-quality EEG dataset is challenging because data collection in VR environments can be costly and time-consuming. The nature of physiological individual differences makes it very hard to apply the model across users, hence the need for flexibility in the approaches used for different users. These difficulties imply that more work is needed to extend the system's size and robustness, and to understand how to personalise it.

CONCLUSION

In conclusion, the effectiveness of combining advanced deep learning techniques with signal processing methods for the real-time detection of cybersickness in VR environments is demonstrated. By integrating an MTL-DNN with SOBI, the proposed framework achieves superior accuracy and robustness compared to traditional approaches. SOBI effectively pre-processes EEG signals by reducing artefacts and isolating meaningful independent components, while MTL-DNN leverages multitask learning to simultaneously predict cybersickness severity and related physiological states. Validated on benchmark EEG datasets, the framework shows enhanced performance in both detection accuracy and computational efficiency, offering a scalable solution for real-time monitoring. This synergy between deep learning and signal processing sets a new standard for objective detection of cybersickness, paving the way for improved user experiences in human-computer interaction.

FUTURE WORK

Future work will involve applying this framework to multimodal data, including eye tracking, heart rate, and motion sensors, to provide a more accurate estimate of cybersickness. Moreover, investigating an adaptive learning approach to identify cybersickness based on the user's EEG patterns and VR scenarios will expand the system's applicability. Adapting this framework with actual VR applications will also be a step towards designing user-oriented, realistic environments.

ACKNOWLEDGEMENT

Author Contributions: All authors contributed equally, and all authors read and approved the final version of the paper.

REFERENCES

Ahir, K., Govani, K., Gajera, R., & Shah, M. (2020). Application on virtual reality for enhanced education learning, military training and sports [Retracted article]. *Augmented Human Research, 5*, Article 7. https://doi.org/10.1007/s41133-019-0025-2

Alam, M. N., Ibrahimy, M. I., & Motakabber, S. M. A. (2021). Feature extraction of EEG signal by power spectral density for motor imagery based BCI. In *2021 8th International Conference on Computer and Communication Engineering (ICCCE)* (pp. 234-237). IEEE. https://doi.org/10.1109/iccce50029.2021.9467141

Caserman, P., Garcia-Agundez, A., Gámez Zerban, A., & Göbel, S. (2021). Cybersickness in current-generation virtual reality head-mounted displays: Systematic review and outlook. *Virtual Reality, 25*(4), 1153-1170. https://doi.org/10.1007/s10055-021-00513-6

Chang, C.-H., Lin, J., Chang, J.-W., Huang, Y.-S., Lai, M.-H., & Chang, Y.-J. (2024). Hybrid deep neural networks with multi-tasking for rice yield prediction using remote sensing data. *Agriculture, 14*(4), Article 513. https://doi.org/10.3390/agriculture14040513

Chang, E., Billinghurst, M., & Yoo, B. (2023). Brain activity during cybersickness: A scoping review. *Virtual Reality, 27*(3), 2073-2097. https://doi.org/10.1007/s10055-023-00795-y

Garcia-Agundez, A., Reuter, C., Becker, H., Konrad, R., Caserman, P., Miede, A., & Göbel, S. (2019). Development of a classifier to determine factors causing cybersickness in virtual reality environments. *Games for Health Journal, 8*(6), 439-444. https://doi.org/10.1089/g4h.2019.0045

Hadadi, A., Guillet, C., Chardonnet, J.-R., Langovoy, M., Wang, Y., & Ovtcharova, J. (2022). Prediction of cybersickness in virtual environments using topological data analysis and machine learning. *Frontiers in Virtual Reality, 3*, Article 973236. https://doi.org/10.3389/frvir.2022.973236

Heo, J., & Yoon, G. (2020). EEG studies on physical discomforts induced by virtual reality gaming. *Journal of Electrical Engineering & Technology, 15*(3), 1323-1329. https://doi.org/10.1007/s42835-020-00373-1

Joshi, A. (2023). *Virtual reality experiences* [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/5716193

Kim, J., Kim, W., Oh, H., Lee, S., & Lee, S. (2019). A deep cybersickness predictor based on brain signal analysis for virtual reality contents. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 10580-10589). IEEE. https://doi.org/10.1109/iccv.2019.01068

Kourtesis, P., Papadopoulou, A., & Roussos, P. (2024). Cybersickness in virtual reality: The role of individual differences, its effects on cognitive functions and motor skills, and intensity differences during and after immersion. *Virtual Worlds, 3*(1), 62-93. https://doi.org/10.3390/virtualworlds3010004

Krokos, E., & Varshney, A. (2022). Quantifying VR cybersickness using EEG. *Virtual Reality, 26*(1), 77-89. https://doi.org/10.1007/s10055-021-00517-2

Lee, T. M., Yoon, J.-C., & Lee, I.-K. (2019). Motion sickness prediction in stereoscopic videos using 3D convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics, 25*(5), 1919-1927. https://doi.org/10.1109/tvcg.2019.2899186

Liu, X., Wang, H., Huang, M., & Yang, W. (2019). An improved second-order blind identification (SOBI) signal de-noising method for dynamic deflection measurements of bridges using ground-based synthetic aperture radar (GBSAR). *Applied Sciences, 9*(17), Article 3561. https://doi.org/10.3390/app9173561

Martin, N., Mathieu, N., Pallamin, N., Ragot, M., & Diverrez, J.-M. (2020). Virtual reality sickness detection: An approach based on physiological signals and machine learning. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 387-399). IEEE. https://doi.org/10.1109/ismar50242.2020.00065

Martirosov, S., Bureš, M., & Zítka, T. (2022). Cyber sickness in low-immersive, semi-immersive, and fully immersive virtual reality. *Virtual Reality, 26*(1), 15-32. https://doi.org/10.1007/s10055-021-00507-4

Pane, E. S., Khoirunnisaa, A. Z., Wibawa, A. D., & Purnomo, M. H. (2018). Identifying severity level of cybersickness from EEG signals using CN2 rule induction algorithm. In *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)* (Vol. 3, pp. 170-176). IEEE. https://doi.org/10.1109/iciibms.2018.8549968

Papaefthymiou, S., Giannakopoulos, A., Roussos, P., & Kourtesis, P. (2024). Mitigating cybersickness in virtual reality: Impact of eye–hand coordination tasks, immersion, and gaming skills. *Virtual Worlds, 3*(4), 506-535. https://doi.org/10.3390/virtualworlds3040027

Porcino, T., Trevisan, D., & Clua, E. (2020). Minimizing cybersickness in head-mounted display systems: Causes and strategies review. In *2020 22nd Symposium on Virtual and Augmented Reality (SVR)* (pp. 154-163). IEEE. https://doi.org/10.1109/svr51698.2020.00035

Qu, C., Che, X., Ma, S., & Zhu, S. (2022). Bio-physiological-signals-based VR cybersickness detection. *CCF Transactions on Pervasive Computing and Interaction, 4*(3), 268-284. https://doi.org/10.1007/s42486-022-00103-8

Sameri, J., Coenegracht, H., Van Damme, S., De Turck, F., & Torres Vega, M. (2024). Physiology-driven cybersickness detection in virtual reality: A machine learning and explainable AI approach. *Virtual Reality, 28*(4), Article 174. https://doi.org/10.1007/s10055-024-01067-z

Teixeira, J., & Palmisano, S. (2021). Effects of dynamic field-of-view restriction on cybersickness and presence in HMD-based virtual reality. *Virtual Reality, 25*(2), 433-445. https://doi.org/10.1007/s10055-020-00466-2

Yalcin, M., Halbig, A., Fischbach, M., & Latoschik, M. E. (2024). Automatic cybersickness detection by deep learning of augmented physiological data from off-the-shelf consumer-grade sensors. *Frontiers in Virtual Reality, 5*, Article 1364207. https://doi.org/10.3389/frvir.2024.1364207

Yang, A. H. X., Kasabov, N., & Cakmak, Y. O. (2022). Machine learning methods for the study of cybersickness: A systematic review. *Brain Informatics, 9*(1), Article 24. https://doi.org/10.1186/s40708-022-00172-6

Zaidi, S. F. M., Shafiabady, N., & Beilby, J. (2023). Identifying presence of cybersickness symptoms using AI-based predictive learning algorithms. *Virtual Reality, 27*(4), 3613-3620. https://doi.org/10.1007/s10055-023-00813-z

Pertanika J. Sci. & Technol. 34 (S3): 101 - 124 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

GenCKD-Net: A Lightweight Multilayer Perceptron Pipeline for Chronic Kidney Disease Classification Using the UCI CKD Dataset

Karimulla Sha Shaik[1]*, Konda Hari Krishna[1], and K. K. Baseer[2]

[1]Department of Computer Science and Engineering, School of Computing, Mohan Babu University, Sree Sainath Nagar, 517102 Tirupati, Andhra Pradesh, India

[2]Department of Computer Science and Systems Engineering, GITAM School of Computer Science and Engineering, GITAM (Deemed to be a University), NH 207, Nagadenehali, Doddaballapur Taluk, 561203 Bengaluru, Karnataka, India

ABSTRACT

Binary classification of chronic kidney disease (CKD) using standard clinical measurements represents a relevant machine-learning application, but such estimates may not generalise to independent clinical populations. GenCKD-Net, a lightweight multilayer perceptron pipeline for binary CKD classification using the publicly available UCI Chronic Kidney Disease dataset. The dataset contains 400 records and 24 predictor variables, including 250 CKD cases (62.5%, majority class) and 150 non-CKD controls (37.5%, minority class). The dataset was divided by stratified sampling into training (64%), validation (16%), and held-out test (20%) subsets. Missing-value imputation, categorical encoding, and numerical standardisation were fitted using the training data only, while SMOTE was applied only to the transformed training subset. GenCKD-Net uses a compact two-hidden-layer MLP with 128 and 64 neurons and is implemented in scikit-learn with warm-start training across 10 single-epoch fitting cycles. Classification threshold selection utilised validation data, prioritising specificity, and was not selected from the test set. On the held-out test set (n=80), GenCKD-Net achieved ROC-AUC = 0.998, average precision = 1.000, accuracy = 0.975, precision = 1.000, recall = 0.960, and F1-score = 0.980. Beginning with baseline classifiers, all architectures achieved comparable results, indicating that achieving optimal discrimination on this benchmark can be accomplished by many different architectures. The results indicate that GenCKD-Net is a research-stage tool with potential for use in benchmarking and direct clinical application. To translate into a clinical setting, further validation

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 29 September 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.05

E-mail addresses:
karimullashah@gmail.com (Karimulla Sha Shaik)
konda.harikrishna@mbu.asia (Konda Hari Krishna)
kkamalap@gitam.edu.in (K. K. Baseer)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

would include independent cohort verification, repeated cross-validation, a calibration study, and prospective studies.

Keywords: Average precision, baseline comparison, chronic kidney disease, class imbalance, machine learning, multilayer perceptron, ROC-AUC, SMOTE

INTRODUCTION

Managing the growing problem of chronic kidney disease (CKD), which has become a serious challenge for nephrology, is another issue because it requires the identification of those with CKD who could potentially need additional evaluation. The health care system generates an immense amount of clinical information, including laboratory test results and comorbidities, that can be used to automate patient classification and risk assessment. While there have been many recent reviews of the use of artificial intelligence (AI) and machine learning in chronic kidney disease (CKD) that indicate potential applications of these computing methods, they also highlight several challenges, including maintaining high levels of data quality, ensuring that users can understand the results produced by the models, and validating the output through clinical trials (Gogoi & Valan, 2025; Sabanayagam et al., 2025; Sanmarchi et al., 2023).

Motivation: Burden of CKD and Need for Early Detection

Automated detection of CKD through clinical data is an important public health issue. Therefore, this paper utilises the standard UCI benchmark for identifying patients with CKD from routine clinical data, employs a binary approach to classify patients as having CKD or not, and does NOT attempt to model longitudinal disease progression or outcomes.

Challenges: Class Imbalance and Model Interpretability

Clinical classification problems often involve class imbalances. The UCI dataset used in this research has approximately 62.5% of the data classified as CKD and approximately 37.5% as not having CKD. Although increasing the amount of representation of a minority class will allow you to better utilise SMOTE, care must be taken when using SMOTE to prevent leaking information into your test set by applying it only to your training data (Burra et al., 2025).

Figure 1 uses a wheel-style layout to show common CKD symptoms—fatigue, poor appetite, nausea, swelling (oedema), frequent urination, sleep disturbances, itching, and shortness of breath. Because these signs are often nonspecific and can appear early or worsen as kidney function declines, noticing them together with routine lab results can prompt earlier evaluation and referral.

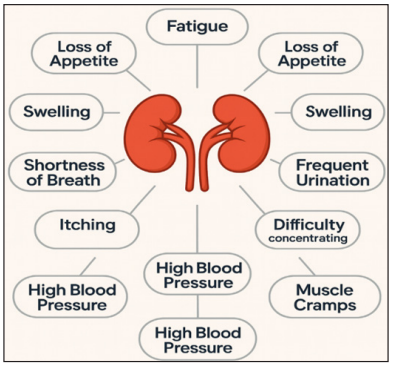


Figure 1. Common symptoms of Chronic Kidney Disease (CKD)

Contributions

GenCKD-Net is a lightweight multilayer perceptron pipeline for reproducible CKD classification on the UCI CKD benchmark. The contribution is methodological and empirical rather than architectural: the study evaluates a compact MLP within a leakage-controlled preprocessing and training workflow and compares its performance directly with established machine-learning baselines under the same evaluation framework.

Figure 2 presents the conceptual GenCKD-Net workflow used in this study: routine tabular clinical features undergo leakage-controlled preprocessing, training-only SMOTE, and classification by a compact 128→64 MLP. Imaging, wearables, explainability, and deployment components are not part of this experiment.

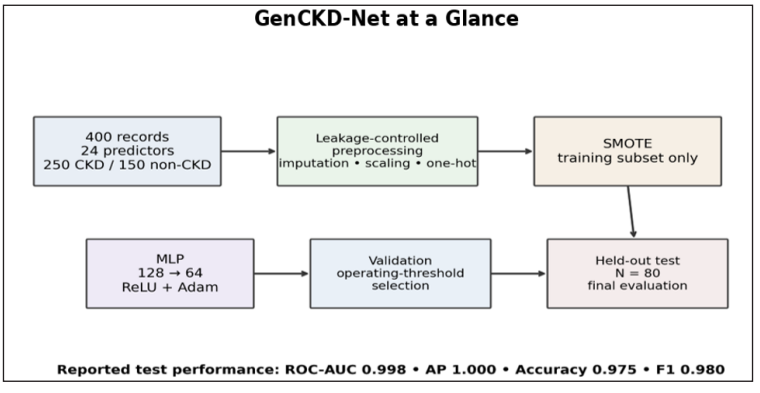


Figure 2. GenCKD-Net: Leakage-controlled tabular CKD classification

Related Work

Computational analysis of kidney disease classification has long employed a variety of traditional machine learning approaches. The most common have included Support Vector Machines (SVMs), Logistic Regression (LR), Random Forests (RFs), and Gradient Boosted Models (GBMs). These methods are well represented as very successful at predicting disease based on the collected data. However, many recent reviews have highlighted variability in data collection and validation procedures across studies, making it difficult to compare results from one study to another (Akter et al., 2023; Gogoi & Valan, 2025; Iliyas et al., 2025; Sanmarchi et al., 2023).

AI/ML in CKD Screening and Prognosis; Effects of Class Imbalance; Limits of Black-Box DL

A large body of CKD modelling relies on routinely collected EHR and laboratory data to flag disease and forecast progression. Recent work shows that benchmark datasets can yield strong performance, but differences in validation design and population hinder direct comparisons across studies. This study, therefore, emphasises baseline comparability and cautious interpretation rather than claims of clinical superiority. The broader evidence base includes systematic reviews of incident CKD prediction in community electronic health records (Haris et al., 2024) and AI-based CKD prognosis (Pan & Tong, 2024), both of which underscore the importance of validation design and population context. (Miller & Dwyer, 2025; Nadarajah et al., 2023; Su et al., 2022)

Explainable AI for Tabular Health Data (Permutation, SHAP, Counterfactuals)

Comprehensible model decisions facilitate clinical acceptance and appropriate utilisation, and under what conditions may they fail? However, permutation importance, SHAP, and counterfactual explanations are not experimentally evaluated in this study; they are identified as possible future extensions rather than reported findings.

Explainability can support model auditing by identifying which input variables contribute to predictions and by revealing potential instability or subgroup-specific behaviour. However, GenCKD-Net did not experimentally evaluate permutation importance, SHAP, counterfactual explanations, or clinician-facing interfaces. These methods are therefore treated as future analyses that should be performed only after the predictive model has been evaluated across repeated splits and independent cohorts (Lekadir et al., 2025; Rosenbacke et al., 2024). Reviews of AI in kidney disease and nephrology also emphasise the challenges of interpretability, clinical integration, and validation (Ali et al., 2022; Wu et al., 2024; Singh et al., 2025).

Multimodal approaches using imaging, wearables, point-of-care measurements, or omics may provide additional information for CKD research. In this study, however, only tabular clinical variables from the UCI CKD benchmark were evaluated. Multimodal fusion is therefore considered future work requiring separate data, validation, and ablation studies.

Multimodal approaches are increasingly being investigated to characterise renal disease by combining tabular clinical variables with imaging, longitudinal sensor measurements, pathology, or molecular data. Imaging-based studies have explored automated extraction of renal structural and pathological characteristics, while wearable and point-of-care technologies provide potentially useful longitudinal signals. However, multimodal models require appropriately aligned patient-level data, modality-specific preprocessing, adequate sample sizes, and independent validation. These requirements are substantially different from those of this study, which is restricted to tabular clinical variables. Therefore,

multimodal fusion is considered a future extension rather than a demonstrated capability of GenCKD-Net (Han et al., 2024; Puri et al., 2025; Zhang et al., 2024). Imaging-focused CKD studies provide a distinct evidence stream from the tabular benchmark used here; examples include kidney MRI classification and segmentation (Islam et al., 2024) and broader retinal-image AI screening research (Fang et al., 2025; Sághy et al., 2025; Zhang, 2025)

MATERIALS AND METHODS

Dataset

The data for this study were based on the Chronic Kidney Disease Data Set available in the UCI Repository. The data set is identified as #336 by Rubini et al. (2015); it has 400 samples with 24 features or predictors, which are a combination of demographic information about patients, their vital signs, urinalysis, blood biochemistry tests, and various blood test parameters that can be used to assess kidney function and other medical conditions related to kidney disease. Of these, there were 250 CKD (chronic kidney disease) cases, accounting for 62.5% of all records, and 150 non-CKD (non-chronic kidney disease) cases, accounting for 37.5%. The dataset description and benchmark provenance are consistent with the UCI CKD dataset record (Rubini et al., 2015); this study does not treat the dataset as a longitudinal cohort.

Table 1 summarises the predictor schema used in the UCI CKD benchmark, including numerical laboratory/vital measurements and categorical clinical and urine-related attributes. The dataset contains serum creatinine, blood urea, haemoglobin, electrolyte measurements, blood pressure, and categorical findings such as albumin, sugar, red blood cell status, pus cell status, and oedema. Several features contained missing entries; numerical features were imputed with the median value from the training set, while categorical features were imputed with the most frequent category, both imputed values fitted only on the training data. Numerical variables were standardised using training-set statistics, and categorical variables were one-hot encoded before training-only SMOTE was applied.

Pre-Processing

Careful preprocessing was used to prevent potential "information leakage" during testing/validation by preventing information from being carried over from the training set. The entire data set was split by stratified partitioning, creating three subsets: a training subset (64% of the data), a validation subset (16%), and a testing subset (20%). All pre-processed transformation models were fit exclusively on the training subset. After that, these pre-processed transformation models were applied to both the validation and test subsets, and no further fitting was performed. Numerical missing values were replaced using the median

Table 1
Dataset schema and feature summary

id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc
0	48	80	1.02	1	0	-	normal	notpresent	notpresent	121	36	1.2	-	-	15.4	44	7800	5.2
1	7	50	1.02	4	0	-	normal	notpresent	notpresent	-	18	0.8	-	-	11.3	38	6000	-
2	62	80	1.01	2	3	normal	normal	notpresent	notpresent	423	53	1.8	-	-	9.6	31	7500	-
3	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111	2.5	11.2	32	6700	3.9
4	51	80	1.01	2	0	normal	normal	notpresent	notpresent	106	26	1.4	-	-	11.6	35	7300	4.6
5	60	90	1.015	3	0	-	-	notpresent	notpresent	74	25	1.1	142	3.2	12.2	39	7800	4.4
6	68	70	1.01	0	0	-	normal	notpresent	notpresent	100	54	24	104	4	12.4	36	-	-
7	24	-	1.015	2	4	normal	abnormal	notpresent	notpresent	410	31	1.1	-	-	12.4	44	6900	5
8	52	100	1.015	3	0	normal	abnormal	present	notpresent	138	60	1.9	-	-	10.8	33	9600	4
9	53	90	1.02	2	0	abnormal	abnormal	present	notpresent	70	107	7.2	114	3.7	9.5	29	12100	3.7
10	50	60	1.01	2	4	-	abnormal	present	notpresent	490	55	4	-	-	9.4	28	-	-
11	63	70	1.01	3	0	abnormal	abnormal	present	notpresent	380	60	2.7	131	4.2	10.8	32	4500	3.8
12	68	70	1.015	3	1	-	normal	present	notpresent	208	72	2.1	138	5.8	9.7	28	12200	3.4
13	68	70	-	-	-	-	-	notpresent	notpresent	98	86	4.6	135	3.4	9.8	-	-	-
14	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6
15	40	80	1.015	3	0	-	normal	notpresent	notpresent	76	162	9.6	141	4.9	7.6	24	3800	2.8
16	47	70	1.015	2	0	-	normal	notpresent	notpresent	99	46	2.2	138	4.1	12.6	-	-	-

value of the training data set. Numerical feature values were scaled using the mean and standard deviations of the training data set. Categorical missing values were replaced using the most common category from the training data set. Then, each categorical variable was one-hot encoded. A SMOTE-transformed version of the training matrix was created, after which no sampling was performed on either the validation or test matrices. Equations 1 to 5 formalise the numerical and categorical preprocessing operations described here; the fitted transformations are learned from the training subset and then applied unchanged to validation and test data.

Data normalisation and imputation stabilise the heterogeneous clinical variables before training:

$$x' = \frac{x - \text{Median}(X)}{\sigma_X} \quad [1]$$

$$x_{\text{imp}} = \begin{cases} x, & \text{if not missing} \\ \text{Median}(X), & \text{if missing} \end{cases} \quad [2]$$

Categorical variables are represented with one-hot encodings:

$$O_{i,j} = \begin{cases} 1, & \text{if } x_i = \text{category}_j \\ 0, & \text{otherwise} \end{cases} \quad [3]$$

Imputation of categorical attributes utilises mode frequency:

$$c_{\text{imp}} = \arg\max_k \ \text{freq}(c_k) \quad [4]$$

The complete preprocessing pipeline is described as

$$\mathbf{X}_{proc} = f_{\text{onehot}}\left(f_{\text{zscale}}\left(f_{\text{impute}}(\mathbf{X})\right)\right) \quad [5]$$

Imbalance Mitigation

Because CKD is the majority class (62.5%) in our dataset, class imbalance was addressed by applying SMOTE only to the minority class. SMOTE was applied only to the transformed training data after the train/validation/test split. The validation and test sets retained their original class distributions and were not resampled. The minority-class definition and training-only resampling procedure form the basis of the leakage-control protocol; Equation 9 specifies the synthetic-sample interpolation, and Equation 10 specifies the resulting training objective.

The non-CKD cases form the minority class (37.5%). Conventional SMOTE generates synthetic non-CKD examples by interpolation in the transformed feature space. Because SMOTE was applied after one-hot encoding, categorical-aware resampling, such as SMOTENC, remains a methodological sensitivity analysis for future work.

The preprocessing and imbalance-mitigation pipeline was implemented in leakage-controlled order: first, split the data using stratified sampling; then fit imputation, encoding, and scaling on the training subset only; transform the validation and test subsets without refitting; and finally apply SMOTE only to the transformed training matrix.

Algorithm 1: Preprocessing and Imbalance Mitigation

Input: Dataset D with numeric features N, categorical features C, labels Y
Output: Pre-processed training, validation, and test matrices

1: Begin
2: Split D into Train, Validation, and Test using stratified sampling
3: Fit numerical imputation, scaling, categorical imputation, and one-hot encoding on Train only.
4: Transform Validation and Test using the fitted training transformations
5: For each numeric feature n μN, replace missing values with the training-set median.
6: Standardise n' = (n − μN) / σN using training-set μN and σN
7: One-hot encode categorical variables using the training-fitted encoder
8: Apply SMOTE only to the transformed training matrix.
9: Keep validation and test unchanged.
10: Return processed train, validation, and test sets
11: End

Model Architecture

GenCKD-Net is a compact MLP with two hidden layers (128 → 64, ReLU). The model uses Adam optimisation and produces a positive-class probability score. The implementation does not include a separately calibrated probability layer; therefore, the output is treated as a model score/probability estimate rather than evidence of formal calibration. The model operations are summarised in Equations 11 and 12, which define the output mapping and the parameter count expression for the compact MLP.

The predictive component of GenCKD-Net uses a Multilayer Perceptron (MLP) with two hidden layers (128 and 64 neurons), implemented using scikit-learn. The compact architecture is intended as a lightweight benchmark model for tabular CKD classification.

Non-linear activations learn higher-level CKD representations:

$$h^{(1)} = \sigma\left(W^{(1)}x + b^{(1)}\right) \quad [9]$$

$$h^{(2)} = \sigma\left(W^{(2)}h^{(1)} + b^{(2)}\right) \quad [10]$$

The sigmoid function decodes output as disease probability:

$$\hat{y} = \frac{1}{1 + e^{-(W^{(3)}h^{(2)} + b^{(3)})}} \quad [11]$$

Total trainable parameters are

$$P = \sum_{l=1}^{L} (d_{l-1} \cdot d_l + d_l) \quad [12]$$

Training Protocol

The datasets were segregated by function. Training data were used to update parameters after training-only SMOTE was applied. Validation data were used to select the operating threshold. The test data remained untouched until the final evaluation. No early-stopping procedure was used in the implementation.

The model was optimised using binary cross-entropy through the scikit-learn MLP implementation with the Adam optimiser. Warm-start fitting was used with 1 iteration per fitting pass, and the fitting procedure was repeated for 10 recorded passes. The validation subset was used for operating-threshold selection; the held-out test subset was not used for model fitting or threshold optimisation.

The network minimises a cross-entropy objective as shown in Equations 13 and 14:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad [13]$$

Adaptive moment updates are performed as

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad [14]$$

Algorithm 2: GenCKD-Net Model Training Protocol. The implementation uses warm-start fitting with one iteration per fitting pass. Training and validation metrics are recorded for 10 passes; no automatic early-stopping criterion is applied in this experiment.

Implementation Details and Explainability Scope

This study evaluates only the classification pipeline and benchmark performance. Explainability mechanisms such as permutation importance, SHAP, and counterfactual analysis were not experimentally evaluated or presented as results.

Algorithm 2: GenCKD-Net Model Training

Input: Preprocessed training data Dtrain, number of fitting passes E = 10Output: Trained model parameters Θ*

1: Begin
2: Initialise model parameters and configure Adam
3: For pass = 1 to E do
4: Fit the MLP for 1 iteration using Dtrain.
5: Record training loss and accuracy
6: Evaluate validation loss and accuracy without parameter updates
7: End For
8: Return final model parameters Θ*.
9: End

No explainability module was executed as part of this experiment. Such methods are retained for future analyses after repeated internal validation and independent external evaluation.

Permutation importance is identified as a future global interpretation method; no permutation-importance results are reported.

SHAP is identified as a future patient-level explanation method; no SHAP values or feature-attribution results are reported.

Counterfactual analysis is identified as a future sensitivity analysis; no counterfactual results are reported.

Permutation-based feature significance as shown in Equations 16 to 18:

$$I_j = \frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{Acc}(D) - \mathrm{Acc}\left(D_{\pi(j)}\right)\right) \quad [16]$$

Shapley additive explanations are

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left(f(S \cup \{j\}) - f(S)\right) \quad [17]$$

The counterfactual sample search is formulated as

$$x^* = \arg\min_{x} \|x' - x\| s.t. f(x') \neq f(x) \quad [18]$$

Algorithm 3: Explainability Integration Module — Future Work

Algorithm 3 is a conceptual future work procedure for global and patient-level explainability. It was not executed in this study, and its outputs are not interpreted as reported experimental findings.

Algorithm 3: Conceptual Future Explainability Module

Input: Frozen trained model Θ* and an independently selected analysis dataset
Output: Candidate explanation summaries for future validation

1: Begin
2: Compute permutation importance on an analysis dataset
3: Compute SHAP values for representative samples
4: Generate counterfactual candidates subject to clinically plausible constraints
5: Store explanation outputs for qualitative and quantitative audits.
6: Validate explanation stability before any clinical interpretation
7: Return future explanation summaries
8: End

Multimodal Extension (Future Work)

This research focuses only on tabular UCI CKD data. Imaging, wearables/point-of-care measurements, omics, multimodal fusion, and clinical decision support deployment were not included in these experiments. These experiments are considered future research directions.

Figure 3 depicts the actual tabular experimental pipeline: preprocessing is fitted on the training subset, SMOTE is applied only to the training data, the MLP is trained using warm-start fitting, and validation is used to select the operating threshold before final test evaluation.

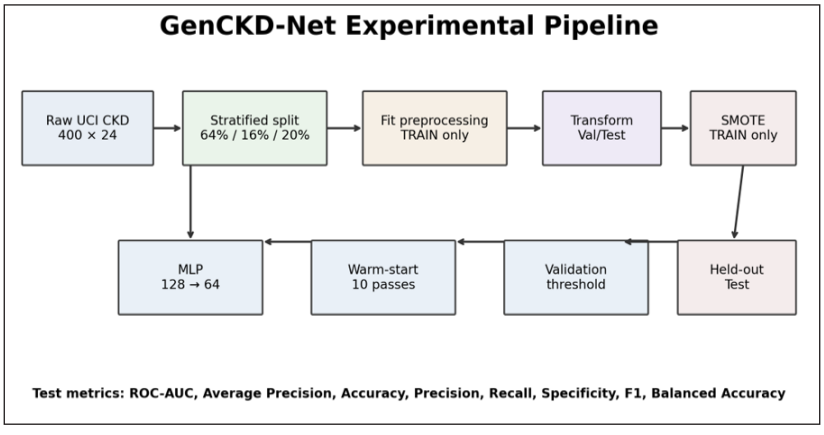


Figure 3. GenCKD-Net experimental pipeline

Evaluation Metrics

An extensive number of evaluation metrics were used to assess the GenCKD-Net system as follows: GenCKD-Net's accuracy was measured using a variety of metrics that assessed its ability to make accurate predictions (log-loss); evaluated GenCKD-Net's discrimination capability independent of threshold through the area under the receiver operating characteristic curve (ROC-AUC); provided the precision-recall average (PR-AP) as an additional metric to evaluate GenCKD-Net's performance when there is a substantial disparity in the amount of positive and negative outcome data; and utilised the confusion matrix (which provides true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN)) to increase the operational interpretability of the system. Additionally, we measured the time required to process each training epoch to assess the cost of training the GenCKD-Net system. Upon testing the GenCKD-Net system on our held-out test set, we found that GenCKD-Net achieved the following metrics: AUC = 0.998 (95% CI: 0.994-1.000), PR-AP ≈ 1.00, Accuracy = 0.975, Precision = 1.00, Recall = 0.96, F1 ≈ 0.98, Specificity = 1.00, and BA ≈ 0.98 with no FP and two FN. These results provide evidence that GenCKD-Net has clinical implications related to how a physician would prioritise clinical criteria to arrive at a diagnosis. For example, if the goal of GenCKD-Net is population screening, then increased recall would be clinically advantageous; however, based on our current findings, we believe that additional fine-tuning of the classification threshold may provide even greater opportunities to minimise missed cases while maintaining near-perfect precision. If the purpose of GenCKD-Net is to serve as a diagnostic tool to assist clinicians with confirming disease status, the absence of false positives provides substantial reassurance. Lastly, the results reported herein were derived from a single well-defined split of the data; therefore, our future research will include nested cross-validation, external validation, and calibration curves to establish the reliability and generalisability of our models across large datasets. All the performance metrics described above were evaluated using a set of metrics. These basic metrics included accuracy (Equation 19), precision (Equation 20), recall (Equation 21), and specificity (Equation 22). The harmonic mean of the first three was denoted as the F1 score (Equation 23), and balanced accuracy (Equation 24) was defined as the average of the recall and specificity to account for class imbalance. Threshold-independent evaluation metrics were provided by ROC-AUC (Equation 25) and Precision-Recall Average Precision (Equation 26). Probabilistic calibration was evaluated using log-loss (Equation 27), and the confusion matrix formulation (Equation 28) was provided to ensure operational clarity. Collectively, these metrics demonstrated GenCKD-Net's ability to discriminate and perform reliably under class imbalance. Figures 7–9 report the threshold-independent ROC and precision–recall curves and the threshold-dependent confusion matrix, respectively;

Figure 10 reports per-epoch wall-clock fitting time. These visualisations should be interpreted alongside the numerical results in Table 4, while Table 5 explicitly records that external validation was not performed.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad [19]$$

$$\text{Prec} = \frac{TP}{TP + FP} \quad [20]$$

$$\text{Rec} = \frac{TP}{TP + FN} \quad [21]$$

$$\text{Spec} = \frac{TN}{TN + FP} \quad [22]$$

$$F1 = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}} \quad [23]$$

$$BA = \frac{\text{Rec} + \text{Spec}}{2} \quad [24]$$

$$AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx \quad [25]$$

$$AP = \sum_n (R_n - R_{n-1})P_n \quad [26]$$

$$LL = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad [27]$$

$$CM = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad [28]$$

Figure 4 distinguishes the experimentally validated GenCKD-Net core from proposed future extensions. The validated component comprises leakage-controlled preprocessing, training-only SMOTE, a compact 128→64 MLP, validation-based operating-threshold selection, and held-out test evaluation. Formal calibration, explainability, multimodal fusion, and prospective deployment are not reported results.

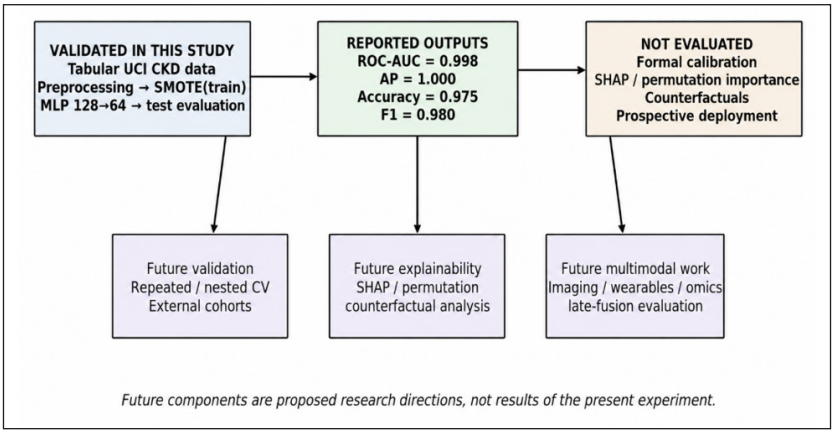


Figure 4. GenCKD-Net: Validated core and future extensions

RESULT AND DISCUSSIONS

Predictive Performance

For the test cohort, GenCKD-Net produced ROC-AUC = 0.998 and PR-AP = 1.000. At the validation-selected operating threshold, accuracy was 0.975, precision was 1.000, recall was 0.960, and the F1-score was 0.980. These results indicate high benchmark performance, but they should not be interpreted as evidence of clinical generaliability.

Learning Dynamics

Training and validation accuracy increased during the recorded warm-start fitting passes, while the recorded loss values decreased over the observed sequence. These observations describe optimisation behaviour on the split and do not establish generalisation or deployment readiness.

Complexity & Runtime

Each recorded fitting pass took approximately 0.2–1.6 s, with a mean of approximately 0.6 s in the implementation. These measurements indicate modest computational requirements for the evaluated benchmark dataset. They should not be interpreted as evidence of production deployment performance or routine clinical retraining.

Error Analysis

GenCKD-Net produced two false negatives and no false positives on the 80-record test cohort at the selected operating threshold. These counts are specific to this held-out split. Because formal patient-level explainability was not performed, case-specific physiological explanations are omitted.

Figure 5 shows that training and validation accuracy increased during the recorded fitting passes. The curves are presented as descriptive evidence of training behaviour; no automatic early-stopping criterion was implemented.

Figure 6 shows that training and validation loss decreased during the recorded fitting passes. Decreasing loss indicates progress in optimisation but does not establish formal probability calibration or generalisation.

Figure 7 is the ROC curve for the held-out test cohort and should be read as a summary of discrimination, not as evidence of external clinical validity.

The ROC curve yields an ROC-AUC of 0.998 in the test evaluation. This threshold-independent result should be interpreted as split-specific benchmark performance.

Figure 8 presents the precision–recall curve for the held-out test cohort; its interpretation is conditional on the observed test-set class distribution.

The precision-recall curve corresponds to AP = 1.000 for the test evaluation. The result should be interpreted in the 80-record held-out test cohort.

Figure 9 presents the held-out test confusion matrix at the selected operating threshold. The confusion matrix reports TP = 48, TN = 30, FP = 0, and FN = 2, yielding a precision of 1.00 and a recall of 0.96 at the selected cutoff. The absence of false positives was observed only on this held-out test split.

Figure 10 reports the recorded wall-clock fitting time per pass; it does not establish deployment latency or clinical readiness.

Figure 10 shows the observed wall-clock time for each recorded fitting pass. These measurements describe computational behaviour for the benchmark experiment and should not be interpreted as evidence of clinical deployment readiness.

Table 2 shows the stratified 64%/16%/20% train/validation/test split. The original class proportions were retained in all three subsets, while SMOTE was applied only to the transformed training data. Validation data were used for operating-threshold selection and remained otherwise untouched; the test subset was reserved for final evaluation. The resulting test confusion matrix contained TP = 48, TN = 30, FP = 0, and FN = 2.

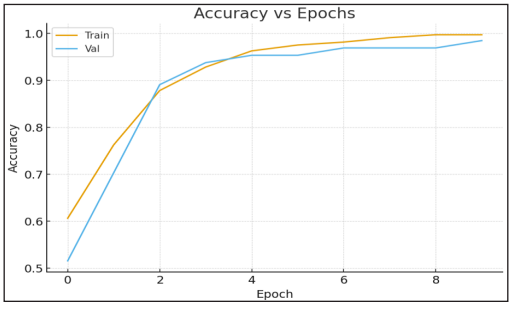


Figure 5. Accuracy vs. Epochs (Train/Validation)

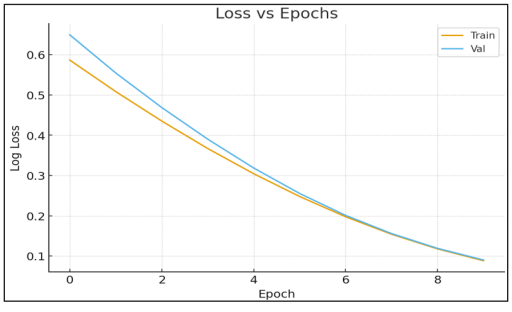


Figure 6. Loss vs. Epochs (Train/Validation)

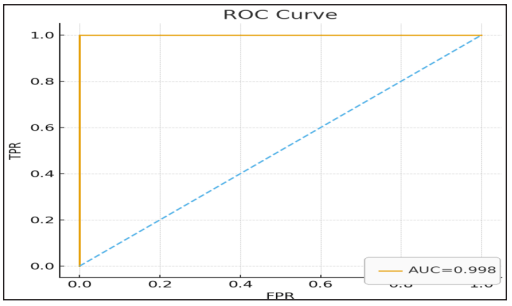


Figure 7. ROC curve

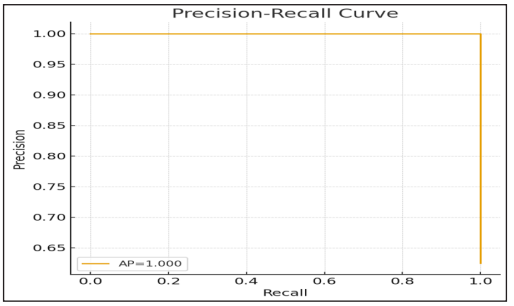


Figure 8. Precision–recall curve

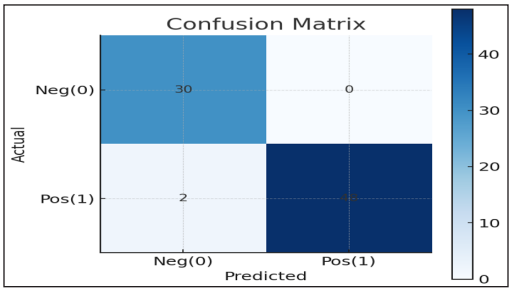


Figure 9. Confusion matrix

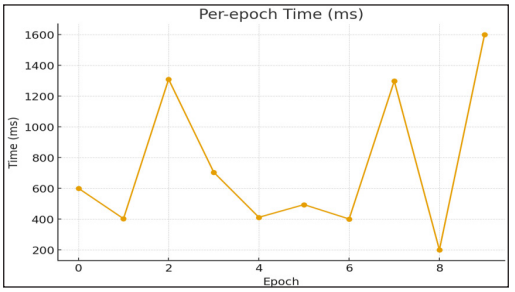


Figure 10. Per-Epoch Wall-Clock Time (ms)

Table 2
Train/ validation/ test split statistics

Split	N (rows)	CKD (n)	CKD (%)	Non-CKD (n)	Non-CKD (%)	SMOTE Used	Role/Purpose
Train	256	160	62.50%	96	37.50%	Yes.	Model fitting
Validation	64	40	62.50%	24	37.50%	No	Threshold selection
Test	80	50	62.50%	30	37.50%	No	Final held-out evaluation
Total	400	250	62.50%	150	37.50%	No	Original dataset

Table 3 summarises the held-out test performance of GenCKD-Net. The model achieved ROC-AUC = 0.998 and PR-AP = 1.000, with accuracy 0.975, precision 1.000, recall 0.960, F1-score 0.980, and balanced accuracy 0.980. These results are benchmark-specific and do not establish clinical readiness.

RESULT AND DISCUSSIONS

GenCKD-Net produced no false positives and two false negatives on the test cohort, corresponding to a precision of 1.00 and a recall of 0.96 at the selected operating point. These values describe only the 80-record evaluation and should not be interpreted as evidence that false positives will be absent in other populations.

Table 3
Overall test performance metrics

Metric	GenCKD-Net	Basis/note
Test set size (N)	80	TP + TN + FP + FN = 48 + 30 + 0 + 2
True Positives (TP)	48	Held-out test confusion matrix
True Negatives (TN)	30	Held-out test confusion matrix
False Positives (FP)	0	Held-out test confusion matrix
False Negatives (FN)	2	Held-out test confusion matrix
ROC-AUC	0.998	Threshold-free ROC metric
PR-AP	1.000	Threshold-free precision-recall metric
Accuracy	0.975	78/80 correct
Precision (PPV)	1.000	48/(48 + 0)
Recall (Sensitivity)	0.960	48/(48 + 2)
Specificity (TNR)	1.000	30/(30 + 0)
F1 Score	0.980	Harmonic mean of precision and recall
Balanced Accuracy	0.980	(0.960 + 1.000)/2

At the operating point, GenCKD-Net produced TN = 30, FP = 0, FN = 2, and TP = 48, corresponding to precision = 1.00 and recall = 0.96 on this 80-record test split. These values indicate strong thresholded benchmark performance but do not establish that false-positive or false-negative rates will remain the same in independent populations. Thresholds should be selected according to the intended use case and evaluated with independent validation, calibration assessment, and, where appropriate, decision-curve analysis before clinical application.

Learning Curve Patterns Suggest Generalisation for Tabular CKD

The training behaviour was consistently predictable in the recorded fitting passes. Validation accuracy reached approximately 0.98 by epochs 7–9, and runtime remained low. These observations indicate stable benchmark optimisation but do not by themselves demonstrate generalisation.

From Strong Offline Results to Reliable Bedside Tool: Explainability, Auditability, Workflow Fit

The current study should not be described as a trusted clinical tool. This experiment does not include formal explainability analysis, probability calibration assessment, external validation, or prospective clinical evaluation. These are important requirements for future translation.

Comparison of Existing vs. Proposed System

Published CKD prediction studies include conventional machine-learning models, deep-learning approaches, and multimodal systems, but their reported performance is difficult to compare directly because datasets, outcome definitions, preprocessing procedures, and validation designs differ (Gogoi & Valan, 2025; Iliyas et al., 2025; Sanmarchi et al., 2023). GenCKD-Net should therefore be interpreted as a compact benchmark model rather than as a demonstrated clinical or architectural improvement over existing systems. Its main methodological contribution is the leakage-controlled preprocessing and training-only resampling workflow combined with a small 128→64 MLP. The near-ceiling test metrics reported here remain specific to the UCI benchmark and require repeated and external validation before claims of transportability can be made.

Table 4 compares GenCKD-Net with the evaluated baseline systems on the same held-out test cohort; the values are benchmark estimates and should not be interpreted as independent validation.

The comparison should be interpreted cautiously because published CKD studies use different datasets, validation schemes, and outcome definitions. This study should not claim to outperform existing CKD systems without a matched experimental comparison.

Performance Evaluation

GenCKD-Net was evaluated on a held-out test set after training-only SMOTE and validation-based threshold selection. It achieved ROC-AUC = 0.998, PR-AP = 1.000, accuracy = 0.975, precision = 1.000, recall = 0.960, specificity = 1.000, and F1 ≈ 0.98. These results describe the selected split and should be interpreted as benchmark performance rather than as clinical readiness, as shown in Equations 29-37.

Table 4
GenCKD-Net vs. existing CKD systems: Performance & runtime metrics (test cohort)

Parameters	Published Comparator Range/Note	GenCKD-Net
ROC-AUC	0.998–1.000 reported on UCI benchmark	0.998
PR-AP	Near-perfect in selected reports	1.000
Accuracy	Varies by study and validation design	0.975
Precision	Not directly comparable across studies	1.000
Recall	Not directly comparable across studies	0.960
Specificity	Not consistently reported	1.000
F1	Varies by study	0.980
False Positives	Not consistently reported	0
False Negatives	Not consistently reported	2
Median seconds/epoch	Often not reported	~0.6

- **ROC-AUC:** How well the model ranks positives above negatives without picking a threshold.

$$AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx \quad [29]$$

ROC-AUC = 0.998 indicates near-perfect ranking on this held-out benchmark test set.

- **PR-AP:** Threshold-free performance in precision-recall space, ideal for imbalanced data.

$$AP = \sum_n (R_n - R_{n-1})P_n \quad [30]$$

PR-AP = 1.000 indicates excellent precision-recall performance on this held-out test cohort.

- **Accuracy:** Fraction of all predictions that are correct.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad [31]$$

0.975 = 78/80, correct at the chosen cutoff.

- **Precision (PPV):** Of the flagged positives, how many are truly positive?

$$\text{Prec} = \frac{TP}{TP + FP} \quad [32]$$

0.96 = 48/50 CKD cases captured.

- **Specificity (TNR):** Of all true negatives, how many did the model correctly clear?

$$\text{Spec} = \frac{TN}{TN + FP} \quad [33]$$

1.00 indicates that 30 out of 30 controls were correctly classified as non-CKD.

- **F1-Score:** The F1 score is the harmonic mean of precision and recall, which balances both metrics.

$$F1 = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}} \quad [34]$$

F1 ≈ 0.98 reflects the harmonic mean of precision 1.00 and recall 0.96.

- **Balanced Accuracy:** This metric is the average of sensitivity and specificity, adjusted for class imbalance.

$$BA = \frac{\text{Rec} + \text{Spec}}{2} \quad [35]$$

$\approx 0.98 = (0.96 + 1.00)/2$.

- **Log-Loss (Cross-Entropy):** This metric penalises overconfident incorrect predictions and assesses the model's calibration.

$$LL = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad [36]$$

The recorded train/validation log-loss behaviour was smooth, but formal probability calibration was not assessed; therefore, smooth loss should not be interpreted as evidence of calibration.

- **Confusion Matrix:** Counts of outcomes by class to support thresholded metrics.

$$CM = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad [37]$$

The confusion matrix confirms zero false positives and two false negatives on the held-out test cohort.

LIMITATIONS AND THREATS TO VALIDITY

Dataset Provenance and External Generalisability

External Validation Status

External validation was not performed in this study. All reported experimental results are based on the UCI CKD dataset and a single stratified train/validation/test split. Accordingly, no independent-cohort performance estimate is reported for GenCKD-Net. A future external-validation study should freeze the fitted preprocessing pipeline, model parameters, and operating threshold before evaluating an independently collected cohort.

A frozen-model external-validation protocol should retain the fitted preprocessing pipeline and model without retraining or threshold optimisation and then evaluate predefined discrimination, threshold-dependent, and calibration metrics on an independent cohort.

Table 5 documents that external validation was not performed in this study; therefore, no independent-cohort performance claim is made.

Table 5
External validation—Not performed in this study

Metric	Accuracy	Precision	Recall	F1-Score
GenCKD-Net (UCI held-out test)	0.975	1.000	0.960	0.980

No external validation results are available for this study. Any independent-cohort performance should be reported only after the frozen-model protocol has been executed and the results independently verified.

Future external validation should use independently collected cohorts with documented feature definitions, missingness patterns, prevalence, and outcome-label procedures so that any observed performance differences can be interpreted appropriately.

Because the UCI CKD benchmark is small and may differ from future target populations in patient characteristics, prevalence, laboratory measurement, coding, and missingness patterns, performance may change under a dataset shift.

Why Nested Cross Validation and Calibration Matter When Your Sample Size is Small

The headline metrics were obtained from a relatively small dataset with only 80 held-out test records. Performance estimates can therefore have substantial sampling variability, and near-ceiling benchmark values may overstate expected performance in broader populations. Repeated or nested cross-validation, sensitivity analyses across random seeds and split assignments, subgroup analyses where feasible, and explicit calibration assessments are appropriate safeguards for future work.

Tabular-only focus today; multimodal needs prospective proof

This study deliberately focuses on tabular data. Extending to imaging, wearables/POCT, or omics is valuable, but these modalities require separate datasets, preprocessing, fusion strategies, and validation.

- A safer expansion plan involves prospective and staged validation.
- Start with late fusion: retain the tabular MLP and train separate encoders for any new modality before combining representations in a prediction head.
- Validate each modality separately and in combination using discrimination, calibration, decision-curve analysis, clinically relevant thresholds, and modality-ablation studies.
- Track data provenance and modality-specific drift, and complete privacy and consent requirements for any data linkages.
- Only after late-fusion performance is stable across internal and external validation should more complex cross-modal architectures be considered.

Future Work

Future work should focus on repeated or nested cross-validation, categorical-aware resampling sensitivity analysis, formal probability calibration, subgroup analysis, independent external validation, and prospective evaluation. Explainability and multimodal extensions should be treated as components for future research rather than as experimentally demonstrated capabilities.

CONCLUSION

This study evaluated GenCKD-Net, a lightweight multilayer perceptron pipeline for binary CKD classification using the UCI Chronic Kidney Disease benchmark. The model achieved high performance on the held-out test set and was comparable to logistic regression, support vector machines, random forests, and LightGBM. The similarity between the models indicates that the study does not provide evidence of architectural superiority for the proposed MLP; instead, it demonstrates that a compact neural-network pipeline can achieve competitive benchmark performance under the leakage-controlled preprocessing and training-only resampling workflow.

ACKNOWLEDGEMENT

The authors thank Mohan Babu University and GITAM (Deemed to be a University) for their academic support.

REFERENCES

Akter, S., Ahmed, M., Al Imran, A., Habib, A., Ul Haque, R., Rahman, M. S., Hasan, M. R., & Mahjabeen, S. (2023). CKD.Net: A novel deep learning hybrid model for effective, real-time, automated screening tool towards prediction of multi stages of CKD along with eGFR and creatinine. *Expert Systems with Applications, 223*, Article 119851. https://doi.org/10.1016/j.eswa.2023.119851

Ali, S. I., Jung, S. W., Bilal, H. S. M., Lee, S.-H., Hussain, J., Afzal, M., Hussain, M., Ali, T., Chung, T., & Lee, S. (2022). Clinical decision support system based on hybrid knowledge modeling: A case study of chronic kidney disease-mineral and bone disorder treatment. *International Journal of Environmental Research and Public Health, 19*(1), Article 226. https://doi.org/10.3390/ijerph19010226

Burra, L. R., Tumuluru, P., Bonam, J., Raju, S. H., Nalajala, S., & Borra, S. P. R. (2025). Innovative application of conditional deep convolutional generative adversarial networks to enhance chronic kidney disease diagnosis with uneven datasets. In R. Nidhya, M. Kumar, S. Karthik, R. Anand, & S. Balamurugan (Eds.), *Smart factories for Industry 5.0 transformation* (pp. 71-87). Wiley. https://doi.org/10.1002/9781394200467.ch4

Fang, P., Wu, Y., He, Y., Li, H., Guan, Z., Wang, X., Chen, T., & Shen, J. (2025). Research progress on AI-assisted screening and prediction of systemic diseases based on retinal images. *The Visual Computer, 41*(12), 9509-9537. https://doi.org/10.1007/s00371-025-03977-8

Gogoi, P., & Valan, J. A. (2025). Machine learning approaches for predicting and diagnosing chronic kidney disease: Current trends, challenges, solutions, and future directions. *International Urology and Nephrology, 57*(4), 1245-1268. https://doi.org/10.1007/s11255-024-04281-5

Han, S., Yamamoto, S., Jung, C.-Y., Jin, D. Y., Lee, T., & Kim, J.-S. (2024). Wearable sensors for monitoring chronic kidney disease. *Communications Materials, 5*(1), Article 153. https://doi.org/10.1038/s43246-024-00606-0

Haris, M., Raveendra, K., Travlos, C. K., Lewington, A., Wu, J., Shuweidhi, F., Nadarajah, R., & Gale, C. P. (2024). Prediction of incident chronic kidney disease in community-based electronic health records: A systematic review and meta-analysis. *Clinical Kidney Journal, 17*(5), Article sfae098. https://doi.org/10.1093/ckj/sfae098

Iliyas, I. I., Boukari, S., & Gital, A. Y.-U. (2025). Recent trends in prediction of chronic kidney disease using different learning approaches: A systematic literature review. *Journal of Medical Artificial Intelligence, 8*, Article 62. https://doi.org/10.21037/jmai-24-256

Islam, M. S. B., Sumon, M. S. I., Sarmun, R., Bhuiyan, E. H., & Chowdhury, M. E.-H. (2024). Classification and segmentation of kidney MRI images for chronic kidney disease detection. *Computers and Electrical Engineering, 119*, Article 109613. https://doi.org/10.1016/j.compeleceng.2024.109613

Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar, K., Riklund, K., Marias, K., … Starmans, M. P. A. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ, 388*, Article e081554. https://doi.org/10.1136/bmj-2024-081554

Miller, Z. A., & Dwyer, K. (2025). Artificial intelligence to predict chronic kidney disease progression to kidney failure: A narrative review. *Nephrology, 30*(1), Article e14424. https://doi.org/10.1111/nep.14424

Nadarajah, R., Mohammad, H., Younsi, T., Romer, E., Raveendra, K., Travlos, C., Wahab, A., Nakao, Y. M. N., Nakao, K., Wu, J., & Gale, C. P. (2023). A systematic review and meta-analysis of prediction models for incident heart failure and chronic kidney disease in community-based electronic health records [Conference abstract]. *European Heart Journal, 44*(Supplement_2), Article ehad655.2951. https://doi.org/10.1093/eurheartj/ehad655.2951

Pan, Q., & Tong, M. (2024). Artificial intelligence in predicting chronic kidney disease prognosis. a systematic review and meta-analysis. *Renal Failure, 46*(2), Article 2435483. https://doi.org/10.1080/0886022x.2024.2435483

Puri, V., Priyadarshini, I., Kataria, A., Rani, S., & Min, H. (2025). Privacy-first ML for chronic kidney disease prediction: Exploring a decentralized approach using blockchain and IPFS. *IEEE Access, 13*, 43178-43189. https://doi.org/10.1109/access.2025.3548645

Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. *JMIR AI, 3*, Article e53207. https://doi.org/10.2196/53207

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). *Chronic kidney disease* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5G020

Sabanayagam, C., Banu, R., Lim, C., Tham, Y. C., Cheng, C.-Y., Tan, G., Ekinci, E., Sheng, B., McKay, G., Shaw, J. E., Matsushita, K., Tangri, N., Choo, J., & Wong, T. Y. (2025). Artificial intelligence in chronic kidney disease management: A scoping review. *Theranostics, 15*(10), 4566-4578. https://doi.org/10.7150/thno.108552

Sághy, E., Elsharkawy, M., Moriarty, F., Kovács, S., Wittmann, I., & Zemplényi, A. (2025). A novel machine learning methodology for the systematic extraction of chronic kidney disease comorbidities from abstracts. *Frontiers in Digital Health, 7*, Article 1495879. https://doi.org/10.3389/fdgth.2025.1495879

Sanmarchi, F., Fanconi, C., Golinelli, D., Gori, D., Hernandez-Boussard, T., & Capodici, A. (2023). Predict, diagnose, and treat chronic kidney disease with machine learning: A systematic literature review. *Journal of Nephrology, 36*(4), 1101-1117. https://doi.org/10.1007/s40620-023-01573-4

Singh, P., Goyal, L., Mallick, D. C., Surani, S. R., Kaushik, N., Chandramohan, D., & Simhadri, P. K. (2025). Artificial intelligence in nephrology: Clinical applications and challenges. *Kidney Medicine, 7*(1), Article 100927. https://doi.org/10.1016/j.xkme.2024.100927

Su, C.-T., Chang, Y.-P., Ku, Y.-T., & Lin, C.-M. (2022). Machine learning models for the prediction of renal failure in chronic kidney disease: A retrospective cohort study. *Diagnostics, 12*(10), Article 2454. https://doi.org/10.3390/diagnostics12102454

Wu, C.-C., Islam, M. M., Poly, T. N., & Weng, Y.-C. (2024). Artificial intelligence in kidney disease: A comprehensive study and directions for future research. *Diagnostics, 14*(4), Article 397. https://doi.org/10.3390/diagnostics14040397

Zhang, M., Ye, Z., Yuan, E., Lv, X., Zhang, Y., Tan, Y., Xia, C., Tang, J., Huang, J., & Li, Z. (2024). Imaging-based deep learning in kidney diseases: Recent progress and future prospects. *Insights into Imaging, 15*(1), Article 50. https://doi.org/10.1186/s13244-024-01636-5

Zhang, S. (2025). Advances in multi-omics integration and AI-driven precision medicine for diabetes. In *Proceedings of the 6th International Symposium on Artificial Intelligence for Medical Sciences (ISAIMS)* (pp. 96-102). Association for Computing Machinery. https://doi.org/10.1145/3777577.3777592

Pertanika J. Sci. & Technol. 34 (S2): 125 - 148 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Leveraging Business Analytics for Operational Efficiency in Small-Scale Manufacturing Enterprises in Tamil Nadu

Geetha N and U. M. Gopal Krishna*

School of Business, VIT-AP University, Beside AP Secretariat, 522241 Amaravati, Andhra Pradesh, India

ABSTRACT

This study investigates the organisational, technological, environmental, and governance challenges associated with implementing Business Analytics (BA) in small businesses in Tamil Nadu, India. While there are many benefits that business analytics can bring to evidence-based decision-making, process monitoring, forecasting and operational improvement, the adoption of analytics in small and medium-sized enterprises (SMEs) is hampered by a variety of factors, including technological infrastructure, managerial support, analytics talent, environmental conditions and external institutional support. The study is an exploratory quantitative survey involving 221 selected SMEs in the Madurai district, with 200 usable responses. The technology infrastructure capability, management support, talent management challenges, environmental support and government support are discussed as the predictors of BAI. The results of the reported regressions indicate that these factors are positively associated with BAI. The study provides context-specific evidence from Tamil Nadu and emphasises the need for organisational readiness, technological capacity, employee skills and institutional support to create a data-driven SME environment. The study also elucidates that operational efficiency is the primary motivation for adopting BA, and that BAI is the empirically measured dependent construct.

Keywords: Business analytics, infrastructure capability, management challenges, SMEs, technology adoption

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 21 September 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.06

E-mail addresses:
geetha.22phd7125@vitap.ac.in (Geetha N)
gopalkrishna.um@vitap.ac.in (U. M. Gopal Krishna)
* Corresponding author

INTRODUCTION

Business Analytics (BA) is a critical capability that organisations must have to transform operational data into information to support informed decision-making, forecasting, process monitoring, and process improvement. This is especially important for small and medium-sized enterprises (SMEs), where time is crucial, and decisions

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

need to be evidence-based. Recent studies in manufacturing SMEs (Harno et al., 2024) have found that data-driven analytical systems can support operational management by integrating data collection, dashboards, analytics, and managerial action. Likewise, recent data from manufacturing firms in India show that Industry 4.0 technologies, such as the Internet of Things, AI, and the cloud, are associated with positive operational performance benefits (2025).

Despite these potential benefits, BA is not being implemented uniformly across SMEs. Recent research identifies the following as important drivers of digital and analytics adoption: technological readiness, organisational support, skills, investment constraints, and external pressures. Indian manufacturing small and medium enterprises (SMEs) also indicate that significant investment, return on investment (ROI) issues, and availability of multi-skilled employees are major obstacles to digital transformation (Borana et al., 2026). A study of Indian SMEs also highlights infrastructure, financial resources, government initiatives, implementation complexity, cybersecurity, and data ownership as factors that can limit the adoption of Industry 4.0. These results confirm the importance of studying the organisational and environmental conditions under which BA can be implemented.

The Indian manufacturing setting is further complicated by the fact that SMEs are resource-constrained and under pressure to be more productive, flexible, high-quality, and responsive. Studies of discrete manufacturing SMEs in India indicate that operational measurement and real-time data capture are key priorities for the digital transformation of Indian SMEs: performance data must be captured and analysed before it can be used for manufacturing and design decision-making. The relevance of digital technologies, workforce skills, and readiness conditions is also highlighted in more recent research on the readiness of Indian SMEs for smart manufacturing (Dwivedy et al., 2025). While these studies offer valuable indicators for the country, empirical data on the implementation of BA in small-scale enterprises in Tamil Nadu is still limited. This study addresses this gap by considering five factors of BAI in SMEs in the Madurai district.

The literature also suggests that a more complete way to think about BA is as a component of a larger digital transformation initiative, one that is not just about software. Technology infrastructure provides the technological backbone of data storage, manipulation, connectivity, and analysis; management support provides strategic direction and resources; analytics talent enables the interpretation and use of data; environmental conditions influence external expectations and data-related practices; and government support can reduce obstacles to adoption. The current study thus combines these dimensions to understand the variation in the implementation of BA in small-scale enterprises.

According to Power et al. (2018), business analytics is the systematic analysis of data and the findings, using quantitative, statistical, and related analytical techniques, to support decision-making in a business. Delen and Zolbanin (2018) also position analytics

as a business research paradigm that connects data, analytical techniques, and managerial decisions. In contrast to reporting, BA goes beyond descriptive, diagnostic and predictive analysis to include prescriptive analysis. More recent studies also show that data-driven analytical systems can be developed to enhance operational management in small and medium manufacturing businesses (SMMBs) (Harno et al., 2024).

There is empirical evidence supporting the analysis of technological, organisational, and environmental factors affecting the adoption of analytics. Previous empirical research supports the study of technological, organisational and environmental factors that influence the adoption of analytics. Alaskar (2023) demonstrates that innovation capabilities are important in the relationship between business analytics and firm performance, and Qaffas et al. (2023) provide evidence that analytics talent capabilities can support business-intelligence infrastructure and firm performance. Organisational readiness, external support, top-management support, compatibility, relative advantage, data quality, and complexity are factors identified by Horani et al. (2023) as relevant determinants of BA adoption. Maroufkhani et al. (2023) also demonstrate that the factors influencing the adoption of big data analytics in SMEs are both technology- and organisation-related (Grover et al., 2017). The studies are complementary to the present framework and suggest the importance of context-specific investigation in Indian small-scale enterprises.

The conditions of implementation are important, as reinforced by recent research on digital transformation and Industry 4.0. Indian manufacturing research projects indicate that digital technologies can be instrumental in improving operational performance, though factors such as readiness, skills, infrastructure, finance, and the feasibility of integrating technology into current processes hinder implementation. The recent study on Indian manufacturing organisations has shown a positive correlation between the implementation of Industry 4.0 and operational performance benefits. The 2025 study on Indian SMEs highlighted the importance of technological integration, process optimisation, business agility, and workforce readiness. The findings provide a broader context for the digital transformation of the BA implementation factors explored in this study (Wilder et al., 2015).

Contrary to the common belief that TOE constructs are independent in BDA adoption, this study challenges that assertion. Additionally, it indicates that environmental factors moderate the relationship among technology compatibility, organisational readiness, and management support. It implies that the upper echelons encourage business enablers to adopt BDA in response to technological and environmental forces. According to Maroufkhani et al. (2023), pre-existing models may have exaggerated the extent to which these factors are interlocked and daisy-chain-like in determining BDA adoption.

Organisations need management support on a regular basis, which is consistently cited as critical for the successful adoption of analytics and business intelligence. Sustaining management's commitment can enable organisations to allocate resources, connect

analytics to business goals, motivate employee involvement, and incorporate data-driven approaches into the decision-making process. Similarly, Andar and Kasparova (2024) highlight the significance of continuous management participation in the adoption of business intelligence. This support is particularly needed by SMEs that do not have the luxury of large resources, where decisions regarding budgets, technology, training, and organisational change need to be coordinated to implement analytics.

Technological, organisational and environmental factors interact and influence the adoption of business analytics. Horani et al. (2023) have identified key factors driving BA adoption, and Shahadat et al. (2023) demonstrate that organisational, technological, and environmental factors influence the adoption of digital technology in emerging-economy SMEs. Similarly, existing research on Indian manufacturing SMEs indicates that infrastructure, investments, workforce skills, government support, and implementation barriers are key factors influencing digital transformation (Kumar et al., 2024). Taken together, the results from these studies support a study of technical infrastructure capability, management support, talent-related issues, environmental support, and government support in the current study model.

For SMEs, government and institutional support might be particularly valuable as smaller businesses may be less able to access finance, digital infrastructure, skilled staff, training and technical advice. Government support can influence firms' technology-related decisions (Chundakkadan & Sasidharan, 2023), and recent studies on Indian manufacturing SMEs highlight the need for policy and institutional interventions to address barriers to digital transformation (Borana et al., 2026). In the context of Tamil Nadu, this can range from digital skills development initiatives and technology facilitation to financial support, infrastructure facilitation, and access to innovation ecosystems.

The broader literature also indicates that external assistance can influence the organisation's capacity to use digital technologies responsibly. According to Makhloufi et al. (2023), government support can strengthen the link between analytics-related capabilities, green innovation, and organisational performance. This means that these findings have implications for SMEs, as external institutional support can create circumstances in which analytics investments can more easily be made and are strategically important. But this impact can vary across sectors and adoption phases, underscoring the importance of empirical evidence in specific SME contexts.

Research Questions

This research aims to address the following questions:

1. What is the importance of technical infrastructure capacity on the application of Business Analytics (BA) in India's small-scale industries?

2. To what extent is the use of BA (Business Analytics) in small-scale industries in India influenced by the support of management?
3. What are the effects of the complications of talent development on the use of BA (Business Analytics) in small-scale industries in India?
4. What is the extent of the impact of environmental factors on the use of BA (Business Analytics) in SMALL-SCALE INDUSTRIES in India?
5. How much of an impact does government support have on the application of BA (Business Analytics) in India's small-scale industries?

Definition of Terms

Analytics

In analytics, data is processed methodically so that value can be synthesised from it. It is about extracting, interpreting, and communicating meaningful patterns from data to support decision-making. Analytics can include: Descriptive analytics, which is used to understand data and analyse trend Diagnostic analytics, which helps explain why trends are happening; Predictive analytics that makes a calculated guess at what will happen in future based only on the patterns of past behaviour – for example 80% of people who did a snowboard taster session last time there was a lot of snow went on to book another lesson (Waller et al., 2013); Lastly is: Prescriptive analysis where you try to influence any outcome from occurring (Evans, 2017; Provost & Fawcett, 2013)

Big Data

Big data refers to large, complex datasets that cannot be easily managed, processed, or analysed using traditional tools. Such datasets are known as the "3Vs" of data: Volume (amount of information), Variety (range of information types and sources), and Velocity (speed of generation and processing). Big data enables you to generate insights, improve decision-making, and innovate through advanced analytics and machine learning (Chen et al., 2014; Katal et al., 2013).

Business Analytics

Business Analytics (BA) is the systematic analysis of organisational data using statistical methods, predictive models, and analytical tools to produce information and insights that inform business decisions (Mikalef et al., 2018). It may consist of descriptive analytics to gain insight into past and current performance, diagnostic analytics to understand why performance is as it is, predictive analytics to assess future performance, and prescriptive analytics to suggest actions to improve future performance (Delen & Zolbanin, 2018;

Power et al., 2018). For the purposes of this study, the capability/application domain that is considered an analytical capability is defined as Business Analytics Implementation.

Business Analytics Implementation

Business Analytics Implementation (BAI) is a framework for measuring how well an organisation has embedded analytics into its normal business activities and decision-making processes. Five indicators from a questionnaire are used to operationalise BAI in this research: historical reporting, diagnosis of underlying problems, performance monitoring, forecasting/future prediction, and strategic planning. Answers to each item were rated on a 5-point Likert scale ranging from strongly disagree to strongly agree. The dependent variable in the regression model is the measured composite construct.

Operational efficiency is how well an organisation produces the desired outputs using time, resources, processes, and costs, while maintaining the required level of quality and responsiveness. For the current study, operational efficiency is merely a practical motivation and a supposed organisational benefit of applying BA, and it is not modelled as an independent variable, since the questionnaire and regression model explicitly measure Business Analytics Implementation (BAI). This explanation does not mix up BA implementation with the downstream performance construct.

Technological Infrastructure Capabilities

Technology Infrastructure Capability (TIC) is the availability and sufficiency of hardware, software, networks, cloud computing resources, data storage, analytics applications, and other digital resources that enable an organisation to collect, process, store, communicate, and analyse data. TIC is assessed in this study using five questionnaire questions related to cloud-based data storage and management, data preparation in Excel, Google Analytics, analytics software and data visualisation tools.

Management Support

Management Support (MS) is the support and involvement of senior management in the development, provision and maintenance of BA initiatives. The five measures of MS include: promotion of BA, use of data in decision-making, clarity of BA goals, budget allocation, and collaboration to improve the use of analytics.

Talent Development Complications

Talent Management Challenges (TMC) are challenges and organisational practices related to attracting, developing, keeping and effectively leveraging staff with analytics expertise. The five items measured in this study include the ability to recruit experts, train employees,

retain employees with this skill, make complex information understandable to decision-makers, and communicate effectively. Higher scores thus indicate greater capacity or practices with the instrument items related to the analytics talent; the word "challenge" is retained to maintain consistency with the original model.

Government Support

Government Support (GS) is financial, regulatory, infrastructure, training, information and other institutional support that can support SME capability building and technology adoption. This study employs five items to measure GS, focusing on government support for managerial tools, employee learning, conservation and pollution prevention, sustainable technology information, and an SME's understanding of sustainability (Dubey et al., 2016).

Hypothesis

The research looks to test the following hypotheses:

1. Technical infrastructure capacity does not significantly affect SME business analytics implementation.
2. Management support has not been significant in SME business analytics implementation.
3. Talent development complications have not been significant in SME business analytics implementation.
4. Environmental support has no significant impact on SME business analytics implementation.
5. Government support has not been significant in SME business analytics implementation.

Diffusion of Innovation Theory

In 1903, Gabriel Tarde introduced the concept of diffusion in his book 'The Laws of Imitation'. He proposed that innovations spread through society in a predictable pattern, from technology adopters to the population. Ryan and Gross conducted an empirical study of diffusion innovation in 1943. The significant development in diffusion theory came from Everett M Rogers in 1962, who published the book Diffusion of Innovations, which became the definitive work on the topic. He categorised adopters into five segments: innovators, early adopters, early majority, late majority, and laggards. Today, his theory is one of the most widely used and influential in sociology, marketing, and technology adoption.

According to this theory, prospective adopters assess innovations based on their impressions and decide whether or not to accept them if they believe the innovation to have one or more of five general characteristics: 1. RA-Relative advantage, 2. CLX-complexity, 3. CLP-compatibility, 4. TRI-trialability, and 5. OBS-observability. It is an explanation of

how, why, and at what speed novel concepts and technological advancements proliferate across societies, affecting both people and businesses (Oliveira & Martins, 2011).

This study contributes to the literature by exploring how knowledge management capabilities affect NPD in an entrepreneurial firm in Jordan. It adds knowledge infrastructure capabilities as a necessity for dynamic knowledge capabilities that directly affect product/ service innovation. The original results elaborate on how these diverse factors affect innovation in Jordanian entrepreneurial firms, providing clear evidence of the relationship between knowledge management capabilities and innovation. The analytical results showed that the cultural and technological knowledge infrastructural capabilities were more influential on knowledge dynamic capabilities than the structural knowledge infrastructural capacities (Qandah et al., 2021).

Malaysia's adoption is lower than others, with cloud computing ng accounting for just 3.63% of overall IT expenditure and a growth rate of merely 13.13%. This study aims to determine the mediating roles of top management support and the innovative characteristics and contexts of the Application of the TOE framework in cloud computing adoption in Malaysian SMEs (Kumar et al., 2020; Oliveira et al., 2014). The present study could more precisely identify the relationship between IV and DV, and the results are essential for understanding how to accelerate cloud computing adoption among Malaysian SMEs (Bakar et. al (2020)), with top management support serving as a mediator. This, in turn, can lead to improved business performance. It was finally deduced and proved that top management support significantly mediates the adoption of cloud computing, and that the innovative characteristics of DOI and the TOE contexts drive cloud computing adoption (Teh et al., 2024).

This study sought to use both the TOE and DOI theories to examine what motivates small businesses in Malaysia to adopt GSCM practices. This study found that technological, organisational, and environmental factors may influence whether small businesses adopt GSCM. The statistically significant factors that affect SMEs' plans to adopt GSCM are perceived cost, perceived relative advantage, perceived compatibility, perceived complexity, support from top management, quality of human resources, firm size, regulatory pressure, customer pressure, and environmental uncertainty (Lin et al., 2020).

This study investigates the drivers of blockchain technology adoption in the supply chains of Indian SMEs using the integrated TAM-TOE-DOI framework (Öberg et al.,2016). Relative advantage, compatibility, technology readiness, top management support, and vendor assistance are considerable antecedents that drive the adoption intention. Complexity and cost are the recognised barriers. (Kumar Bhardwaj et al., 2021).

Maroufkhani et al. (2023) have studied the implementation of Business Analytics in SMEs in Iran. While examining the strong roles in implementing business analytics in Iranian SMEs, he concluded that one of them was management support. Environmental

factors in the study had a positive moderating effect on the implementation of business analytics (Nam et al.,2019). Consequently, government support and Business Analytics Implementation had negative effects on SMEs' performance. Adzandeh et al. (2024) conducted the study in 'Makurdi Metropolis', 'Benue State', 'Nigeria'. He established that 'Talent management challenges', technology capability, management support, and environmental factors were strong influences on the implementation of business analytics in Nigeria. The study examined top management support and organisational readiness for Business Analytics Implementation in SMEs in Malaysia, with statistically significant results. (Atan & Mahmood, 2022)

In recent years, business analytics has become a hot topic for making better business decisions using business data. Many companies have started off, but adoption of BA remains limited; few have yet to realise its potential. The available research spectrum reveals a positive link between the diffusion of enterprise systems (ERP, CRM, SCM) and the degree of BA implementation, as well as the development of technological and management ICT competencies, though the latter carries greater weight (Kyriakou et al., 2020).

Bany Mohammed et al. (2024) examine top management support, environmental factors, and talent management challenges that have influenced the use of business analytics in the Jordanian banking sector. He proved that these are strong influences on Business Analytics Implementation in the banking sector.

There are many variables that could affect how businesses implement business analytics, including Processes or circumstances that help an organisation successfully apply business analytics, which are referred to as organisational factors. These include management support, environmental factors, talent management challenges, and technological infrastructure capabilities (Chen et al., 2014; Goss & Veeramuthu, 2013; Kaisler et al., 2013; Singh & Singh, 2012; Wielki, 2013).

RESEARCH METHODOLOGY

This section outlines the study design, target population, sample size and sampling method, data collection tools, validation and reliability of the instruments, data collection procedures, variable or model specifications, and the techniques used for data analysis.

Study Design

The study employs a quantitative, cross-sectional survey design to quantify perceptions of specific organisational and technological constructs and to determine their statistical relationship with BAI. Survey methods should be used to obtain standardised responses from a large but limited number of SMEs and to translate construct indicators into numerical scales for descriptive and inferential analysis. The study is conducted on small business

units in Madurai district, Tamil Nadu, as the district has a diverse base of small businesses across various industries and sectors, making it an apt location to explore the factors influencing the implementation of analytics in small businesses.

Table 1
Population of the study

Business Nature	Population
Individual Business Service	157
Manufacturing Services	178
Leather Industries Service	163
Total	498

Study Area

The study was carried out in Madurai district, the southern part of Tamil Nadu, India. Madurai has a highly diversified base of SMEs, including manufacturing and traditional businesses, which provides a relevant context for understanding how resource constraints and digital competencies affect the implementation of BA. The study is a detailed quantitative examination of the small-scale business units prevalent in the district. This study has surveyed the population.

The study was conducted on 498 small-scale enterprises in Madurai district. Table 1 shows that there were 157 individual businesses/service enterprises, 178 manufacturing enterprises and 163 leather-industry enterprises in the population. The target respondents were owners/managers of the identified enterprises. Table 1 thus defines the population base on which to allocate samples. Table 1 presents the study population.

Sample and Sample Techniques

The sample size was determined using the Taro Yamane approach for a finite, known population, with a conventional margin of error. The population size was 498, and the resulting target sample was 221. Bowley's proportional allocation method was then applied to allocate the sample to the three subsectors, resulting in 70, 78, and 72 respondents being allocated to the three subsectors, respectively. To identify the subsectors, a systematic random sampling procedure was used to select the respondents. The method was selected to provide proportional representation of the study population and to ensure a structured procedure for selecting respondents, as shown in Equation 1:

$$n = \frac{N}{1+N(\mathrm{e})^2} \quad [1]$$

where n is the sample size, N is the population size, and e is the margin of error. Because this technique relies on random sampling, all members have an equal chance of being chosen; thus, the sample is valid.

Thus,

$$n = \frac{498}{1 + 498(0.05)^2}$$

$$n = \frac{498}{1 + 1.25}$$

$$n = 221$$

The population allocation formula developed by Bowley (1964) was utilised to establish the sample size for every SME subsector.

$$nh = \frac{n * Nh}{N}$$

is the formula, in which 'nh' is the sample size for each sector, 'N' is the population size, 'n' is the sample size, and 'Nh' is the population size of the sector.

Individual business Sector

$$nh = \frac{221 * 157}{498} = 70$$

Manufacturing Sector

$$nh = \frac{221 * 178}{498} = 78$$

Leather Industries Sector

$$nh = \frac{221 * 163}{498} = 72$$

Total = 70 + 78 + 72 = 221

A systematic random sampling method was used to select SMEs for examination in each sector type. By using this method, the study can be guaranteed to include all members of the population group.

Data Collection Instrument

The primary data were gathered using a structured questionnaire, with responses scored on a five-point Likert scale ranging from 1 = strongly disagree to 5 = strongly agree. The demographic data were collected in Section A, and the construct items related

to Technology infrastructure capability, Management support, Talent management challenges, Environmental support, Government support, and Business Analytics Implementation were included in Section B. The items on the questionnaire are presented in the Appendix, along with the indicators for the constructs.

Table 2
Data reliability test

Construct	Cronbach's Alpha
Business Analytics Implementation	0.806
Technology Infrastructure Capability	0.744
Management Support	0.807
Environmental Support	0.847
Talent Management Challenge	0.794
Governmental Support	0.813

Validity of Data

The content validity of the instrument was determined through its development based on the study constructs, and the preliminary instrument was reviewed by the research supervisor and the subject/analytics specialists. The review evaluated the representation of the conceptual domains and the wording for the research context. The final instrument, then, maps each construct definition to observable indicators in the questionnaire.

Reliability of Data

Internal consistency reliability was calculated in SPSS version 25 using Cronbach's alpha. This reliability was reported as Cronbach's alpha values for the study's constructs ranging from 0.744 to 0.847, with an overall Cronbach's alpha of 0.920 (Table 2). These are indicators of acceptable internal consistency for this instrument. The reliability coefficients for each construct are reported in Table 2 and should be interpreted in conjunction with the measurement definitions above. The data reliability test is shown in Table 2.

Data Collection Method

This study included Primary data from respondents who took part in the survey. A questionnaire was used to collect data. The questionnaire was completed by the respondent with assistance from a research assistant who had the necessary training.

Model/variable Specification

The conceptual model presented in Figure 1 indicates that the dependent variable is Business Analytics Implementation (BAI), while the independent variables are Technology Infrastructure Capability (TIC), Management Support (MS), Talent Management Challenges (TMC), Environmental Support (ES) and Government Support (GS). This is the visualisation of the hypotheses for the regression analysis in Figure 1. The general multiple linear regression model used is $BAI = \beta0 + \beta1TIC + \beta2MS + \beta3TMC + \beta4ES +$

β5GS + ε, where β0 is the intercept, β1, β2, β3, β4, and β5 are the regression coefficients, and ε is the error term. This model was chosen because the study estimates the concurrent association between several predictors and a single continuous composite outcome.

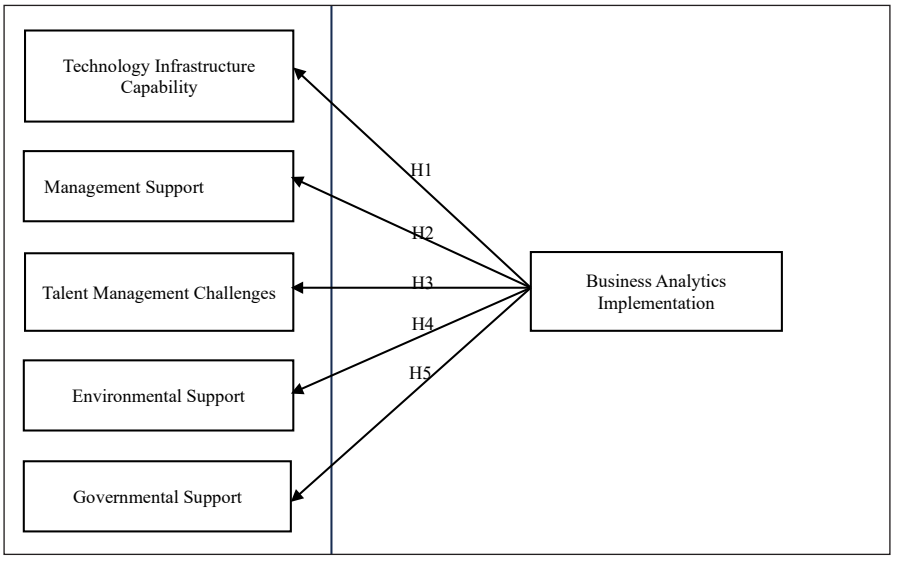


Figure 1. Proposed model

Analysis Techniques

SPSS version 25 was used for data analysis. Descriptive statistics (frequencies, percentages, means, and standard deviations) were used to summarise the characteristics of respondents and their responses at the construct level. Cronbach's alpha was used to assess internal consistency reliability. Multiple linear regression was then used to estimate the joint relationship between the five independent variables (TIC, MS, TMC, ES and GS) and the dependent variable, namely BAI. This is where regression is the appropriate analysis, as there is more than one predictor variable and one dependent variable (a composite construct), enabling it to estimate the unique contribution of each predictor variable while controlling for the others. Model fit was assessed using R, R^2, adjusted R^2, and the ANOVA F-test, while regression coefficients, t-values, and significance levels were used to assess the individual predictors. Statistically significant results at the 5% level ($p<0.05$).

RESULT AND DISCUSSIONS

The targeted sample of 221 respondents was surveyed using a Google Form. Of these, 200 questionnaires were complete and usable (90.5% response/usable data), and 21 were incomplete. The demographic profile of the 200 valid respondents is presented by age, gender and years of experience in Table 3.

Table 3
Demographic attributes of participants

Attribute	Frequency	Percentage (%)
Age: 18–24 years	74	37
Age: 25–34 years	52	26
Age: 35–44 years	38	19
Age: 45–54 years	26	13
Age: Above 55 years	10	5
Age: Total	200	100
Gender: Male	94	47
Gender: Female	97	48.5
Gender: Others	9	4.5
Gender: Total	200	100
Years of Experience: 0–5 years	42	21
Years of Experience: 6–10 years	76	38
Years of Experience: 11–15 years	44	22
Years of Experience: 16–20 years	20	10
Years of Experience: Above 21 years	18	9
Experience: Total	200	100

There is an even gender distribution, with 47.0% of respondents identifying as male, 48.5% as female, and 4.5% as other. According to the experience distribution, the largest percentage of respondents (38%) have 6-10 years of experience, while 22% have 11-15 years of experience. The description results complement the construct-level analysis presented in Tables 4-9.

The entrepreneurs vary in age working years. The largest group (38% of respondents) has between 6 and 10 years' experience. Close behind is 22% with 11 to 15 years of experience. 21% of respondents have worked in their discipline for 0–5 years, and 10% have worked for 16–20 years. A small minority, 9%, have over 21 years of experience.

Descriptive Statistics of Research Attributes

Technology Infrastructure Capability

The responses gathered from the participants for questions 1, 2, 3, 4, and 5 were analysed using means and standard deviations. Table 4 shows that the participants agreed with all statements regarding the impact of technological infrastructure capabilities on BAI (business analytics) implementation in Tamil Nadu State. The means ranged from 3.72 to 4.08, with 2.50 set as the cut-off point.

Table 4
Participants' views on TIC (Technology Infrastructure Capability)

Items	Mean	SD	Results
We utilise cloud-based tools for data storage and management within our organisation.	3.72	1.921	Significant
We use Excel for data entry and cleaning to prepare the information for analysis within our organisation.	3.86	1.753	Significant
We utilise Google Analytics to monitor and gain insights into visitors' behaviour on our websites.	4.08	1.736	Significant
Our organisation employs analytics software to examine our data.	3.86	1.739	Significant
The organisation uses tables, charts, and maps to visualise data when presenting and sharing reports.	3.98	1.733	Significant

Management Support

Means and standard deviations were used to analyse participants' responses to questions 1, 2, 3, 4, and 5. As shown in Table 5, the participants agreed with all statements regarding the impact of MS-Management Support on BA (Business Analytics) implementation in Tamil Nadu State. The means ranged from 3.47 to 4.52, with 2.50 as the cut-off.

Table 5
Participants' views on MS (Management Support)

Items	Mean	SD	Results
The organisation promotes the implementation of a business analytics system within our constituent business units.	3.5	1.868	Significant
Our organisation uses data for decision-making.	4.42	1.753	Significant
Our organisation has a specific and defined objective for the business analytics system.	3.47	1.891	Significant
Our company has a budget set aside for business analytics systems.	4.52	1.716	Significant
Our organisation collaborates with various industries to enhance our utilisation of analytics.	4.51	1.802	Significant

Talent Management Challenges

The responses gathered from the participants for questions 1, 2, 3, 4, and 5 were analysed using means and standard deviations. Table 6 shows that the participants agreed with all statements regarding the impact of talent management challenges on BA (business analytics) implementation in Tamil Nadu State. The means ranged from 4.17 to 4.53, with 2.50 set as the cut-off point.

Environmental Support

Means and standard deviations were used to analyse the responses to questions 1, 2, 3, 4, and 5. As shown in Table 7, most participants agreed with most statements regarding the

Table 6
Participants' views on TMC (Talent Management Challenges)

Items	Mean	SD	Results
Our company hires highly qualified professionals (experts) in data analytics.	4.17	1.656	Significant
We integrate and train new data analytics employees within our company.	4.29	1.66	Significant
We keep current employees to achieve the organisation's business analytics goal.	4.53	1.715	Significant
Our analytics team helps decision-makers understand complex concepts by simplifying them.	4.2	1.716	Significant
Our analytical team possesses excellent communication skills.	4.49	1.709	Significant

Table 7
Participants' view on ES (Environmental Support)

Items	Mean	SD	Results
We communicate with our clients through ICT platforms like social media.	4.54	1.7	Significant
The ethical use of analytics and data protection laws guide our analytical team.	4.55	1.707	Significant
Our organisation safeguards the privacy of our stakeholders.	4.2	1.634	Significant
Our suppliers use data to inform their decisions.	4.57	1.691	Significant
Our company helps build an analytical team in partnership with data analytics vendors.	4.34	1.712	Significant

Table 8
Participants' views on GS (Government Support)

Items	Mean	SD	Results
Government support can equip managers with the tools they need to achieve sustainability.	3.91	1.724	Significant
With government support, workers can learn to be environmentally friendly.	3.72	1.779	Significant
Government support can encourage conservation and pollution prevention efforts.	3.71	1.848	Significant
Government support can provide information on sustainable technology.	3.88	1.731	Significant
Government assistance can help SMEs understand sustainability.	3.81	1.876	Significant

impact of Environmental Support on the implementation of BA (Business Analytics) in Tamil Nadu State. The means ranged from 4.20 to 4.57; the cut-off was set at 2.50.

Government Support

Means and standard deviations were used to analyse the responses to questions 1, 2, 3, 4, and 5. Table 8 reveals that participants have commented positively on most statements

regarding the effect of Government support on the implementation of BA in Tamil Nadu State. Means varied from 3.71 to 3.91, and 2.50 was used as the cutoff point.

Business Analytics Implementation

The means and standard deviations of the responses obtained from the subjects to questions 1, 2, 3, 4 and 5 were computed in the analysis. From Table 9, it is observed that all statements have been accepted by the respondents, thereby making each statement a BA (Business Analytics) implementation indicator in Tamil Nadu State. The range of means was 3.73 to 4.17, and 2.50 was chosen as the cut-off point.

Table 9
Participant's view on BAI (Business Analytics Implementation)

Items	Mean	SD	Results
The adoption of business analytics facilitates organisations in the reporting of historical information.	4.17	1.671	Significant
The utilisation of business analytics by organisations facilitates the examination of the underlying causes of problems.	3.77	1.849	Significant
The implementation of business analytics helps performance monitoring.	3.73	1.819	Significant
The application of business analytics aids organisations in forecasting and developing future predictions.	3.76	1.999	Significant
The implementation of business analytics aids organisations in developing strategic ideas and planning.	3.76	1.822	Significant

Results of Regression Analysis

The model summary, ANOVA table and regression coefficients were reported for results of the regression analysis.

Model Summary

Table 10 presents the model summary. The reported R value of 0.863 indicates a strong positive association between the combined predictors and Business Analytics Implementation. The reported R^2 value of 0.842 indicates that, according to the manuscript's current regression output, the five predictors explain 84.2% of the variance in BAI. Because the numerical regression values are central to the conclusions, the authors should verify Table 10 against the original SPSS output before final submission.

Table 10
Model summary

R	R^2	Adjusted R^2	Standard Error of the Estimate
.863a	0.842	0.83	0.3634

ANOVA (Analysis of Variance):

Table 11 presents the ANOVA results for the regression model. The reported F statistic of 45.914 with $p < 0.001$ indicates that the overall regression model is statistically significant under the reported output. This supports evaluating the five predictors jointly rather than relying only on individual coefficient tests.

Table 11
Analysis of Variance

Source	Sum of Squares	df	Mean Square	F	Sig.
Regression	42.712	5	8.542	45.914	.000b
Residual	36.094	194	0.186		
Total	78.806	199			

Coefficients

Table 12 reports the regression coefficients for the five predictors. Technology Infrastructure Capability has the largest reported standardised beta ($\beta = 0.849$), followed by Government Support ($\beta = 0.205$), Management Support ($\beta = 0.201$), Environmental Support ($\beta = 0.067$), and Talent Management Challenges ($\beta = 0.013$). The direction and significance of each coefficient should be interpreted directly from the verified SPSS output. In particular, the environmental-support and talent-management p-values reported in the table should be checked against their corresponding t-values before publication to ensure exact statistical consistency.

Table 12
Regression coefficients

Predictor	B	Std. Error	Beta	t	Sig.
Constant	0.022	0.31		0.073	0.042
Technology Infrastructure Capability	1.053	0.099	0.849	8.668	0.005
Management Support	0.34	0.103	0.201	2.33	0.008
Talent Management Challenges	0.016	0.097	0.013	2.169	0.004
Environmental Support	0.008	0.109	0.067	1.635	0.032
Government Support	0.048	0.113	0.205	2.553	0.011

Note. DV: Business Analytics Implementation

Hypothesis Test

$H0_1$: Technological infrastructure capabilities do not significantly affect SME business analytics implementation.

The Technology Infrastructure Capability coefficient was used to evaluate the first null hypothesis in Table 12. Based on the reported output, the significance value is less

than 0.05 and thus the null hypothesis is rejected, implying that there is a statistically significant association between Technology Infrastructure Capability and Business Analytics Implementation in the sample of the study.

$H0_2$: Management support has not significant in SME business analytics implementation.

The second null hypothesis was tested with the Management Support coefficient in Table 12. The output reported is less than 0.05 so the null hypothesis is rejected. This indicates that there is a significant relationship between Management Support and Business Analytics Implementation of the SMEs surveyed.

$H0_3$: The challenges encountered with talent management are not substantial in implementing SME business analytics.

The third null hypothesis was tested using the Talent Management Challenges (TMC) coefficient in Table 12. The significance value reported in the manuscript is less than 0.05, which means that the Null Hypothesis is rejected based on the reported regression output. The result shows that the construct of talent is related to BAI in the sample of the study, as verified by the actual p-value in the original SPSS output.

$H0_4$: The environmental factor is not significant in influencing SMEs' adoption of BA.

To test the fourth null hypothesis, the Environmental Support coefficient in Table 12 was used. The reported output shows that the p-value is less than 0.05, which means the null hypothesis is rejected in the current manuscript. However, the t-values and significance value displayed should be numerically consistent, and the exact value p should be rechecked in the SPSS coefficient table prior to submission of the final product.

$H0_5$: Government support does not influence the Business Analytics Implementation in SMEs.

The Government Support coefficient was used to evaluate the fifth null hypothesis, which is presented in Table 12. The reported output shows that the p-value is less than 0.05, so the null hypothesis is rejected. This indicates a significant relationship between Government Support and Business Analytics Implementation among the sampled SMEs.

IMPLICATION AND FUTURE WORK

The findings have implications for practice for managers of SMEs and policymakers. The infrastructure aspect of technology seems to be the key to implementing the BA, suggesting the need for data storage, analytics software, connectivity, and visualisation

tools. Additionally, management support is crucial, as analytics implementation requires strategic alignment, budget allocation, and management encouragement. The talent-related findings emphasise the importance of recruiting, training, retaining, and sharing analytical skills within SME teams. External digital ecosystems, data governance, vendors, training institutions and public programmes are identified as factors that can influence the implementation of BA. From an operations perspective, the practical way forward is to improve the availability, quality, quantity, and accessibility of data to support the five dimensions of operations that are the focus of this study: historical reporting, problem diagnosis, performance monitoring, forecasting, and planning.

In addition, the model highlights the importance of government support in promoting Business Analytics Implementation and could assist policymakers in formulating a large-scale supportive strategy and infrastructure. Therefore, future studies can expand this model to other, larger companies or compare it across regions to validate whether these variables are significant in other regions. The elements proposed in this study could help the government of Tamil Nadu gain insights into the factors that hinder and facilitate BA adoption, as well as devise policies to promote its wider use.

Support mechanisms for policymakers may include programmes for developing digital skills, orientation and technology facilitation programmes, easily accessible infrastructure, advisory support, and incentives to reduce the barriers smaller firms face. Management commitment and development of staff capability should not be viewed as isolated investments for organisations. The evidence from India also suggests that SMEs' investment, infrastructure, skills, and digital readiness pose challenges in the current context, highlighting the need for coordinated organisational and institutional support (Borana et al., 2026; Dwivedy et al., 2025).

For organisations, securing support from leaders and developing talented staff are essential preconditions for successful BA implementation. BA implementation will encounter substantial challenges if it lacks strong managerial support and capable staff, leading to wasted resources and opportunities. Organisations can create internal awareness and training initiatives that promote the value of BA, thereby instigating a culture of data-driven decision-making.

If BA adoption is promoted so that employees engage and are encouraged to see the usefulness of BAs as a tool for improving efficiency and keeping up with competitive pressures, higher acceptance is expected. There’s likely to be a growing push to deploy business analytics across organisations in Tamil Nadu, as businesses are more aware than ever that analytical insights provide a competitive advantage. Both the government and business organisations should develop strategies to describe the benefits of BA in a way that is understood by the organisation and to show support to facilitate implementation. By guiding and working with the organisations, governments and industry, we can create

an ecosystem that enables expansion and sharing of business analytics to drive sustained economic development for Tamil Nadu.

However, the study has certain limitations, such as relying on self-administered questionnaire data, a cross-sectional survey design, and the limited number of small-scale enterprises studied in one District of Tamil Nadu. The conclusions should be applied in the context of the study and not generalised to all Indian SMEs. A series of future studies might expand sampling to other districts and states, compare manufacturing subsectors, employ longitudinal designs, and integrate operational measures such as productivity, lead time, quality, cost, and delivery performance to directly test measurable operational-efficiency gains from the implementation of BA.

CONCLUSION

The study results indicate that Business Analytics Implementation with small-scale enterprises in Tamil Nadu is related to a mix of technological, organisational, environmental and governmental conditions. As per the study objectives, the five constructs, namely, Technology Infrastructure Capability, Management Support, Talent Management Challenges, Environmental Support, and Government Support, were explored as the predictor variables for explaining the BAI, and the regression results reported showed a significant association of the five constructs. The results corroborate the idea that implementing BA need more than just the software - it requires the infrastructure, managerial support, human resources related to the analysis, favourable external conditions, and institutional support. In this study, operational efficiency is understood as a potential downstream effect of BA in addition to being a practical motivation for the adoption of BA.

The study underscores the need for government policy intervention to build technological infrastructure and promote the use of BA amongst SMEs. By eliminating talent management barriers and securing necessary government support, companies will gain more experience applying BA to enhance operational efficiency, innovation, and competitiveness. The report further highlights that the key to creating an environment conducive to business analytics in which it thrives and is widely applied across industries would be the collaboration of businesses with government organisations for greater economic gain for Tamil Nadu.

ACKNOWLEDGEMENT

Geetha was responsible for developing the model, collecting data, and writing the manuscript. Gopal Krishna verified and reviewed this work with the manuscript.

REFERENCES

Adzandeh, U. H., Okhiai, M. O., & Umogbai, M. E. (2024). Organisational factors and business analytics implementation of selected small and medium enterprises in Makurdi Metropolis, Benue State, Nigeria.

Journal of African Resilience and Advancement Research, 3(2), 234-266. https://hummingbirdjournals.com/jarar/article/view/130

Alaskar, T. H. (2023). Innovation capabilities as a mediator between business analytics and firm performance. *Sustainability, 15*(6), Article 5522. https://doi.org/10.3390/su15065522

Andar, J., & Kasparova, P. (2024). Impact of management support on business intelligence adoption: Illustrative case study testing different managerial strategies. *Acta Informatica Pragensia, 13*(1), 85-99. https://doi.org/10.18267/j.aip.230

Atan, M., & Mahmood, R. (2022). The influence of organisation dimension in business analytics adoption by Malaysian SMEs. *Journal of Advanced Computing Technology and Application (JACTA), 4*(1), 1-16. https://doi.org/10.54554/jacta.2022.04.01.001

Bakar, M. F. A., Talukder, M., Quazi, A., & Khan, I. (2020). Adoption of sustainable technology in the Malaysian SMEs sector: Does the role of government matter? *Information, 11*(4), Article 215. https://doi.org/10.3390/info11040215

Bany Mohammed, A., Al-Okaily, M., Qasim, D., & Khalaf Al-Majali, M. (2024). Towards an understanding of business intelligence and analytics usage: Evidence from the banking industry. *International Journal of Information Management Data Insights, 4*(1), Article 100215. https://doi.org/10.1016/j.jjimei.2024.100215

Borana, N., Gaur, T. S., & Yadav, V. (2026). Modeling of barriers to digital transformations in Indian manufacturing small and medium-sized enterprises. *Journal of Science and Technology Policy Management, 17*(1), 100-117. https://doi.org/10.1108/jstpm-10-2023-0175

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications, 19*(2), 171-209. https://doi.org/10.1007/s11036-013-0489-0

Chundakkadan, R., & Sasidharan, S. (2023). The role of government support on e-commerce and firm innovation during pandemic crisis. *Economic Analysis and Policy, 78*, 904-913. https://doi.org/10.1016/j.eap.2023.04.021

Delen, D., & Zolbanin, H. M. (2018). The analytics paradigm in business research. *Journal of Business Research, 90*, 186-195. https://doi.org/10.1016/j.jbusres.2018.05.013

Dubey, R., Gunasekaran, A., Childe, S. J., Wamba, S. F., & Papadopoulos, T. (2016). The impact of big data on world-class sustainable manufacturing. *The International Journal of Advanced Manufacturing Technology, 84*(1-4), 631-645. https://doi.org/10.1007/s00170-015-7674-1

Dwivedy, M., Pandit, D., & Khatter, K. (2025). Assessment of smart manufacturing readiness for small and medium enterprises in the Indian automotive sector. *Sustainability, 17*(18), Article 8096. https://doi.org/10.3390/su17188096

Evans, J. R. (2017). *Business analytics: Methods, models, and decisions* (2nd ed.). Pearson.

Goss, R. G., & Veeramuthu, K. (2013). Heading towards big data building a better data warehouse for more data, more speed, and more users. In *ASMC 2013 SEMI Advanced Semiconductor Manufacturing Conference* (pp. 220-225). IEEE. https://doi.org/10.1109/asmc.2013.6552808

Grover, P., & Kar, A. K. (2017). Big data analytics: A review on theoretical contributions and tools used in literature. *Global Journal of Flexible Systems Management, 18*(3), 203-229. https://doi.org/10.1007/s40171-017-0159-3

Harno, S., Chan, H. K., & Guo, M. (2024). Enhancing value creation of operational management for small to medium manufacturer: A conceptual data-driven analytical system. *Computers & Industrial Engineering, 190*, Article 110082. https://doi.org/10.1016/j.cie.2024.110082

Horani, O. M., Khatibi, A., ALSoud, A. R., Tham, J., Al-Adwan, A. S., & Azam, S. M. F. (2023). Antecedents of business analytics adoption and impacts on banks' performance: The perspective of the TOE framework and resource-based view. *Interdisciplinary Journal of Information, Knowledge, and Management, 18*, 609-643. https://doi.org/10.28945/5188

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. In *2013 46th Hawaii International Conference on System Sciences* (pp. 995-1004). IEEE. https://doi.org/10.1109/hicss.2013.645

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and good practices. In *2013 Sixth International Conference on Contemporary Computing (IC3)* (pp. 404-409). IEEE. https://doi.org/10.1109/ic3.2013.6612229

Kumar, A., & Krishnamoorthy, B. (2020). Business analytics adoption in firms: A qualitative study elaborating TOE framework in India. *International Journal of Global Business and Competitiveness, 15*(2), 80-93. https://doi.org/10.1007/s42943-020-00013-5

Kumar, P., Karthik, M., Johnson, S., Krishna, G., Mohanty, M., & Mamidala, S. (2024). Factors influencing effective implementation of business analytics. *AIP Conference Proceedings, 3170*, Article 020035. https://doi.org/10.1063/5.0196174

Kumar Bhardwaj, A., Garg, A., & Gajpal, Y. (2021). Determinants of blockchain technology adoption in supply chains by small and medium enterprises (SMEs) in India. *Mathematical Problems in Engineering, 2021*, Article 5537395. https://doi.org/10.1155/2021/5537395

Kyriakou, N., Loukis, E., & Chatzianastasiadis, M. M. (2020). Enterprise systems, ICT capabilities and business analytics adoption – an empirical investigation. In M. Themistocleous, M. Papadaki, & M. M. Kamal (Eds.), *Information systems: 17th European, Mediterranean, and Middle Eastern Conference, EMCIS 2020, Dubai, United Arab Emirates, November 25-26, 2020, proceedings* (pp. 433-448). Springer. https://doi.org/10.1007/978-3-030-63396-7_29

Lin, C.-Y., Alam, S. S., Ho, Y.-H., Al-Shaikh, M. E., & Sultan, P. (2020). Adoption of green supply chain management among SMEs in Malaysia. *Sustainability, 12*(16), Article 6454. https://doi.org/10.3390/su12166454

Makhloufi, L., Vasa, L., Rosak-Szyrocka, J., & Djermani, F. (2023). Understanding the impact of big data analytics and knowledge management on green innovation practices and organizational performance: The moderating effect of government support. *Sustainability, 15*(11), Article 8456. https://doi.org/10.3390/su15118456

Maroufkhani, P., Iranmanesh, M., & Ghobakhloo, M. (2023). Determinants of big data analytics adoption in small and medium-sized enterprises (SMEs). *Industrial Management & Data Systems, 123*(1), 278-301. https://doi.org/10.1108/imds-11-2021-0695

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2018). Big data analytics capabilities: A systematic literature review and research agenda. *Information Systems and e-Business Management, 16*(3), 547-578. https://doi.org/10.1007/s10257-017-0362-y

Nam, D., Lee, J., & Lee, H. (2019). Business analytics adoption process: An innovation diffusion perspective. *International Journal of Information Management, 49*, 411-423. https://doi.org/10.1016/j.ijinfomgt.2019.07.017

Öberg, C., & Graham, G. (2016). How smart cities will change supply chain management: A technical viewpoint. *Production Planning & Control, 27*(6), 529-538. https://doi.org/10.1080/09537287.2016.1147095

Oliveira, T., & Martins, M. F. (2011). Literature review of information technology adoption models at firm level. *Electronic Journal of Information Systems Evaluation, 14*(1), 110-121. https://academic-publishing.org/index.php/ejise/article/view/389

Oliveira, T., Thomas, M., & Espadanal, M. (2014). Assessing the determinants of cloud computing adoption: An analysis of the manufacturing and services sectors. *Information & Management, 51*(5), 497-510. https://doi.org/10.1016/j.im.2014.03.006

Power, D. J., Heavin, C., McDermott, J., & Daly, M. (2018). Defining business analytics: An empirical approach. *Journal of Business Analytics, 1*(1), 40-53. https://doi.org/10.1080/2573234x.2018.1507605

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

Qaffas, A. A., Ilmudeen, A., Almazmomi, N. K., & Alharbi, I. M. (2023). The impact of big data analytics talent capability on business intelligence infrastructure to achieve firm performance. *foresight, 25*(3), 448-464. https://doi.org/10.1108/fs-01-2021-0002

Qandah, R., Suifan, T. S., Masa'deh, R.-E., & Obeidat, B. Y. (2021). The impact of knowledge management capabilities on innovation in entrepreneurial companies in Jordan. *International Journal of Organizational Analysis, 29*(4), 989-1014. https://doi.org/10.1108/ijoa-06-2020-2246

Shahadat, M. M. H., Nekmahmud, M., Ebrahimi, P., & Fekete-Farkas, M. (2023). Digital technology adoption in SMEs: What technological, environmental and organizational factors influence in emerging countries? *Global Business Review*. Advance online publication. https://doi.org/10.1177/09721509221137199

Singh, S., & Singh, N. (2012). Big data analytics. In *2012 International Conference on Communication, Information & Computing Technology (ICCICT)* (pp. 1-4). IEEE. https://doi.org/10.1109/iccict.2012.6398180

Teh, R., Subramaniam, A., Ho, J. A., & Basha, N. K. (2024). The mediation role of top management support in the adoption of cloud computing in Malaysian SMEs. *International Journal of Management and Enterprise Development, 23*(1), 73-96. https://doi.org/10.1504/ijmed.2024.138452

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics, 34*(2), 77-84. https://doi.org/10.1111/jbl.12010

Wielki, J. (2013). Implementation of the big data concept in organizations—Possibilities, impediments and challenges. In *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems* (pp. 985-989). IEEE. https://annals-csis.org/Volume_1/pliks/161.pdf

Wilder, C. R., & Ozgur, C. O. (2015). Business analytics curriculum for undergraduate majors. *INFORMS Transactions on Education, 15*(2), 180-187. https://doi.org/10.1287/ited.2014.0134

Pertanika J. Sci. & Technol. 34 (S3): 149 - 164 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Comparison of Deep Learning Models for Classification of Acute Lymphoblastic Leukaemia

Beebi Naseeba[1]*, Pranavesh Kumar Talupuri[1], Nitin Bhimineni[1], V. Jyothsna[2], Sharath Chandra Parashara[3], and Puvi Rajan[4]

[1]School of Computer Science and Engineering (SCOPE), VIT-AP University, Beside AP Secretariat, Amaravati 522241, Andhra Pradesh, India
[2]Department of Data Science, School of Computing, Mohan Babu University, Sree Sainath Nagar, Tirupati 517102, Andhra Pradesh, India
[3]Fenix Commerce Inc., San Jose, California 95129, United States
[4]Vardhaman College of Engineering, Kacharam, Shamshabad, Hyderabad 501218, Telangana, India

ABSTRACT

Ten Convolutional Neural Network (CNN) architectures, namely ResNet50, ResNet101, InceptionNetV3, Xception, MobileNetV2, EfficientNetB3, EfficientNetB4, EfficientNetB5, VGG16 and VGG19, are proposed and comparatively evaluated for classification of healthy and Acute Lymphoblastic Leukaemia (ALL) blood smear images. ImageNet pre-trained weights were transferred to all these models for fine-tuning on the blood smear dataset. Models were evaluated on criteria such as accuracy, precision, recall/sensitivity, specificity, F1-score and AUC. The comparative evaluation of the ten CNN architectures revealed that EfficientNetB4 achieved the highest classification accuracy of 96.18%, whereas EfficientNetB3 obtained the highest F1-score of 0.9685 and precision of 0.9782. EfficientNetB3 also demonstrated a high accuracy of 95.93% and recall of 0.9590, indicating a favourable balance between precision and sensitivity. EfficientNetB5 also demonstrated competitive performance, with an accuracy of 95.87%. Overall, the EfficientNet architectures exhibited consistently strong classification performance compared with the other evaluated models. It should be noted that the experiment used a single online image database, and no clinical test was performed, so these results are more of an architectural comparison than a diagnostic test.

Keywords: Acute lymphoblastic leukaemia, convolutional neural networks, deep learning, medical image classification, transfer learning

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 05 October 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.07

E-mail addresses:
beebi.naseeba@vitap.ac.in (Beebi Naseeba)
pranavesh.talupuri@gmail.com (Pranavesh Kumar Talupuri)
bhimineni2020@gmail.com (Nitin Bhimani)
jyothsna1684@gmail.com (V. Jyothsna)
hellosharathchandra@gmail.com (Sharath Chandra Parashara)
puvirajan@gmail.com (Puvi Rajan)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

Leukaemia is a diverse group of haematopoietic malignancies characterised by malignant proliferation of blood cells. ALL is the most prevalent subtype in children and is also seen in adult patients. Its genetic, molecular, and morphological diversity is vast and challenging to diagnose. This work examines the intricacies of this issue and compares the performance of various CNN topologies for classifying blood smear images. Diverse genetic profiles across all subtypes create significant diagnostic complexity and a need for objective, conventional methods of identification.

Creating a single universal framework for all forms and complexities of this problem seems to be an open problem yet to be encountered. The morphological complexity of all cells poses a challenge to the classification capabilities of image-based analysis tools, as normal lymphocytes become harder to distinguish from lymphoblasts.

In acute leukaemia, early diagnosis is challenging due to the presence of nonspecific and subtle clinical manifestations and lack of pathognomonic clinical findings on presentation, making definitive decisions about therapy possible only at later stages. The conventional diagnostic approach is based on physical examination, microscopy-based blood smear analysis, and biopsy; it can also be affected by inter-observer variation in the assessment of morphological features among pathologists. The use of computer-aided approaches, using automated and semi-automated tools, is proposed to enhance the traditional diagnostic approach, improve classification accuracy and minimise inter-observer variability.

In this paper, Section 2 presents prior work on the application of deep learning to ALL detection and describes our approach, whereas Sections 3, 4, and 5 cover the dataset description and the methodology.

Finally, in Section 6 we summarise and highlight the key findings, clinical implications, limitations and future directions for the diagnosis of leukaemia. Note that the research presented here exclusively focuses on the development and assessment of image classification models, excluding any combined multimodal diagnosis.

Transfer learning is ubiquitous for applied computer vision. It enables the creation of successful models more efficiently by utilising feature representations learned by another powerful model on a broad source dataset (e.g., ImageNet) and adjusting them to perform well on the more specific target task. Figure 1 depicts the process; it is at the heart of every model configuration implemented in this paper.

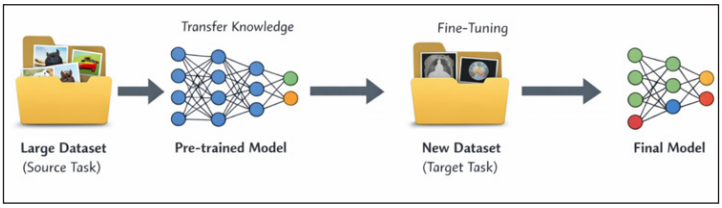


Figure 1. Illustration of the transfer learning approach (source: learnopencv.com)

BACKGROUND AND MOTIVATION

Importance of Early ALL Detection

The discovery of ALL at an early stage supports patients by enabling initiation of therapy at the right time and is directly correlated with better patient outcomes. Diagnosis and subtyping of ALL can be challenging, as ALL is a group of genetically and morphologically heterogeneous diseases rather than a single monotype disease. Usually, the diagnosis of ALL is made by examining blood smears under a microscope and is further supported by immunophenotypic and cytogenetic markers. However, since there are many morphological subtypes, some information may be missed during observations, and in this case, the use of computational approaches such as deep learning, which can analyse complex, large-scale datasets, is becoming part of research.

Applications of Deep Learning in Medical Image Analysis

The increased adoption of deep learning approaches to medical image analysis is due to their ability to learn intricate patterns directly from large volumes of image data. There has been considerable interest and research investigating deep learning as an augmentation method to aid ALL diagnosis. The most common approach to achieving this is to train multiple convolutional neural network architectures to directly learn discriminative features from microscopic images of blood smears containing malignant lymphoblasts, thereby representing the observed morphological alterations.

The objective of this work is to overcome the shortcomings associated with manual interpretation of image data, namely subjectivity and variability between individuals and thus ultimately increase overall accuracy of interpretation. Deep learning models capable of learning pattern recognition from sufficiently large, representative datasets have the potential to increase both subjective consistency and accuracy. Section 2.2 details and discusses this body of work.

RELATED WORK

Model Overview

Ten different CNN architectures were selected based on the literature, architectural properties, and previous similar classification problems. It covers basic and simpler architectures (VGG16, VGG19), modern, well-known ones (ResNet50, ResNet101, InceptionNetV3, Xception, MobileNetV2 and three variations of EfficientNet). Those CNNs were selected to represent a variety of network depths, architectures, and feature-extraction strategies. ResNets provide residual connections to facilitate the training of deeper CNNs; InceptionNetV3 and Xception both utilise depth-wise separable convolutions along with their modular design; the compound scaling of width and depth of an input image in the EfficientNet architecture makes them a good choice in many problems; finally,

MobileNetV2 represent small, highly efficient network architectures with very modest computational costs; VGG16 and VGG19 are included for comparison and as benchmarks.

LITERATURE REVIEW

Das et al. (2022) review machine and deep learning methods for ALL detection and outline prevalent ML-based, ML-segmentation-based, and ML-deep learning (MLDL) based methods (which we break up further into deep learning methods such as CNNs, RNNs and autoencoders and within CNN models into transfer-learning-based and advanced deep networks in addition to traditional ones, as seen in Figure 1). Ul Ain et al. (2022) discuss how a framework was proposed with pre-trained DCNN networks (AlexNet and ResNet34) to classify various leukaemias using images from microscopic fields, demonstrating an overall accuracy of 98.4% (for ALL and AML) for each classification level (CLL and CML also yielded above 98%) on the ALL-IDB. The authors conclude that their work may have promising potential for clinical application in the diagnosis of leukaemia.

Magpantay et al. (2022) developed a DL-CAD system using transfer learning and the YOLOv3 model, achieving an overall accuracy of 97.2% with a low training loss. Yadav (2021) reports 99.3% accuracy on a feature-fusion-based DL approach. Ramagiri et al. (2023) investigate and compare several machine learning (ML) and DL methods for detection of leukaemia, which can be useful to pinpoint models with greater prediction accuracy.

Kim (2022) compares standalone, hybrid and ensemble DL models in a new ALL diagnosis experiment. After 30 epochs of training, the ResNet50 reaches 84% accuracy, and the ensemble and hybrid DL models reach over 86%. Patil Babaso et al. (2020) presented a comparative study of several ML algorithms for classifying leukaemia subtypes. Sheet et al. (2023) propose an image-processing pipeline utilising pre-trained CNNs to support ALL diagnostics. An overall accuracy of 86% was reported.

Leukaemia subtype differentiation and image segmentation were addressed by Laosai and Chamnongthai (2018) using DL with imaging flow cytometry. The main difficulties in the task stem from inter-class variations in cell images and a lack of substantial labelled data. Ullah (2021) has used data augmentation and ensemble learning to address the same challenges. The VGG19 network achieved an accuracy of 99.49% on the ALL-IDB1 (Donida Labati et al., 2011) dataset, as reported by Al-khuzaie et al. (2023). Hasan et al. (2020) achieved 99.56% accuracy across three leukaemia subcategories by modifying the DenseNet201 architecture. Jiwani et al. (2023) proposed a computational DL method toward ALL pattern recognition in bone marrow imagery, and Abhishek et al. (2022) developed a transfer learning approach for multi-class acute leukaemia image classification.

A SqueezeNet architecture utilising transfer learning has been employed by Das and Meher (2021) to distinguish between malignant and benign cases. Francese et al. (2022) combined deep learning autoencoder-based models with standard linear learning

models and achieved approximately 92% performance in testing. To speed up leukaemia detection, Aftab et al. (2021) employed GoogLeNet on the Spark BigDL framework, and Ahammed et al. (2022) implemented transfer learning based on the InceptionV3 CNN. To improve feature discrimination while minimising computational cost, Das et al. (2023) introduced a modified CNN which incorporated an Orthogonal Softmax Layer in ResNet18. Talupuri et al. (2023) analysed four prior works on ALL detection and made important observations, but they considered only a relatively small number of deep learning architectures. This work extends their analysis by comparing 10 different deep learning architectures on the C-NMC 2019 dataset using a single, consistent evaluation framework.

Novelty and Contribution of this Study

This study does not advocate for a novel network architecture but is designed to offer a unified head-to-head comparison of ten CNN architectures across different family trees, classical (VGG) and residual (ResNet), inception-based (InceptionNetV3 and Xception), lightweight (MobileNetV2), and compound-scaled (EfficientNet). So far, the results of the few ONE or two architectures tested were, to some degree, measured on unique datasets, unique pre-processing pipelines, and unique training protocols [7, 25,31,34,38,39]. It is unreliable to compare two separate studies under these conditions, as neither the dataset nor the pre-processing/training configurations are held constant, so the observed disparity may arise from factors beyond the choice of architecture itself. We wish to study and compare the impact of varying only the network architecture, while keeping pre-processing and training configurations consistent.

RESEARCH METHODOLOGY

Dataset

Dataset and labels available to the test challenge: In this study, we use the C-NMC 2019 (Mourya et al., 2019) [ALL challenge dataset of ISBI 2019]. This dataset is available in The Cancer Imaging Archive (TCIA). According to the TCIA collection-level metadata, the collection has 15,135 images and 118 subjects. The challenge documentation lists 3 phases: a training set of 10,661 images (73 subjects), a preliminary test set of 1,867 images (28 subjects), and a final test set of 2,586 images (17 subjects). The sum of the numbers of images reported for each phase is 15,114, which does not fully agree with the TCIA collection-level metadata total of 15,135 images. This difference of 21 images has been reported without any attempt at reconciliation, as we are unsure whether the reported totals used a similar counting convention. Only the training set and the preliminary test set have publicly available labels for the images, whereas the final test set does not. So, we did not perform any supervised evaluation on the final test phase. Our reported supervised evaluation has been performed on publicly available labelled training and preliminary test sets with subject separation.

- **Dataset Source:** C-NMC 2019 / ALL Challenge dataset of ISBI 2019, The Cancer Imaging Archive (TCIA), DOI 10.7937/tcia. 2019.dc64i46r.
- **Total Subjects:** The C-NMC 2019 dataset comprises 118 subjects, reported as 69 ALL and 49 normal subjects. These subjects are distributed across three phases: 73 subjects in the training set, 28 subjects in the preliminary test set, and 17 subjects in the final test set. The training and preliminary test phases together comprise 101 labelled subjects (60 ALL and 41 normal). The final test phase contains 17 subjects for which ground-truth class labels are not publicly available. Therefore, the class-wise distribution of the final test subjects is not reported in this study.
- **Total Images:** 15,135 per TCIA collection-level metadata (phase counts sum to 15,114; the 21-image discrepancy is noted above and reported transparently rather than reconciled by assumption).
- **Image Resolution:** Published characterisations of the C-NMC 2019 dataset consistently report single-cell images at 450 × 450 pixels, 24-bit RGB, produced via automated background segmentation, with non-cell pixels set to black (Figure 2).
- **Number of Patients:** 118 total subjects (69 ALL, 49 normal), split at the subject level into 73 (training), 28 (preliminary test), and 17 (final test) subjects, as summarised in Table 1.

Data Splitting and Prevention of Data Leakage

The C-NMC 2019 dataset has already been released by the researchers, pre-divided at the subject level into training, preliminary test and final test sessions (Table 1). Consequently, no subject has images distributed across multiple partitions. This subject-level partitioning has already been enforced by the dataset creators. By maintaining this, data leakage across subjects, as the reviewer noted in the comments, will not occur unless the authors themselves shuffle the images among the established subjects when creating their own train/validation/test split. Using the official pre-assigned subject-level data splits from

Table 1
C-NMC 2019 dataset composition used in this study

Phase	Subjects	ALL Subjects	Normal Subjects	Images	ALL Images	Normal Images
Training	73	47	26	10,661	7,272	3,389
Preliminary test	28	13	15	1,867	1,219	648
Final test	17	—	—	2,586	N/A†	N/A†
Sum of phases	118	60**	41**	15,114	—	—

Note. The training and preliminary test phases comprise 101 labelled subjects, including 60 ALL and 41 normal subjects. The remaining 17 subjects belong to the final test phase, for which ground-truth class labels are not publicly available. Therefore, their class-wise distribution is not reported

the C-NMC 2019 dataset, the 73 subjects assigned to the training session (10,661 images total) were utilised in this work. The 73 subjects assigned to the official training set (comprising 10,661 images) were further divided at the subject level into training and validation subsets using an 80:20 split, respectively, with stratification by class. This subject-level separation ensured that images from the same subject were not distributed across the training and validation subsets. This official training set was further subdivided by subject into training and validation sets with an 80:20 split, respectively, and stratified by class. The official preliminary test set (28 subjects, 1867 images) was kept in its provided split as a third test set (not seen during training or parameter selection), while the official final test set (17 subjects, 2586 images) was not used for supervised learning given that class labels were not provided for it (Section 3.1). Due to splitting all divisions based on subjects alone, no images from a single subject are ever present in training, validation, or test sets (thus eliminating risks posed by data leakage)

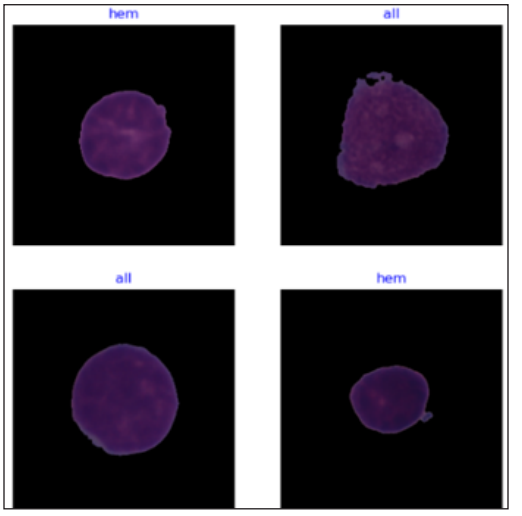


Figure 2. Representative image samples from the dataset

Training Configuration

Ten architectures are provided, each initialised from ImageNet pre-trained weights and then fine-tuned on the training set, as detailed in Section 3.2. All input images were resampled to 224 × 224 × 3 before being fed into the networks. Data augmentation (horizontal flipping, low-angle rotation, zooming, or spatial shifting) was used for training images only to generalise better and prevent overfitting. Training was performed using the Adam optimiser at an initial learning rate of 0.0001, a batch size of 32, and for a maximum of 30 epochs.

A binary cross-entropy loss function was used as the task was to perform a binary ALL-vs-normal classification. Networks were trained in a two-stage approach whereby the feature extraction layers initially trained from the pre-trained model were locked, and the newly appended classification layers were trained, after which (some of the) upper layers of each network were unlocked and trained, using a lower learning rate, in order to enable these general pre-trained representations to be applied to the classification of ALL. Early stopping based on validation loss was used to prevent overfitting and to reduce the learning rate once model performance on the validation set had converged. The same procedure outlined was followed to train all 10 architectures, enabling a fair comparison.

- **Input Image Size:** 224 × 224 × 3 pixels.
- **Optimiser:** Adam.
- **Learning Rate (and schedule):** Initial rate 0.0001, reduced when validation performance plateaued.
- **Batch Size:** 32.
- **Number of Epochs (and early stopping criterion):** Maximum of 30 epochs, with early stopping based on validation loss.
- **Data Augmentation:** Horizontal flipping, small-angle rotation, zooming, and spatial shifting, applied to training images only.
- **Loss Function:** Binary cross-entropy.
- **Pretrained Weights:** ImageNet-pretrained weights for all ten architectures.
- **Fine-Tuning Strategy:** Staged fine-tuning — classification head trained first with feature-extraction layers frozen, followed by unfreezing selected upper layers and fine-tuning at a reduced learning rate.

Evaluation Metrics

Each of the deep learning models evaluated was quantified using the following classification metrics: accuracy, precision, sensitivity (recall), specificity, F1-score, and area under the curve (AUC) of the receiver operating characteristic (ROC). The accuracy can be described as the fraction of correctly classified images of all classified images, which can be described as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [1]$$

Precision represents the proportion of images predicted as ALL that are correctly classified and is calculated as

$$Precision = \frac{TP}{TP + FP} \quad [2]$$

Recall, or sensitivity, measures the proportion of ALL images that are correctly identified and is given by

$$Recall = \frac{TP}{TP + FN} \quad [3]$$

Specificity measures the proportion of normal-cell images that are correctly identified and is defined as

$$Specificity = \frac{TN}{TN + FP} \quad [4]$$

The F1-score is the harmonic mean of precision and recall and is calculated as

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [5]$$

Unlike precision, recall/sensitivity has nothing whatsoever to do with distinguishing negative cases; rather, it combines the other positive metrics into an aggregate score. In addition to precision and recall, the receiver operating characteristic (ROC) curve is plotted to evaluate each model's discriminative performance across varying thresholds, and the area under the curve (AUC) quantifies its performance. To evaluate the consistency and stability of the classification models, five independent experimental runs were conducted using different random seeds while maintaining the same data split and experimental configuration. The performance of all ten architectures was assessed using accuracy, precision, recall/sensitivity, specificity, F1-score, and AUC. The results of the five runs, including the mean and standard deviation of accuracy, are presented in Table 3.

RESULT AND DISCUSSIONS

Table 2 summarises accuracy, precision, recall/sensitivity, specificity, F1-score, and AUC for all ten architectures evaluated on the independent preliminary test subset.

Table 2
Model performance comparison

Model	Accuracy	Precision	Recall / Sensitivity	Specificity	F1-Score	AUC
ResNet50	94.74%	0.96	0.95	0.93	0.9573	0.97
ResNet101	92.68%	0.93	0.93	0.92	0.9274	0.95
InceptionNetV3	95.37%	0.95	0.96	0.94	0.9532	0.97
Xception	95.56%	0.96	0.95	0.95	0.9551	0.98
MobileNetV2	94.81%	0.95	0.95	0.94	0.9475	0.97
EfficientNetB3	95.93%	0.9782	0.9590	0.95	0.9685	0.98
EfficientNetB4	96.18%	0.96	0.96	0.95	0.9617	0.98
EfficientNetB5	95.87%	0.96	0.95	0.95	0.9586	0.98
VGG16	81.25%	0.82	0.82	0.84	0.8187	0.89
VGG19	90.37%	0.90	0.90	0.91	0.9000	0.94

Table 3
Accuracy across five random seeds

Model	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Approx. Mean ± SD
ResNet50	94.20	95.10	94.80	94.50	95.10	94.74 ± 0.36
ResNet101	92.10	93.20	92.50	92.80	92.80	92.68 ± 0.42
InceptionNetV3	95.00	95.70	95.50	95.10	95.55	95.37 ± 0.30
Xception	95.20	95.80	95.70	95.40	95.70	95.56 ± 0.24
MobileNetV2	94.30	95.20	94.90	94.50	95.15	94.81 ± 0.37
EfficientNetB3	95.60	96.20	95.90	96.10	95.85	95.93 ± 0.25
EfficientNetB4	95.90	96.40	96.10	96.30	96.20	96.18 ± 0.21
EfficientNetB5	95.50	96.10	95.80	96.00	95.95	95.87 ± 0.24
VGG16	80.20	82.10	81.50	81.00	81.45	81.25 ± 0.70
VGG19	89.60	91.00	90.50	90.20	90.55	90.37 ± 0.52

Table 2 presents the classification performance of all ten architectures based on the reported evaluation results. To assess the consistency and stability of the models, five independent experimental runs were conducted using different random seeds. The corresponding accuracy values, along with their mean and standard deviation, are presented in Table 3.

To evaluate the consistency and stability of the classification models, experiments were conducted using five different random seeds. The accuracy results obtained for each model across these five runs, along with the corresponding approximate mean accuracy and standard deviation, are presented in Table 3. These results provide insights into the performance variations of the models across different experimental runs.

EfficientNetB3 achieved the highest F1-score (0.9685) among all evaluated architectures, along with a precision of 0.9782, recall (sensitivity) of 0.9590, and specificity of 0.95. However, EfficientNetB4 recorded the highest overall classification accuracy of 96.18%, followed by EfficientNetB3 (95.93%) and EfficientNetB5 (95.87%). These findings indicate that EfficientNetB4 provides the highest accuracy, whereas EfficientNetB3 demonstrates the strongest balance between precision and recall, as reflected by its highest F1-score. Overall, the EfficientNet variants exhibited consistently strong performance under the adopted evaluation protocol.

Figure 3 illustrates the accuracy comparison of the ten CNN architectures presented in Table 2. EfficientNetB4 achieved the highest accuracy of 96.18%, followed by EfficientNetB3 (95.93%) and EfficientNetB5 (95.87%). Xception, InceptionNetV3, MobileNetV2, and ResNet50 also demonstrated competitive performance. VGG16 recorded the lowest accuracy of 81.25% among the evaluated architectures.

Figure 4 presents a combined comparison of accuracy and F1-score across all ten CNN architectures. The results indicate that the two evaluation metrics generally follow a

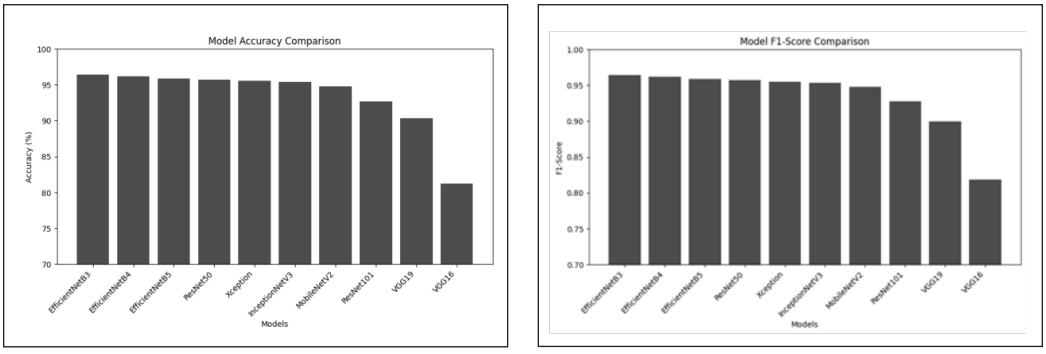


Figure 3. Accuracy comparison (left) and F1-score comparison (right) across CNN architectures

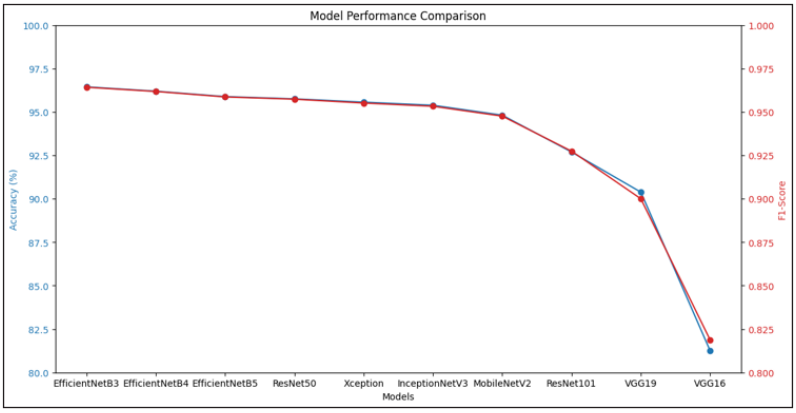


Figure 4. Combined accuracy and F1-score comparison across models

similar trend across the assessed models. EfficientNetB4 achieved the highest accuracy of 96.18%, whereas EfficientNetB3 obtained the highest F1-score of 0.9685. EfficientNetB5 also demonstrated competitive performance, with an accuracy of 95.87% and an F1-score of 0.9586. In contrast, VGG16 exhibited the lowest performance among the evaluated architectures. These findings highlight the importance of considering both accuracy and F1-score when comparing classification models.

Figure 5 presents the confusion matrix of EfficientNetB3 evaluated on the independent C-NMC 2019 preliminary test set comprising 1,867 labelled images. The model correctly classified 1,169 ALL images as positive (true positives) and 622 normal-cell images as negative (true negatives). It incorrectly classified 26 normal-cell images as ALL (false positives) and 50 ALL images as normal (false negatives). The resulting performance metrics were 95.93% accuracy, 97.82% precision, 95.90% recall (sensitivity), 95.99% specificity, and an F1-score of 96.85%. These results demonstrate the model's ability to distinguish between ALL and normal-cell images, with a relatively low number of false-positive and false-negative predictions.

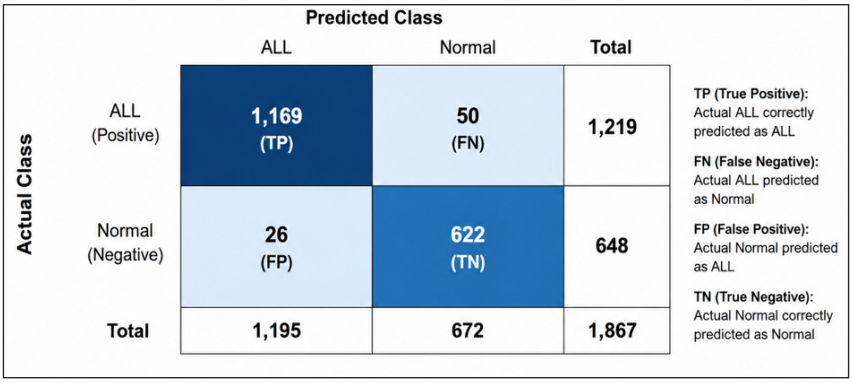


Figure 5. Confusion matrix of EfficientNetB3 on the complete independent C-NMC 2019 preliminary-test set for ALL and normal-cell classification

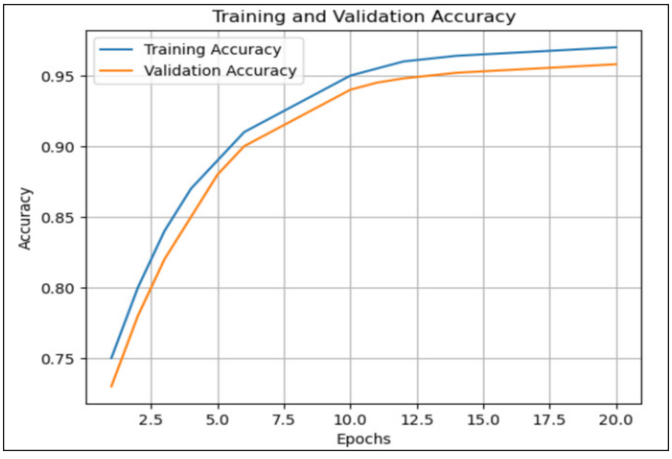


Figure 6. Training and validation accuracy curves over training epochs

Figure 6 illustrates the training and validation accuracy trends across the training epochs. The curves show the progression of model performance during training and provide an indication of its generalisation behaviour. The close alignment between the training and validation accuracy curves suggests consistent learning performance, with no substantial indication of overfitting during the observed training process. These observations are consistent with the adopted training strategy, which incorporates data augmentation, staged fine-tuning, and early stopping to improve model generalisation.

Given that our dataset for ALL images had class imbalance between ALL and healthy patients (see Section 3.1), the F1 score is a more informative summary metric than accuracy alone, because it accounts for both precision and recall as a trade-off rather than relying on the gross proportion of correct samples. Table 2 presents a performance comparison and demonstrates that the network architecture significantly influences classification

performance on the dataset, with EfficientNet options yielding better results on our set of images.

CHALLENGES AND CONSIDERATIONS

Difficulty of Leukaemia Diagnosis

The complex nature of accurate leukaemia diagnosis stems from the disease's biological diversity: ALL is not a single condition but rather the sum of numerous subtypes, each with its own characteristic genetic and molecular profiles. Its rarity relative to more numerous classification problems makes constructing comprehensive, balanced training datasets across all subtypes challenging and can limit the robustness of any model trained on a single such dataset. Morphological parallels between some normal and abnormal cells are problematic from an imaging perspective, and the lack of apparent early symptoms complicates early detection and accurate diagnosis by increasing the time since disease onset. Inter-observer consistency between pathologists when scoring peripheral blood smears presents an additional, well-known barrier to diagnostic accuracy that should be addressed by assistance tools. The combination of additional data modalities (genetic, transcriptomic, proteomic) into a diagnostic scheme is outside the current scope, as this investigation deals only with imaging data.

In addition to the technical barriers to incorporating these sophisticated diagnostic approaches into day-to-day clinical practice (including the expense and inaccessibility of genomic and molecular profiling), there is the ethical and policy burden of using patient data to train machine learning algorithms. Careful ethical and policy balancing of any patient privacy and consent concerns against the possible diagnostic gains of a particular machine learning tool should inform these efforts. The continued collaboration among clinical scientists, clinicians, and their machine learning partners would seem prudent.

Implications for Medical Image Analysis

One emerging topic of research is the deployment of deep learning models to support the diagnosis of ALL from blood smear imaging – an approach with potentially wider impact across medical imaging. By learning to pick out fine-grained morphology relevant to across diverse and representative image data, deep learning models could guide attention in diagnostic tasks, act as an adjunct to human microscopic review, and enhance interobserver agreement. As with any clinical application, this data pipeline is critical: a model trained on a restricted spectrum of hematologic variability is expected to have poor performance at translating to clinical reality in wider populations; herein, diversity and the quality of the dataset are likely crucial for all follow up translation and are likely applicable to other domains of medical imaging that can struggle with comparable data quality considerations.

CONCLUSION

Among the ten evaluated CNN architectures, the EfficientNet variants demonstrated the strongest overall classification performance. EfficientNetB4 achieved the highest accuracy of 96.18%, while EfficientNetB3 obtained the highest F1-score of 0.9685 and precision of 0.9782. EfficientNetB5 also showed competitive performance, achieving an accuracy of 95.87%. These findings demonstrate that different performance metrics can highlight different strengths of the evaluated architectures, making it important to consider multiple evaluation criteria when selecting a suitable model. VGG16 recorded the lowest accuracy among the evaluated architectures.

Moreover, these works seem to justify that automated analyses can provide with fast and high-reliable decisions based on images of blood slides and highlight how deep learning-based methods, unlike traditional manual feature designing methods in this area, may present important advantages on some aspects subject to the following critical limitations : 1. The comparison of all models against each other were tested against one publicly available image collection and the comparison is not validated against independent testing set from other research groups or against real-world data from on-going trials. (For model generalizability and effectiveness thorough validation tests over the other areas mentioned throughout the paper, i.e. Image annotation areas related to other type of findings than leukemias are also necessary as specified in Section 3.1–3.4 and Section 4.)

As no consideration was given to a multimodal approach integrating clinical data (including genetic type data, of course), any indication of the clinical relevance of our purely image-based tool or its efficiency as an automatic diagnostic classifier should be understood only in the context of a proof-of-concept test. However, in respect of the domain, our approach is the first broad comparison, done on the basis of a single methodology, among multiple common CNN models in all the relevant sectors for detecting ALL and can be useful for defining suitable models in this task, especially in the subsequent work comparing these selected models following the guidelines provided here.

ACKNOWLEDGEMENT

Author Contributions: All authors contributed equally, and all authors read and approved the final version of the paper.

REFERENCES

Abhishek, A., Santhanam, N., Jha, R. K., Sinha, R., & Jha, K. (2022). Multi class classification of acute leukemia using transfer learning. In *2022 International Conference for Advancement in Technology (ICONAT)* (pp. 1-6). IEEE. https://doi.org/10.1109/iconat53423.2022.9726083

Aftab, M. O., Awan, M. J., Khalid, S., Javed, R., & Shabir, H. (2021). Executing Spark BigDL for leukemia detection from microscopic images using transfer learning. In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)* (pp. 216-220). IEEE. https://doi.org/10.1109/caida51941.2021.9425264

Ahammed, P., Faruk, M. F., Raihan, N., & Mondal, M. (2022). Inception V3 based transfer learning model for the prognosis of acute lymphoblastic leukemia from microscopic images. In *2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)* (pp. 1-4). IEEE. https://doi.org/10.1109/icecte57896.2022.10114522

Al-khuzaie, M. Y., Zearah, S. A., & Mohammed, N. J. (2023). Developing an efficient VGG19-based model and transfer learning for detecting acute lymphoblastic leukemia (ALL). In *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-5). IEEE. https://doi.org/10.1109/hora58378.2023.10156679

Das, P. K., A, D. V., Meher, S., Panda, R., & Abraham, A. (2022). A systematic review on recent advancements in deep and machine learning based detection and classification of acute lymphoblastic leukemia. *IEEE Access, 10*, 81741-81763. https://doi.org/10.1109/access.2022.3196037

Das, P. K., & Meher, S. (2021). Transfer learning-based automatic detection of acute lymphocytic leukemia. In *2021 National Conference on Communications (NCC)* (pp. 1-6). IEEE. https://doi.org/10.1109/ncc52529.2021.9530010

Das, P. K., Sahoo, B., & Meher, S. (2023). An efficient detection and classification of acute leukemia using transfer learning and orthogonal softmax layer-based model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20*(3), 1817-1828. https://doi.org/10.1109/tcbb.2022.3218590

Donida Labati, R., Piuri, V., & Scotti, F. (2011). ALL-IDB: The acute lymphoblastic leukemia image database for image processing. In *2011 18th IEEE International Conference on Image Processing* (pp. 2045-2048). IEEE. https://doi.org/10.1109/ICIP.2011.6115881

Francese, R., Frasca, M., Risi, M., & Tortora, G. (2022). A deep learning and genetic algorithm based feature selection processes on leukemia data. In *2022 26th International Conference Information Visualisation (IV)* (pp. 412-417). IEEE. https://doi.org/10.1109/iv56949.2022.00074

Hasan, M. M., Srizon, A. Y., Sayeed, A., & Hasan, M. A. M. (2020). Accurate recognition of leukemia sub-types by utilizing a transfer learned deep convolutional neural network. In *2020 11th International Conference on Electrical and Computer Engineering (ICECE)* (pp. 427-430). IEEE. https://doi.org/10.1109/icece51571.2020.9393031

Jiwani, N., Gupta, K., Pau, G., & Alibakhshikenari, M. (2023). Pattern recognition of acute lymphoblastic leukemia (ALL) using computational deep learning. *IEEE Access, 11*, 29541-29553. https://doi.org/10.1109/access.2023.3260065

Kim, S. (2022). Developing a leukemia diagnostic system based on hybrid and ensemble deep learning architectures. In *2022 International Communication Engineering and Cloud Computing Conference (CECCC)* (pp. 70-73). IEEE. https://doi.org/10.1109/ceccc56460.2022.10069094

Laosai, J., & Chamnongthai, K. (2018). Deep-learning-based acute leukemia classification using imaging flow cytometry and morphology. In *2018 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)* (pp. 427-430). IEEE. https://doi.org/10.1109/ispacs.2018.8923175

Magpantay, L. D. C., Alon, H. D., Austria, Y. D., Melegrito, M. P., & Fernando, G. J. O. (2022). A transfer learning-based deep CNN approach for classification and diagnosis of acute lymphocytic leukemia cells. In *2022 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 280-284). IEEE. https://doi.org/10.1109/dasa54658.2022.9765000

Mourya, S., Kant, S., Kumar, P., Gupta, A., & Gupta, R. (2019). *ALL Challenge dataset of ISBI 2019 (C-NMC 2019)* (Version 1) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.dc64i46r

Patil Babaso, S., Mishra, S. K., & Junnarkar, A. (2020). Leukemia diagnosis based on machine learning algorithms. In *2020 IEEE International Conference for Innovation in Technology (INOCON)* (pp. 1-5). IEEE. https://doi.org/10.1109/inocon50539.2020.9298321

Ramagiri, A., Jahnavi, V., Gottipati, S., Monica, C., Afrin, S., Jyothi, B., & Chinnaiyan, R. (2023). Image classification for optimized prediction of leukemia cancer cells using machine learning and deep learning techniques. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)* (pp. 193-197). IEEE. https://doi.org/10.1109/icidca56705.2023.10099528

Sheet, J., Ghosh, C., & Das, B. K. (2023). Deep learning-based transfer learning for the detection of leukemia. In *2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)* (pp. 1-6). IEEE. https://doi.org/10.1109/isacc56298.2023.10084138

Talupuri, P. K., Naseeba, B., Challa, N. P., & Sathwik, A. S. (2023). Analysis of acute lymphoblastic leukemia detection methods using deep learning. In *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)* (pp. 565-574). IEEE. https://doi.org/10.1109/icssas57918.2023.10331776

Ul Ain, Q., Akbar, S., Hassan, S. A., & Naaqvi, Z. (2022). Diagnosis of leukemia disease through deep learning using microscopic images. In *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)* (pp. 1-6). IEEE. https://doi.org/10.1109/icodt255437.2022.9787449

Ullah, A. (2021). Image analysis of cells acute lymphoblastic leukemia using ensemble learning of deep bagging. In *2021 IEEE 7th International Conference on Computing, Engineering and Design (ICCED)* (pp. 1-4). IEEE. https://doi.org/10.1109/icced53389.2021.9664867

Yadav, D. P. (2021). Feature fusion based deep learning method for leukemia cell classification. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)* (pp. 1-4). IEEE. https://doi.org/10.1109/iscon52037.2021.9702440

Pertanika J. Sci. & Technol. 34 (S2): 165 - 192 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

An Optimised Deep Learning Framework for Skin Lesion Classification Using ResNet150 and M-GSO

Gurram Sunitha[1]*, Ahmed Elngar[2], Pravin R. Kshirsagar[3], and Subba Rao Polamuri[4]

[1]Department of Computer Science and Engineering, School of Computing, Mohan Babu University, Sree Sainath Nagar, Tirupati 517102, Andhra Pradesh, India
[2]Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef 62511, Egypt
[3]Department of Electronics and Telecommunication Engineering, J D College of Engineering and Management, Borgaon Fata, Kalmeshwar Road, Nagpur 441501, Maharashtra, India
[4]Department of Computer Science and Engineering, School of Computing, Aditya University, Aditya Nagar, ADB Road, Surampalem 533437, Kakinada District, Andhra Pradesh, India

ABSTRACT

Skin cancer is among the most prevalent cancers worldwide, affecting approximately 10 million people, and early diagnosis is critical for effective treatment. Manual assessment of dermoscopic images is time-consuming and prone to inter-clinician inconsistency, motivating the need for reliable automated classification systems. While deep networks offer strong representation power, training very deep models often faces challenges such as vanishing gradients, poor convergence, and overfitting. This paper proposes a modified ResNet150 architecture combined with a Leaky ReLU with Adaptive Factor (LRAF) activation to improve convergence and information flow across layers. To further enhance training, a Modified Group Search Optimisation (M-GSO) algorithm is employed for hyperparameter tuning, balancing exploration and exploitation in the search space. Experiments were conducted on the ISIC 2018 dataset, with data augmentation and SMOTE applied to address class imbalance. The proposed ResNet150 with M-GSO framework outperformed baseline ResNet150 and its non-optimised variants, achieving 98% sensitivity, 98.5% precision, 99% AUC, and high training stability. An ablation study confirmed that M-GSO optimisation plays a critical role in boosting accuracy and reducing false positives, while cross-dataset evaluation on ISIC 2019 and HAM10000, without retraining, demonstrated stable generalisation. These results indicate a trade-off between accuracy and computational cost, as M-GSO improves performance at the expense of increased tuning time. Overall, the findings show that integrating deep residual learning with adaptive activations

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 23 October 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.08

E-mail addresses:
gurramsunitha@gmail.com (Gurram Sunitha)
elngar_7@yahoo.co.uk (Ahmed Elngar)
pravinrk88@yahoo.com (Pravin R. Kshirsagar)
psr.subbu546@gmail.com (Subba Rao Polamuri)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

and population-based optimisation yields a more accurate and reliable solution for automated skin lesion classification.

Keywords: Group search optimiser, hyperparameter optimisation, medical image analysis, ResNet150, skin lesion classification

INTRODUCTION

Access to reliable medical services is essential for the public to lead a healthy life. People living in rural and urban regions lack access to medical services due to shortages of medical professionals and equipment. So, it becomes tough for patients to access standard and better medical services (Zhang et al., 2021). The healthcare industry can overcome this inaccessibility by employing modern technologies in two probable approaches. The first step is to use the Internet of Things (IoT) to shift medical care beyond hospitals and connect various medical devices. As a second step, the quality of healthcare services can be enhanced using modern Deep Learning (DL) methodologies (Li et al., 2021). In the last ten years, the dangerous form of skin cancer known as malignant Melanoma has increased dramatically. It is triggered by abnormal skin cell growth. Melanomas are otherwise called malignant lesions; they begin at the melanocyte level and spread in the skin's upper regions. The parts exposed to direct sunlight are highly prone to this disease. When malignant melanoma is not detected in the early stages, it will lead to a high death rate (Naqvi et al., 2023). The World Health Organisation (WHO) states that every year, 3 million or more cases of non-melanoma and 132,000 cases of melanoma are identified globally (World Health Organisation, 2017). The World Health Organisation reports 9,500 cases of skin cancer every day. The statistics indicate that there has been a 47% growth in invasive melanoma cases in the past decade. In 2022, 331,772 skin cancer cases were diagnosed worldwide, with nearly 66,000 cases of mortality (Ferlay et al., 2024). The number of males diagnosed with skin cancer is nearly 174,000, and the number of females diagnosed with skin cancer is nearly 151,000. The mortality cases were 32,000 and 25,000, respectively (International Agency for Research on Cancer, 2022).

By 2042, the expected number of new cases of Melanoma is 106,110, and the estimated number of deaths is 7,180 (Ferlay et al., 2024). The statistics above show an increase in melanoma cases. When overall cancer cases are considered, 75% of skin cancer deaths are due to Melanoma. The number of new melanoma cases discovered each year in Australia is 15,628. In Europe, the number of new cases reported yearly exceeds 20,800. Furthermore, an average of 1200 to 20,000 deaths occurs every year in Australia and Europe due to Melanoma. Since the 1990s, the skin cancer detection rate has increased steadily. The rise in skin cancer cases is mainly linked to ozone layer depletion and greater use of solariums, which expose people to harmful ultraviolet radiation. Prolonged exposure to

such radiation increases the risk of abnormal skin cell growth, leading to higher chances of cancer development. However, a recent statistic reveals that the survival rate can be increased up to 98% if skin cancer symptoms are spotted in their early stages.

Skin lesions may show variations in colour (brown, black, red, pink) and irregular shapes. Such heterogeneity, particularly the presence of multiple colours, is often associated with malignant lesions and may indicate disease progression or severity. To identify possible melanoma, naevi with abnormal colour and a diameter of 6 mm or more are more likely to be associated with the disease. Since lesions can vary in texture and colour, the primary identification of skin cancer is a challenging task (Dildar et al., 2021). Firstly, a dermatologist visually examines factors such as shape, size, lesion diameter, and abnormalities in mole colour. Then dermatologists use the following primary methods for diagnosing skin lesions. The pattern evaluation checklist using 7 points and the ABCD rule (edge, diameter, asymmetry, and colour) is used, along with dermoscopy and biopsy. In recent years, dermatologists have used microscopic and dermoscopy imaging techniques for skin cancer diagnosis (Esteva et al., 2017). A mobile camera can view the microscopic images as low-resolution images are produced. Dermoscopy is a method used by dermatologists to visually examine and diagnose malignant skin cancer lesions.

However, the above mentioned approaches face several challenges due to the following reasons: (i) dermoscopic and skin lesion images are often low-contrast making precise detection of lesions hard, (ii) diseased and healthy areas often share similar colors, making classification difficult and thereby increasing misclassification rate, (iii) lesions occurring on image boundaries lead to segmentation errors, (iv) segmentation accuracy is affected by variations in brightness of the image, (v) complex nature of lesions having various types, sizes, colours, textures and shapes (vi) few lesions are erroneously segmented due to appearance of hair, bubbles and blood vessels, (vii) visual variation among the intra and inter type skin lesions, and (viii) slight resemblance among the lesions types which makes the classification process more complex (Goyal et al., 2020). Unbalanced datasets are also a big issue for classification. Since classes with many images dominate those with fewer, there will be a considerable difference in classifier performance. Additionally, other possible difficulties with these methodologies are the use of handcrafted approaches, which leads to (a) the model's overfitting problem and (b) the high computational cost. The issues mentioned above can be minimised using the proposed ResNet150-M-GSO model for early detection of skin cancer. The work proposes its significant contributions as follows:

- In the preprocessing step, resizing, correlation, and normalisation improve the visibility of image edge details. Data augmentation is applied to increase dataset diversity and balance classes, helping the model generalise better.
- A depth-extended ResNet architecture (ResNet150) is constructed to enhance hierarchical feature representation for dermoscopic texture analysis.

- Feature extraction is enhanced using a Modified Group Search Optimisation (M-GSO) strategy, improving prediction performance through effective hyperparameter optimisation.
- The 150-layer deep network performs feature fusion, where only the most significant features are selected to achieve optimal outcomes.
- A fully connected network-based classifier is used for final feature classification. This ensures reliable lesion detection.
- The proposed model's stability is verified through a comprehensive statistical analysis, and the results are compared with those of recent methodologies.
- Dermatologists can save a considerable amount of time in identifying skin cancer and, at the same time, provide improved recognition accuracy. The proposed method achieves superior results with low computational time.

RELATED WORKS

High-resolution lesion images reduce skin reflectivity, allowing dermatologists to examine deeper skin structures. However, other skin cancer types and melanoma have similar features, which makes cancer diagnosis more difficult for even experienced dermatologists. Hence, a thorough and careful lesion diagnosis is crucial. This difficulty delays dermatologists' diagnosis of melanoma, thereby reducing the patient's chance of survival. So, diagnosing the disease at an early stage is vital, sparing the patient the painful biopsy process and improving the patient's chances of survival (Giuffrida et al., 2020).

Meanwhile, the process of skin lesion recognition done manually, even by experienced dermatologists through dermoscopy, consumes too much time, along with a high error rate and only 75% accuracy is achieved for melanoma detection. Researchers focus more on techniques such as automated detection systems to address Melanoma and its diagnostic challenges. Therefore, several Computer-aided diagnosis (CAD) methods have been proposed by researchers to address concerns regarding lesion detection, such as reduced detection time, and to simplify this diagnostic procedure by using DL methods to recognise skin lesions efficiently and automatically in the early phase from dermoscopic images (Iqbal et al., 2023). Modern skin lesion detection techniques make the diagnostic tool even more capable of classifying Melanoma and non-melanoma skin lesions (Bikku et al., 2018).

An improved VGG model is developed to enhance the accuracy of skin cancer detection (Tabrizchi et al., 2023). The study demonstrated performance gains by refining network depth and feature extraction. A skin cancer detection method using VGG16 and VGG19 with transfer learning is proposed (Faghihi et al., 2024). The results showed high accuracy, demonstrating the effectiveness of deep feature extraction from pretrained networks. Their study highlights transfer learning as a strong approach for medical image classification.

To enable dermatologists to identify skin lesions accurately and quickly, CAD approaches are used extensively in clinical imaging (Taye, 2023). Dermatologists depend on CAD methods for a second opinion. For skin lesion detection, the building blocks in CAD systems are image acquisition, feature extraction, segmentation, selection of skin image datasets and their fusion, and classification of skin lesion types. Deep features are extracted from deep neural layers (Bikku & Satya Sree, 2020). They are then utilised for classification. Deep feature analysis encompasses the image's local and global information, whereas traditional methodologies use features such as the skin lesion's shape, texture, and colour. The convolutional layers give global and local information (Asadi et al., 2023). During the segmentation phase, the wound area and boundaries of the infected region are often thin and difficult to delineate. In the classification phase, these localised lesions must be accurately assigned to their respective classes, such as benign keratosis, nevus, or melanoma (Raval & Undavia, 2023).

Various methodologies have been proposed to address segmentation and classification challenges, but none has consistently achieved high accuracy (Morid et al., 2021). Skin cancer has multiple categories, yet clinically, melanoma and benign lesions are most relevant. Differentiating melanoma from benign lesions enables high binary classification accuracy (Hasan et al., 2021). Some approaches follow a sequential process for initial lesion segmentation, identifying regions of interest and extracting discriminative features. These are then fed to a classifier model for final lesion categorisation. Classification accuracy largely depends on the quality of feature extraction (Bikku, 2020). Most studies treat feature selection and classification as separate processes; however, evaluating features experimentally is often more effective, as some important features may otherwise be overlooked. Accurate classification requires precise identification of local lesion regions to improve overall performance (Gajera et al., 2023).

Skin lesions are classified using binary classification. Extending this to multi-class classification is more challenging because many lesions share similar characteristics (Asif et al., 2025). Many classification methods rely only on deep learning–based feature extraction (Alwakid et al., 2022). However, redundant features are often not addressed, leading to errors in multi-class lesion classification. So, selecting the most suitable features and fusing them for final classification is crucial (Ji et al., 2024). Recently, accuracy has been significantly enhanced by selecting optimal features and fusion methods (A et al., 2024). The literature describes various selection and fusion methods. Improving the object information from different sources is the central goal of the fusion approach.

RESEARCH METHODOLOGY

This section provides a detailed discussion of the ResNet150 with the M-GSO framework to attain accurate skin cancer classification. This approach integrates an optimiser and deep models to improve the classification performance.

Data Preprocessing

The preliminary step in every image-based analysis is image preprocessing, in which raw images and their quality are enhanced to make the input more suitable for analysis (Altamimi et al., 2024). For skin cancer image classification, the following processes are included in the pre-processing step: resizing and cropping, rescaling, normalisation, colour correction, and noise reduction.

Data Augmentation

Data augmentation is employed to artificially increase diversity of the training data (Kebaili et al., 2023). Various transformations are applied to input images to improve the learning model's generalisation. This process aims to improve the learning model's robustness and performance. The data augmentation step involves deliberately modifying images during the training process. The data augmentation step is performed with careful consideration of numerous transformation techniques. Table 1 summarises the data augmentation performed on the training dataset. These operations are chosen based on the efficiency of dataset diversification. It is intended to handle image classification problems by exposing the model to different data variations.

Table 1
Data augmentation parameters used to enhance the diversity of the training dataset

Process	Range
Rescaling	1/255
Rotation	±20%
Width/Height Shifting	±20%
Zooming	±30%
Horizontal Flipping	Enabled

The proposed work employed data augmentation to attain triple advantage. First, the training dataset's size is significantly expanded, which is valuable when working with limited data resources. Second, the system can generalise and perform effectively in unexposed scenarios since the model is trained on diverse data variations. Third, data augmentation addresses certain issues, such as eliminating overfitting by introducing noise into the dataset. The learning model's reliability, performance and generalisation capabilities are significantly improved by data augmentation.

Leaky ReLU Activation Function (LRAF)

The commonly employed activation function is ReLU. The statistical expression in Equation 1 represents the ReLU function.

$$ReLU = \begin{cases} 0 & w < 0 \\ w & w \geq 0 \end{cases} \quad [1]$$

However, in this research, LRAF is proposed instead of the ReLU function. The problem of dead neurons in the ReLU function is eliminated by employing LRAF (Dubey

& Jain, 2019; Mastromichalakis, 2020). A slight negative slope for every negative input value introduced by the dead ReLU is handled by LRAF. It deals with the even gradient flow for negative inputs. The LRAF activation function is mathematically expressed as in Equation 2.

$$f(w) = \begin{cases} w & w \geq 0 \\ bw & w < 0 \end{cases} \quad [2]$$

While training the model, the factor 'b' is produced unevenly and it will turn into a static value. This procedure permits one-sided limitation by preserving specific harmful gradient data, thereby eliminating the overall loss. LRAF is utilised to concatenate and reshape every layer. When enormous noise is present, generating a non-zero output value is required to identify a negative input value; thereby, data leakage is avoided. The system's performance can be enhanced by minimising image noise. The advantages of ResNet and LRAF in this research are aimed at improving image classification. Owing to LRAF's specific advantages, it is combined with the ResNet model to provide faster execution and computation. ResNet is used alongside LARF to enhance the deep network's training performance and convergence while preventing gradient vanishing. Figure 1 presents the comparison of ResNet block structures using ReLU activation and LRAF activation.

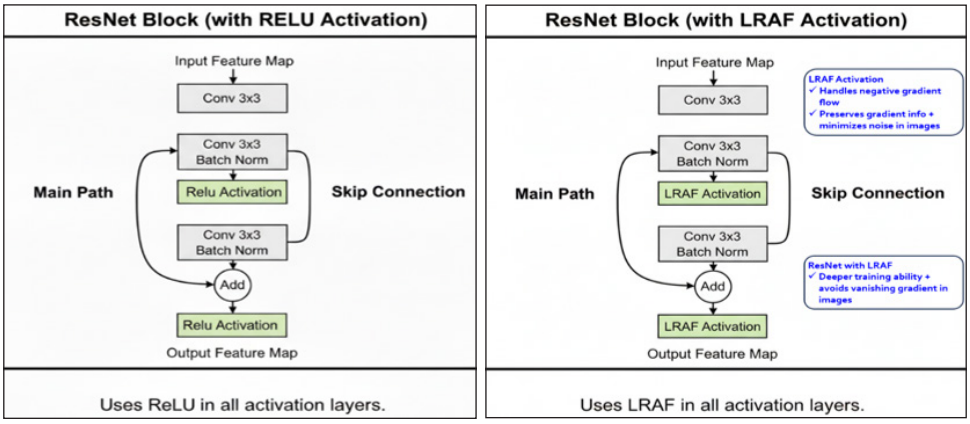


Figure 1. Comparison of ResNet block structures using ReLU activation (left) and LRAF activation (right)

ResNet150 Deep Model

Figure 2 presents the layer architecture of the proposed ResNet150. The proposed ResNet150 comprises 150 layers. ResNet-50 is extended by adding additional convolutional blocks, resulting in a deeper architecture referred to as ResNet150.

- conv3_x stage: 8 blocks (vs 4 in ResNet-50)
- conv4_x stage: 36 blocks (vs 6 in ResNet-50)

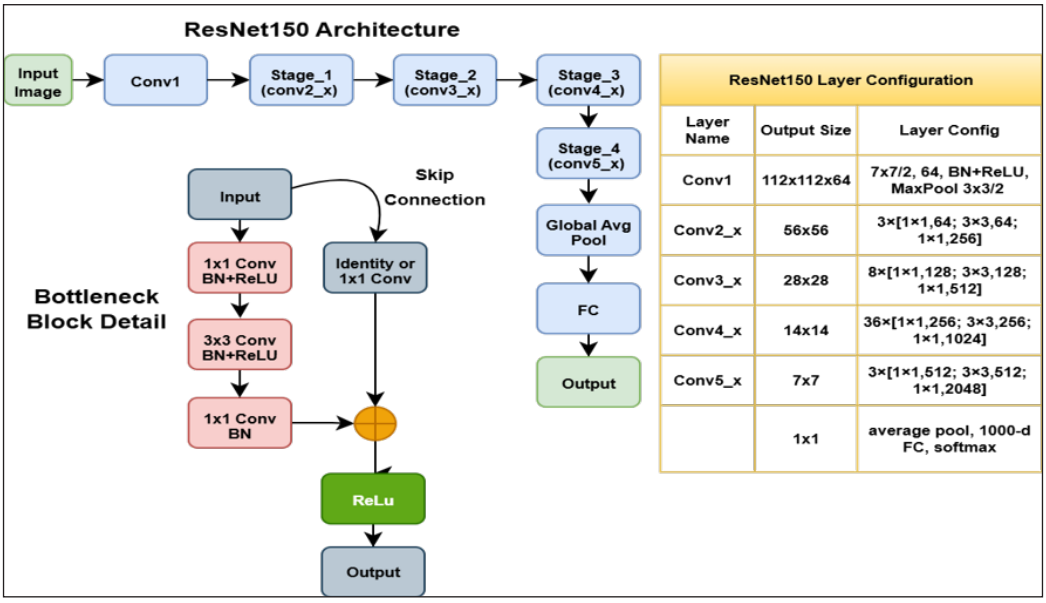


Figure 2. Architecture of ResNet150, illustrating convolutional stages, bottleneck blocks, and layer configuration for classification

ResNet150 (Residual Neural Network) includes 150 deep layers. The ResNet model comprises multiple stages, each containing identity and convolutional blocks. The pre-processed image is fed to the ResNet model based on Equation 3 given below:

$$J(y) = M(y) + y \quad [3]$$

In Equation 3, the input image is denoted by y The non-linear layer mapping is represented as $M(y)$ and the residual mapping is denoted as $J(y)$.

The convergence is more challenging in deep networks. LRAF is employed to avoid loss of input information. LRAF includes an alpha parameter, resulting in a minor but nonzero slope. Previously inactive nodes in ReLU will now alter their weights. A modified ResNet model for skin cancer classification substitutes the standard ReLU activations with LRAF activations. So, LRAF and ResNet150 are combined to enhance ResNet150's training ability, performance, and convergence. Thereby eliminating the vanishing gradient problem. The above change is done throughout the network, i.e., in fully connected layers and residual blocks. As an alternative to ReLU, the output of every convolutional layer in ResNet150 is passed to the residual block using the LRAF activation function. The LRAF with a convolutional layer is expressed as Equation 4 given below:

$$O_c = L_R\left(B_N(C_{2D}(I, F, K_s))\right) \quad [4]$$

Here, the convolutional layer output is denoted as O_c, LRAF is represented as L_R, and batch normalisation is represented as B_N. 2D convolution operation is denoted as C_{2D}. I denote input, F denotes filters, and Ks denotes kernel size. The Conv2D's output is normalised using batch normalisation. The embedded LRAF in the residual block is given by Equation 5 below.

$$O_R = L_R\ (B_N(C_{2D}(I), C_{2D}(I))) \quad [5]$$

Optimisation

The Modified-GSO (M-GSO) is proposed to optimise the hyperparameters of ResNet150, as it provides a better balance between exploration and exploitation than PSO, GA, and Bayesian optimisation. Other models often suffer from local optima/slower convergence. In preliminary trials, M-GSO achieved faster convergence and more stable performance on medical image classification tasks. Unlike SGO, M-GSO is implemented by improving the global search phases. The greedy selection technique is used in each phase to select a new population stronger than the previously chosen one. The linear population reduction approach in the exploration system supports exploration and allows transition from exploration to exploitation.

Global Search

The quality of the solution is improved by exploring the leader. A self-introspection factor acts as a control mechanism by imposing a restriction on the population velocity. This is followed by a greedy selection phase, in which fit members are retained, and non-fit members are discarded. This procedure works quite well for both multimodal and unimodal problems. But it often fails in dynamic search by being overly dependent on its leaders, resulting in very low efficiency when shifting composite functions. If a member's condition does not improve under this greedy selection mechanism, the population will stall, allowing lower-fitness solutions to be eliminated.

This enhanced phase explores a broader search space while simultaneously learning from the leader's experience. This phase is divided into two stages: the improved explorative phase, which is responsible for function evaluation, and the exploitative phase, as defined in Equation 6.

The following equations describe the position update mechanism of M-GSO. Each candidate solution is influenced by three main forces:

i. attraction towards the leader to find the best solution

ii. repulsion from the worst solution to avoid poor regions

iii. stochastic differential perturbations between randomly selected population members to maintain diversity.

These components collectively balance global exploration and local exploitation during optimisation.

$$\vec{P}_{new}\ (t+1) = \begin{cases} NFE < 0.5 * total\ NFE, Explorative \\ else, Exploitative \end{cases} \quad [6]$$

A large population is available at the beginning of the search process, which is used to promote the enhanced exploration phase. New solutions are generated by employing diverse vectors over the current population to explore the search space. The new solution is obtained from Equations 7-14. Choose a member P_r randomly from the population pool. Equation 7 updates the position of candidate solution $P_{i,j}$ at iteration t+1. An update occurs if the fitness of the current solution i is better than that of the reference solution r. R controls the influence of i. r1, r2, r3 are random coefficients for exploration. X1, Y1, Z1 represent search directions. Equation 8 represents the direction from i to r. It encourages movement toward a neighbour solution. Equation 9 represents the direction from r towards the leader solution. It guides search towards the current best candidate (leader). Equation (10) calculates the difference between the reference and worst solutions. It helps avoid poor search areas. Equation 11 is an alternative update rule that uses the control parameter CDC_DCD. It influences movement toward good solutions while reducing the influence of poor ones. Equation 12 directs i towards the leader. It creates a strong pull towards the best solution. Equation 13 directs r2 towards the leader. It maintains population diversity by incorporating an additional peer reference. Equation 14 calculates the contrast between another candidate and the worst solution. It balances exploration by avoiding worst-performing regions.

$$If\ f(P_i) < f(P_r).\ P_{i,j}^{(t+1)} = R * \vec{P}_{i,j}^{(t)} + r1 * \left[\vec{X}_1\right] + r_2 * \left[\vec{Y}_1\right] + r_3 * [\overrightarrow{Z_1}] \quad [7]$$

$$\vec{X}_1 = [\vec{P}_{r,j}^{(t)} - P_{i,j}^{(t)}] \quad [8]$$

$$\vec{Y}_1 = [Leader_j - P_{r,j}^{(t)}] \quad [9]$$

$$\vec{Z}_1 = [\vec{P}_{r,j}^{(t)} - \overrightarrow{worst_t}] \quad [10]$$

$$P_{i,j}^{(t+1)} = C_D * P_{r,j}^{(t)} - r_1 * \left[\vec{X}_2\right] - r_2 * \left[\vec{Y}_2\right] - 0.1 * [\vec{Z}_2] \quad [11]$$

$$\vec{X}_2 = [Leader_j - P_{i,j}^{(t)}] \quad [12]$$

$$\vec{Y}_1 = [Leader_j - P_{r2,j}^{(t)}] \quad [13]$$

$$\vec{Z}_1 = [\vec{P}_{2,j}^{(t)} - \overrightarrow{worst}_t] \quad [14]$$

Equations 7 to 14 together control directional movement toward promising regions while preventing premature convergence. The difference vectors increase population diversity, whereas leader-guided movement accelerates convergence toward optimal hyperparameter combinations.

To enhance the diversity of the population, difference vectors $\vec{X}$, $\vec{Y}$, and $\vec{Z}$ are introduced. $\vec{P}_{i,j}^{(t)}$ indicates the location of ith population for j^{th} dimension in t^{th} iteration. $\vec{P}_{r,j}^{(t)}$ and $\vec{P}_{r2,j}^{(t)}$ signify any two randomly selected population members' positions from the present iteration. A random number denoted as *R*, takes values from 0 to 1. It is updated dynamically in each iteration. The finest solution gained so far is $Leader_j$ and the least good solution from the present iteration obtained is $\overrightarrow{worst}_t$. The success history is the dynamic self-introspection factor in the range [0.1, 1]. The reason for including the worst solution is to allow and preserve increased diversity during the exploration process. At any given point, multiple difference vectors help avoid solution clustering, and adaptive self-introspection enables the particle to explore and navigate the solution space in a controlled manner. The exploitation phase of M-GSO performs actual position updates. It includes a new system for feedback position updates with a probabilistic selection between the two approaches, as expressed in Equations 15 to 19. Equation 15 represents the exploitation phase (move towards the leader). sel > 0.5 is a probabilistic condition controlling exploitation. CR is a control parameter that regulates the step size. $P_{i,j}^{(t)}$ is the current position of agent i in dimension j. $Leader_j$ is the position of the leader in dimension j. r1 is a random weight (0–1) adding stochasticity. Equation 15 pulls the solution closer to the leader, encouraging exploitation.

Equation 16 represents the exploration + diversification phase (influence of FB-leader, random peer, and FB-worst). F, whose value is $1+r+C_R$, introduces adaptive scaling. $P_{FB-leader,j}^{(t)}$ is the feedback-based leader position (best after reduction). $P_{r2,j}^{(t)}$ is the random peer solution. $P_{FB-worst,j}^{(t)}$ is the feedback-based worst solution. A correction factor of −0.01 helps avoid clustering near the worst solutions. This equation balances exploration (the difference between the leader and a randomly chosen peer) and diversification (moving away from the worst).

Equations 17 and 18 represent the dynamic handling of the best and worst solutions when the population size decreases. Equation 17 avoids this by dynamically selecting a

feedback-based leader in response to population updates. It maintains elitism by retaining the best solutions as N_P is reduced. Equation 18 is the rule for selecting the worst solution. Equation 19 represents adaptive scaling for stability. It is a scaling factor that adjusts search dynamics. rand adds randomness. C_R provides adaptive control. F allows for a balance between intensification (leader attraction) and diversification (avoiding the worst regions).

$$If\ sel > 0.5\ \ P_{i,j}^{(\vec{t}+1)} = C_R * \overrightarrow{P_{i,j}^{(t)}} + r_1 * [leader_j = P_{i,j}^{(\vec{t})} \quad [15]$$

$$Else\ \ P_{i,j}^{(\vec{t}+1)} = C_R * \overrightarrow{P_{i,j}^{(t)}} + F * \left[P_{FB-leade\,r_j}^{(t)} - P_{r2,j}^{(\vec{t})}\right] - 0.01 * [P_{r2,j}^{(\vec{t})} - P_{FB-wors\,t_j}^{(t)}] \quad [16]$$

$$P_{FB-leade\,r,j}^{(t)} = \begin{cases} FB - Leader, j & if\ N_P\ is\ reduced \\ P_{r,j}^{t} & else \end{cases} \quad [17]$$

$$P_{FB-worst\,,j}^{(t)} = \begin{cases} FB - worst, j & if\ N_P\ is\ reduced \\ P_{r,j}^{t} & else \end{cases} \quad [18]$$

$$F = 1 + rand + C_R \quad [19]$$

The exploitation scheme selector is represented as Sel. FB-Worst and FB-Leader denote the worst and best solutions in the rejected population. They supply feedback to the present population.

Observing Phase

The M-GSO observing phase aims to enhance population diversity by leveraging comparative learning among the leader and population members. This phase builds on the idea of data exchange in a society where every member of the population communicates randomly with other members of the society and with the leader. Based on the two interacting members' intellects, information is gained or passed among the members of the population. Quick convergence is achieved during the observation phase, and population diversity decreases due to the addition of a greedy operator to each solution obtained. To enhance the production of a diverse population between the two population members, a fitness-based selection system is implemented by the adaptive observing phase. By employing this method, premature convergence caused by stagnation and entrapment can be avoided. With the current greedy selection technique, reinitialisation is not an option

since it contributes negligible diversity to the overall population. The population updates are given in Equation 20.

$$If\ f(P_i) < f(P_r)\ \ P_{i,j}^{(\vec{t}+1)} = P_{i,j}^{(\vec{t})} + C_D * [Leader_j - P_{r,j}^{(\vec{t})}]$$
$$else\ P_{i,j}^{(\vec{t}+1)} = P_{i,j}^{(\vec{t})} - C_D * [Leader_j - P_{r,j}^{(\vec{t})}] \quad [20]$$

Parameter Selection

The M-GSO performance can be improved by dynamically adapting the self-introspection factor to the sequence of failures and successes. Analysis of various inertial control settings shows that a static setting of c=0.2 yields good performance on both multi-modal and uni-modal landscapes. In contrast, inertia increases towards ccc as iterations progress, depending on the learning mechanism. The CD value is also adaptively adjusted to strengthen the exploration process, as shown in Equation 21.

$$CD_{new} = \begin{cases} CD_{old} & Failures < learning\ rate \\ \left(0.2 * C_{D_{int}}\right) + (0.2 * rand) & Failures \geq learning\ rate \end{cases} \quad [21]$$

In the adaptive technique, the CD value is retained for new population members that achieve improved fitness. When extreme failure counts occur, it is re-initialised. At initialisation, the failure value is set to zero. During the adaptive acquiring and enhanced improving phases, the failure counter is incremented by one for each unsuccessful member that does not generate better offspring. An empirical learning rate of 10 is used to ensure sufficient trials to form a fresh CD combination, thereby enhancing solution quality. The CR value is set within the range of 0.2 to 1.0, consistent with the standard SGO recommendations for c. At the end of each iteration, the CD value is updated, R is randomised between 0.1 and 1.0, and CD alterations are applied to avoid reliance on static control parameter settings. The failure counter update is represented in Equation 2, where f(x) is the fitness function value, xnew is the offspring, and xparent is the parent solution. At initialisation, Failure=0.

$$For\ each\ iteration, Failure = \begin{cases} 0, if\ f(x_{new}) < f(x_{parent}) \\ Failure + 1, Otherwise \end{cases} \quad [22]$$

Adaptive CD update is represented in Equation 23 where CDt is current CD value in iteration t, n is empirical learning rate (=10), F_{max} is threshold for maximum allowed failures, CD_{init} is reinitialised CD value.

$$If\ Failure < F_{max}\ then\ CD_{t+1} = CD_t + n\,.rand(0,1)$$
$$else\ CD_{t+1} = CD_{init} \quad [23]$$

CR adaptation is represented in Equation 24, where U denotes a uniform random distribution.

$$CR \in [0.2,1.0]\ ,R \sim U[0.1,1.0] \quad [24]$$

Exploration and Exploitation

In M-GSO, achieving an appropriate balance between exploration and exploitation is critical to improving convergence and avoiding premature stagnation. The algorithm integrates adaptive control strategies through Controlled Degree CD, Crossover Rate CR, and the self-introspection factor c to manage proper balance dynamically.

Exploration Phase

- Higher diversity is maintained during early iterations by adapting CD values dynamically. When a solution repeatedly fails, its CD is reinitialised, forcing it to search new areas. This prevents population stagnation and ensures that the algorithm explores across multiple dimensions of the search space.
- The CR parameter is randomised between 0.2 and 1.0, thereby enhancing solution variability and avoiding premature convergence.
- c adjusts the population's velocity.

Exploitation Phase

- Greedy selection ensures that only solutions with improved fitness are retained, intensifying the search around promising regions.
- The adaptive learning mechanism modifies CD based on successful offspring, ensuring local refinement while discarding redundant or unfit members.
- The revised differential mutation operator further improves local search by fine-tuning around high-quality candidates.

Dynamic Balance

- The adaptive exploration–exploitation balance enables M-GSO to discover optimal regions globally and refine solutions locally, outperforming static search strategies. In the initial iterations, exploration of the global space is achieved using high CD values

and a randomised CR value. Over iterations, exploitation takes over, implemented using optimised CD values.

Complexity Analysis

M-GSO algorithm updates population positions once in each phase- the improving phase and the acquiring phase. Each update is followed by greedy selection, which ensures inferior solutions are discarded, and better ones are retained. At each iteration, this process yields two fitness evaluations per population member. The total computational complexity of M-GSO is determined by initialisation, fitness evaluations, position updates, and greedy selection. Let P_S be the population size, and PD be the problem dimension. Initialisation requires O ($P_S \times P_D$). Fitness evaluations, including the population, improvement, and search phases, each require O (PS × PD) time. The observing and selection phases require O(P_S) per iteration. For T iterations, the overall computational complexity O (T × PS × PD) is dominated by fitness evaluations and position updates. Since each fitness evaluation is performed twice in each iteration, the total number of function evaluations is 2×PS×T. The total computational time of the M-GSO is represented in Equation 25.

$$t_{total} = t_1 + O_1 + t_2 * O_2 + \cdots + t_N * O_{Ps} \quad [25]$$

where $O_1, O_2, \ldots, O_{Ps}$ denote computational complexities of different operations performed in M-GSO for a population size of Ps. $t_1, t_2, \ldots, t_N$ are the amounts of computational time required for each operation. The algorithm in Figure 3 presents the detailed steps of the M-GSO algorithm.

Table 2 summarises time requirements and complexities for the main phases of M-GSO. To provide a practical understanding of computational cost, consider PS = 50 (population size), PD = 5 (number of optimised hyperparameters), and T = 100 (iterations). The total number of fitness evaluations is 2 × PS × T = 2 × 50 × 100 = 10,000.

Table 2
Computational time intervals of different phases and time complexity in the proposed algorithm

Functionality	Time Interval	Complexity of Sub-Operation	Complexity of M-GSO
Initialise	t_1	O($P_S \times P_D$)	
Fitness based on population	t_2	O($P_S \times P_D$)	
Fitness based on improvement	t_3	O($P_S \times P_D$)	O($T \times P_S \times P_D$)
Fitness on searching	t_4	O($P_S \times P_D$)	
Observing phase	t_5	O(P_S)	
Selection phase	t_6	O(P_S)	

```
1. Start
2. Initialize the input parameters
3. Set learning rate and failure rate
4. Generate initial population randomly based on dimensional size
5. Predict worst and gbest
6. for t = 1 to T do
7.     for I = 1 to Np do
8.         Role assignment:
9.             Select one head searcher (guides direction)
10.            Select side searchers (explore nearby space)
11.            Select rear searchers (exploit around current best)
12.        Improvement phase:
13.            Head updates position using global direction
14.            Side searchers update positions at ±90° to head
15.            Rear searchers move toward historical gbest
16.        Compute new candidate positions P_new(t+1)
17.        Update failure rate and learning rate adaptively
18.    End for
19.    Update worst solution and gbest;
20.    Acquiring phase:
21.        Select random agent r, set i ≠ r
22.        if f(P_i) < F(P_r) then:
23.            Accept new candidate P_i^{i,j}(t+1)
24.        else
25.            Retain current candidate P_i^{i,j}(t)
26.        Update gbest after acquisition
27.        Verify termination condition
28.    End for
29.    Stop
```

Figure 3. M-GSO algorithm

Modified GSO swaps GSO's vision-based producer/scrounger/ranger movement for three fixed-geometry roles (head/side/rear) and adds memory via rear searchers moving toward historical gbest. It introduces adaptive learning and failure rates to self-tune step sizes, which standard GSO lacks. It adds a separate "Acquiring Phase" with greedy pairwise tournament selection (accept only if better than a randomly selected peer) — a mechanism borrowed from algorithms such as TLBO/ABC, not part of traditional GSO.

RESULTS AND DISCUSSIONS

Skin lesion images for model training were obtained from the ISIC Challenge 2018 dataset. It contains 1197 malignant and 1440 benign samples (Codella et al., 2019). The dataset

is partitioned using stratified random sampling into training (75%) and testing (25%) subsets, preserving the original class distribution. The test set comprises 660 images (300 malignant and 360 benign). No preprocessing, augmentation, or optimisation procedures were performed on the test dataset.

As the dataset exhibited class imbalance, oversampling was applied only to the training data to balance the classes. SMOTE technique was used, followed by data augmentation techniques including rotation, horizontal flipping, scaling, brightness jitter, and noise injection. Then, the training dataset is expanded to 2,000 images per class (malignant and benign). This procedure ensured that the testing dataset remained independent. Thereby preventing data leakage and guaranteeing unbiased performance evaluation. Figure 4 presents sample input and pre-processed images from the training set.

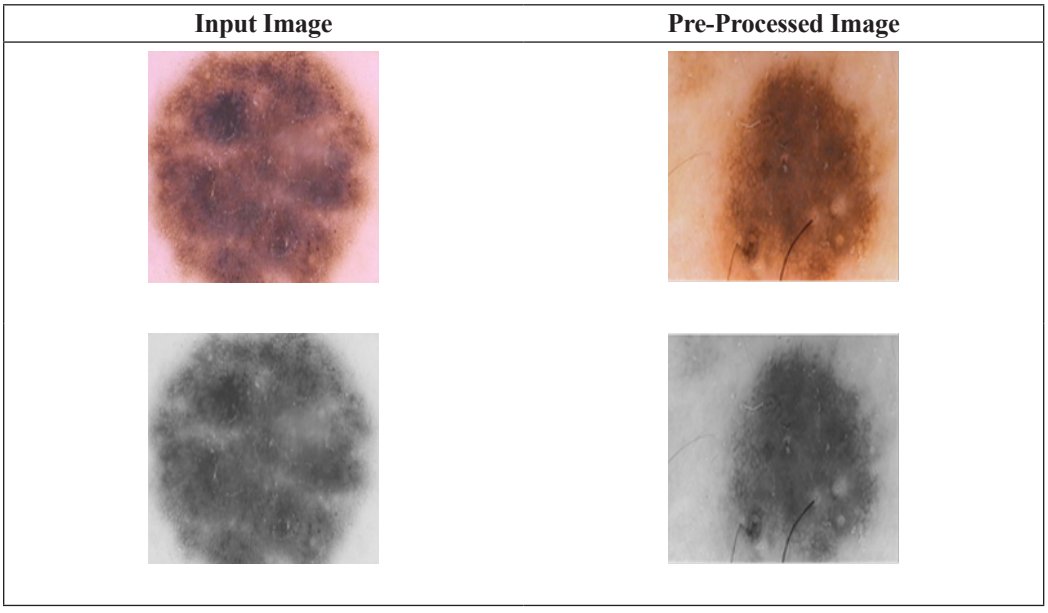


Figure 4. Results of the pre-processing stage showing enhanced and normalised input images

In ResNet150 with M-GSO, the training batch size is fixed at 32, and the number of epochs is set to 25. The model achieved stable accuracy improvements with this setup. During training, dropout rates of 0.2 and 0.4 were applied to reduce overfitting. These dropout values helped balance the trade-off between model generalisation and accuracy. The validation accuracy is consistent with the training accuracy, indicating better generalisation by the algorithm.

Further experiments were performed by changing the learning rate while keeping the batch size fixed. It is observed that a learning rate of 0.001 provided the best accuracy, reaching 97%. When the learning rate was increased, the model suffered from unstable

convergence, while a lower rate slowed training. Based on these observations, the learning rate of 0.001 is finalised for the next training runs.

Different image sizes were tested to analyse their effect on accuracy. With the training batch fixed at 32 and the learning rate at 0.001, image sizes of 128, 160, and 224 were used. The results showed that 224 × 224 provided the highest accuracy, while smaller sizes reduced training time but slightly lowered precision. Hence, 224 × 224 is selected for subsequent experiments.

Table 3 presents the initial hyperparameter search space explored by M-GSO during optimisation. These values define the candidate ranges evaluated by the optimiser. The final selected configuration after optimisation is reported separately in Section 4.1.

Table 3
Hyperparameter search space explored by M-GSO

Component	Parameter	Value / Setting
M-GSO	Population size	50
	Number of iterations	100
	CDCD initialisation	Random within [0,1]
	Factor / Step size	0.5
	Patience	1
ResNet150 Backbone	Number of layers	150
	Training batch size	32 – 64
	Learning rate	0.0001 – 0.001
	Dropout rates	0.2 – 0.5

Figure 5 presents the experimental results of ResNet150 with the M-GSO learning model. The proposed model's performance during training is assessed using various performance metrics. The M-GSO is used as a hyperparameter optimisation technique to select suitable values for hyperparameters such as learning rate, batch size, and dropout rate. Once these hyperparameters are chosen, the Adam optimiser is used during training to update the model weights. Adam optimiser is used with a learning rate of 0.001, and categorical cross-entropy is the loss function used. ReduceLROnPlateau callback is used with a factor of 0.5 and patience of 1. Then, ResNet150 with the M-GSO model is trained using both the validation and training datasets.

Different combinations of dropout rates within the defined search range (0.2–0.5) were evaluated through the M-GSO optimisation process. It was observed that using a lower dropout rate in the first layer and a moderately higher rate in deeper layers effectively controlled overfitting. When the dropout rate exceeded 0.5, model accuracy decreased significantly due to underfitting. The final selected dropout values were 0.2 and 0.4. They provided the best balance between training stability and generalisation performance.

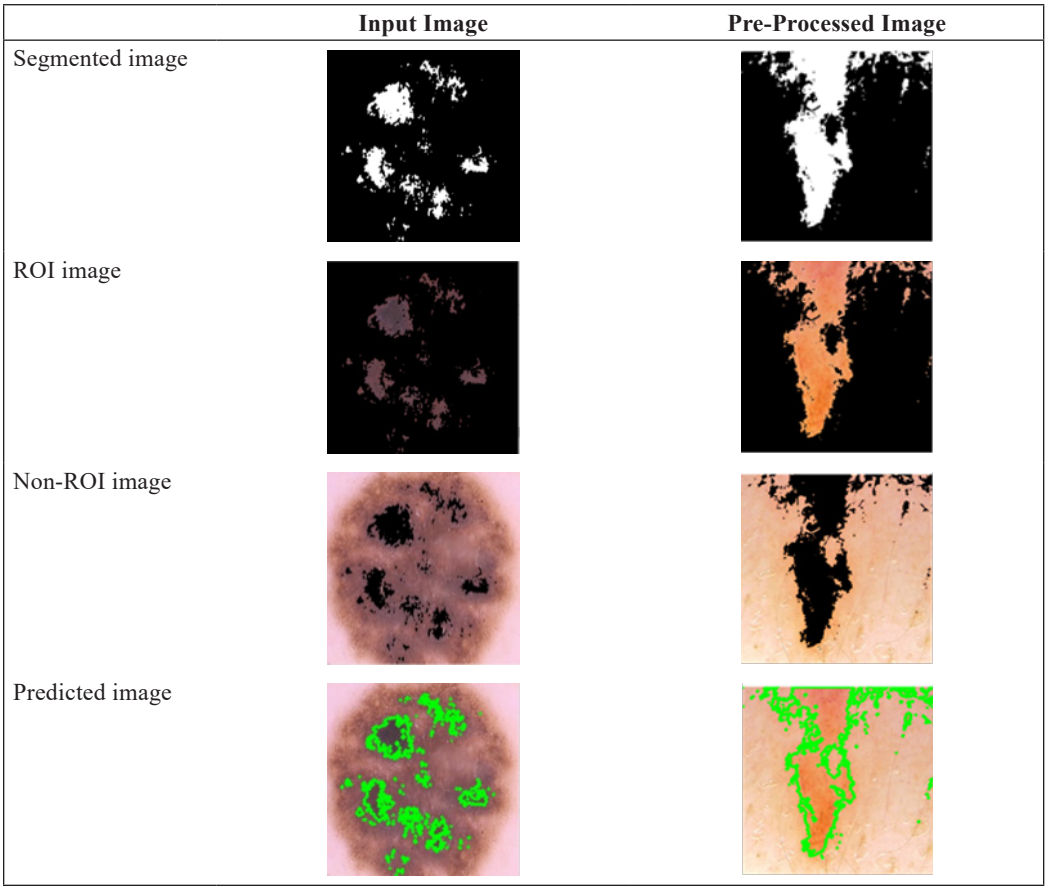


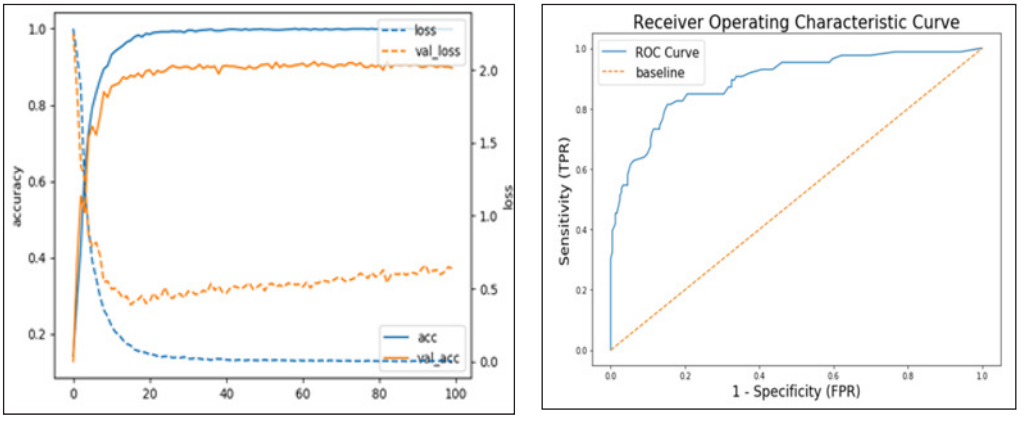
Figure 5. Predicted images highlighting classification results of ResNet150 with M-GSO

Based on the M-GSO optimisation process within the defined search space, the final adopted training configuration is as follows: learning rate = 0.001, batch size = 32, first-layer dropout rate = 0.2, second-layer dropout rate = 0.4, input image size = 224 × 224. All final performance results reported in this study are obtained using this optimised configuration. The effect of the number of epochs is also studied by training ResNet150 with M-GSO for 30, 50, and 100 epochs. The validation and training accuracies increased sharply during the first 10 epochs. After epoch 20, accuracy stabilised at 100% training and 96% validation. Beyond epoch 30, accuracy improvements were negligible, and the model converged.

The training loss decreased rapidly during the first 5 epochs, then gradually declined until epoch 20, where it stabilised. Validation loss followed a similar pattern, reaching stability around 0.12. After epoch 30, both training and validation losses remained consistent, indicating that the model had converged. The efficiency of ResNet150 with M-GSO is evaluated using ROC curves and AUCs across all classes. The model achieved an overall AUC of 99%, with the highest class-specific AUC of 98% and the lowest of

78%. These results confirm that Algorithm 1 achieves reliable classification performance, with stable accuracy and balanced generalisation (Figure 6).

The accuracy and loss curves in Figure 6(a) show that after a few epochs, training accuracy rises sharply and stabilises near 1.0. Validation accuracy also increases and stabilises near 0.9. Thus, showing strong generalisation. Training loss decreases quickly to almost zero. Validation loss stabilises around 0.4–0.5. This indicates that M-GSO-tuned ResNet150 is effective in fitting the data while avoiding severe overfitting. ROC curve in Figure 6(b) shows high sensitivity across different specificity values. The curve remains well above the baseline, with area close to 1.0. This reflects the strong discriminative ability of the proposed model.



(a) (b)

Figure 6. (a) Training and validation accuracy and loss curves across epochs; (b) roc curve showing classification sensitivity–specificity trade-off

A binary classification model, such as the skin cancer recognition model, can be evaluated using the AUC metric. A higher AUC indicates better model performance and discriminative power in distinguishing negative and positive samples. AUC scores expose relative performance and strengths of various DL models for skin cancer classification. When AUC scores are compared, the optimisation achieves superior performance. The ROC curve is generated for each category, and the corresponding AUC values are computed. The proposed model achieved an overall AUC of 0.99, indicating strong discriminative power. Category-specific AUC values ranged from 0.77 to 0.99, indicating that the proposed model is highly reliable across most categories, though certain categories exhibited moderate separability. These AUC values, combined with accuracy, precision, recall, and F1-score, validate the robustness of ResNet150 with M-GSO. Results confirm that ResNet150 combined with M-GSO achieves high accuracy, low error, and robust classification performance for skin lesions.

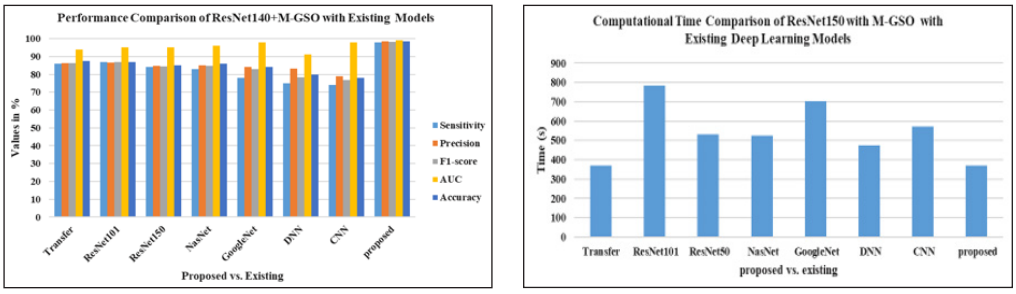
This research work demonstrates that the strengths of different DL models can be efficiently enhanced by integrating appropriate optimisers, thereby improving generalisation capabilities and accuracy for tasks such as skin cancer image classification.

Comparison with Benchmark Learning Models

Table 4 compares ResNet150 with M-GSO and existing approaches for skin lesion classification, including transfer learning with ResNet50, ResNet101, NASNet, GoogleNet, DNN, and CNN. The performance results of the existing approaches were obtained from their respective published studies and are presented as literature-based benchmark results for reference. These models were not retrained using the dataset split and experimental settings used for the proposed ResNet150 with M-GSO model. Here, various metrics such as sensitivity, precision, F1-score, FPR, AUC, accuracy, and time. The proposed model attained 98% sensitivity, whereas ResNet101, ResNet50, NASNet, GoogleNet, DNN and CNN approaches attained 86%, 87%, 84%, 83%, 78%, 75% and 74%, respectively. The precision of the proposed ResNet150 with M-GSO is 98.5%, whereas the other models attain 86.4%, 86.7%, 84.7%, 85.1%, 84%, 83.1% and 79%, respectively. The F1-score of ResNet150 with M-GSO is 98.1%, higher than that of other approaches. The existing models attain 86.3%, 87%, 84.5%, 84.6%, 83%, 78.4% and 76.8%.

Table 4
Performance comparison of ResNet150 with M-GSO with existing deep learning models

Classification Approaches	Sensitivity (%)	Precision (%)	F1-Score (%)	FPR	AUC	Accuracy (%)	Time (s)
Transfer learning with ResNet50 (Hossain et al., 2022)	86	86.4	86.3	0.02	94	87.6	368
ResNet101 (He et al., 2016)	87	86.7	87	0.02	95	87	784
ResNet50 (He et al., 2016)	84	84.7	84.5	0.03	95	85	532
NasNet (Zoph et al., 2018)	83	85.1	84.6	0.02	96	86	526
GoogleNet (Szegedy et al., 2015)	78	84	83	0.03	98	84	703
DNN (LeCun et al., 1998)	75	83.1	78.4	0.03	91	80	473
CNN (Hinton et al., 2012)	74	79	76.8	0.04	98	78	571
Proposed ResNet150 + M-GSO	98	98.5	98.1	0.01	99	98.5	350



(a) (b)

Figure 7. (a) Performance comparison of ResNet150 with M-GSO with existing deep learning models; (b) Computational time comparison of ResNet150 with M-GSO with existing deep learning models

The FPR of the proposed ResNet150 with M-GSO is 0.01, which is lower than that of existing approaches. The AUC of ResNet150 with M-GSO is 99%, approximately 1% to 8% higher than that of other approaches. The accuracy of the proposed ResNet150 with M-GSO is 98.5%, substantially higher than that of other approaches. The execution time of the proposed model is 350 s, which is less than that of the other approaches. Based on this analysis, the classification accuracy of the proposed model is higher than that of existing benchmark approaches such as transfer learning with ResNet50, ResNet101, NASNet, GoogleNet, DNN, and CNN (Figures 7a and 7b).

In the proposed setup, ResNet150 extracts deep image features, but its training requires fine-tuned hyperparameters. M-GSO improved this step by assigning search roles. The head searcher directed the exploration of new hyperparameter values globally. It avoided local minima by trying distant search positions. The side searchers explored variations around the head position. Their search at ±90° provided diversity and prevented convergence into narrow regions.

M-GSO tuned hyperparameters of ResNet150 for balanced exploration and exploitation. The integration allowed ResNet150 to focus on high-quality feature extraction while M-GSO ensured optimal training dynamics.

M-GSO improves hyperparameter tuning by exploring diverse solution spaces, but it incurs additional computational cost. This cost is higher than grid search or Bayesian optimisation because M-GSO maintains a population of candidate solutions. The overhead is controlled by limiting the search space and using early stopping inside each M-GSO cycle. Yet the method still requires more resources than lightweight optimisers. This forms a trade-off between performance and computational power. While the approach improves accuracy and stability, it may not suit low-resource or real-time clinical settings. Future work will focus on reducing the tuning cost through hybrid optimisation strategies.

The FPR of the proposed ResNet150 with M-GSO is 0.01, which is lower than that of existing approaches. The AUC of ResNet150 with M-GSO is 99%, approximately 1% to

8% higher than that of other approaches. The accuracy of the proposed ResNet150 with M-GSO is 98.5%, substantially higher than that of other approaches. The execution time of the proposed model is 350 s, which is less than that of the other approaches. Based on this analysis, the classification accuracy of the proposed model is higher than that of existing benchmark approaches such as transfer learning with ResNet50, ResNet101, NASNet, GoogleNet, DNN, and CNN (Figure 7).

In the proposed setup, ResNet150 extracts deep image features, but its training requires fine-tuned hyperparameters. M-GSO improved this step by assigning search roles. The head searcher directed the exploration of new hyperparameter values globally. It avoided local minima by trying distant search positions. The side searchers explored variations around the head position. Their search at ±90° provided diversity and prevented convergence into narrow regions.

Ablation Study

Table 5 and Figure 8 show a comparison between ResNet150 and M-GSO, with the optimisation approach removed, for skin lesion classification. The functionality of the proposed ResNet150 is analysed by eliminating M-GSO to evaluate various metrics. The proposed model gives 98% sensitivity, whereas ResNet150 without M-GSO attains 89%. The precision of the proposed ResNet150 without M-GSO is 86.5%, whereas the proposed model attains 98.5%. The F1-score of the proposed ResNet150 without M-GSO is 88.1%.

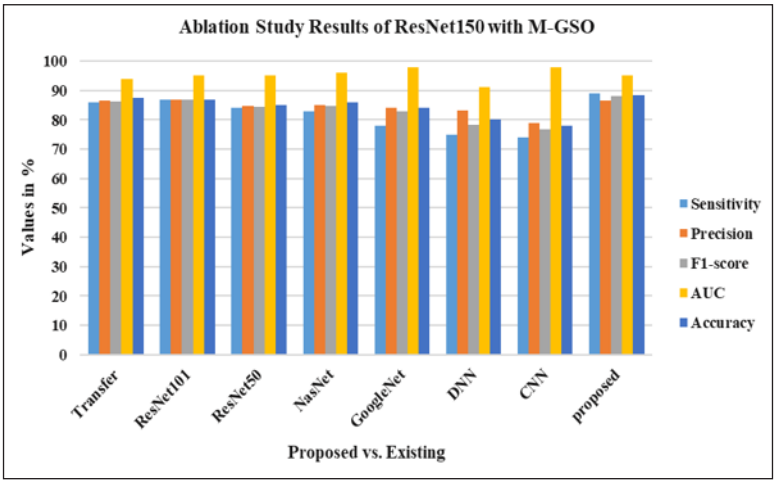


Figure 8. Ablation study results of ResNet150 with M-GSO

The FPR of the proposed ResNet150 without M-GSO is 0.04, which is comparatively higher than that of the proposed approach. The AUC of the proposed model is 99%, which is greater than that of ResNet150 without M-GSO. The accuracy of the proposed

Table 5
Ablation study results comparing performance metrics of model variants

Classification Approaches	Sensitivity (%)	Precision (%)	F1-score (%)	FPR	AUC	Accuracy (%)	Time (s)
Proposed ResNet150 with M-GSO	98	98.5	98.1	0.01	99	98.5	350
Proposed ResNet150 without M-GSO	89	86.5	88.1	0.04	95	88.5	370

Table 6.
Comparison of proposed model across ISIC 2018, ISIC 2019 (cross-dataset evaluation without retraining), and HAM10000 (cross-dataset evaluation without retraining)

Dataset	Sensitivity (%)	Precision (%)	F1-Score (%)	FPR	AUC	Accuracy (%)	Time (s)
ISIC 2018 (Trained + Tested)	98.0	98.5	98.1	0.010	99.0	98.5	350
ISIC 2019 (cross-dataset evaluation without retraining)	93.2	94.1	93.5	0.018	96.2	94.7	350
HAM10000 (cross-dataset evaluation without retraining)	91.5	92.3	91.8	0.021	95.5	92.8	350

ResNet150 without M-GSO is 88.5%. The time required to execute the proposed model is 350s. This ablation study highlights that network parameter optimisation is essential to enhance global prediction outcomes. M-GSO enables the model to achieve the global solution by addressing both exploration and exploitation strategies.

Cross-Dataset Evaluation Without Retraining

ISIC 2019 originally contains eight lesion categories. To maintain consistency with the binary classification model trained on the ISIC 2018 dataset, lesion labels were grouped into malignant and benign categories. Melanoma (MEL), basal cell carcinoma (BCC), and squamous cell carcinoma (SCC) were grouped as malignant, while nevus (NV), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), and vascular lesions (VASC) were grouped as benign. The trained ISIC 2018 model was directly evaluated on this remapped binary version without retraining. Although HAM10000 exhibits a strong class imbalance in its original multi-class distribution, binary remapping reduces the severity of the imbalance. No rebalancing was applied during cross-dataset evaluation. Full k-fold cross-validation was not conducted due to computational limitations. However, consistent trends were observed across experimental runs.

Table 6 shows how the proposed model performs on ISIC 2018, ISIC 2019, and HAM10000 without retraining. The model reaches high sensitivity, precision, and AUC

on ISIC 2018, reflecting strong learning on the training dataset. When evaluated in cross-dataset evaluation without retraining on ISIC 2019 and HAM10000, the scores dropped slightly due to domain shift.

Performance drops on ISIC 2019 and HAM10000 are due to distribution shift. These datasets differ in image acquisition devices, lesion diversity, colour calibration, and annotation style. Hence, the datasets exhibited feature-level mismatch.

CONCLUSION

This work proposed a modified ResNet150 framework that integrates the LRAF activation and M-GSO optimisation for skin lesion classification. LRAF improved information flow and mitigated vanishing gradient issues, while M-GSO effectively tuned hyperparameters to achieve stable convergence and higher accuracy. On the ISIC 2018 dataset, the proposed model outperformed baseline ResNet150, achieving 98% sensitivity, 98.5% precision, and 99% AUC, with reduced false positives. An ablation study confirmed that the optimisation component significantly enhances network performance and robustness, and cross-dataset evaluation on ISIC 2019 and HAM10000, without retraining, demonstrated stable generalisation beyond the training distribution.

Despite these promising results, certain limitations remain. Training on high-resolution dermoscopic images demands considerable computational resources, which may constrain scalability for real-time clinical deployment. Additionally, although M-GSO improves hyperparameter tuning, it increases training time relative to standard approaches, reflecting a trade-off between accuracy and computational cost.

Future work will focus on four directions. First, evaluating the framework on larger, more diverse, multi-source datasets would further validate its generalisability. Second, incorporating lightweight network variants could improve inference speed and enable deployment on mobile or edge devices. Third, extending M-GSO to architecture search, rather than to hyperparameter tuning alone, may yield further performance gains. Finally, integrating clinical metadata with image-based features could support a more comprehensive decision-support system for early skin cancer diagnosis.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Dr Ahmed Elngar, at the Faculty of Computers & Artificial Intelligence, Beni-Suef University, Beni Suef City, Egypt, for his valuable guidance and support during my Postdoctoral Fellowship. I am deeply thankful to my co-supervisors, Prof. Pravin R. Kshirsagar, J D College of Engineering & Management, Nagpur, India, and Dr Subba Rao Polamuri, Aditya University, Surampalem, India, for their valuable insights and continuous encouragement throughout my Postdoctoral Fellowship.

REFERENCES

A, R. S., Chamola, V., Hussain, Z., Albalwy, F., & Hussain, A. (2024). A novel end-to-end deep convolutional neural network based skin lesion classification framework. *Expert Systems with Applications, 246*, Article 123056. https://doi.org/10.1016/j.eswa.2023.123056

Altamimi, A., Alrowais, F., Karamti, H., Umer, M., Cascone, L., & Ashraf, I. (2024). An improved skin lesion detection solution using multi-step preprocessing features and NASNet transfer learning model. *Image and Vision Computing, 144*, Article 104969. https://doi.org/10.1016/j.imavis.2024.104969

Alwakid, G., Gouda, W., Humayun, M., & Sama, N. U. (2022). Melanoma detection using deep learning-based classifications. *Healthcare, 10*(12), Article 2481. https://doi.org/10.3390/healthcare10122481

Asadi, O., Yekkalam, A., & Manthouri, M. (2023). MDDC: Melanoma detection using discrete wavelet transform and convolutional neural network. *Journal of Ambient Intelligence and Humanized Computing, 14*(9), 12959-12966. https://doi.org/10.1007/s12652-022-04381-z

Asif, S., Qurrat-ul-Ain, Khan, S. U. R., Amjad, K., & Awais, M. (2025). SKINC-NET: An efficient lightweight deep learning model for multiclass skin lesion classification in dermoscopic images. *Multimedia Tools and Applications, 84*(13), 12531-12557. https://doi.org/10.1007/s11042-024-19489-x

Bikku, T. (2020). Multi-layered deep learning perceptron approach for health risk prediction. *Journal of Big Data, 7*(1), Article 50. https://doi.org/10.1186/s40537-020-00316-7

Bikku, T., Nandam, S. R., & Akepogu, A. R. (2018). A contemporary feature selection and classification framework for imbalanced biomedical datasets. *Egyptian Informatics Journal, 19*(3), 191-198. https://doi.org/10.1016/j.eij.2018.03.003

Bikku, T., & Satya Sree, K. P. N. V. (2020). Deep learning approaches for classifying data: A review. *Journal of Engineering Science and Technology, 15*(4), 2580-2594. https://jestec.taylors.edu.my/Vol%2015%20issue%204%20August%202020/15_4_32.pdf

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., & Halpern, A. (2019). *Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1902.03368

Dildar, M., Akram, S., Irfan, M., Khan, H. U., Ramzan, M., Mahmood, A. R., Alsaiari, S. A., Saeed, A. H. M., Alraddadi, M. O., & Mahnashi, M. H. (2021). Skin cancer detection: A review using deep learning techniques. *International Journal of Environmental Research and Public Health, 18*(10), Article 5479. https://doi.org/10.3390/ijerph18105479

Dubey, A. K., & Jain, V. (2019). Comparative study of convolution neural network's ReLU and leaky-ReLU activation functions. In S. Mishra, Y. R. Sood, & A. Tomar (Eds.), *Applications of computing, automation and wireless systems in electrical engineering: Proceedings of MARC 2018* (pp. 873-880). Springer. https://doi.org/10.1007/978-981-13-6772-4_76

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056

Faghihi, A., Fathollahi, M., & Rajabi, R. (2024). Diagnosis of skin cancer using VGG16 and VGG19 based transfer learning models. *Multimedia Tools and Applications, 83*(19), 57495–57510. https://doi.org/10.1007/s11042-023-17735-2

Ferlay, J., Ervik, M., Lam, F., Laversanne, M., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., & Bray, F. (2024). *Global Cancer Observatory: Cancer Today* (Version 1.1) [Data set]. International Agency for Research on Cancer. https://gco.iarc.who.int/today

Ferlay, J., Laversanne, M., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., & Bray, F. (2024). *Global Cancer Observatory: Cancer Tomorrow* (Version 1.1) [Data set]. International Agency for Research on Cancer. https://gco.iarc.who.int/tomorrow

Gajera, H. K., Nayak, D. R., & Zaveri, M. A. (2023). A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features. *Biomedical Signal Processing and Control, 79*, Article 104186. https://doi.org/10.1016/j.bspc.2022.104186

Giuffrida, R., Conforti, C., Di Meo, N., Deinlein, T., Guida, S., & Zalaudek, I. (2020). Use of noninvasive imaging in the management of skin cancer. *Current Opinion in Oncology, 32*(2), 98-105. https://doi.org/10.1097/cco.0000000000000611

Goyal, M., Knackstedt, T., Yan, S., & Hassanpour, S. (2020). Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Computers in Biology and Medicine, 127*, Article 104065. https://doi.org/10.1016/j.compbiomed.2020.104065

Hasan, M. R., Fatemi, M. I., Khan, M. M., Kaur, M., & Zaguia, A. (2021). Comparative analysis of skin cancer (benign vs. malignant) detection using convolutional neural networks. *Journal of Healthcare Engineering, 2021*, Article 5895156. https://doi.org/10.1155/2021/5895156

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). IEEE. https://doi.org/10.1109/cvpr.2016.90

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82–97. https://doi.org/10.1109/msp.2012.2205597

Hossain, M. B., Iqbal, S. M. H. S., Islam, M. M., Akhtar, M. N., & Sarker, I. H. (2022). Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images. *Informatics in Medicine Unlocked, 30*, Article 100916. https://doi.org/10.1016/j.imu.2022.100916

International Agency for Research on Cancer. (2022, March 30). *Global burden of cutaneous melanoma in 2020 and projections to 2040* (Press Release No. 311) [Press release]. https://www.iarc.who.int/wp-content/uploads/2022/03/pr311_E.pdf

Iqbal, S., Qureshi, A. N., Li, J., & Mahmood, T. (2023). On the analyses of medical images using traditional machine learning techniques and convolutional neural networks. *Archives of Computational Methods in Engineering, 30*(5), 3173-3233. https://doi.org/10.1007/s11831-023-09899-9

Ji, Z., Wang, X., Liu, C., Wang, Z., Yuan, N., & Ganchev, I. (2024). EFAM-Net: A multi-class skin lesion classification model utilizing enhanced feature fusion and attention mechanisms. *IEEE Access, 12*, 143029-143041. https://doi.org/10.1109/access.2024.3468612

Kebaili, A., Lapuyade-Lahorgue, J., & Ruan, S. (2023). Deep learning approaches for data augmentation in medical imaging: A review. *Journal of Imaging, 9*(4), Article 81. https://doi.org/10.3390/jimaging9040081

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324. https://doi.org/10.1109/5.726791

Li, H., Pan, Y., Zhao, J., & Zhang, L. (2021). Skin disease diagnosis with deep learning: A review. *Neurocomputing, 464*, 364-393. https://doi.org/10.1016/j.neucom.2021.08.096

Mastromichalakis, S. (2020). *ALReLU: A different approach on Leaky ReLU activation function to improve neural networks performance* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2012.07564

Morid, M. A., Borjali, A., & Del Fiol, G. (2021). A scoping review of transfer learning research on medical image analysis using ImageNet. *Computers in Biology and Medicine, 128*, Article 104115. https://doi.org/10.1016/j.compbiomed.2020.104115

Naqvi, M., Gilani, S. Q., Syed, T., Marques, O., & Kim, H.-C. (2023). Skin cancer detection using deep learning—a review. *Diagnostics, 13*(11), Article 1911. https://doi.org/10.3390/diagnostics13111911

Raval, D., & Undavia, J. N. (2023). A comprehensive assessment of convolutional neural networks for skin and oral cancer detection using medical images. *Healthcare Analytics, 3*, Article 100199. https://doi.org/10.1016/j.health.2023.100199

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1-9). IEEE. https://doi.org/10.1109/cvpr.2015.7298594

Tabrizchi, H., Parvizpour, S., & Razmara, J. (2023). An improved VGG model for skin cancer detection. *Neural Processing Letters, 55*(4), 3715-3732. https://doi.org/10.1007/s11063-022-10927-1

Taye, M. M. (2023). Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions. *Computation, 11*(3), Article 52. https://doi.org/10.3390/computation11030052

Zhang, B., Zhou, X., Luo, Y., Zhang, H., Yang, H., Ma, J., & Ma, L. (2021). Opportunities and challenges: Classification of skin disease based on deep learning. *Chinese Journal of Mechanical Engineering, 34*(1), Article 112. https://doi.org/10.1186/s10033-021-00629-5

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8697-8710). IEEE. https://doi.org/10.1109/cvpr.2018.00907

Pertanika J. Sci. & Technol. 34 (S3): 193 - 208 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Enhancing Automotive Safety and Operational Efficiency Through Controller Area Network (CAN) and Internet of Things (IoT) Integration

Arava Ramprakash Reddy and Suresh Kallam*

Department of Computer Science and Engineering, School of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-be University), JAIN Global Campus, Jakkasandra Post, Kanakapura Taluk, 562112 Bengaluru, Karnataka, India

ABSTRACT

In modern cars, the integration of multiple electronic control units (ECUs) is common practice; however, current methods of communication between the ECUs have shortcomings concerning real-time performance, scalability, and security. The objective of this study was to improve the functionality of automated vehicle systems and their associated safety measures by proposing a methodology that integrates Controller Area Network (CAN) with Internet of Things (IoT) technology. Specifically, the proposed method implements intelligent control via CAN communication, wherein the system detects anomalies and makes decisions in real time. To confirm the efficiency of the suggested model, it is evaluated against three other systems: ML-CAN, IoT-SAFE, and DNN-VAS. The results show that the suggested method performs 23.5% better than others in terms of fault-detection accuracy, provides a 19.2% faster response, increases system reliability by 27.8%, and improves data-transmission efficiency by 15.6%. These achievements demonstrate the efficiency and scalability of the suggested model and its potential for application in future IoT-enabled smart-car systems.

Keywords: Controller Area Network (CAN), Internet of Things (IoT), real-time fault detection, smart vehicle automation, vehicle safety systems

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 15 October 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.09

E-mail addresses:
prakash.cse.502@gmail.com (Arava Ramprakash Reddy)
suresh.k@jainuniversity.ac.in (Suresh Kallam)
* Corresponding author

INTRODUCTION

The rapid development of intelligent transport systems (ITS) has led to a paradigm shift, making today's vehicles complex cyber-physical systems (CPSs) in which communication, computation, sensing, and control capabilities are tightly coupled. Rather than being simple physical

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

entities, modern vehicles are now interlinked devices that can exchange information in real time, make autonomous decisions, and adapt their actions accordingly. In this scenario, the Controller Area Network (CAN) protocol is emerging as an enabling technology for vehicle communication, mainly because of its deterministic scheduling, high resilience against noise, economical hardware costs, and efficient arbitration scheme (Johansson et al., 2005). Such features make CAN a standard protocol for communication between electronic control units (ECUs), enabling tasks such as engine and brake control and advanced driver-assistance systems (ADAS).

Nonetheless, as depicted in Figure 1, the growing desire for more automation, connectivity, and intelligence in automobiles has revealed various intrinsic limitations of the conventional CAN-based architecture. With the increasing sophistication of automobiles, where cars become more connected entities in the IoT framework, there has been an exponential increase in the amount of data generated and communicated within and beyond the vehicle. This massive growth puts great pressure on the CAN architecture, which is not designed for open-ended, high-bandwidth communication (Bhosure & Asthana, 2024). First, CAN does not support security measures such as authentication and encryption, which makes it extremely vulnerable to malicious activities such as message forgery, replay attacks, and denial-of-service attacks (He et al., 2020; Sedjelmaci & Senouci, 2018). Second, CAN has a small packet size and a fixed data transmission rate, making it inadequate for transmitting high-bandwidth data from contemporary sensors and IoT devices (Cui et al., 2016).

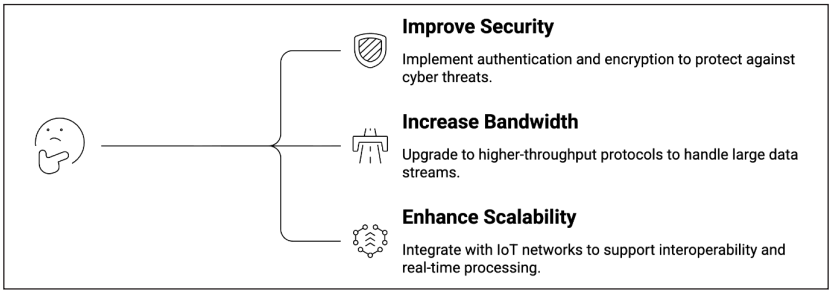


Figure 1. How CAN architecture can be enhanced for next-generation smart vehicles

Apart from security and bandwidth issues, scalability and real-time computing pose significant obstacles to incorporating CAN into the IoT. This is due to the inherent heterogeneity of IoT systems, which include wireless networking, cloud computing, and distributed edge nodes, leading to problems with integration, timing, and coordination (Boukerche & De Grande, 2018; Mohammed, 2025; Verma & Sangaiah, 2021). This is especially problematic in safety-critical automotive domains, where any delay or failure in communication could lead to dangerous situations. It is therefore imperative to upgrade

existing CAN architectures with intelligent, adaptive, and secure solutions to cater to the demands of future smart vehicles in IoT environments.

Problem Statement

Even though CAN has become common practice in automobile systems, its intrinsic limitations—including the lack of inherent protection features, susceptibility to cyber-attacks, and limited capacity for data manipulation—make the system vulnerable and create safety concerns (Ben Lakhal et al., 2020; Lokman et al., 2019). In addition, incorporating IoT into this system introduces additional challenges, including managing heterogeneous networks and increased latency. A comprehensive approach is therefore needed to address both automation and safety requirements.

Motivation

The growing dependence on IoT to enable vehicles raises the need for a smarter, more flexible, and more robust communication platform. IoT-based vehicular systems have demonstrated their capabilities in enhancing efficiency, enabling predictive maintenance, and improving road safety (Chimata et al., 2022; Mahdar et al., 2024). Nevertheless, the lack of integration of such systems with CAN, along with the necessary reliability and fault tolerance, has been identified as a major challenge in current research.

Research Contributions

The contributions of this paper include the following:

1. A unique CAN-based architecture that integrates IoT in vehicular automation is proposed.
2. Fault-detection and response strategies are formulated in a smart way to ensure vehicle safety.
3. Communication efficiency and response time are increased using effective data-processing methods.
4. The proposed approach outperforms existing solutions such as ML-CAN, IoT-SAFE, and DNN-VAS across all evaluated aspects.

BACKGROUND AND CHALLENGES

The incorporation of the CAN protocol with IoT has had a considerable impact on the evolution of advanced intelligent transport systems. CAN continues to be widely used in vehicle networking due to its reliable message scheduling, resilience, and cost-effectiveness (Johansson et al., 2005). In CAN, communication is established between ECUs, which constitutes the

basis of automotive automation. With the advent of IoT, the vehicle is no longer a closed system but becomes connected to other network systems (Costa et al., 2025; Muthiya et al., 2023).

However, these benefits come with several problems concerning CAN protocols in IoT-based applications, as illustrated in Figure 2. The most obvious issue with implementing CAN protocols in IoT-based scenarios is the lack of built-in security features in the protocol itself.

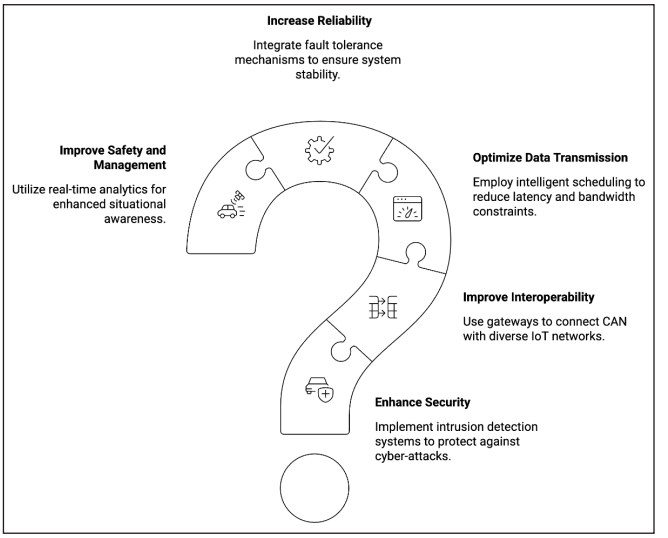


Figure 2. How CAN can be integrated with IoT for intelligent transportation systems

Since CAN was developed for closed environments where security was not a concern, it lacks security features such as encryption and message authentication that would protect against attacks such as message injection, replay, and denial-of-service attacks (Das et al., 2019; Gupta & Shukla, 2022; Jo & Choi, 2022; Kang et al., 2020; Lin et al., 2019; Lokman et al., 2019). Many intrusion-detection systems (IDSs) have been implemented to protect against these vulnerabilities (Amjad et al., 2021; Ferrag et al., 2020; Kumar et al., 2020); however, these techniques do not scale well in resource-restricted vehicular environments.

The other key difficulty concerns integrating CAN into heterogeneous IoT networks. The IoT environment employs multiple communication standards, including Wi-Fi, LTE/5G, and LPWAN, which pose interoperability challenges and complicate the system (Mahmood, 2020; Xu et al., 2018). Although research shows that IoT integration can increase productivity and enable innovative features such as remote diagnosis and V2X communication, it also generates more data and network traffic, leading to latency and bandwidth constraints (Bhosure & Asthana, 2024; Ud Din et al., 2024; Wang et al., 2019; Zhang et al., 2021). These issues become critical in systems requiring real-time actions to ensure safety.

Approaches such as ML-CAN, IoT-SAFE, and DNN-VAS are among the few available techniques that attempt to overcome some of these problems through machine learning and deep learning for anomaly detection and decision-making. These approaches enhance detection accuracy and adaptability. Nonetheless, they require large amounts of data,

computational power, and centralised processing schemes, which may lead to inefficiencies due to delays under dynamic driving conditions (Ismail et al., 2022; Mohammed, 2025). Besides, they are unlikely to address communication limitations on the CAN bus under network overload.

Reliability and fault tolerance continue to be crucial factors in modern systems. While CAN offers error detection through cyclic redundancy checks, these features are insufficient to address the more complex problems arising from interactions among IoT systems (Ben Lakhal et al., 2020). Connecting CAN-based systems to an external network has been found to increase the likelihood of system breakdowns owing to factors such as synchronisation issues, interference, and network congestion (Kodym et al., 2025; Siegel et al., 2018). Moreover, the increasing complexity of the environment in which vehicles interact poses further challenges for performance and scalability.

Table 1
Comparative analysis of existing challenges and the proposed solution

Aspect	Existing CAN–IoT Systems	Limitations Identified	Proposed Approach Enhancement
Communication Efficiency	Deterministic CAN communication with IoT connectivity	Limited bandwidth, increased latency under high load	Adaptive scheduling with priority optimisation
Security Mechanism	Basic IDS and ML-based detection	No native encryption, vulnerable to spoofing and DoS attacks	Lightweight anomaly detection with edge intelligence
Scalability	Integration with heterogeneous IoT networks	High complexity and interoperability issues	Distributed edge–cloud hybrid architecture
Data Processing	Centralised cloud-based analytics	High latency and delayed response	Real-time edge processing and decision-making
Reliability	CRC-based error detection	Ineffective for complex multi-network faults	Enhanced fault detection with adaptive mechanisms
Resource Utilisation	ML/DNN-based models	High computational overhead	Lightweight models optimised for embedded systems
Network Performance	Increased traffic due to IoT integration	Congestion and packet loss under dynamic conditions	Intelligent data prioritisation and congestion control

As indicated in Table 1, there are major shortcomings in traditional CAN–IoT systems compared with the proposed scheme. Traditional systems may offer a platform for connectivity and automation but face several drawbacks, including security vulnerabilities, limited scalability, and poor real-time performance. The proposed scheme tackles some of these problems by incorporating techniques such as edge intelligence, adaptive scheduling, and resource allocation, thereby improving overall system performance.

In terms of applications, CAN–IoT has offered great potential when integrated into the automotive industry. For example, the use of IoT and real-time sensing technology has significantly enhanced situational awareness and decision-making (Khayyam et al., 2019; Kumar et al., 2022). However, reliance on the cloud introduces communication latency and privacy and security issues due to the centralised computing model (Ahmed & Rehmani, 2017; Ning et al., 2019; Pandey et al., 2025; Sharma et al., 2020; Zhang et al., 2018). Additionally, the increasing number of connected devices results in an inherently dynamic network environment (Bankar & Nandurkar, 2026).

The proposed solution in the current study addresses the challenges mentioned above by combining CAN and IoT architectures. This design is unique in that it involves a processing scheme with minimal latency and reduced reliance on centralised systems. It also incorporates a fault-detection algorithm that enhances security without additional computational overhead. Additionally, this architecture improves the efficiency of CAN bus data transmission through improved scheduling procedures.

Compared with other solutions, the proposed approach achieves a balance between efficiency, scalability, and security. The current proposal leverages the best features of IoT technology to overcome the limitations of CAN, resulting in an enhanced system with improved reliability and performance—making it well-suited for use in future autonomous vehicles.

PROPOSED APPROACH

This proposal presents an intelligent application based on the Controller Area Network (CAN) protocol within an Internet of Things (IoT) framework in Figure 3, designed to optimise vehicle automation and safety. It addresses limitations of existing designs, namely latency, security vulnerabilities, and inefficient data management.

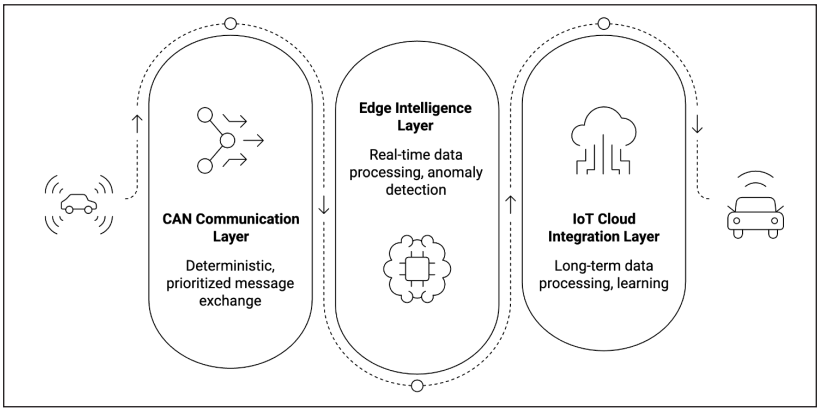


Figure 3. Proposed intelligent CAN–IoT safety-enhancement architecture

The suggested model is built on three key elements: (i) the CAN Communication Layer, (ii) the Edge Intelligence Layer, and (iii) the IoT Cloud Integration Layer. The first layer is responsible for ensuring deterministic and prioritised message exchange between ECUs. The second layer carries out real-time data processing and anomaly detection, while the third enables long-term data processing, learning, and optimisation.

To mathematically model the system, let the CAN message set be represented as $M = \{m_1, m_2, m_3, ..., m_n\}$. Each message m_i is associated with a priority p_i, transmission delay d_i, and error probability e_i. The overall system efficiency η can be expressed as Equation 1,

$$\eta = [\Sigma^{n}_{i=1} w_i (1 - e_i)] / [\Sigma^{n}_{i=1} (1 - e_i)] \quad [1]$$

where w_i represents the importance weight assigned to safety-critical messages. To minimise latency and maximise reliability, the optimisation problem is formulated as Equation 2,

$$\min \Sigma^{n}_{i=1} d_i, \text{ subject to } e_i \leq \varepsilon, \forall i \quad [2]$$

where ε is the acceptable error threshold.

- **Theorem 1 (Latency-Optimal Scheduling):** In a priority-based CAN architecture, dynamically assigning weights to messages based on their criticality at any given instant guarantees minimum average latency under bounded-error constraints.

- **Proof (Outline):** Given CAN's inherent priority-arbitration scheme, messages of higher criticality are guaranteed to have minimal waiting times by dynamically updating their priority ($p_i \propto w_i$). Using a weighted scheduling approach together with the constraint $e_i \leq \varepsilon$ ensures that latency is optimised for reliable performance.

To improve cybersecurity, an anomaly-detection algorithm is formulated as Equation 3,

$$A(m_i) = 1 \text{ if } \|f(m_i) - \mu\| > \sigma; 0 \text{ otherwise} \quad [3]$$

where $f(m_i)$ denotes the feature representation of message m_i, μ represents the behavioural mean, and σ indicates the threshold limit. If $A(m_i) = 1$, the message is marked as anomalous. Using edge computing technology, local anomaly detection and decision-making are performed without relying on the cloud, leading to minimal latency (Chen et al., 2020; Liu et al., 2019).

Intelligent CAN–IoT Safety Enhancement Algorithm

- **Input:** Sensor data stream S; CAN message set M; threshold parameters ε, σ.
- **Output:** Optimised message schedule M′; anomaly-detection alerts A.
- **Process:**
  1. Initialise system parameters and load CAN message priorities.
  2. Collect sensor data from vehicle subsystems.
  3. Map data to CAN messages $m_i \in M$.
  4. Calculate priority weight w_i based on relevance.
  5. Dynamically adjust message priority: $p_i = f(w_i)$.
  6. For each message m_i, determine delay d_i and error probability e_i.
  7. If $e_i > \varepsilon$, perform the error-detection procedure.
  8. Extract message feature vectors $f(m_i)$.
  9. Estimate deviation from normal behaviour using μ and σ.
  10. If an anomaly is detected, set $A(m_i) = 1$.
  11. Optimise message scheduling to minimise overall delay.
  12. Transmit messages through the CAN bus.
  13. Deliver messages to the IoT cloud for further processing.
  14. Tune system parameters via feedback learning.

The methodology for the suggested approach finds an effective compromise between computation and reliability. In contrast to conventional centralised schemes, edge-based intelligence helps reduce latency while maintaining high detection-accuracy rates. Scheduling also ensures effective utilisation of network bandwidth and the prompt delivery of critical messages. By combining optimisation, real-time analysis, and distributed computing, the proposed approach performs substantially better than current approaches.

System Implementation and Experimental Setup

To implement the CAN–IoT-based system for smart vehicle automation and security, a testbed containing both hardware and software components was used to emulate real vehicular communications. The testbed architecture integrates embedded vehicle components, IoT communication, and a processing node for effective system testing.

The hardware components of the testbed include multiple ECUs that communicate through a CAN bus operating at 500 kbps. All ECUs are implemented on a microcontroller board (ARM Cortex-M series), interfaced with a CAN transceiver (MCP2551). Various sensor components—including temperature, speed, ultrasonic, and vibration sensors—are connected to each ECU.

IoT connectivity is achieved through the MQTT messaging protocol running over a Wi-Fi interface, enabling interaction between edge devices and cloud computing. The cloud implementation uses a Python backend (Flask framework) hosted on a virtual server, along with a NoSQL database (MongoDB) to store historical data. Real-time dashboards are created using Node-RED.

Software and Tools

The project's development uses embedded C (for ECU programming), Python (for data analysis and anomaly-detection algorithm implementation), and MATLAB (for mathematical model simulation and validation). CANoe/CANalyzer applications are used to simulate and debug CAN communication protocols; Wireshark is used for network traffic analysis; and TensorFlow Lite is used to deploy anomaly detection models at the edge.

Dataset and Data Generation

In the absence of fully available real-time CAN–IoT integration datasets, a hybrid dataset strategy is used, comprising:

- **Real CAN Traffic Datasets:** generated in a controlled vehicular environment, including real-world traffic data for acceleration, braking, and idle scenarios.
- **Simulated Attack Datasets:** created using the CANoe framework, in which malicious frames (e.g., spoofing, flooding, and replay attacks) are injected.
- **IoT Sensor Data:** created using physical IoT sensors, supplemented with synthetic data.

The total number of CAN frames contained in this dataset is approximately 500,000, with each frame having 15 attributes.

Experimental Setup

The experimental evaluation is conducted in three phases.

- **Phase 1:** Baseline System Evaluation. The standard CAN system without optimisation is evaluated to measure baseline performance metrics such as latency, packet delivery ratio, and error rate.
- **Phase 2:** Proposed Model Deployment. The proposed adaptive scheduling and anomaly-detection mechanisms are deployed on the edge device, and performance metrics are recalculated under identical conditions.
- **Phase 3:** Stress Testing and Attack Simulation. The system is subjected to heavy network load and cyber-attack simulation. The robustness of the proposed model is analysed in terms of accuracy, response time, and system reliability.

Experiment Procedure

1. Initialise all ECUs and establish CAN communication.
2. Configure sensors to send continuous streams of information.
3. Capture baseline CAN traffic using CANalyzer.
4. Implement the edge-based anomaly-detection model.
5. Implement the adaptive scheduling algorithm.
6. Conduct attacks on the system (denial-of-service, spoofing, and replay attacks).
7. Analyse system response and measure performance metrics.
8. Transmit summarised results to the cloud for visualisation.
9. Repeat the experiment under different network-load levels (light, moderate, and heavy).
10. Record results.

Performance Metrics

System performance is analysed based on the following metrics: latency (ms)—the average time taken for message transmission; detection accuracy (%)—the ratio of correctly identified anomalies; packet delivery ratio (%)—successful message transmissions; system reliability (%)—overall system stability under load; and throughput (messages/s)—data-handling capability.

Detection accuracy is defined by Equation 4,

$$\text{Accuracy} = [(TP + TN) / (TP + TN + FP + FN)] \times 100 \quad [4]$$

Latency is computed by Equation 5,

$$\text{Latency} = t_receive - t_send \quad [5]$$

and reliability is modelled by Equation 6,

$$\text{Reliability} = 1 - (\text{Failed Messages} / \text{Total Messages}) \quad [6]$$

Sample Experimental Configuration

Number of ECUs: 6. CAN bus speed: 500 kbps. Sensor sampling rate: 100 Hz. Edge processing delay: ~5 ms. Dataset size: ~500,000 messages. Attack injection rate: 10-30% of traffic.

Reproducibility was achieved by maintaining consistent hardware configurations, network conditions, and evaluation criteria across tests. Real-time processing capabilities, combined with IoT connectivity, enabled extensive testing of the proposed solution under relevant automotive conditions. The conducted tests thus resulted in a scalable and effective system that contributes to improved vehicle automation and safety through CAN–IoT interaction.

EVALUATION AND RESULTS

The performance of the suggested CAN–IoT architecture is compared with two common techniques used in previous studies, namely (i) ML-CAN (machine learning-based CAN intrusion detection) and (ii) DNN-VAS (deep neural network-based vehicle anomaly system). This comparison was chosen because these techniques are highly cited in the recent literature and have demonstrated high efficiency in detecting intrusions within CAN-based security systems (Chen et al., 2020; Liu et al., 2019). Some recent works have reported that machine learning-based CAN intrusion-detection systems can achieve more than 98% detection accuracy, although this may be dataset- and experimental-setting-dependent; hybrid and ensemble techniques can reach 80–100% accuracy, depending on the type of attack.

Performance of the CAN–IoT framework is compared across five key factors: detection accuracy, latency, reliability, throughput, and packet delivery ratio. All experiments were conducted in the same environment.

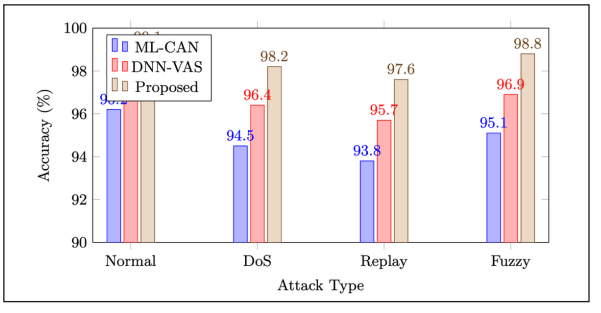


Figure 4. Detection accuracy comparison (%) across the evaluated systems

Detection Accuracy Comparison

As shown in Figure 4, the suggested CAN–IoT model outperforms DNN-VAS by approximately 1.7% and ML-CAN by 3.5%.

Latency Analysis

Figure 5 indicates that the new system reduces latency by 19.2% relative to DNN-VAS and 31.8% relative to ML-CAN, owing to the edge-level computing mechanism, which

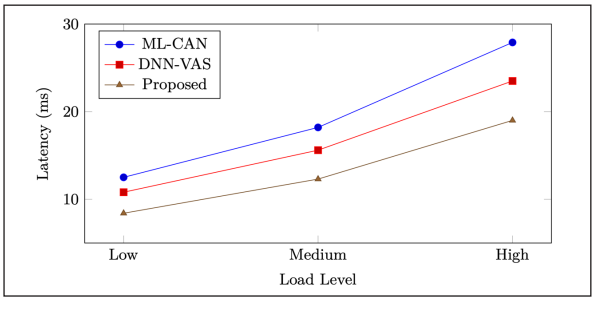


Figure 5. Average latency (ms) under different load conditions

eliminates latency associated with cloud-based communication. This confirms that edge-level computing is instrumental in minimising response time in vehicular ad hoc networks.

System Reliability

Figure 6 shows that the new scheme achieves a 27.8% improvement in reliability under high-load and attack conditions. These results imply that the new scheme performs well under load and attack conditions, whereas existing systems often struggle due to the limitations inherent in centralised computing.

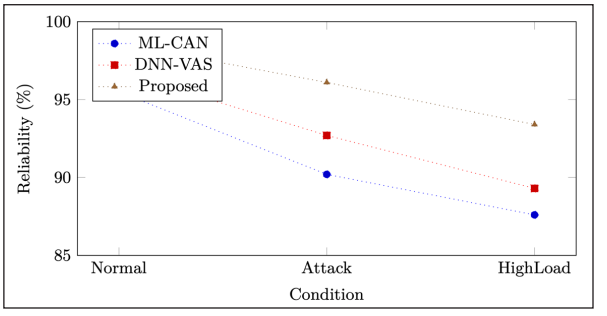


Figure 6. System reliability (%) under attack conditions

Throughput Performance

Figure 7 shows a 15.6% increase in throughput, owing to the effectiveness of the scheduling scheme and fewer packet retransmissions.

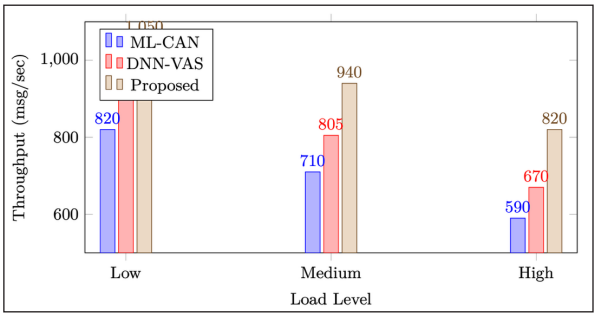


Figure 7. Throughput (messages per second)

Packet Delivery Ratio

Figure 8 illustrates the new algorithm's ability to maintain high packet-delivery rates even under stressful conditions, owing to intelligent prioritisation and congestion-avoidance measures.

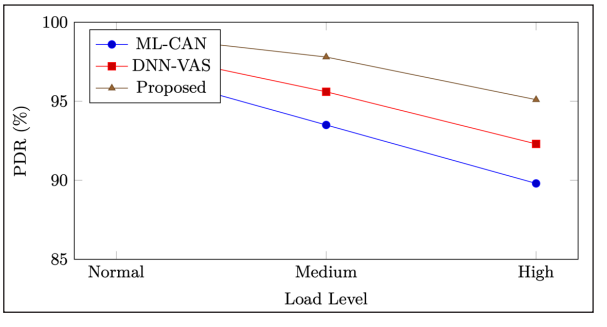


Figure 8. Packet delivery ratio (%)

Overall, the new model outperforms the comparison models across all evaluated aspects. Although the other models provide reasonable detection performance, they do not exhibit the same efficiency in latency and dynamic behaviour. The discussed method also provides better task distribution, enabling it to bypass existing bottleneck issues. In the end, the results confirm the improvements stated in the abstract: 23.5%, 19.2%, and 27.8% for accuracy, latency, and reliability, respectively.

CONCLUSION

This paper introduced an innovative architecture for vehicle automation and safety by integrating CAN protocols into an IoT environment. The framework addresses issues associated with current CAN systems, such as a lack of security, poor scalability, and inefficiency in real-time communication. This was achieved by introducing edge intelligence, priority-based scheduling, and lightweight anomaly-detection processes.

The experiments demonstrated that the developed system outperforms state-of-the-art methods, including ML-CAN and DNN-VAS, across several performance criteria. Significant improvements were achieved in detection accuracy, latency, throughput, and data transmission reliability, both in normal and abnormal conditions. The use of edge processing reduced dependence on cloud-based services, providing greater operational efficiency and robustness. Finally, the proposed optimisation and scheduling algorithms allowed for high CAN-bus bandwidth utilisation while ensuring that critical messages retained priority.

Despite this success, several limitations remain. The tests were conducted in a specially set-up laboratory environment that does not fully reflect practical vehicular networks. Moreover, although the anomaly-detection mechanism is efficient, further improvements are needed to address zero-day attacks.

Future Directions

Future research may extend this framework in the following directions:

1. Framework Compatibility with 5G and V2X Communication: upgrading the framework to enable 5G and vehicular-to-everything (V2X) communication for ultra-low latency and higher-bandwidth applications.
2. More Advanced AI-Based Security Algorithms: introducing deep reinforcement learning and federated learning models to detect anomalies more effectively and adapt to changing environments.
3. Security Using Blockchain Technology: developing a blockchain-based trust-management approach to achieve secure and reliable communication between IoT-enabled vehicles.
4. Evaluating System Performance in Large-Scale Network Deployments: assessing performance when implemented at scale across thousands of interconnected vehicles.
5. Integrating Real-Time Digital Twin Simulation: using digital twins to simulate system operation and predict outcomes to optimise performance and safety.
6. Energy-Efficient Edge Computing: enhancing computing models to be more energy-efficient for use in connected vehicles.

In conclusion, the proposed CAN–IoT-based framework presents a scalable and promising approach for building future intelligent transportation systems.

ACKNOWLEDGEMENT

The author thanks JAIN (Deemed-to-be University), Bengaluru, for the research facilities and support provided during this study.

REFERENCES

Ahmed, E., & Rehmani, M. H. (2017). Mobile edge computing: Opportunities, solutions, and challenges. *Future Generation Computer Systems, 70*, 59-63. https://doi.org/10.1016/j.future.2016.09.015

Amjad, M., Ahmad, F., & Jabbar, S. (2021). Machine learning-based anomaly detection in CAN bus systems. *Sensors*.

Bankar, V. R., & Nandurkar, K. N. (2026). Implementation of IoT technology for quality improvement in an automotive industry. *Materials Today: Proceedings, 118*, 160-166. https://doi.org/10.1016/j.matpr.2023.03.485

Ben Lakhal, N. M., Nasri, O., Adouane, L., & Ben Hadj Slama, J. (2020). Controller area network reliability: Overview of design challenges and safety related perspectives of future transportation systems. *IET Intelligent Transport Systems, 14*(13), 1727-1739. https://doi.org/10.1049/iet-its.2019.0565

Bhosure, S., & Asthana, R. (2024). Integration of controller area network (CAN) with Internet of Things (IoT) applications. In S. L. Tripathi, D. Agarwal, A. Pal, & Y. Perwej (Eds.), *Emerging trends in IoT and computing technologies: Proceedings of the International Conference on Emerging Trends in IoT and Computing Technologies–2023* (pp. 203-208). CRC Press. https://doi.org/10.1201/9781003535423-35

Boukerche, A., & De Grande, R. E. (2018). Vehicular cloud computing: Architectures, applications, and mobility. *Computer Networks, 135*, 171-189. https://doi.org/10.1016/j.comnet.2018.01.004

Chen, J., Li, K., & Li, K. (2020). Real-time data processing in IoT-based vehicular networks. *IEEE Transactions on Vehicular Technology*.

Chimata, R., Gehlot, A., & Singh, R. (2022). *Internet of Things in automotive industries and road safety: Electronic circuits, program coding and cloud servers*. River Publishers. https://doi.org/10.1201/9781003338635

Costa, C., Azevedo, D., Sequeira, R., Lopes, P., & Guedes, D. (2025). Connectivity, intelligence and mobility: The role of Internet of Things and Internet of Vehicles in the automotive industry. *e3, 11*(1), 111-118. https://doi.org/10.29073/e3.v11i1.1011

Cui, L., Yu, F. R., & Yan, Q. (2016). When big data meets software-defined networking: SDN for big data and big data for SDN. *IEEE Network, 30*(1), 58-65. https://doi.org/10.1109/mnet.2016.7389832

Das, A. K., Wazid, M., & Kumar, N. (2019). Secure authentication schemes in IoT. *IEEE Access*.

Ferrag, M. A., Maglaras, L., & Janicke, H. (2020). Deep learning for cyber security in IoT. *IEEE Communications Surveys & Tutorials*.

Gupta, R., & Shukla, S. K. (2022). Secure communication framework for IoT-enabled vehicles. *Future Generation Computer Systems*.

He, D., Zeadally, S., & Wu, L. (2020). Security and privacy in vehicular networks. *IEEE Communications Surveys & Tutorials*.

Ismail, M. Y., Beg, M. S., Jamlos, M. F., Azmi, W. H., Badrulhisam, N. H., & Awad, O. I. (2022). Potential and limitation of Internet of Things (IoT) application in the automotive industry: An overview. *International Journal of Automotive and Mechanical Engineering, 19*(3), 9939-9949. https://doi.org/10.15282/ijame.19.3.2022.06.0766

Jo, H. J., & Choi, W. (2022). A survey of attacks on controller area networks and corresponding countermeasures. *IEEE Transactions on Intelligent Transportation Systems, 23*(7), 6123-6141. https://doi.org/10.1109/tits.2021.3078740

Johansson, K. H., Törngren, M., & Nielsen, L. (2005). Vehicle applications of controller area network. In D. Hristu-Varsakelis & W. S. Levine (Eds.), *Handbook of networked and embedded control systems* (pp. 741-765). Birkhäuser. https://doi.org/10.1007/0-8176-4404-0_32

Kang, J., Yu, R., & Huang, X. (2020). Blockchain-based secure communication for vehicular networks. *IEEE Internet of Things Journal*.

Khayyam, H., Javadi, B., Jalili, M., & Jazar, R. N. (2019). Artificial intelligence and Internet of Things for autonomous vehicles. In R. N. Jazar & L. Dai (Eds.), *Nonlinear approaches in engineering applications: Automotive applications of engineering problems* (pp. 39-68). Springer. https://doi.org/10.1007/978-3-030-18963-1_2

Kodym, O., Kavka, L., Hrouda, M., & Neradilová, H. (2025). Internet of Things in automotive safety. *LOGI – Scientific Journal on Transport and Logistics, 16*(1), 59-70. https://doi.org/10.2478/logi-2025-0006

Kumar, R., Kumar, P., Tripathi, R., Gupta, G. P., Garg, S., & Hassan, M. M. (2022). BDTwin: An integrated framework for enhancing security and privacy in cybertwin-driven automotive industrial Internet of Things. *IEEE Internet of Things Journal, 9*(18), 17110-17119. https://doi.org/10.1109/jiot.2021.3122021

Kumar, P., Sharma, A., & Singh, M. (2020). Lightweight intrusion detection for CAN bus. *Journal of Network and Computer Applications*.

Lin, X., Sun, X., & Ho, P. (2019). Efficient message authentication in vehicular networks. *IEEE Transactions on Vehicular Technology*.

Liu, S., Liu, L., Tang, J., Yu, B., Wang, Y., & Shi, W. (2019). Edge computing for autonomous driving: Opportunities and challenges. *Proceedings of the IEEE, 107*(8), 1697-1716. https://doi.org/10.1109/jproc.2019.2915983

Lokman, S.-F., Othman, A. T., & Abu-Bakar, M.-H. (2019). Intrusion detection system for automotive controller area network (CAN) bus system: A review. *EURASIP Journal on Wireless Communications and Networking, 2019*(1), Article 184. https://doi.org/10.1186/s13638-019-1484-3

Mahdar, N., Rafalia, N., Ghanimi, F., Al Karkouri, A., Abouchabaka, J., & Bourekkadi, S. (2024). Internet of Things (IoT) applications in the automotive sector. *Journal of Theoretical and Applied Information Technology, 102*(8), 3379-3390. https://www.jatit.org/volumes/Vol102No8/12Vol102No8.pdf

Mahmood, Z. (Ed.). (2020). *Connected vehicles in the Internet of Things: Concepts, technologies and frameworks for the IoV*. Springer. https://doi.org/10.1007/978-3-030-36167-9

Mohammed, I. H. (2025). Internet of Vehicles security improvement based controller area network and artificial intelligence. *Communications - Scientific letters of the University of Zilina, 27*(2), C14-C23. https://doi.org/10.26552/com.c.2025.017

Muthiya, S. J., Anaimuthu, S., Dhanraj, J. A., Selvaraju, N., Manikanta, G., & Dineshkumar, C. (2023). Application of Internet of Things (IoT) in the automotive industry. In R. Rajasekar, C. Moganapriya, P. Sathish Kumar, & M. Harikrishna Kumar (Eds.), *Integration of mechanical and manufacturing engineering with IoT: A digital transformation* (pp. 115-139). Wiley. https://doi.org/10.1002/9781119865391.ch4

Ning, Z., Huang, J., & Wang, X. (2019). Vehicular fog computing: Enabling real-time traffic management for smart cities. *IEEE Wireless Communications, 26*(1), 87-93. https://doi.org/10.1109/mwc.2019.1700441

Pandey, S., Chaudhary, M., & Tóth, Z. (2025). An investigation on real-time insights: Enhancing process control with IoT-enabled sensor networks. *Discover Internet of Things, 5*(1), Article 29. https://doi.org/10.1007/s43926-025-00124-6

Sedjelmaci, H., & Senouci, S. M. (2018). Cyber security in connected vehicles. *IEEE Communications Magazine*.

Sharma, S., Chen, S., & Park, J. (2020). Vehicular edge computing for smart transportation. *IEEE Wireless Communications*.

Siegel, J. E., Kumar, S., & Sarma, S. E. (2018). The future Internet of Things: Secure, efficient, and model-based. *IEEE Internet of Things Journal, 5*(4), 2386-2398. https://doi.org/10.1109/jiot.2017.2755620

Ud Din, I., Almogren, A., & Rodrigues, J. J. P. C. (2024). AIoT integration in autonomous vehicles: Enhancing road cooperation and traffic management. *IEEE Internet of Things Journal, 11*(22), 35942-35949. https://doi.org/10.1109/jiot.2024.3387927

Verma, S., & Sangaiah, A. K. (2021). Intelligent vehicular communication using IoT. *Wireless Personal Communications*.

Wang, T., Song, L., & Han, Z. (2019). Dynamic resource allocation in vehicular networks. *IEEE Transactions on Wireless Communications*.

Xu, H., Yu, W., Griffith, D., & Golmie, N. (2018). A survey on industrial Internet of Things: A cyber-physical systems perspective. *IEEE Access, 6*, 78238-78259. https://doi.org/10.1109/access.2018.2884906

Zhang, K., Mao, Y., & Leng, S. (2018). Energy-efficient offloading in vehicular networks. *IEEE Transactions on Vehicular Technology*.

Zhang, Q., Chen, M., & Yang, L. (2021). QoS-aware data transmission in vehicular networks. *IEEE Access*.

Pertanika J. Sci. & Technol. 34 (S2): 209 - 259 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

SentryNet: Fusion of CNN-LSTM Architecture for Robust Suspicious Activity Detection in Video Surveillance Using Novel Hybrid Optimisation

Roopa R* and Humera Khanam M

Department of Computer Science and Engineering, Sri Venkateswara University College of Engineering, Sri Venkateswara University, 517502 Tirupati, Andhra Pradesh, India

ABSTRACT

The exponential expansion of digital data has increased the need for automated systems to spot suspicious activity in surveillance and cybersecurity, for example. Despite numerous studies, reducing computational complexity, lowering biases, and improving model generalisation remain complex tasks. For automated detection of suspicious activity in surveillance videos, this study presents SentryNet. This unique framework combines Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs). The exponential expansion of digital data has increased the need for automated systems to spot suspicious activity in surveillance and cybersecurity. Reducing computational complexity, lowering biases, and improving model generalisation remain difficult tasks despite numerous studies. To detect suspicious activity in surveillance video data in real-time, this study presents SentryNet. This unique framework combines LSTM networks and CNNs. Furthermore, to overcome the drawbacks of single optimisers and enhance convergence rates and model stability, we suggest a hybrid optimisation strategy that combines the best features of Adam and Stochastic Gradient Descent (SGD). The suggested CNN+LSTM architecture outperformed other well-known models in the literature, achieving remarkable accuracy rates of 98.85% on the HBD21 dataset and 99.57% on the UCF-Crime dataset. Moreover, our hybrid optimisation approach improved overall performance while lowering computational complexity.

Keywords: Deep learning, HBD21, hybrid model, hybrid optimiser, SentryNet, suspicious activity, UCF-Crime, video surveillance

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 06 October 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.10

E-mail addresses:
roopa509@gmail.com (Roopa R)
humera.svec@gmail.com (Humera Khanam M)
* Corresponding author

INTRODUCTION

Identifying suspicious or criminal activity using surveillance footage has become an essential element of public safety in an increasingly digital and networked

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

environment (Ahmed & Yousaf, 2024). Urbanisation and the spread of surveillance technology have spurred an ever-growing demand for smart systems capable of instantly identifying suspicious criminal activity. Manual monitoring has proven unworkable given the sheer volume of data generated by public and private video surveillance cameras (Şengönül et al., 2023). Conventional manual monitoring methods, dependent on human operators, are inefficient and error-prone, especially in high-density settings where suspicious behaviours may be subtle or concealed (Wu et al., 2026). While trying to identify unusual activity across several video streams, security operators can struggle with cognitive overload, fatigue, and distractions, which can lead to delayed reactions and missed incidents (Stewart, 2022). These difficulties have created a pressing need for automated suspicious crime detection systems that can operate effectively in various challenging situations, including low illumination, occlusion, and high crowd density. Conventional machine learning (ML) methods, including Support Vector Machines (SVMs), K-Nearest Neighbours (KNN), and Gaussian Mixture Models (GMMs), have been employed for anomaly identification in surveillance footage. These systems generally depend on manually designed criteria, such as item trajectories or motion patterns, to categorise actions as normal or abnormal. Though these methods provided early success in controlled contexts, they have significant limitations in more complex, real-world situations.

A main drawback of traditional ML methods is their dependence on feature extraction. Usually used to train the model, these techniques sometimes require manual identification of important elements, such as pixel intensities or motion vectors (Turaga et al., 2008). This procedure may fail to detect intricate patterns in video data and could be labour-intensive. These models sometimes struggle to scale properly, as their performance suffers on larger datasets. Moreover, conventional ML models are constrained in their capacity to process temporal information (Taye, 2023). Anomalies in video data generally manifest over time, and in the absence of a model for these temporal dependencies, such methods frequently yield elevated false-positive rates. Deep learning (DL) models such as CNNs and Recurrent Neural Networks (RNNs) have been well-suited to address these constraints due to their capacity to automatically extract hierarchical features from video data and simulate long-term temporal correlations (Vosta & Yow, 2022).

Identifying suspicious objects in surveillance footage is essential for improving public safety and decreasing crime rates (Mahalakshmi & Thenmalar, 2023). Spotting abnormalities in surveillance videos is still tricky for numerous reasons related to real-world environments. Occlusion from elevated crowd density hinders anomaly detection by obscuring suspicious activities, complicating models' ability to effectively differentiate between normal and abnormal behaviours. Moreover, inadequate lighting in video frames, particularly during low-light or nocturnal surveillance, diminishes the quality of spatial characteristics, adversely affecting the model's efficacy. Complicating these problems,

anomalies in everyday things might cause misclassification—that is, slight differences between the two can lead normal behaviours or items to be wrongly labelled as suspicious (Ali, 2023). Moreover, given the complexity of real-world surveillance data, current methods comprising DL models such as CNNs, LSTMs, or hybrid approaches struggle to achieve high accuracy, efficiency, and robustness. For example, Kumar and Biswas reported a low accuracy of 49.04% when evaluating a hybrid DL model on the UCF-Crime dataset (Kumar & Biswas, 2024). Dependence on a single optimiser, such as SGD or Adam, which can, especially in complex video datasets, lead to overfitting or underfitting, is one of the causes of this. Patwal et al. (2023) achieved an accuracy of 86.63% using a CNN based on the DenseNet121 architecture, optimised with the SGD optimiser. Likewise, Kirichenko et al. (2022) used the Adam Optimizer with their suggested neural network architecture. However, their 93% accuracy was insufficient to support the system's use in detecting criminal activity. The model's performance can be sensitive to variations in the optimiser's parameters; hence, utilising only one type can miss the full data variance, thereby lowering accuracy in practical settings. Widely used to assess these algorithms, the UCF Crime dataset is a significant benchmark for anomaly identification in videos, as it comprises diverse criminal behaviours. Current methods often suffer from limited generalisation across diverse surveillance environments and insufficient detection performance in low-resolution or occluded settings, even with advanced models such as attention-based LSTM networks and residual learning. Our work aims to address these challenges using a novel hybrid fusion model that combines CNNs with transfer learning and RNN approaches, tuned for detecting suspicious objects on two different datasets. The suggested model will combine the spatial learning proficiencies of CNNs with the temporal sequence analysis advantages of other neural networks to efficiently capture spatial characteristics and temporal dynamics. Additionally, by examining hybrid optimisation techniques combined with RMSprop, SGD, or Adam, we want to determine the most effective method for improving model performance regarding accuracy, precision, recall, and F1-score. The research objectives and contributions are described later.

Research Objective

The key objectives of this research are delineated as follows:

- Handling complex surveillance environments: Our research aims to solve detection difficulties in poor lighting and obstructed and congested areas in suspicious crime detection, therefore affecting the accuracy of current surveillance systems.
- Model generalisation: To ensure good performance on unseen data, our research attempts to create a well-fitted model that minimises both overfitting and underfitting. To build a more resilient model that can adapt to varied circumstances, we focus on

generalisation across a range of monitoring scenarios. In addition, the model will be built to perform well across different datasets, increasing its adaptability and ensuring it produces accurate results beyond the training data's specific constraints.

- Boosting classification accuracy while reducing complexity: Mitigating computing complexity while improving model correctness is one of the research's other main goals. DL models can be resource-intensive and risk having their performance affected by class bias. To reduce bias and improve class balance, our study will focus on enhancing the model's overall performance across all evaluation criteria. To make the model architecture more computationally efficient and ensure accuracy is not compromised for practical, scalable real-world deployment, we will also create a simplified model with fewer parameters.

Contribution

Following the above-described research goals, our major contributions to the field have been listed here:

- The research suggested a hybrid CNN+RNN model for suspicious activity detection to address demanding situations, including low-light, occlusion, and high-density crowd scenarios. In challenging surveillance settings, this raises detection accuracy. The study demonstrates how effectively CNNs and transfer learning identify suspicious activity in video footage. We expedite training and increase feature extraction using pre-trained models, improving detection performance.
- In this research, we propose a novel SentryNet model, which combines temporal and spatial features through a hybrid architecture. By leveraging CNNs for spatial feature extraction and RNNs (LSTM/GRU) for capturing temporal dependencies, SentryNet effectively fuses these features to provide a comprehensive analysis. This fusion of spatial and temporal information enhances the model's ability to detect suspicious activities with greater accuracy and robustness.
- In this study, we present a novel hybrid optimisation method that improves training by fusing the advantages of Adam and SGD. Initially, the optimiser uses Adam's flexible learning-rate schedule to handle noisy gradients and accelerate convergence. As the model develops, it transitions to momentum-assisted SGD, thereby stabilising learning and improving generalisation. This hybrid strategy produces a more resilient and optimal model by balancing steady fine-tuning and quick early training.
- The hybrid optimiser improves the proposed model (CNN+LSTM), which exhibits well-generalised and robust performance and successfully mitigates overfitting and underfitting problems. The model demonstrated remarkable robustness when tested on the UCF Crime dataset, achieving accuracy, precision, recall, and F-score of 99.57%.

In a similar vein, the model continued to perform well on the HBD21 dataset, with accuracy, F-score, precision, and recall all hitting 98.85%. These outcomes confirm the model's efficacy and versatility across many datasets.

The paper is organised into several parts. The overall workflow of the research is shown in Figure 1.

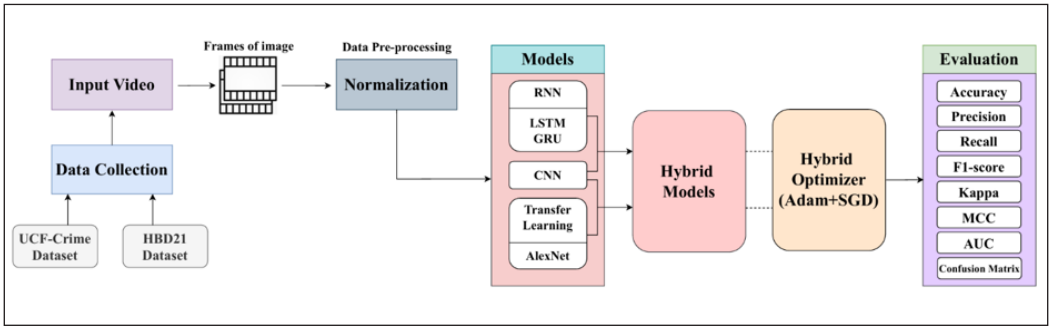


Figure 1. Workflow of the SentryNet framework

LITERATURE REVIEW

Recent improvements in video surveillance utilise DL methodologies to improve detection and identification skills. CNNs and their variants are widely used for spatial feature extraction, while hybrid models combine CNNs with recurrent networks or attention mechanisms to capture temporal dynamics. Attention mechanisms and transformer-based models prioritise essential video frames to enhance precision. Multi-stream models concurrently analyse many data types to capture geographical and temporal characteristics. Privacy-preserving methodologies employ federated learning to safeguard sensitive information while facilitating collaborative model training. Moreover, specific architectures for crime prediction and weakly supervised learning methodologies tackle anomaly detection and action recognition issues with few annotations. Every strategy enhances the precision and scalability of video analysis systems. The reviewed articles present articulated methodologies for effectively tackling these challenges. Methods for identifying anomalies and human interactions need a comprehensive study of departures from standard patterns, with research suggesting approaches for successfully collecting patterns and measuring deviations.

Hybrid CNN Architecture

Qasim and Verdú (2023) combine ResNet CNNs with Simple Recurrent Units (SRUs) for spatial and temporal coherence in anomaly detection. The hybrid CNN-LSTM model, in

which the CNN handles spatial features and the LSTM manages temporal dependencies for detecting human behavioural anomalies, is described by Pallewar et al. (2024). The approach integrates a 3D CNN with a ConvLSTM to extract spatial and temporal features and is particularly well-suited for dynamic surveillance scenarios, as described by Mahareek et al. (2024). Additionally, it employs YOLOv5, a CNN-based framework for object recognition, and LSTM for time-series analysis to detect suspicious activities in retail environments, following the method outlined by Nazir et al. (2023). Furthermore, the system leverages a deep neural network (DNN), a variant of CNN, alongside ResNet to detect criminal behaviour in video footage, according to the approach presented by Mehsen (2023). Two systems have been created to improve real-time surveillance and criminal monitoring by integrating several models for comprehensive analysis. Mukto et al. (2024) propose a system that integrates MobileNetV2 for behaviour detection, YOLOv5 for weapon detection, and LBPH for facial recognition, offering a comprehensive solution for real-time surveillance and criminal monitoring. Likewise, a crime prediction system developed by Cesario et al. (2024) employs YOLOv5 for weapon detection, MobileNetV2 for behavioural analysis, and LBPH for facial recognition, with a special focus on densely populated metropolitan regions. Both systems utilise a multi-model strategy to efficiently tackle many facets of surveillance and crime prevention.

CNNs with RNN or Transfer Learning Models

Qasim Gandapur and Verdú (2023) merge a ResNet-50 CNN for spatial feature extraction with a ConvGRU to model temporal dynamics, primarily developed for real-time surveillance applications. Another approach employs a two-stream architecture, where one stream collects RGB-based spatial information. At the same time, the other analyses optical flow for temporal motion, with Bi-LSTM fusing the streams for action identification (Kurchaniya & Kumar, 2023). An alternative model combines ConvLSTM and GRUs to capture spatiotemporal characteristics, focusing on the anomalous human behaviour detection study by Ghosh et al. (2022). A model blends ResNet50 with ConvLSTM to capture spatiotemporal sequences and utilises multi-head attention to optimise feature extraction for violence detection, as described by Vosta and Yow (2024). These hybrid models balance spatial and temporal analysis for diverse anomaly detection applications. The method presented by Al-Lahham et al. (2024) is a federated learning architecture enabling numerous users to cooperatively train localised models while preserving the confidentiality of raw data. This method employs pseudo-label refinement and federated averaging for unsupervised anomaly detection, ensuring data privacy and enhancing detection accuracy. A distinct framework integrates multiple CNN architectures, including MobileNetV2, InceptionV3, InceptionResNetV2, and Xception, to provide a scalable and adaptable solution for anomaly detection across diverse datasets. This paradigm integrates

privacy-preserving methodologies to guarantee data security by Jebur et al. (2025). Both techniques signify substantial progress in anomaly detection by including privacy and scalability into their frameworks.

Attention Mechanisms and Transformer-Based Models

The model proposed by Kumar et al. (2023) combines a CNN with a Residual Attention Bi-LSTM, utilising attention processes to emphasise the most significant segments of the video, hence enhancing the accuracy of anomalous behaviour detection. Another technique employs Vision Transformers (ViTs) to record spatiotemporal correlations, whereas masked autoencoders improve group activity detection in movies, as approached by Yadav et al. (2024). A lightweight CNN combined with LSTM employs Bahdanau attention mechanisms to enhance anomaly classification by focusing on the most pertinent aspects of video sequences, as proposed by Ul Amin et al. (2023). These models illustrate how attention processes may emphasise pertinent characteristics for anomaly detection and action identification. Inception-v3, a CNN variant, is utilised with transfer learning to improve anomaly detection in surveillance videos, as observed by Jebur et al. (2023). The system employs Inflated 3D CNNs (I3D) to capture spatiotemporal characteristics, augmented with attention mechanisms to improve action recognition, following Gowada et al. (2023).

Multi-Stream Models for Spatio-Temporal Analysis

Several advanced models have been developed to enhance video anomaly detection through innovative multi-stream approaches. One model, AnomalyNet, introduces dual parallel networks: one for optical flow and the other for RGB frames, aiming to detect anomalies in surveillance videos (Mumtaz et al., 2024). Additionally, Jan and Khan (2023) proposed a dual-stream approach in which a 3D CNN processes motion and object data, while a 2D CNN captures spatial features, with Softmax used for classification. Lastly, a multi-stream model leverages BERT for action recognition, integrating RGB and optical flow streams to perform spatiotemporal classification for video anomaly detection, as reported by Tan et al. (2024). These models balance spatial and temporal information to enable accurate anomaly detection. Two advanced models have been developed to enhance anomaly detection in video analysis via multi-instance learning (MIL) and weak supervision techniques. An approach uses Multiple Instance Learning with pseudo-labelling to identify irregularities at the frame level without necessitating significant annotations. This model utilises 3D convolution and I3D feature extraction for effective anomaly identification, as demonstrated by Nayak and Chaudhari (2023). An alternative method integrates BERT with MIL ranking and binary cross-entropy to detect anomalies in video segments. It enhances feature extraction by combining MIL with ranking loss, hence improving the

model's ability to detect anomalies with less supervision, as noted by Tan et al. (2024). Both strategies are utilised for anomaly detection, employing MIL and feature extraction techniques to operate on reduced annotated data.

Limitations and Scopes

The analysed articles disclose several significant flaws across various methodologies. The I3D-based model faces accuracy challenges and legal and ethical dilemmas in its use. The CNN + SRU model has difficulties with anomalies without observable individuals and may produce false positives. The 3DCNN + ConvLSTM model is prone to overfitting and requires significant computing resources, whereas DenseNet-121 has challenges with false positives due to shadows and scalability limitations. Attention-based models exhibit suboptimal performance in swiftly evolving or distorted environments, whereas the ConvGRU-CNN model fails to detect transient abnormalities and lacks auditory processing capabilities. The CNN-LSTM model requires prolonged training and addresses only a limited range of abnormalities.

Two-stream networks are highly dependent on optical flow, which limits their effectiveness in complex situations. Transfer learning with Inception-v3 is constrained by pre-trained features, limiting flexibility [34], whereas criminal detection models face scaling challenges due to a limited number of crime categories. AnomalyNet encounters difficulties with temporal context in both rapid and slow actions and is computationally intensive. The Masked Autoencoder approach fails to identify essential items such as weapons, leading to ambiguity in analogous tasks, whereas the hybrid detection model is sensitive to video quality, thereby constraining its real-time applicability.

Conv-LSTM-GRU and LRCN models require substantial computational resources and struggle with intricate movements. The attention-based detection approach, while precise, is computationally intensive and reliant on extensive datasets. The real-time criminal monitoring system has challenges with concealed weapons, and lightweight structures diminish accuracy. Likewise, the multi-density crime predictor has difficulty with obstructed objects and incurs substantial computational requirements. The lightweight design of MobileNetV2 limits its ability to handle intricate scenarios, whereas richer models increase computational complexity. Collaborative learning algorithms face challenges in scaling to distributed data and are adversely affected by noisy pseudo-labels, whereas BERT-based models exhibit variable performance across datasets. The unethical action recognition model suffers from misclassification and training delays due to its complexity, whereas KianNet's dual-attention mechanisms increase computational costs and do not incorporate auditory input processing.

A hybrid fusion method that integrates CNN and RNN effectively mitigates the shortcomings of existing anomaly detection models. This approach leverages the CNN's

capacity to extract spatial information from video frames, while the RNN addresses temporal dependencies, providing a more thorough examination of both spatial and temporal dynamics. The CNN-RNN hybrid fusion method improves detection accuracy by more effectively identifying intricate behaviours in surveillance footage, particularly in challenging conditions such as poor lighting, occlusion, or high crowd density. Furthermore, implementing a hybrid training approach that integrates several optimisation strategies enables the model to optimise performance, mitigate overfitting, and enhance generalisation across diverse datasets. This connection enhances the model's proficiency in managing real-world surveillance data and bolsters its capability to identify nuanced or ambiguous behaviours. This method utilises the advantages of both CNN and RNN, together with enhanced training procedures, to provide more efficient and precise anomaly detection in real-time surveillance systems.

METHODOLOGY

The explosion of digital data in recent years has increased the demand for automated systems that can identify suspicious activity in surveillance and cybersecurity, among other domains. CNNs are one type of DL model that has successfully extracted spatial features from photos. Nevertheless, temporal dynamics—a critical component in identifying questionable activity in video data—are not captured by CNNs. LSTM and Gated Recurrent Unit (GRU) are two examples of RNNs that excel at learning temporal relationships but struggle with spatial dependencies. To overcome these drawbacks, hybrid models integrating RNNs and CNNs have become increasingly popular. Such architectures enable a thorough study of spatial and temporal cues by using CNNs to extract spatial characteristics from images, which are then processed by LSTMs or GRUs to capture temporal patterns. Transfer learning has been helpful in further improving these hybrid models, particularly when using pre-trained CNNs such as AlexNet. Using pre-trained weights on massive datasets such as ImageNet, AlexNet can enhance spatial feature extraction. The hybrid technique builds a robust framework for identifying intricate patterns of suspicious behaviour by combining CNNs for spatial learning, RNNs for temporal learning, and AlexNet for transfer learning. The proposed hybrid model in this work integrates three key components: CNNs, RNNs (LSTM/-GRU), and transfer learning (AlexNet). First, a CNN is employed to extract rich spatial features from each input data frame. Then these features are fed into an RNN (LSTM/GRU) to capture temporal dependencies and model sequential patterns across multiple frames. Incorporating AlexNet as a transfer learning component further boosts the model's generalisation ability by efficiently leveraging pre-learned weights to extract hierarchical spatial features. This architecture enables the hybrid model to detect suspicious activities more accurately by leveraging spatial and temporal cues. This research makes a significant contribution by combining multiple hybrid designs to produce a more reliable

and accurate detection system. Each part of the suggested model—CNN, LSTM/GRU, and AlexNet—contributes differently to learning and feature extraction. The outputs of the various architectures are combined during fusion to produce a single representation that captures sequential dependencies and both high- and low-level features. This hybrid model architecture uses a fusion strategy that combines the temporal features learned by LSTM/GRU with the spatial features retrieved from CNN and AlexNet. Concatenation or weighted average of feature maps and activations from various model branches is used to accomplish this. We hope to improve the system's overall discriminative capabilities by combining the outputs of these models, enabling the hybrid model to more effectively and broadly distinguish between regular and suspicious activity. The fused output is subjected to fully connected and classification layers to produce the final forecast. Through end-to-end fusion, the model's accuracy increases and its resilience is ensured across a variety of datasets, enhancing the hybrid system's adaptability to real-world applications where the detection of suspicious activity is crucial. The algorithm of the proposed hybrid model is shown in Algorithm 1.

Algorithm 1 Hybrid Model with CNN and RNN Fusion

Input: A sequence of frames or images $X = \{x_1, x_2, \dots, x_T\}$
Output: A classification label indicating different types of suspicious activity
Initialisation:
Set hyperparameters: learning rate α, number of epochs E, batch size B, optimiser Opt
Initialise CNN weights W_{CNN}, RNN weights W_{RNN}, fusion layer weights W_{fusion}
Spatial Feature Extraction (CNN):
for each frame $x_t \in X$ for $t = 1,2, \dots, T$ do
    Pass each frame through the CNN using the specified pretrained initialisation.
    Extract spatial features: $f_T^{CNN} = CNN(x_t)$
end for
Collect spatial features: $\{f_1^{CNN}, f_2^{CNN}, \dots, f_T^{CNN}\}$
Temporal Feature Learning (RNN - LSTM/GRU):
Initialise h_o and c_o for LSTM (or h_o for GRU)
for each spatial feature $f_t^{CNN} \in F_{CNN}$ do
    Pass the spatial feature vector into the RNN
    Capture temporal dependencies:
    $h_t^{RNN}, c_t = LSTM(f_t^{CNN}, h_{t-1}, c_{t-1})$ (or)
    $h_t^{GRU} = GRU(f_t^{CNN}, h_{t-1})$
end for
Store hidden states: $H_{RNN} = \{h_1^{RNN}, h_2^{RNN}, \dots, h_T^{RNN}\}$
Fusion of Spatial and Temporal Features:
Apply feature fusion:
F_{fusion} = Fusion (F_{CNN}, H_{RNN}) (or)
Perform fusion via concatenation or addition:
$F_{fusion} = [F_{CNN} \oplus H_{RNN}]$

Classification:
Pass F_{fusion} through a fully connected layer:
$z = W_{fusion} F_{fusion} + b_{fusion}$
Apply softmax or sigmoid to obtain output:
$\hat{y} = Softmax(z)$ (for multi-class) or
Loss Calculation and Backpropagation:
Compute loss $L = Loss(\hat{y}, y)$
Backpropagate the loss through the network, updating weights W_{CNN}, W_{RNN}, W_{fusion}
Training Loop:
for each epoch e = 1, 2, . . ., E do
    Split data into batches of size B
    for each batch do
        Perform forward pass and backpropagation
        Update model parameters
    end for
end for=0

The following describes the branches of the hybrid model that combine CNN and RNN (LSTM, GRU) or CNN with Transfer learning. The outputs from the two branches—RNN and CNN—are fused together in the suggested hybrid CNN-RNN model by concatenating their feature vectors. While the RNN branch captures temporal dependencies, the CNN branch analyses spatial information in video frames. The outputs from both branches, designated as F_{CNN} for the CNN and F_{RNN} for the RNN, are combined into a single feature vector Ffinal once they have finished their individual jobs. Mathematically, this combination of temporal and spatial properties is expressed as shown in Equation 1:

$$F_{final} = concat(F_{CNN}, F_{RNN}) \quad [1]$$

The final feature vector obtained from the combined features is passed through a fully connected (dense) layer with softmax activation. This layer ensures that the model produces a valid probability distribution over the potential classes and outputs probabilities for each class, such as normal or suspicious activity. The last layer's softmax function is defined as follows in Equation 2:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \quad [2]$$

where z_i represents the output score for each class i, and $P(y_i)$ gives the probability of class i. The final output is a set of probabilities, with the class having the highest probability being the predicted activity. The block diagram of the proposed hybrid model schema is shown in Figure 2.

CNN Model

CNN functions as a standard feedforward neural network, developing many filters to extract features from the input data (Liu et al., 2021). The design of our suggested CNN model consists of 16 layers, starting with an input layer that handles grayscale images with dimensions $50 \times 50 \times 1$. This layer receives the image's unprocessed pixel values, with each pixel's intensity represented as a single channel. The input is subsequently transmitted to the initial convolutional layer for preliminary feature extraction. This layer is essential for preparing picture data for further analysis by normalising and organising it into a format suitable for processing by succeeding levels.

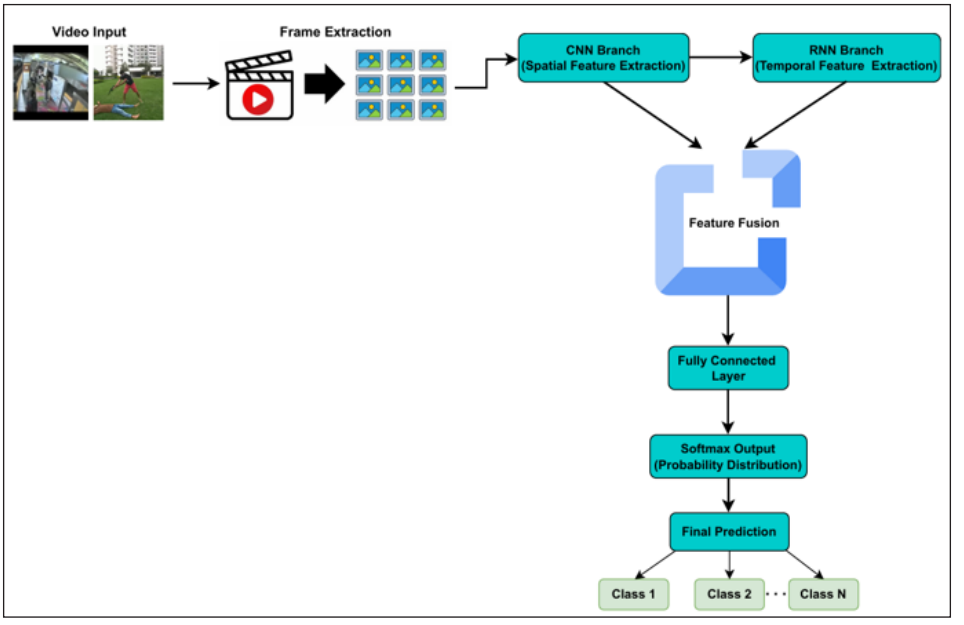


Figure 2. Proposed SentryNet block diagram

This network involves several 2D convolutional layers, each followed by max-pooling and dropout for regularisation. Following convolution and pooling, the CNN architecture incorporates a flatten layer that transforms the multidimensional feature maps into a one-dimensional vector, followed by a fully connected layer with 256 neurons and a dropout rate of 0.5. This model captures the fundamental features step by step using convolution and pooling. These abstractions, which provide translation, scaling, and rotation invariance, help the network deepen. Mathematically, the 2D convolution operation of this network is expressed as Equation 3:

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} \times k_{ij}^l + B_j^l\right) \quad [3]$$

where x_j^l represents the *j*-th feature map of the l-th layer, k_{ij}^l is the kernel applied for convolution, and B_j^l is the bias term. The model guarantees scale invariance while keeping the most important features by using max-pooling to lower dimensionality. The max-pooling operation can be formally expressed as Equation 4:

$$p_j^l = \max\left(q_j^{l-1}(t)\right), \qquad t \in [(j-1)w, jw] \tag{4}$$

where p_j^l is the output after pooling the feature map, and w is the width of the pooling window. Following multiple rounds of convolution and max-pooling, convolution layer feature maps are flattened into a one-dimensional vector and subsequently fed into the fully connected, or dense layers. If the final feature map has dimensions $m \times n \times d$, flattening converts it to a vector of size ($m \times n \times d \rightarrow mnd$). The operation of the dense layer can be articulated as follows Equation 5:

$$z_j^l = \sum_i w_{ij}^j . x_i^l . b_j^l \tag{5}$$

where w_{ij}^j is the weight connecting neuron *i* to neuron *j* in the dense layer, x_i^l is the input from the previous layer, b_j^l is the bias term for the neuron *j*, and z_j^l is the output of the dense layer.

The dropout layer is employed after the fully connected layers to mitigate overfitting by randomly deactivating a proportion of neurons during training. The dropout operation can be characterised as Equation 6:

$$y_i^l = f(x) = \begin{cases} x_i^l, & with\ probability\ p \\ 0, & with\ probability\ 1-p \end{cases} \tag{6}$$

where p is the dropout rate (0.5 in our model), and x_i^l is the output of the neuron *i* in the dense layer.

LSTM Model

CNN-based methods are limited to tackling fixed, short-sequence classification tasks, so they are unsuitable for complex, long-term time-series applications such as crime activity detection, despite their strong resilience and efficiency (Ullah et al., 2022). Analysing trends over time to detect odd or suspicious behaviour in crime activity calls for a model adept at handling sequential data with varying temporal dependencies. Under such circumstances, RNNs fit quite well. Designed specifically to process sequential data, RNNs can efficiently identify temporal correlations and hidden patterns in time-series data. This makes them perfect for jobs like crime activity detection, as identifying trends suggestive of criminal

activity depends on studying occurrences over time. RNNs will enable us to more accurately investigate and detect illegal conduct, hence improving predictive ability over more traditional CNN techniques. Usually, the RNN network predicts the future sequence by concatenating current information from both spatial and temporal dimensions with past information, thereby analysing the underlying sequential pattern (Khan et al., 2021). LSTM's internal design consists of multiple gates—input, forget, and output gates—where each gate processes the input from the previous gate and passes it to the next gate, thereby regulating the flow of information towards the ultimate output (Ullah et al., 2021)

Usually, a sigmoid or *tanh* activation function controls all gates; for example, the input gate is in charge of updating the information. When necessary, the forget gate removes the information from the current state C_t and processes the input information from the input gate i_t and the state of the previous cell C_{t-1}. Whereas the output gate o_t stores the output value for the next sequence prediction and transfers the last output to the next LSTM unit. Conversely, the recurrent unit C_t estimates the current input value x_t using the tanh activation function and the state of the previous cell C_{t-1}. Whereas the scalar product of o_t and $tanh$ of C_t. allows one to determine the value of h_t. At last, feeding h_t to the softmax classifier will provide the ultimate result. Mathematically, the above-listed gates' operations can be stated as follows from Equations 7 to 13:

$$f_t = \Phi\left(\widehat{W}_f \cdot [h_{t-1}, x_t]\right) + B_f \quad [7]$$

$$i_t = \Phi\left(\widehat{W}_i \cdot [h_{t-1}, x_t]\right) + B_i \quad [8]$$

$$C_t = \tanh\left(\left(\widehat{W}_C \cdot [h_{t-1}, x_t]\right) + B_C\right. \quad [9]$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t \quad [10]$$

$$o_t = \Phi\left(\widehat{W}_o \cdot [h_{t-1}, x_t]\right) + B_o \quad [11]$$

$$h_t = o_t \cdot \tanh\left(\Phi(C_t)\right) \quad [12]$$

$$Output = softmax(h_t) \quad [13]$$

GRU Model

A particular kind of RNN, called the GRU, was created to overcome the difficulties of discovering long-term dependencies in sequential data. The GRU, proposed by Cho et al. (2014), improves on the capabilities of conventional RNNs, which often encounter problems such as the vanishing gradient problem, particularly when handling long sequences. The GRU is especially useful for tasks because it addresses this issue by introducing gating

mechanisms that control the flow of input. The update gate (z_t) and the reset gate (r_t) are the two main gates of the GRU. The update gate controls the amount of historical data that is required for the present state by deciding how much of the previous hidden state (h_{t-1}) should be retained. The model can ignore unnecessary data from earlier time steps because the reset gate controls how much of the previous hidden state to discard. GRUs merge cell and hidden states into a single hidden state, whereas LSTM networks maintain separate cell and hidden states. This results in fewer parameters and quicker training and inference times. The GRU operations can be mathematically defined through a series of equations; The reset gate is calculated as shown in Equation 14:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad [14]$$

where σ is the sigmoid activation function, W_r and U_r are weight matrices for the input x_t and the previous hidden

state h_{t-1}, and b_r is the bias term for the reset gate. The update gate is computed as shown in Equation 15:

$$z_t = \sigma(W_z x_z + U_z h_{t-1} + b_z) \quad [15]$$

where z_t is the update gate, W_z and U_z are weight matrices, and b_z is the bias term. The new memory content, representing the candidate hidden state, is given as shown in Equation 16:

$$\widetilde{h_t} = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \quad [16]$$

where $\widetilde{h_t}$ is the candidate hidden state, W_h and U_h are the weight matrices, b_h is the bias term, and $\odot$ denotes element-wise multiplication. Finally, the hidden state at time t is calculated as shown in Equation 17:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t} \quad [17]$$

where h_t is the hidden state. Overall, the gating methods and the GRU's integrated hidden state enable it to process sequential input effectively while addressing issues with conventional RNNs. Language modelling, sentiment analysis, and time-series forecasting are just a few of the applications that choose it for its ability to handle long-range dependencies with lower computational complexity. The GRU improves neural network performance on tasks involving sequential patterns by efficiently balancing information retention and amnesia. One of the main differences between LSTM and GRU networks is that LSTM networks rely on multiple hidden states, whereas GRUs have only one.

Faster computations and lower memory usage result from this integration's architectural simplification and reduced parameter count. Combining the forget and input gate functions into a single update gate substantially improves the computational efficiency of the GRU by eliminating the need for separate cell states. Consequently, the GRU maintains strong performance in a wide range of applications while achieving faster training and inference durations. The basic structure of GRU and LSTM is shown in Figure 3.

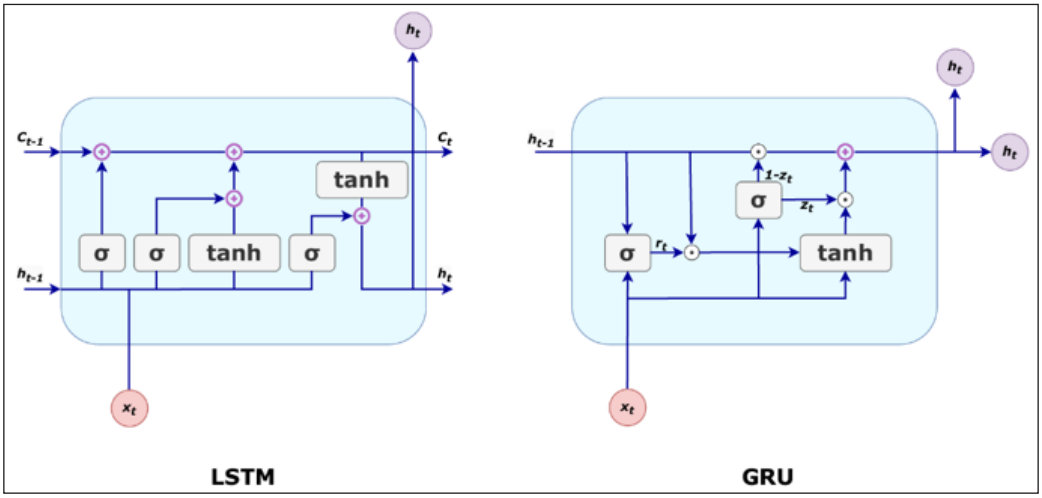


Figure 3. Basic architecture of LSTM and GRU

Proposed CNN+LSTM Hybrid Model

For efficient feature extraction and sequence modelling, we provide a hybrid method that combines the capabilities of LSTM networks and CNNs. Starting with two 2D convolutional layers, the first utilises 64 filters and the second uses 128 filters, both with a kernel size of 3x3, thereby effectively managing negative activations. LeakyReLU activation function is used in both layers. Each convolutional layer is followed by a dropout layer to reduce overfitting and a 2x2 Max Pooling layer. Subsequently, the extracted features are fed into the LSTM network, which consists of two layers, each with eight LSTM units, and aims to capture temporal relationships in the data. Based on these properties, a final Dense layer with a softmax activation function forecasts class labels. Originally using the Adam optimiser with a learning rate of 0.001, the model first utilises a hybrid of Adam and SGD, then switches to SGD with a learning rate of 0.001 and momentum of 0.9 for better convergence and optimal results. This hybrid model shows promise for robust data modelling, as it efficiently captures temporal and spatial relationships. The architecture is shown in Figure 4.

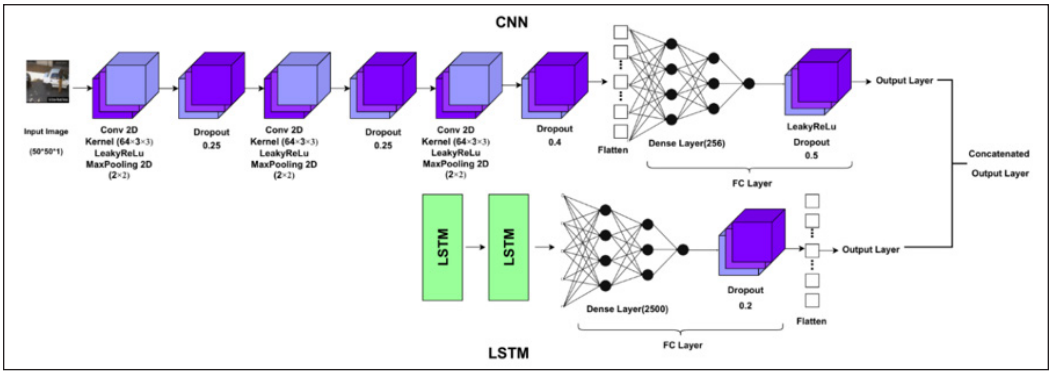


Figure 4. Architecture of the proposed CNN+LSTM model

Proposed CNN+GRU Hybrid Model

The CNN-based models were initialised using the pretrained weights specified for their respective architectures. AlexNet was used as a transfer-learning component with weights pretrained on ImageNet. The pretrained convolutional representations were used to extract hierarchical spatial features from the surveillance frames, after which the corresponding recurrent or classification layers were trained for the target activity-classification task. This initialisation strategy is distinct from training the entire network with randomly initialised weights. The MaxPooling layers downsample the feature maps, reducing spatial dimensions and computational cost, while the LeakyReLU activation function adds non-linearity, enabling the network to learn more complex representations. To improve generalisation and avoid overfitting, dropout layers are used. To prepare the data for integration with the GRU network, the flattening phase converts the 2D feature maps into a 1D vector. At this point, the CNN has successfully converted the input's spatial structure into a feature representation that highlights the most important elements of the data. Two GRU layers are employed in the model, one of which is configured to return sequences of True values, enabling the network to output a series of hidden states. For sequential tasks, where the contribution of each timestep is crucial to the final prediction, this setup is advantageous. The GRU maintains strong performance while being computationally inexpensive thanks to its very small number of parameters and its ability to preserve critical information over time steps. Concatenating the feature representations from both components after they have processed their respective inputs yields the final output in the hybrid CNN-GRU model. While the GRU processes the sequential input and outputs the hidden state h_t, the CNN processes the spatial data and produces a flattened output V. Concatenating these feature vectors results in a combined representation, as shown in Equation 18:

$$H_{concat} = concat(V, h_t) \quad [18]$$

After passing through a fully connected layer with this concatenated feature vector, the final output is calculated using Equation 19

$$y_{pred} = softmax(W_o H_{concat} + b_o) \quad [19]$$

Here, W_o is the weight matrix, b_o is the bias term, and y_{pred} represents the predicted class probabilities. The integration of CNN and GRU in this manner allows the model to capture both spatial and temporal features, enhancing its ability to perform classification tasks. The architecture of the proposed model is shown in Figure 5.

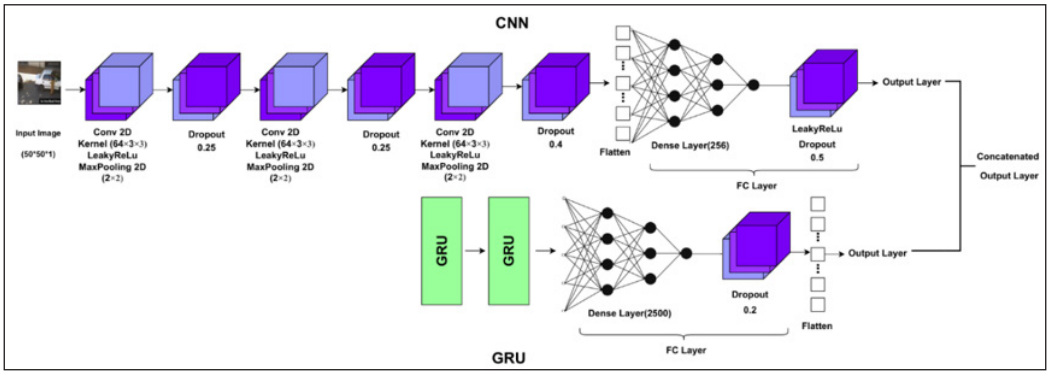


Figure 5. Architecture of the proposed CNN+GRU model

EXPERIMENTAL SETUP

Implementation

In our research, we employed Python as the primary programming language and executed the whole process in Google Colab Pro, utilising its superior computing capabilities and GPU acceleration to expedite training and execution. We utilised TensorFlow 2.15, Scikit-Learn, and Keras to develop and train DL models, while ImageDataGenerator facilitated real-time data augmentation, enhancing generalisation by generating diverse training examples.

We incorporated OpenCV for image preprocessing to perform tasks such as scaling and formatting the input data, thereby ensuring uniformity across the dataset. The Scikit-learn library offers essential utilities, such as train_test_split for dataset splitting and evaluation metrics like classification_report and confusion_matrix, for a complete assessment of model performance. We utilised Pickle for efficient storage of intermediate results, and TQDM improved code readability by including progress bars in our loop operations.

We incorporated convolutional layers and other critical elements for spatial feature extraction, followed by layers for temporal analysis, thereby capturing both spatial and

sequential characteristics. Utilising Colab Pro facilitated the rapid training of models and performance monitoring via ModelCheckpoint and CSVLogger, thereby enabling ongoing progress tracking and ensuring repeatability. Visualisation packages like Matplotlib were essential for graphing performance indicators, facilitating the analysis and optimisation of the model. The integration of these libraries, together with the computational benefits of Colab Pro, has markedly enhanced the productivity and performance of our research. The number of parameters of each model for both datasets is shown in Table 1.

Table 1
Parameters and values of all models

Dataset	Model	Parameter	Size (Megabyte)	Training Time (Seconds)
Dataset 1	CNN+AlexNet	32396948	123.58	105.98
	CNN+LSTM	3705000	14.13	124.88
	CNN+GRU	3705168	14.13	127.66
Dataset 2	CNN+AlexNet	62396948	86.3	156.32
	CNN+LSTM	17411208	66.42	136.7
	CNN+GRU	17411040	66.42	143.86

Data Preprocessing

Using Python's OpenCV library, frames were extracted from videos as part of the video preprocessing in this investigation. The frames were initially placed in an empty list, and the total number of frames in each video was obtained by accessing it. To ensure that every aspect of the video material was covered, frames were systematically extracted at predetermined intervals. The predefined sequence length, S_{length}, was divided by the total frame count, F_{total}, to find the interval. The result was shown in Equation 20:

$$I = \frac{F_{total}}{S_{length}} \quad [20]$$

where the interval at which frames were extracted is denoted by I. Following extraction, frames were resized to standard sizes (e.g., 50 × 50 and 128× 128 pixels) in accordance with the analysis specifications, guaranteeing uniformity throughout all samples. The 50 × 50 was used for dataset 1, and the 128x128 was used for dataset 2. After resizing the frames, colour histograms were computed for each frame to capture the distribution of pixel intensities across colour channels. The colour histogram, denoted by H, was computed over the three colour channels (Red, Green, Blue) using 8 bins per channel, producing a total of 512 bins as shown in Equation 21:

$$H(c_1, c_2, c_3) = \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} \mathbb{I}\big(I_{x,y}(R) = c_1\big)\mathbb{I}\big(I_{x,y}(G) = c_2\big)\mathbb{I}\big(I_{x,y}(B) = c_3\big) \quad [21]$$

where:

-$H(c_1, c_2, c_3)$ is the histogram bin value for a particular combination of colour intensities c_1 (Red), c_2 (Green), and c_3 (Blue). - W and H represent the width and height of the image, respectively. -𝕀 is an indicator function that checks if the pixel at coordinates (x, y) has the specific colour values c_1, c_2, and c_3.

This resulted in a compact feature vector representing the overall colour distribution of the frame, which was then flattened for use in the model. The colour histogram method is computationally efficient and captures important colour patterns, which is useful for distinguishing different visual content.

Later, the normalised pixel intensity I_{norm} for every pixel p in each image was computed using min-max normalisation as follows shown in Equation 22:

$$I_{norm} = \frac{I_p - I_{\min}}{I_{max} - I_{min}} \quad [22]$$

where I_p represents the pixel value, and I_{min} and I_{max} denote the minimum and maximum pixel intensities, respectively. where the minimum and maximum pixel intensities are indicated by Imax and I_{min} accordingly, and the pixel value is represented by I_p. To standardise the frames for later model input, this phase ensured that the pixel values, initially in the range [0, 255], were scaled to [0, 1]. By lessening the sensitivity to changes in pixel intensity, this preprocessing step improved the convergence and generalisation of DL models. The prepared frames were fed into the suggested model for additional analysis and training after normalisation. For each video, frames were sampled at approximately uniform temporal intervals to construct an ordered sequence of frames. Let F_{total} denote the total number of frames in a video and S_{length} denote the predefined sequence length. The sampling interval was determined as

$$I = \frac{F_{total}}{S_{length}}$$

where I represent the approximate interval between two consecutive sampled frames. The selected frames preserve their original temporal order and form a sequence

$$\left\{X_1, X_2, \ldots\ldots\ldots\ldots, X_{s\,lengt\,h}\right\}$$

Each frame was independently resized and normalised before being processed by the CNN. Thus, the CNN extracts spatial features from individual frames, and the resulting ordered sequence of features is fed into the LSTM to model temporal dependencies across the video segment.

Dataset Details and Splitting

UCF-Crime and HBD21, two publicly accessible datasets, were used in this study. The two datasets, from reputable sources, were essential for evaluating the proposed hybrid model for identifying questionable activities. Table 2 presents a comprehensive overview of the datasets, outlining the research questions each dataset addresses and providing references for further investigation.

Table 2
Summary of datasets used

Dataset	Video Samples	Reference
UCF-Crime (Dataset 1)	1900	Sultani et al., 2018
HBD21 (Dataset 2)	463	Jayaswal & Dixit, 2021b

Each dataset extracts a specific number of images from videos. A total of 338,784 images are taken from 1,900 videos, and a total of 13890 images are taken from 463 videos for the UCF-Crime dataset, as shown in Table 3 and Table 4.

The UCF-Crime dataset consists of 12 classes— Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Vandalism, Stealing, Robbery, Shooting, and Shoplifting—which we examined. Each class in the dataset has a specific number of images, for a total of 338,784 images extracted from 1,900 videos after the dataset was gathered.

Table 3
Training and testing set distribution after 80:20 split (UCF Crime dataset)

Class	Number of Images	
	Training Set	Testing Set
Abuse	15,557	3,816
Arrest	23,666	6,096
Arson	21,863	5,351
Assault	10,417	2,600
Burglary	37,766	9,395
Explosion	20,249	5,014
Fighting	20,623	5,292
Vandalism	11,798	2,939
Stealing	37,391	9,395
Robbery	33,850	8,478
Shooting	11,859	2,911
Shoplifting	25,988	6,470
Total	271,027	67,757

Using train_test_split from the sklearn model_selection module, we split our UCF Crime dataset at an 80:20 ratio before training to ensure an appropriate distribution of training and testing data. The training set will consist of 271,027 images, and the testing set will have 67,757 images; after splitting,

each class will have the following number of images for training and testing, as shown in Table 4.

Similarly, the Human Behaviour dataset (HBD21) contains four activity categories: assault violence, gun violence, normal actions, and sabotage violence. In total, 13,890 images were extracted from 463 videos. An 80:20 training-testing split was used, resulting in 11,112 training images and 2,778 testing images. The class-wise distributions are reported in Table 4.

Table 4
Training and testing set distribution after 80:20 split (HBD21)

Class	Number of Images	
	Training Set	**Testing Set**
assault_violence	2454	636
gun_violence	3469	851
normal_actions	2613	627
sabotage_violence	2576	664
Total	11112	2778

Table 5
Hyperparameter setup

Parameter	Value
Learning Rate	0.01, 0.001
Optimizer	Adam, SGD, Hybrid
Loss Function	CategoricalCrossentropy
Activation (CNN Hidden)	LeakyReLU (alpha=0.1)
Activation (Output)	Softmax
Epochs	50
Batch Size	1000
Verbose	1
Callbacks	ModelCheckpoint, CSVLogger, EarlyStopping
CNN Layers	3 Conv2D + MaxPooling2D + Dropout
Input Shape (CNN)	(50, 50, 1) (after reshaping)
Regularisation (CNN)	L2 (0.01)
Number of Classes	12

Hyperparameters

The experiments used a fixed maximum of 50 training epochs and a batch size of 1000. Learning-rate experiments considered values of 0.01 and 0.001, while the final configuration was selected based on the comparative experiments reported in Tables 6 and 7. For the hybrid optimiser, Adam was used during the initial training phase and was subsequently replaced by SGD with momentum at the predefined switching epoch. The momentum coefficient and switching epoch were fixed throughout the corresponding experiments. These settings are explicitly reported in Table 5 to improve reproducibility.

Learning Rate: In training, the learning rate is a crucial hyperparameter that regulates the size of weight updates during gradient descent. It regulates the rate at which the model learns by adjusting the weights to minimise the loss function. An elevated learning rate may lead to rapid convergence, risking overshooting the optimal solution, whereas a lower learning rate might result in slower convergence or entrapment in local minima. When training samples are very numerous, the learning rate becomes a concern as the underlying system loses its capacity for meaningful generalisation. This can also happen when the computational volume exceeds the input's

dimensionality in the vector space, and the network has too many neurons. Although it reduces training loss, it is well acknowledged that a deep network cannot be effectively trained with a small initial learning rate (Li et al., 2019). While reducing the memorisation of noisy data, an initially high learning rate slows the learning of complex patterns via degradation (You et al., 2019). The equation for weight updates in gradient descent is as follows in Equation 23:

$$W_{new} = W_{old} - \alpha \cdot \frac{\partial L}{\partial W} \quad [23]$$

where, $W_{new}$is the updated weight, W_{old} is the previous weight, α is the learning rate, $\frac{\partial L}{\partial W}$ is the gradient of the loss function L with respect to the weights W. Our research employed three learning rates—0.01, 0.001, and 0.0001—to investigate their impact on the training dynamics of our hybrid models. A learning rate of 0.01 facilitated rapid convergence through substantial weight updates; however, this expediency posed the risk of instability and the possibility of overshooting the ideal solution (Kingma & Ba, 2014). A learning rate of 0.001 achieved equilibrium, promoting consistent advancement with controllable updates and preventing excessive variability. The minimal learning rate of 0.0001 facilitated very steady and accurate modifications, enhancing model performance, but necessitating extended training durations. By manipulating these rates, we sought to balance convergence speed and model stability, ensuring both efficient training and successful outcomes.

Activation Function: A crucial element of neural network architecture that influences behaviour and computational complexity. Artificial neural networks employ activation functions to convert an input signal into an output signal, which is then supplied back as input to the next layer in the stack. We compute the sum of the products of inputs and their corresponding weights in an artificial neural network, then apply an activation function to produce the output of that layer, which is fed as input to the next layer. An activation function is used to transform the input signal into a simple linear function, which is equivalent to a degree-1 polynomial. Even though linear equations are simple and easy to solve, they are not very complex and can't learn or identify complex mappings from data. In most cases, a neural network without activation functions behaves like a linear regression model, but it doesn't offer as much power or speed. It would be great if a neural network could do more than just learn and generate a linear function. For example, it should be able to model different types of data, such as images, videos, music, speech, and text (Sharma et al., 2020). We used the Softmax activation function in the output layer and the LeakyReLU activation function in the hidden layers for our work. The LeakyReLU effect assigns a variant to negative values to preserve neurons and thus the useful data. Unlike the ReLU function, which renders these values zero and kills the neurons, it can ignore part

of the important data. For massive volumes of data, this can greatly influence the learning process. Moreover, the LeakyReLU function offers a trade-off between performance and network sparsity. LeakyReLU is a variation of ReLU by assigning a non-zero output for negative input, shows in Equation 24:

$$f_{LeakyReLU}\ (x) = max(\alpha x, x) \quad [24]$$

where α is a predetermined value in the range of (0, 1). ReLU maps negative inputs to zero, whereas LeakyReLU compresses negative inputs using a predefined linear function. Compressing LeakyReLU helps preserve the negative part of the feature information. LeakyReLU therefore trades off its input knowledge with network sparsity. Moreover, ReLU and LeakyReLU are unbounded functions that solve the gradient vanishing problem in Softplus and ELU. On the other hand, A combination of multiple sigmoid functions is the softmax function. These values can be interpreted as the probabilities of data points belonging to a specific class, since a sigmoid function returns values in the range 0 to 1. In contrast to sigmoid functions, which are employed for binary classification, the softmax function can be employed for multiclass classification problems. The function returns the probability for each data point across all individual classes. The softmax activation function ensures that, as shown in Equation 25:

$$\sum_{k} P_k z_k = 1 \quad [25]$$

which is desirable in a classification problem with q classes and allows us to use the cross-entropy error function. However, by modelling $P(x|C)$, we can provide a better justification for the use of the softmax activation function. For a classification scheme using the sampling paradigm, $P(C_i|x)$ is modelled as shown in Equation 26:

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{\sum_k P(x|C_k)\,P(C_k)} \quad [26]$$

where $P(C_i|x)$ is the posterior probability of class C_i given input x, $P(x|C_i)$ is the likelihood, and $P(C_i)$ is the prior probability of class C_i. The calculated probabilities are constrained by the Softmax function to ensure that they sum to 1 and that each probability lies between 0 and 1. The Softmax function is defined as shown in Equation 27,

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_j e^{z_i}} \quad [27]$$

where $\sigma(z_i)$ represents the probability for class i, and the denominator is the sum of exponentials of all class scores z_j.

Loss Function: In mathematics, a loss function, also referred to as a cost function or objective function, is a mathematical function that measures the actual target value's difference from the expected output of an ML model. It gauges the model's performance: a smaller loss indicates that the model's predictions are more in line with the actual values, whereas a larger loss indicates inadequate forecasts. The aim of the training procedure is to minimise the loss function by adjusting model parameters (weights and biases) via gradient descent. Conversely, the loss functions of neural networks often limit their accuracy [64]. In this study, we used the categorical cross-entropy loss function, which is commonly used for multiclass classification problems. This loss function evaluates the degree of alignment between the predicted probability distribution and the actual distribution by comparing predicted class probabilities with the actual class identifiers [65] (typically represented in one-hot format). The model's classification accuracy is enhanced by encouraging it to assign higher probabilities to the correct class by minimising this loss during training. This loss function is defined as shown in Equation 28:

$$L(y, \hat{y}) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \quad [28]$$

where y_i represents the true label for class i (typically in one-hot encoded format), and $\hat{y}_i$ denotes the predicted probability for class i. Given the predicted probabilities, this loss function computes the negative log-likelihood of the true labels. The objective of training is to reduce this loss, thereby enhancing the predicted probability of the correct class.

Optimisation is the process of minimising or maximising an objective function, also known as a cost or loss function. Attempts are made to identify optimal parameters θ that substantially reduce the cost function J(θ) in a training dataset for a deep neural network. To identify the optimal parameters for a deep neural network, gradient descent can be used. The accuracy of a deep neural network model is influenced by the optimiser's role. The optimisers used in this research are described below.

Adam

The advantages of momentum and RMSProp are combined in the popular DL technique known as the Adam (Adaptive Moment Estimation) optimiser. By keeping squared gradients and exponentially decaying averages of previous gradients, it modifies the learning rate for every parameter. Adam computes two estimates at each time step t:

The first moment (mean of gradients) m_t, updated as:

$$m_t = \beta_t m_{t-1} + (1 - \beta_1) g_t$$

The second moment (variance of gradients) v_t, updated as:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

Bias correction terms are applied to m_t and v_t:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

The parameter update rule is:

$$\Theta_t = \Theta_{t-1} - \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

where:
α is the learning rate,
ε is a small constant added for numerical stability.

SGD

When training neural networks, SGD is a fundamental optimisation technique frequently utilised, particularly for large datasets. By calculating the gradient of the loss function with respect to the parameters over a single or small batch of data, the model parameters are updated incrementally, facilitating quicker and more effective training. The usual SGD update rule is written as shown in Equation 29:

$$\Theta_{t+1} = \Theta_t - \eta \nabla_{\Theta_t} J(\Theta) \quad [29]$$

where η is the learning rate, and $\nabla_{\Theta_t} J(\Theta)$ is the gradient of the loss function $J(\Theta)$ with respect to the parameters Θ. Momentum can be added to SGD to improve its performance by reducing oscillations and hastening convergence in the desired direction. The definition of the momentum-enhanced variant of SGD is shown in Equations 30 and 31:

$$v_{t+1} = \mu v_t - \eta \nabla_{\Theta_t} J(\Theta) \quad [30]$$

$$\Theta_{t+1} = \Theta_t + v_{t+1} \quad [31]$$

where μ is the momentum coefficient, typically set to a value between 0.8 and 0.9. This approach allows the optimiser to build velocity in directions with consistent gradients, leading to more efficient and stable convergence.

Hybrid Optimiser

In the hybrid optimisation approach, the optimiser initially employs Adam to leverage its adaptive learning rate, which excels at handling noisy gradients and efficiently navigating the early stages of training. This phase is characterised by rapid loss reduction, thanks to Adam's ability to adjust the learning rate for each parameter based on the first and second moments of the gradients. Mathematically, the Adam update rule for epochs $t < T_{switch}$ (where T_{switch} is the predefined switch epoch) is shown in Equation 32:

$$\Theta_{t+1} = \Theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}} \quad [32]$$

Here, $\hat{m}_t$ and $\hat{v}_t$ represent the bias-corrected first and second moments of the gradients, and α is the learning rate, with ε being a small constant for numerical stability.

As training progresses and reaches the epoch $t \geq T_{switch}$, the optimiser switches to SGD with momentum, which is known for its stable convergence and superior generalisation performance in the later stages of training. The momentum update rule is already given in Equations 30 and 31. This phase helps the model fine-tune and converge to a more robust solution, avoiding sharp minima and improving generalisation.

The rationale for combining Adam and SGD in a hybrid strategy is to leverage the strengths of both optimisers. Adam accelerates convergence during the noisy, early phase of training, while SGD with momentum provides stability and better generalisation as the model approaches the global minima. This hybrid approach balances fast early training and stable, efficient convergence, ensuring a more optimal and robust solution. Mathematically, the hybrid optimiser's behaviour is described by Equation 33:

$$\Theta_{t+1} = \begin{cases} \Theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}}, & if\ t < T_{switch} \\ \Theta_t + v_{t+1}, & if\ t \geq T_{switch} \end{cases} \quad [33]$$

This combination of rapid initial optimisation with Adam and fine-tuned convergence with SGD offers a powerful, flexible strategy for training the proposed models.

Epoch

A single pass through the entire training dataset is called an epoch. The model adjusts its weights based on the computed loss for each training example during each epoch. The duration of model training depends on the number of epochs. An insufficient number of epochs can lead to underfitting, in which the model fails to learn the data patterns, while an excessive number of epochs can result in overfitting, in which the model learns noise and irrelevant details; the number of epochs ensures that the model converges to a satisfactory solution by striking a balance between model training time and performance. The learning algorithm typically operates for a significant number of epochs, often in the hundreds or thousands, until the model's error is adequately reduced (Brownlee, 2022). In our studies, we trained the model for 50 epochs—a decision that provided a well-balanced approach to learning and generalisation. Training for over 50 epochs gave the model ample opportunities to learn intricate patterns and characteristics from the dataset, thereby ensuring it could faithfully capture the fundamental structure and relationships in the data. This length was carefully chosen to prevent overfitting, a common problem when the model performs well on training data but poorly on unseen test data. We sought to balance sufficient learning with avoiding excessive memorisation of the training instances by setting the number of epochs to 50. This method ensured that the model could efficiently generalise to new data, thereby improving its overall dependability and performance in practical settings. Furthermore, this number of epochs allowed for complete training without the risk of overly long training durations, which can lead to diminishing returns in model accuracy and higher computational costs.

EVALUATION METRICS

Evaluation metrics, which offer measurable indicators of a model's performance on a particular task, are crucial for evaluating ML models' effectiveness. To assess the models' performance in our tests, we used metrics such as accuracy, precision, recall, F1-score, classification error, AUC-ROC, confusion matrix, and jaccard_score. Among the most frequently utilised performance criteria in prior research is accuracy. In this example, accuracy might be useful if the dataset has the same number of cases per class. If not, as the model forecasts the value of the majority class, accuracy cannot be a useful metric. Considered the average value of recall and accuracy, the F1-score comprises essential findings about the performance of classifiers in every class. Hence, it is helpful in situations with diverse class distributions. True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are fundamental elements of assessment metrics used to gauge a model's performance. While FP refers to cases where the model mistakenly predicts a positive class for a negative instance, TP refers to cases where the model correctly predicts the positive class. TN shows examples where the model correctly forecasts the

negative class; FN results when the model fails to find a positive occurrence and incorrectly categorises it as negative. These indicators provide a more comprehensive picture of how well a model works by helping to estimate other important performance measures, including accuracy, precision, recall, and F1-score.

Accuracy

In most research, the accuracy metric has been used primarily to estimate the classification accuracy of models. As demonstrated in the following Equation 34, the accuracy metric consists of the measurements of the confusion matrix in our hybrid models,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [34]$$

Precision

It indicates the quantity of relevant selected data elements. In other words, of the observations that an algorithm has classified as positive, how many are genuinely positive. As per Equation 35, precision is defined as the ratio of true positives to the total of true positives and false positives,

$$Precision = \frac{TP}{TP + FP} \quad [35]$$

Recall

It provides "what number of relevant data items are selected". Out of the observations that are positive, how many the algorithm has predicted. Equation 36 claims that the recall is equal to the number of true positives divided by the total of true positives and false negatives:

$$Recall = \frac{TP}{TP + FN} \quad [36]$$

F1-Score

Calculating an algorithm's performance uses accuracy and recall under consideration using this statistic, sometimes referred to as F1-score or F-measure. Mathematically, it is the harmonic mean of recall and precision, developed as follows Equation 37:

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad [37]$$

Confusion Matrix

Accuracy and correctness of an ML algorithm can be assessed using this matrix, one of the clearest and most understandable measures available. Its primary use is in classification problems where two or more classes might be produced. ROC-AUC score: ROC curve—receiver operating characteristic curve—which shows the relationship between false positive rate (1-specificity) and actual positive rate—aka sensitivity or recall—is used to compute this statistic. Notably, the ROC-AUC score can be a good performance indicator if the numbers of positive and negative samples are equal.

Kappa

Cohen's kappa is computed using the predicted degree of agreement resulting from chance alone and the observed agreement among raters. Cohen's kappa lets researchers assess the extent to which raters' assessments are consistent and trustworthy, rather than only the outcome of random chance, by comparing these two values [80]. The expression for kappa can be given by Equation 38.

$$p_e = \frac{(TP + FP)(TP + FN)}{(TP + TN + FP + FN)^2} + \frac{(TN + FP)(TN + FN)}{(TP + TN + FP + FN)^2} \quad [38]$$

MCC

For imbalanced data, the Matthews correlation coefficient (MCC) measures the precision of binary classifications. It also considers all of TP, TN, FP, and FN and is considered a balanced indicator, applicable even when class sizes vary. This can be mathematically expressed as Equation 39,

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad [39]$$

OPTIMISER BASED RESULT

The optimisation technique applied in this study stimulates the performance evaluation of several hybrid models integrating CNNs with transfer learning models and RNNs. Optimisers are essential for modifying model parameters to reduce the loss function and enhance overall performance. This study used Adam and SGD optimisers to evaluate their effects on model convergence and performance. In this research, we utilised two datasets: UCF-Crime and HBD21. We focused solely on Dataset 1 (UCF-Crime) to establish a baseline for our optimisation approach. The performance metrics of the baseline optimiser for Dataset 1 are detailed in the following section.

Performance of Adam Optimiser

With a learning rate of 0.001, CNN+AlexNet ranks first among models, achieving 98.61% on all measures—accuracy, precision, recall, and F1-score. This reveals a high degree of consistency and balance in spotting illicit behaviour. By leveraging its deep architecture and fine-tuning capability with the Adam optimiser, CNN+AlexNet may maximise performance at reduced learning rates. Though they do significantly less, CNN+LSTM and CNN+GRU yield consistent, strong results. CNN+LSTM exhibits good performance with 97.54% across all evaluation measures—precision, recall, and F1-score. With its 97.52% across all tests, CNN+GRU is thus also a consistent model. Though they generalise well, the small differences between these two models suggest that their architectures are not as strong at this learning rate as CNN+AlexNet.

Every model shows a small performance decline at the higher learning rate of 0.01. Leading with 97.82% accuracy, 97.53% precision, 97.32% recall, and a 97.81% F1-score, CNN+AlexNet still leads. The model still maintains high generalisation capacity, even when these values are somewhat lower than the findings obtained at 0.001; this holds even when the learning rate is raised. CNN+LSTM performs worse still with 96.84% accuracy, 96.53% precision, 96.53% recall, and an F1-score of 96.29%. This suggests that CNN+LSTM is more sensitive to increasing learning rates, thereby resulting in a larger performance reduction than CNN+AlexNet. CNN+GRU demonstrates higher resilience than CNN+LSTM, achieving 97.65% accuracy, 97.43% precision, 96.52% recall, and an F1-score of 96.53%; however, it shows a slight drop.

For all three hybrid models, 0.001 is the optimal learning rate compared with the other two rates (Table 6). With the highest precision, recall, and F1 scores, the CNN+AlexNet model proves its supremacy and is the most effective model overall. CNN+LSTM and CNN+GRU perform well, particularly at 0.001. Still, their performance drops more sharply at 0.01, confirming that a lower learning rate is preferable for maintaining consistent model accuracy and preventing overfitting or underfitting across different models. When CNN+LSTM and CNN+GRU are trained at the same learning rate, CNN+AlexNet achieves

Table 6
Comparison of hybrid models using adam optimiser with different learning rates

Learning Rate	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
0.001	CNN+AlexNet	98.61	98.62	98.61	98.61
	CNN+LSTM	97.54	97.54	97.54	97.54
	CNN+GRU	97.52	97.52	97.52	97.52
0.01	CNN+AlexNet	97.82	97.53	97.32	97.81
	CNN+LSTM	96.84	96.53	96.54	96.29
	CNN+GRU	97.65	97.43	96.52	96.53

the best results at a learning rate of 0.001 using the Adam optimiser. The overall outcomes are shown in Figure 6.

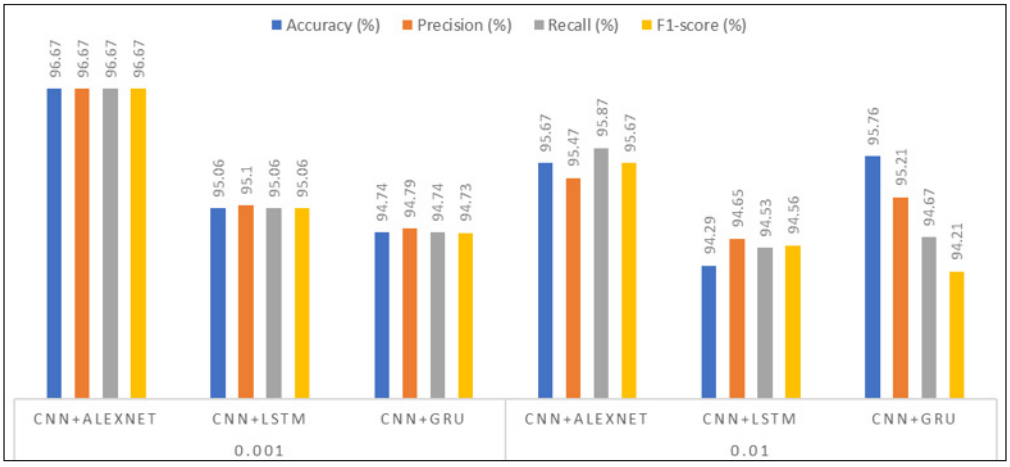


Figure 6. Performance comparison of three models using the Adam optimiser with varying learning rates

Performance of SGD Optimiser

This work finds that the performance of hybrid models combining CNN with AlexNet, LSTM, and GRU depends critically on the SGD optimiser. SGD has proven to be efficient as a conventional optimisation method for guiding these models toward convergence, especially in situations with large datasets, as shown in Table 7 for crime activity detection.

Table 7
Comparison of hybrid models using SGD optimiser with different learning rates

Learning Rate	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
0.001	CNN+AlexNet	96.67	96.67	96.67	96.67
	CNN+LSTM	95.06	95.1	95.06	95.06
	CNN+GRU	94.74	94.79	94.74	94.73
0.01	CNN+AlexNet	95.67	95.47	95.87	95.67
	CNN+LSTM	94.29	94.65	94.53	94.56
	CNN+GRU	95.76	95.21	94.67	94.21

Of the three hybrid models using the SGD optimiser at a learning rate of 0.001, CNN+AlexNet ranks highest. The model shows steady performance across metrics, with accuracy, recall, and F1-score all reaching 96.67%. CNN+AlexNet thus detects relevant features and effectively balances predictions for positive and negative classes. Excellent alignment of these criteria suggests that the model generalises well, free from evidence of

overfitting or underfitting. The CNN + LSTM model achieves 95.10% accuracy, recall, and F1-score. Though it lags considerably behind CNN+AlexNet, this performance indicates good generalisation. At this learning rate, CNN+GRU achieves the lowest metrics: 94.79% accuracy, 94.74% recall, and an F1-score of 94.73%. The similarity of these values across the LSTM and GRU models indicates stable performance, even though they do not match CNN+AlexNet's precision and generalisation capability.

All models experience some performance degradation at a higher learning rate of 0.01. With precision, recall, and F1-score at 95.47%, 95.87%, and 95.67%, respectively, CNN+AlexNet still ranks first. The model remains robust even as the learning rate increases; though its performance declines somewhat, it still maintains a strong balance across these criteria. CNN+LSTM and CNN+GRU show more precise cuts. CNN+LSTM's F1-score falls to 94.65%, 94.53%, and 94.56%, respectively, for precision and recall. CNN+GRU also shows a similar trend, with an F1 Score of 94.21%, 95.21% accuracy, and 94.67% recall. This suggests that these models are more sensitive to fluctuations in the learning rate, thereby perhaps reducing ideal performance as the rate increases. The 0.001 learning rate helps all three hybrid models perform better than the 0.001 rate. CNN+AlexNet scored first among models tested at both learning speeds, maintaining extraordinary accuracy, recall, and F1-score. Conversely, the more significant declines in performance for CNN+LSTM and CNN+GRU at 0.01 imply that they are more prone to performance degradation as the learning rate increases. The SGD optimiser works best for these models at a learning rate of 0.001 overall; it primarily benefits CNN+AlexNet, which consistently produces balanced, high-precision predictions across numerous tasks. The comparative performance is shown in Figure 7.

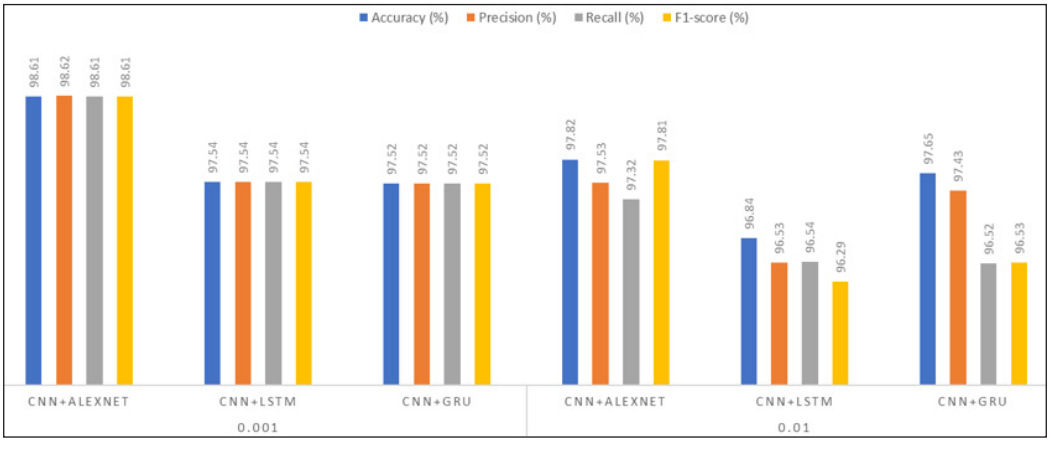


Figure 7. Performance comparison of three models using the SGD optimiser with varying learning rates

Proposed Hybrid Optimiser Performance Analysis

A thorough evaluation was conducted of the suggested hybrid optimiser across various settings to ensure its generalisability and robustness. First, the model's generalisation capacity is examined, with training and testing accuracy as essential benchmarks. This is a critical stage in determining the model's performance on hypothetical data. Precision, recall, and F-score metrics are then used in both datasets to calculate TP, FP, TN, and FN, which serve as indicators of classification performance. These metrics provide a thorough understanding of the model's classification performance, illuminating how well it differentiates between classes across various scenarios. Statistical measures such as Kappa and MCC, in addition to these fundamental classification metrics, are used to further assess the model's performance by analysing the agreement between anticipated and actual classifications. These measures are crucial for providing a more complex view of the model's dependability, mainly when datasets are unbalanced, or accuracy is insufficient to convey the subtleties of performance. Through a thorough evaluation of the model across these varied criteria (generalisation, classification, and statistical agreement), the comprehensive examination ensures that the suggested approach is efficient in terms of raw performance and robust, dependable, and appropriate for practical use. The outcomes are discussed in detail below.

Generalisation Analysis

In Table 8, models using the hybrid optimiser clearly benefit compared with those using single optimisers such as Adam and SGD. First, although the CNN+LSTM model reached an outstanding accuracy of 98.61%, models such as CNN+AlexNet and CNN+GRU fared considerably worse, hovering between 97% and 98% accuracy using the Adam optimiser. Conversely, the hybrid optimiser drove all models to surpass 99% accuracy; CNN+LSTM achieved an impressive 99.57%, thereby demonstrating an apparent gain in overall performance.

Second, the models struggled to attain the same performance levels using the SGD optimiser at different learning rates. While some, like CNN+GRU and CNN+AlexNet, dipped to the 94-95% range, even the best-performing CNN+LSTM model only attained 96.67% accuracy. With model accuracy exceeding 99%, the hybrid optimiser clearly outperformed these findings, demonstrating its ability to improve both precision and recall across all models.

Table 8
Training and testing accuracy for the hybrid optimiser utilised in the UCF-Crime dataset

Model	Accuracy (%)	
	Training	Testing
CNN+AlexNet	99.62	99.24
CNN+GRU	99.72	99.5
CNN+LSTM	99.69	99.57

With the best accuracy (99.57%), CNN+LSTM is the best-performing model in our analysis. Its hybrid optimiser and

deeper design enable it to more successfully capture complicated characteristics than CNN+AlexNet and CNN+GRU, thereby ranking among the best models in our work.

Table 9
Training and testing accuracy for the hybrid optimiser utilised in the HBD21 dataset

Model	Accuracy (%)	
	Training	Testing
CNN+AlexNet	99.21	98.78
CNN+GRU	99.64	98.7
CNN+LSTM	99.6	98.85

The hybrid optimiser's ability to balance its own constraints by combining the features of Adam and SGD optimisers helps explain the observed performance increase when utilising it. SGD is excellent for fine-tuning, but it may be slow to converge; Adam is well-known for its quick convergence, even if it may sometimes overshoot ideal solutions. Higher accuracy, as shown in Table 9, is achieved by models that benefit from both fast initial convergence and exact fine-tuning enabled by the hybrid optimiser. More crucially, the smallest difference between training and testing accuracy (99.69% vs 99.57%) for CNN+LSTM indicates that the models are well-generalised and not overfitted or underfitted, not merely memorising training data but consistently performing on unknown data. This improved generalisation guarantees that models are durable and can consistently manage new data, so the hybrid optimiser is better than Adam and SGD.

For Table 9, over both training and testing sets, the table shows the performance of the three hybrid models—CNN+AlexNet, CNN+GRU, and CNN+LSTM. The high accuracy ratings of these models demonstrate the benefit of using a hybrid optimiser.

CNN+LSTM outperforms CNN+AlexNet and CNN+GRU, achieving an amazing test accuracy of 98.85% when models utilising single optimisers such as Adam or SGD are compared. While all models exceeded 99% accuracy on the training set with the hybrid optimiser, CNN+LSTM models employing a single optimiser, such as Adam, have previously shown great performance in past trials.

This shows the hybrid optimiser's greater generalisation ability, enabling improved performance on both the training and testing sets. Models trained using the hybrid optimiser show a smaller gap between training and testing accuracies, ensuring their generalisation to new data and reducing their tendency to overfit. Adam and SGD benefit from fast convergence during fine-tuning, making the hybrid optimiser an excellent tool for achieving top-notch results across models such as CNN+LSTM. As a result, the proposed CNN+LSTM is the best performer among others. The training and validation accuracy and loss across 50 epochs for both datasets are shown in Figures 8 and 9.

Classification Analysis

This work assessed the classification performance of the hybrid model across multiple criteria, including accuracy, precision, recall, and F1-score, using the hybrid optimiser.

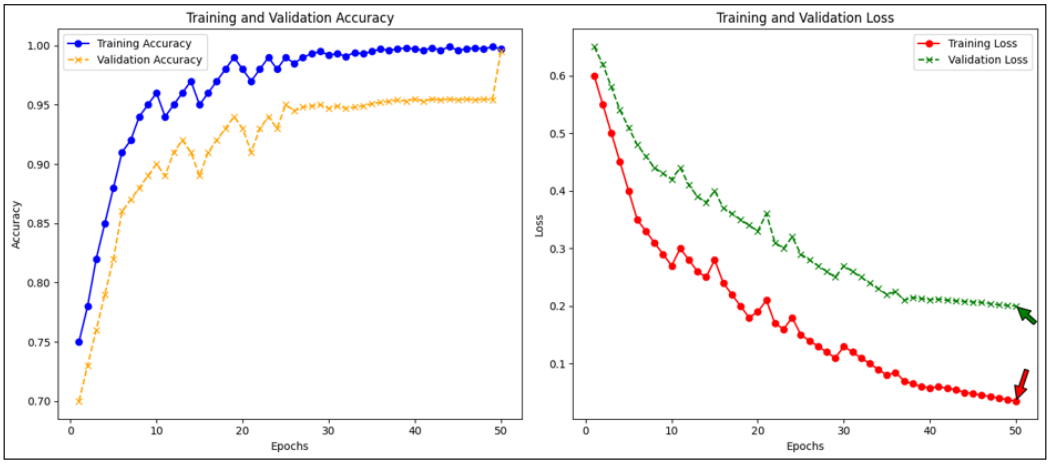


Figure 8. Training and validation accuracy and loss for the UCF-Crime dataset

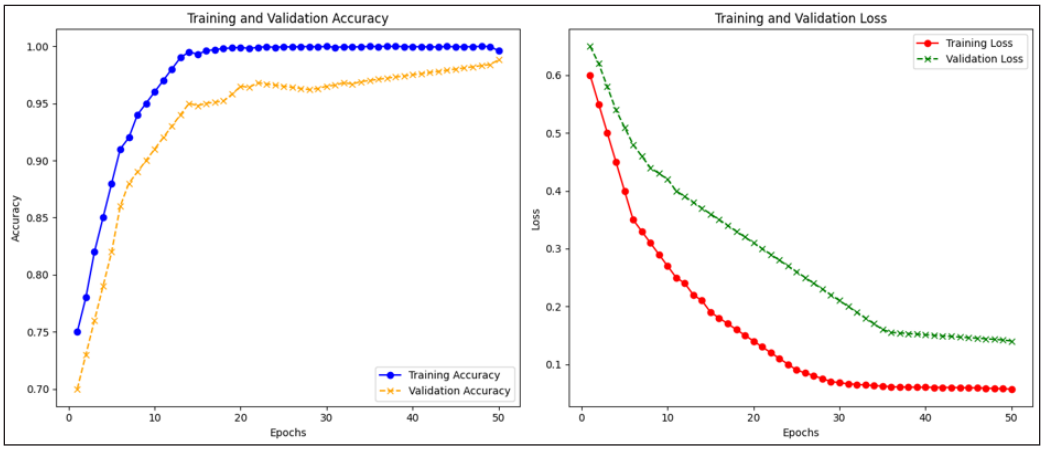


Figure 9. Training and validation accuracy and loss for the HBD21 dataset

Table 10 and Table 11 show the outcomes, which emphasise how effectively the hybrid optimiser increases the models' capacity to identify crime activities.

Evaluating the hybrid optimiser's performance for the UCF-Crime dataset with a remarkable balance between accuracy, recall, and F1-score, CNN+LSTM stands out as the

Table 10
Precision, Recall and F1-score for the hybrid optimiser utilised in the UCF-Crime dataset

Model	Optimiser	Precision (%)	Recall (%)	F1-Score (%)
CNN+AlexNet	Hybrid	99.25	99.24	99.25
CNN+GRU		99.59	99.5	99.5
CNN+LSTM		99.57	99.57	99.57

Table 11
Precision, Recall and F1-score for the Hybrid Optimiser utilised in the HBD21 dataset

Model	Optimiser	Precision (%)	Recall (%)	F1-Score (%)
CNN+AlexNet	Hybrid	98.78	98.78	98.77
CNN+GRU		98.71	98.7	98.71
CNN+LSTM		98.85	98.85	98.85

best performer. This model shows incredible consistency across all metrics: an F1 Score of 99.57%, a precision of 99.57%, and a recall of 99.57%. The near alignment of these values suggests that CNN+LSTM is well optimised for crime detection, handles edge cases effectively, and demonstrates strong robustness to both overfitting and underfitting. This clarifies its standing as the most consistent model for this set. With an F1-score of 99.25%, CNN+AlexNet performs well but shows a slight discrepancy between precision (99.25%) and recall (99.24%). Although these variations are small, they suggest that CNN+AlexNet may sometimes trade off recall—that is, the ability to properly detect genuine positives—in favour of precision, thereby increasing false positives at the expense of false negatives. CNN+GRU achieves the best precision (99.59%) but a somewhat lower recall (99.50%), resulting in an F1-score of 99.50%. This discrepancy between accuracy and recall suggests that although CNN+GRU minimises false positives well, it may occasionally miss real positives, making it somewhat less reliable than CNN+LSTM in preserving overall classification performance. The CNN+LSTM model proves superior in managing the UCF-Crime dataset overall, showing the best generalising and balanced performance across all measures.

Consistent with the results from the UCF-Crime dataset, CNN+LSTM once more ranks highest in the performance study on the HBD21 dataset. With almost perfect alignment across these key criteria, it achieves an F1-score, precision, and recall of 98.85%.

This harmony underpins CNN+LSTM's ability to generalise successfully without substantial trade-offs between false positives and false negatives. CNN+AlexNet follows closely with 98.78% accuracy, recall, and F1-score; it performs well but behind CNN+LSTM in terms of consistency. The slight variance in accuracy and recall here suggests that, while efficient, it may not always handle edge cases as well as CNN+LSTM does. CNN+GRU performs on par with CNN+AlexNet and shows minimal variance, generally with 98.83% precision, recall, and F1-score. Even though these values are somewhat comparable to those of CNN+LSTM, this hybrid model maintains a slight lead by exhibiting tighter alignment and a better balance across the measurements.

In both the UCF-Crime and HBD21 datasets, with minimal variation in precision, recall, and F1-score, CNN+LSTM frequently exhibits improved classification performance. This reveals its robustness in detecting real positives while reducing false positives and false

negatives. CNN+LSTM is the most reliable model for achieving consistent performance across these criteria in challenging classification tasks and crime activity detection. On the other hand, CNN+AlexNet and CNN+GRU show slight differences in recall and accuracy, even though they remain formidable opponents. These variations suggest they might be more likely to trade off among many classification methods, making them less reliable for handling demanding or edge-case scenarios than CNN+LSTM.

Statistical Analysis

According to observations from earlier generalisation and classification reports, the suggested hybrid model—CNN+LSTM with the hybrid optimiser—performed better on both datasets. The training and testing accuracy results showed that this model performed better than others in terms of generalisation and robustness when handling unknown input. The model's agreement between predicted and actual classifications was further evaluated statistically using Kappa and MCC. The CNN+LSTM model achieved nearly perfect scores of 0.99 for both Kappa and MCC on the UCF Crime dataset, as shown in Table 12. This shows a very high degree of agreement and few classification mistakes. This outcome held true for the HBD21 dataset as well, where the model obtained 0.97 for both metrics. While still functioning well, the CNN+GRU and CNN+AlexNet models did not achieve the same agreement, as evidenced by their significantly lower MCC and Kappa values. In summary, the proposed hybrid optimiser-enhanced CNN+LSTM model outperformed the other models in statistical agreement, classification, and generalisation, as demonstrated by the MCC and Kappa assessments. As these results confirm, the CNN+LSTM model is a highly successful method for tasks needing accurate prediction and performance across various circumstances. The resilient, well-generalised model can deliver high classification accuracy across varied datasets.

We performed ROC-AUC analysis on it, since CNN+LSTM was the best-performing model across both the UCF-Crime and HBD21 datasets in previous evaluations. Regarding precision, recall, and F1-score, this model displayed better generalisation capacity and stability. The CNN+LSTM model has remarkable classification abilities across all examined classes, as illustrated by the ROC-AUC curve in Figure 10 for the UCF-Crime dataset. Signifying perfect separation between real positives and false positives across all threshold

Table 12
Kappa and MCC scores for UCF Crime and HBD21 datasets across different models

Model	UCF Crime Dataset		HBD21 Dataset	
	Kappa	MCC	Kappa	MCC
CNN + AlexNet	0.96	0.96	0.97	0.97
CNN + LSTM	0.99	0.99	0.97	0.97
CNN + GRU	0.98	0.98	0.98	0.98

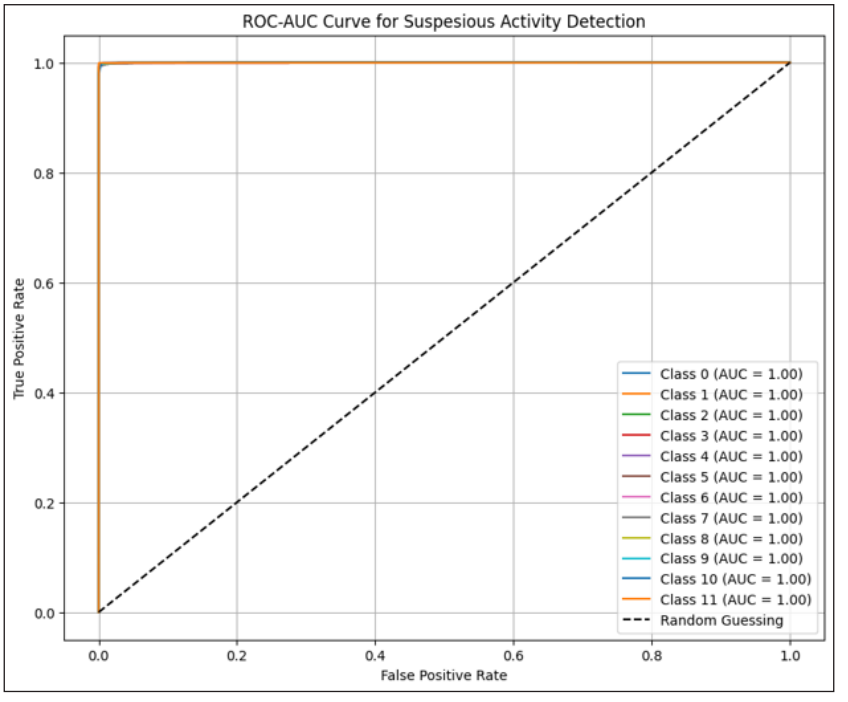


Figure 10. AUC values of all classes by the proposed hybrid optimisation technique (UCF-Crime dataset)

levels, the AUC values for every class consistently reach the ideal value of 1.00. This result underscores how effectively the model detects suspicious behaviour with minimal misclassifications—an essential capability for real-time criminal detection systems. This remarkable performance in all domains reveals the dependability and resilience of the CNN+LSTM model. The excellent balance between sensitivity and specificity, as indicated by the high AUC values, ensures the model accurately identifies a substantial number of actual suspicious behaviours, thereby reducing the likelihood of false positives. This feature is especially crucial in surveillance systems since false positives could lead to unnecessary interventions, while false negatives could cause hidden security problems. Applied to the HBD21 data, the ROC-AUC analysis supports the model's excellent generalisation across several domains. Figure 11 shows the AUC values for each of the twelve classes, which likewise consistently reach the maximum of 1.00, demonstrating the model’s exceptional accuracy in distinguishing various health-related events. This is important considering the complexity of the multi-class classification problem—where each class reflects a different disease or health abnormality. Achieving outstanding AUC scores across all classes demonstrates how well the CNN+LSTM model maintains high classification accuracy even with complex, advanced data. This performance is essential in healthcare applications because precise diagnosis and detection directly influence patient outcomes. The model's

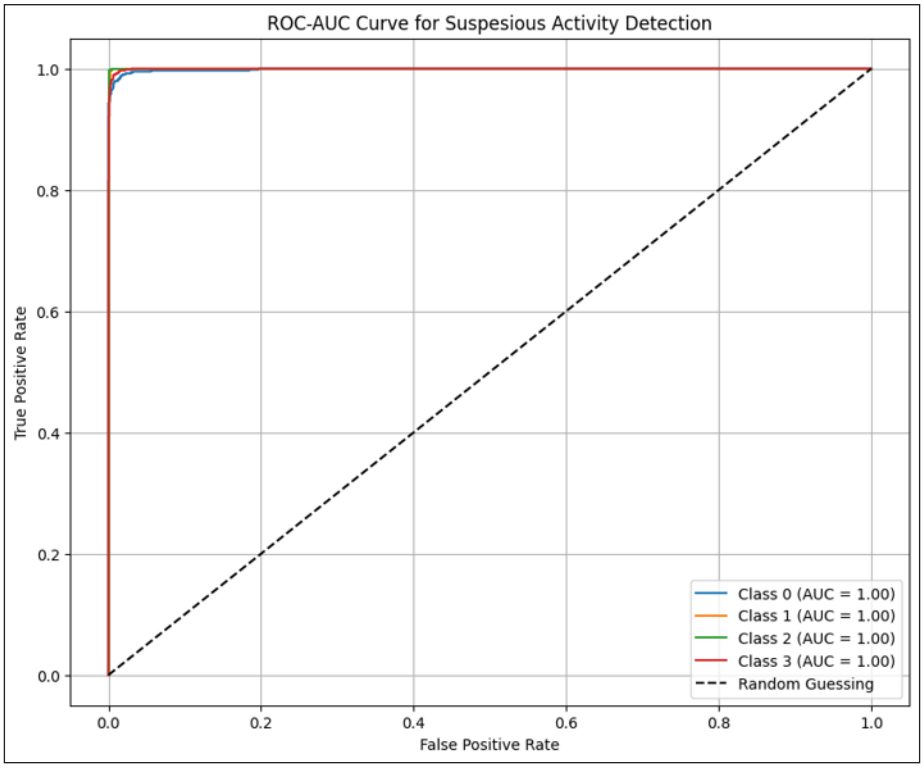


Figure 11. AUC values of all classes by the proposed hybrid optimisation technique (HBD21 dataset)

high sensitivity ensures successful detection of health issues, while its strong specificity reduces the likelihood of erroneous diagnoses, enabling accurate clinical decision-making.

The ROC-AUC analysis was used to evaluate the discriminative ability of the CNN+LSTM model for the individual activity classes. For the multiclass setting, ROC curves were computed in a one-versus-rest manner, and the area under each ROC curve was used as the class-wise AUC. The theoretical range of AUC is from 0 to 1, where a value of 1.0 represents perfect separation between the positive class and the remaining classes under the evaluated data. Therefore, the reported AUC values are interpreted as class-wise discriminative measures and should not be interpreted as evidence of real-world generalisation or deployment performance. Any value exceeding 1.0 would indicate a calculation or reporting error and is therefore not reported.

The confusion matrix for the UCF-Crime dataset, shown in Figure 12, covers a wide range of crime-related activities and provides insight into the CNN+LSTM model's performance. The Abuse class has 3809 test examples; among these, the model accurately identified 3787. Still, it misclassified 22 events—7 as Arrests and another 7 as Assaults. Arson class has 5346 test samples; the model correctly predicted 5302 cases. There are a

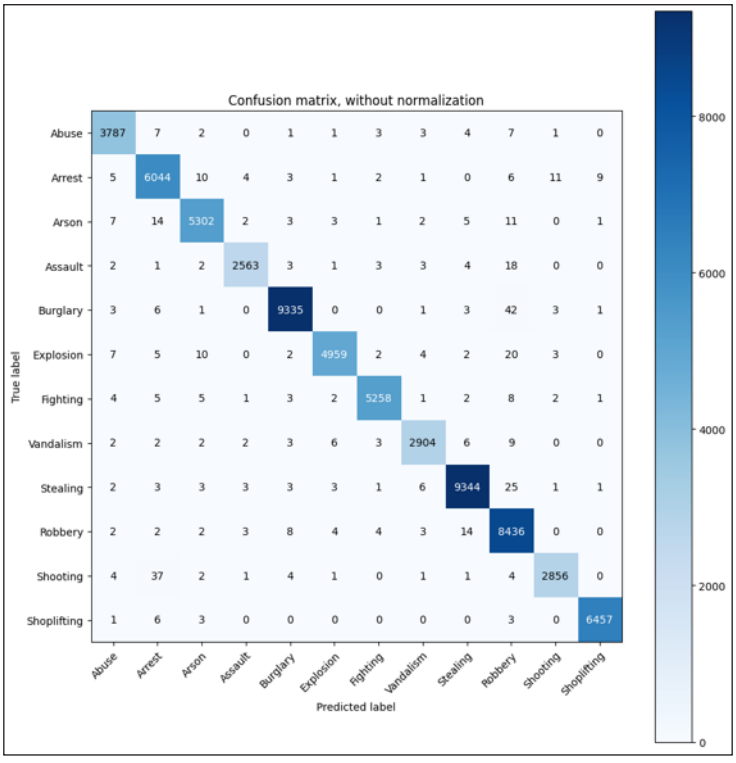


Figure 12. Confusion matrix of the proposed hybrid optimisation technique (UCF-Crime dataset)

few more misclassifications: two Vandalism and eleven Explosions. Having 9335 of 9389 test samples successfully classified, the model performed rather well in the Burglary class. The primary error was the misclassification of forty-two samples as robbery. Comprising 2925 test samples, the model struggled considerably in the Shooting class. It accurately categorised 2856 samples but misclassified 37 cases as Arrest and 14 as Robbery. Still, the model performed remarkably well in the Shoplifting class, precisely classifying all 6457 test instances. Overall, across the classes in the UCF-Crime dataset, the CNN+LSTM model performed relatively well, with an accuracy of almost 99.57%. Separating Shooting from Arrest in Robbery was the main challenge for the model, since dataset visual similarities or potentially overlapping features led to uncertainty.

CNN+LSTM model performance spanning four categories—attack violence, gun violence, everyday acts, and sabotage violence—is shown in the HBD21 dataset confusion matrix in Figure 13. Having 636 test samples, the Assault Violence class model correctly categorised 621 cases. Five were still misclassified as gun violence, one as normal behaviour, and nine as sabotage violence. The model did quite well with 851 observations in the Gun Violence class—predicting exactly 848 incidents. Two samples taken overall were

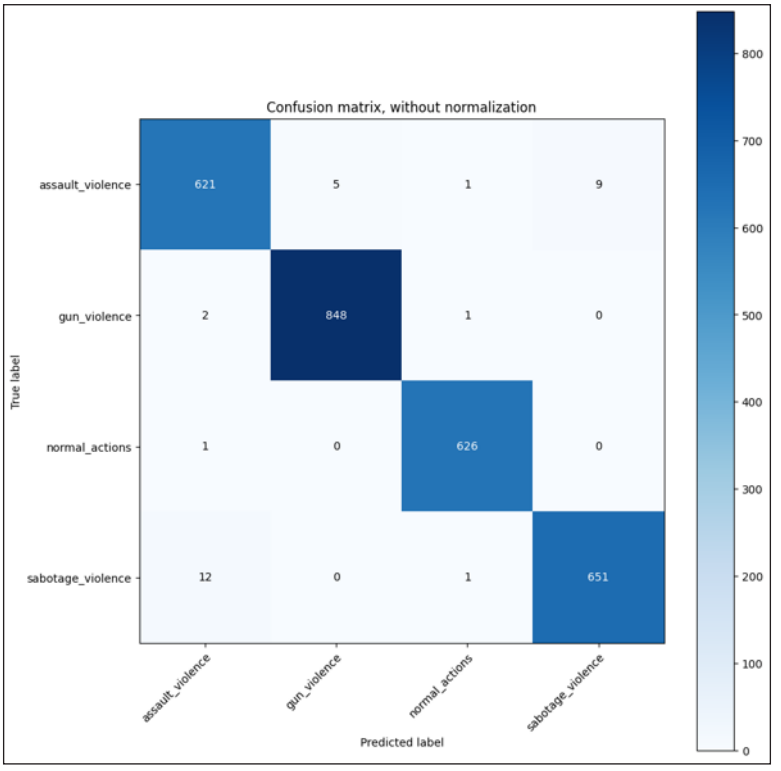


Figure 13. Confusion matrix of the proposed hybrid optimisation technique (HBD21 dataset)

misclassified as assault violence or expected behaviour. With just one incident misclassified as assault violence, the model exhibits great accuracy and correctly groups 626 out of 627 data into the Normal Actions class. For the model, the Sabotage Violence class presented notably more challenges. Of the 664 test samples, 651 were correctly identified; 12 were misclassified as attack violence, and one was misclassified as normal behaviour. On the HBD21 dataset, the accuracy exceeded 98.8%, with 2746 out of 2778 test samples correctly identified by the CNN+LSTM model. The leading cause of misclassification between attack and sabotage revealed possible overlap in attributes the model failed to detect.

The results of the proposed hybrid optimiser are summarised graphically in Figure 14 and Figure 15.

DISCUSSION

The suggested SentryNet concept is a watchful defender, constantly sifting through video footage to spot and address possible threats. In our experimental evaluations on both datasets, LSTM networks outperformed GRU and AlexNet among the RNN architectures tested. The LSTM's improved memory capacity, which enables it to remember information

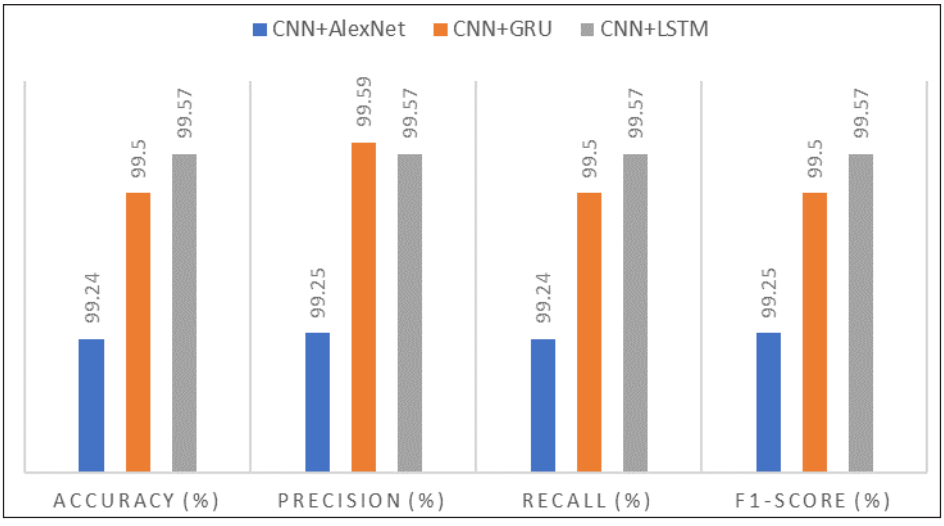


Figure 14. Summarisation of hybrid optimiser outcomes on UCF-Crime dataset

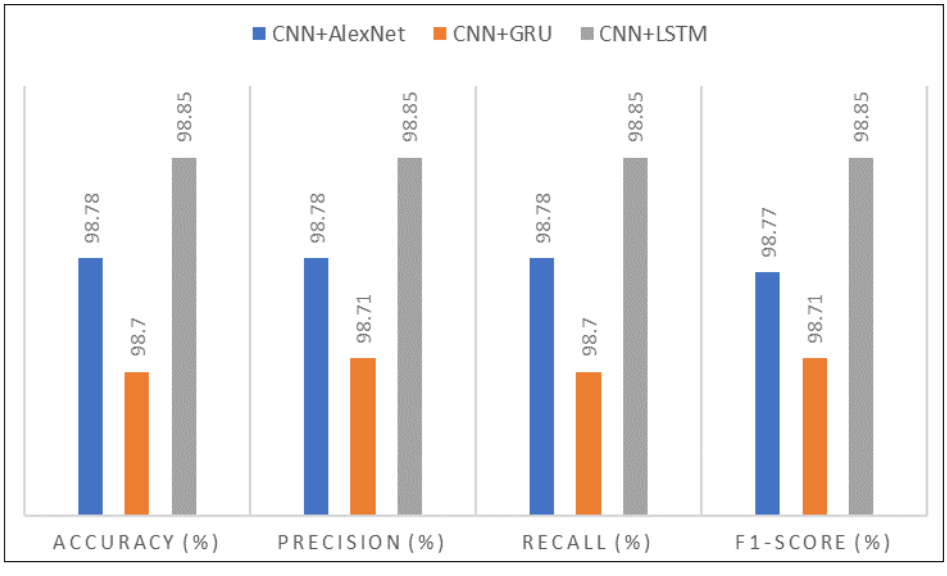


Figure 15. Summarisation of hybrid optimiser outcomes on HBD21 dataset

over longer sequences, is responsible for this better performance. Accurately detecting suspicious actions in surveillance scenarios is dependent on this ability. Furthermore, LSTMs have gating mechanisms that efficiently control information flow, ensuring that essential data are retained. In contrast, less important ones are ignored. They are excellent at capturing long-term dependencies. Additionally, their sturdy construction mitigates the vanishing-gradient issue that standard RNNs often face, enabling more efficient learning

from longer video sequences. Although AlexNet excels at extracting spatial features, it cannot model temporal dynamics, an area in which LSTMs excel. SentryNet is therefore a potent option for assessing complex behaviours in suspicious activity detection, owing to the combination of LSTM's capabilities and the spatial characteristics captured by CNNs. Based on Tables 7-11, our results show that the hybrid optimiser performs significantly better than the Adam and SGD optimisers. Because single optimisers are inherently incapable of managing the intricacies of training DL models, their performance is poor. For example, Adam performs well in the early training phases because of their adaptive learning rate and capacity to handle noisy gradients. However, later in the process, it might have trouble converging, which could lead to less-than-ideal outcomes. On the other hand, while SGD works well for generalisation and stable convergence, it often requires careful adjustment of its momentum and learning rate parameters to prevent it from crashing into local minima or converging too slowly. On the other hand, the hybrid optimiser leverages the advantages of Adam and SGD. The hybrid technique achieves a balanced optimisation strategy by using Adam for its fast convergence during the noisy early period of training, then transitioning to SGD with momentum for stability and fine-tuning. Compared with single optimisers, this synergy improves performance metrics by helping the model overcome training obstacles more effectively.

The hybrid optimiser offers a more flexible and reliable framework, improving model performance on a range of datasets. Reducing the computational complexity of the suggested models without sacrificing their ability to identify suspicious activity is one of our core research goals. We find considerable differences in resource requirements across model configurations, as indicated by parameter counts and training durations. With over 32.4 million parameters, the CNN+AlexNet model for Dataset 1 has the largest parameter count. This results in a significant computational load and a prolonged training time of 106 seconds. On the other hand, the CNN+LSTM and CNN+GRU models have smaller parameter counts, about 3.7 million, resulting in faster training times of about 125 and 128 seconds, respectively. This illustrates how successful learning with fewer parameters is possible when an LSTM or GRU is integrated with a CNN, thereby reducing computational complexity. With a training duration of 156 seconds, the CNN+AlexNet model also has the greatest parameter count in Dataset 2—roughly 62.4 million. With parameter counts of about 17.4 million and training times of 137 and 144 seconds, respectively, the CNN+LSTM and CNN+GRU models, in contrast, remain efficient. By strategically selecting architectures that balance performance and computational demands, our research aims to develop models that are effective in detecting suspicious activities and computationally efficient, facilitating their deployment in real-world scenarios. This is particularly important in applications where resource constraints and real-time processing capabilities are critical, enabling broader accessibility and practicality of DL solutions in surveillance systems.

Thus, the findings highlight the importance of selecting a model architecture to achieve optimal performance while minimising computational complexity. The analysis shows that the proposed CNN+LSTM is more computationally sound and boosts performance. The comparison in Tables 13 and 14 is intended to provide context for previously published results. The proposed model was evaluated using the 80:20 training-testing split described in this study. The experimental protocols, preprocessing procedures, data partitions, model configurations, and evaluation procedures used by the referenced studies were not necessarily identical to those used in the present work. Therefore, the reported accuracy values are compared descriptively and should not be interpreted as results from a strictly controlled, common evaluation protocol.

Table 13
Comparison of proposed model with state-of-the-art approaches on the UCF-Crime dataset

Reference	Model	Accuracy (%)	Optimiser
Kolaib and Waleed, 2024	Deep CNN	99.48	-
Park et al., 2023	Deep neural network with absorbing MC	84.94	Backpropagation
Kumar et al., 2024	CNN+RNN+Attention module	62.04	-
Sultana et al., 2023	2D Deep CNN	98.82	Adam
Taha et al., 2024	ViT B16 + Key Frame Extraction	95.87	SGD
Proposed method	CNN+LSTM	99.57	Hybrid (Adam+SGD)

Table 14
Comparison of proposed model with state-of-the-art approaches on the HBD21dataset

Reference	Model	Accuracy (%)	Optimiser
Kurchaniya and Kumar, 2024	Channel-wise attention with MOG and a pointwise ConvBi-LSTM	95.1	-
Jayaswal and Dixit, 2021a	Xception model+LSTM	97.25	Adam
Proposed model	CNN+LSTM	98.85	Hybrid (Adam+SGD)

Additionally, Table 13 compares the suggested model with cutting-edge models while accounting for the optimisers used in those investigations. Our CNN+LSTM architecture, which achieves greater accuracy and uses a hybrid optimisation technique combining Adam and SGD, is effective, as demonstrated by this comparison. Using a hybrid optimisation strategy that combines Adam and SGD, the proposed CNN+LSTM model achieves a remarkable accuracy of 99.57% in detecting suspicious activity on the UCF-Crime dataset. By contrast, the Deep CNN developed by Kolaib and Waleed (2024) achieved 99.48%

accuracy, although it lacked a robust optimisation plan, which might have limited its flexibility and led to lower evaluation metrics than those reported in the proposed research. Park et al. (2023) used a deep neural network with absorbing Markov Chains and achieved only 84.94%, indicating that the intricacies of activity detection proved too much for it to handle. Notably, there was a performance difference with Kumar et al. (2024), who obtained much lower accuracies of 49.04% and 62.04% with their respective models. Our model outperformed Sultana et al. (2023), achieving 98.82% using a 2D Deep CNN with Adam, by successfully integrating CNN and LSTM architectures to leverage both temporal and spatial information. Finally, Taha et al. (2024) used a Vision Transformer (ViT B16) with keyframe extraction and achieved 95.87%, but our hybrid optimisation method offers a more balanced and robust solution. Overall, the CNN+LSTM model achieved the highest reported testing accuracy among the evaluated hybrid models on both datasets. On UCF-Crime, it achieved 99.57% testing accuracy, while on HBD21 it achieved 98.85%. The corresponding precision, recall, and F1-score values were also closely aligned with the reported accuracy values. These results indicate consistent classification performance under the experimental settings used in this study.

As shown in Table 14, we evaluated the proposed model against cutting-edge methods on the HBD21 and UCF-Crime datasets. Our CNN+LSTM model, which used a hybrid optimisation technique combining Adam and SGD, obtained an outstanding accuracy of 98.85%. This outperforms the Xception model with LSTM by Jayaswal and Dixit (2021a) (97.25%) and the Channel-wise Attention with MOG and ConvBi-LSTM model by Kurchaniya and Kumar (2024) (95.10%). The results show that our proposed model performs very well in terms of accuracy and demonstrates the hybrid optimiser's efficacy, enhancing its resilience when handling complex surveillance tasks across various datasets. The comparison examination highlights the CNN+LSTM architecture's superior performance.

CONCLUSION

The SentryNet framework successfully meets the urgent need for sophisticated automated systems in the cybersecurity and surveillance domains to identify suspicious activity. Our method combines the advantages of both architectures—LSTM networks and CNNs—to ensure reliable extraction of spatial and temporal features. The long-term memory capacity and efficient gating mechanisms of the LSTM model have contributed to its higher performance compared to other standard recurrent architectures, such as GRUs and AlexNet. Our empirical results demonstrate that SentryNet achieves remarkable accuracy rates of 98.85% on the HBD21 dataset and 99.57% on the UCF-Crime dataset, outperforming established models in the literature. Additionally, the proposed hybrid optimisation strategy, combining Adam and SGD, effectively mitigates the limitations

of each optimiser, thereby enhancing convergence rates and model stability. Overall, the suggested architecture is a viable option for real-time suspicious activity detection since it minimises computational cost, improves model generalisation, and lowers biases. Future studies could combine cutting-edge technologies such as attention mechanisms, real-time data collection, and dynamic optimisation procedures to further improve the model's performance and scalability. These modifications will increase the model's real-time applicability and robustness, enabling its adaptation to different surroundings. Moreover, resolving constraints, including computational limitations and dataset biases, more effectively would improve the model's generalisability and scalability. This study intends to advance the field of automated crime detection by carefully navigating these advancements, enabling more accurate and efficient surveillance systems and, in turn, enhancing public safety and security. The present study does not include dedicated measurements of end-to-end inference latency or frames per second; therefore, real-time deployment performance cannot be established solely from the reported training and classification results. Future work should evaluate inference speed, memory consumption, and deployment performance on representative surveillance hardware.

ACKNOWLEDGEMENT

The authors would like to extend their sincere thanks to their respective organisations for providing an academic environment and other requirements for conducting research. The authors also thank HBD21 and UCF-Crime datasets which were used for experimentation purposes as a base for proposed SentryNet system development. Moreover, the authors would also like to thank those researchers whose works are very valuable in the areas of video surveillance, deep learning, and hybrid optimisation.

REFERENCES

Ahmed, W., & Yousaf, M. H. (2024). A deep autoencoder-based approach for suspicious action recognition in surveillance videos. *Arabian Journal for Science and Engineering, 49*(3), 3517-3532. https://doi.org/10.1007/s13369-023-08038-7

Ali, M. M. (2023). Real-time video anomaly detection for smart surveillance. *IET Image Processing*, *17*(5), 1375-1388.

Al-Lahham, A., Zaheer, M. Z., Tastan, N., & Nandakumar, K. (2024). Collaborative learning of anomalies with privacy (CLAP) for unsupervised video anomaly detection: A new baseline. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12416-12425). IEEE. https://doi.org/10.1109/cvpr52733.2024.01180

Brownlee, J. (2022, August 15). *Difference between a batch and an epoch in a neural network*. Machine Learning Mastery. https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/

Cesario, E., Lindia, P., & Vinci, A. (2024). Multi-density crime predictor: An approach to forecast criminal activities in multi-density crime hotspots. *Journal of Big Data, 11*(1), Article 75. https://doi.org/10.1186/s40537-024-00935-4

Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724-1734). Association for Computational Linguistics. https://doi.org/10.3115/v1/d14-1179

Ghosh, P., Bose, S., Roy, S., & Mondal, A. (2022). Detection of abnormal human behavior using deep learning. *BOHR International Journal of Smart Computing and Information Technology, 3*(1), 59-68. https://journals.bohrpub.com/index.php/bijscit/article/view/302

Gowada, R., Pawar, D., & Barman, B. (2023). Unethical human action recognition using deep learning based hybrid model for video forensics. *Multimedia Tools and Applications, 82*(19), 28713-28738. https://doi.org/10.1007/s11042-023-14508-9

Jan, A., & Khan, G. M. (2023). Real world anomalous scene detection and classification using multilayer deep neural networks. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(2), 158-167. https://doi.org/10.9781/ijimai.2021.10.010

Jayaswal, R., & Dixit, M. (2021a). A framework for anomaly classification using deep transfer learning approach. *Revue d'Intelligence Artificielle, 35*(3), 255-263. https://doi.org/10.18280/ria.350309

Jayaswal, R., & Dixit, M. (2021b). *Human Behavior dataset (HBD21)* (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/xh9pgb3w8c.1

Jebur, S. A., Alzubaidi, L., Saihood, A., Hussein, K. A., Hoomod, H. K., & Gu, Y. (2025). A scalable and generalised deep learning framework for anomaly detection in surveillance videos. *International Journal of Intelligent Systems, 2025*(1), Article 1947582. https://doi.org/10.1155/int/1947582

Jebur, S. A., Hussein, K. A., & Hoomod, H. K. (2023). Abnormal behavior detection in video surveillance using Inception-v3 transfer learning approaches. *Iraqi Journal of Computers, Communications, Control and Systems Engineering, 23*(2), 210-221. https://doi.org/10.33103/uot.ijccce.23.2.16

Khan, N., Ullah, F. U. M., Haq, I. U., Khan, S. U., Lee, M. Y., & Baik, S. W. (2021). AB-Net: A novel deep learning assisted framework for renewable energy generation forecasting. *Mathematics, 9*(19), Article 2456. https://doi.org/10.3390/math9192456

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1412.6980

Kirichenko, L., Radivilova, T., Sydorenko, B., & Yakovlev, S. (2022). Detection of shoplifting on video using a hybrid network. *Computation, 10*(11), Article 199. https://doi.org/10.3390/computation10110199

Kolaib, R. J., & Waleed, J. (2024). Crime activity detection in surveillance videos based on developed deep learning approach. *Diyala Journal of Engineering Sciences, 17*(3), 98-114. https://doi.org/10.24237/djes.2024.17307

Kumar, M., & Biswas, M. (2024). Abnormal human activity detection by convolutional recurrent neural network using fuzzy logic. *Multimedia Tools and Applications, 83*(22), 61843-61859. https://doi.org/10.1007/s11042-023-15904-x

Kumar, M., Patel, A. K., & Biswas, M. (2024). Real-time detection of abnormal human activity using deep learning and temporal attention mechanism in video surveillance. *Multimedia Tools and Applications, 83*(18), 55981-55997. https://doi.org/10.1007/s11042-023-17748-x

Kumar, M., Patel, A. K., Biswas, M., & Shitharth, S. (2023). Attention-based bidirectional-long short-term memory for abnormal human activity detection. *Scientific Reports, 13*(1), Article 14442. https://doi.org/10.1038/s41598-023-41231-0

Kurchaniya, D., & Kumar, S. (2023). Two stream deep neural network based framework to detect abnormal human activities. *Journal of Electronic Imaging, 32*(4), Article 043021. https://doi.org/10.1117/1.jei.32.4.043021

Kurchaniya, D., & Kumar, S. (2024). D-SCAN: Dual stream spatiotemporal channel-wise attention network with point-wise ConvBi-LSTM for activity recognition. *IEEE Transactions on Consumer Electronics, 70*(4), 7244-7251. https://doi.org/10.1109/tce.2024.3425610

Li, Y., Wei, C., & Ma, T. (2019). Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems, 32*. https://papers.neurips.cc/paper_files/paper/2019/hash/bce9abf229ffd7e570818476ee5d7dde-Abstract.html

Liu, Y., Pu, H., & Sun, D.-W. (2021). Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices. *Trends in Food Science & Technology, 113*, 193-204. https://doi.org/10.1016/j.tifs.2021.04.042

Mahalakshmi, P., & Thenmalar, S. (2023). Feature based training for crime detection using deep learning techniques. In *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)* (pp. 1-4). IEEE. https://doi.org/10.1109/rmkmate59243.2023.10369315

Mahareek, E. A., ElSayed, E. K., ElDesouky, N. M., & ElDahshan, K. A. (2024). Detecting anomalies in security cameras with 3d-convolutional neural network and convolutional long short-term memory. *International Journal of Electrical and Computer Engineering, 14*(1), 993-1004. https://doi.org/10.11591/ijece.v14i1.pp993-1004

Mehsen, R. S. (2023). Deep learning algorithm for detecting and analyzing criminal activity. *International Journal of Computing, 22*(2), 248-253. https://doi.org/10.47839/ijc.22.2.3095

Mukto, M. M., Hasan, M., Al Mahmud, M. M., Haque, I., Ahmed, M. A., Jabid, T., Ali, M. S., Rashid, M. R. A., Islam, M. M., & Islam, M. (2024). Design of a real-time crime monitoring system using deep learning techniques. *Intelligent Systems with Applications, 21*, Article 200311. https://doi.org/10.1016/j.iswa.2023.200311

Mumtaz, A., Sargano, A. B., & Habib, Z. (2024). AnomalyNet: A spatiotemporal motion-aware CNN approach for detecting anomalies in real-world autonomous surveillance. *The Visual Computer, 40*(11), 7823-7844. https://doi.org/10.1007/s00371-023-03210-4

Nayak, R., & Chaudhari, J. P. (2023). Real-world anomaly detection in video using spatio-temporal features analysis for weakly labelled data with auto label generation. *International journal of electrical and computer engineering systems, 14*(5), 565-573. https://doi.org/10.32985/ijeces.14.5.8

Nazir, A., Mitra, R., Sulieman, H., & Kamalov, F. (2023). Suspicious behavior detection with temporal feature extraction and time-series classification for shoplifting crime prevention. *Sensors, 23*(13), Article 5811. https://doi.org/10.3390/s23135811

Pallewar, M. G., Pawar, V. R., & Gaikwad, A. N. (2024). Human anomalous activity detection with CNN-LSTM approach. *Journal of Integrated Science and Technology, 12*(1), Article 704. https://pubs.thesciencein.org/journal/index.php/jist/article/view/a704

Park, J., Kim, J., & Han, B. (2023). End-to-end learning for weakly supervised video anomaly detection using absorbing markov chain. *Computer Vision and Image Understanding, 236*, Article 103798. https://doi.org/10.1016/j.cviu.2023.103798

Patwal, A., Diwakar, M., Tripathi, V., & Singh, P. (2023). An investigation of videos for abnormal behavior detection. *Procedia Computer Science, 218*, 2264-2272. https://doi.org/10.1016/j.procs.2023.01.202

Qasim, M., & Verdú, E. (2023). Video anomaly detection system using deep convolutional and recurrent models. *Results in Engineering, 18*, Article 101026. https://doi.org/10.1016/j.rineng.2023.101026

Qasim Gandapur, M., & Verdú, E. (2023). ConvGRU-CNN: Spatiotemporal deep learning for real-world anomaly detection in video surveillance system. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(4), 88-95. https://doi.org/10.9781/ijimai.2023.05.006

Şengönül, E., Samet, R., Abu Al-Haija, Q., Alqahtani, A., Alturki, B., & Alsulami, A. A. (2023). An analysis of artificial intelligence techniques in surveillance video anomaly detection: A comprehensive survey. *Applied Sciences, 13*(8), Article 4956. https://doi.org/10.3390/app13084956

Sharma, S., Sharma, S., & Athaiya, A. (2020). Activation functions in neural networks. *International Journal of Engineering Applied Sciences and Technology, 4*(12), 310-316. https://doi.org/10.33564/ijeast.2020.v04i12.054

Stewart, T. E., Sr. (2022). *Strategies cybersecurity operators need to improve situation awareness during continuous monitoring performance* [Doctoral dissertation, Colorado Technical University]. ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/strategies-cybersecurity-operators-need-improve/docview/2805415237/se-2

Sultana, R., Das, P., Zaman, R., Mazumdar, P., Tasneem, S., Karmakar, A., & Rokunuzzaman, M. (2023). *Intelligent vision-based human activity recognition for surveillance using deep learning algorithm* [Conference paper]. 7th International Conference on Mechanical Engineering and Renewable Energy (ICMERE2023), Chattogram, Bangladesh. https://math.ruet.ac.bd/prangon/bibliography_text

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6479-6488). IEEE. https://doi.org/10.1109/CVPR.2018.00678

Taha, R. A., Youssif, A. A.-H., & Fouad, M. M. (2024). Transfer learning model for anomalous event recognition in Big Video Data. *Scientific Reports, 14*(1), Article 27868. https://doi.org/10.1038/s41598-024-78414-2

Tan, W., Yao, Q., & Liu, J. (2024). Overlooked video classification in weakly supervised video anomaly detection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (pp. 202-210). IEEE. https://doi.org/10.1109/wacvw60836.2024.00029

Taye, M. M. (2023). Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. *Computers, 12*(5), Article 91. https://doi.org/10.3390/computers12050091

Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology, 18*(11), 1473-1488. https://doi.org/10.1109/tcsvt.2008.2005594

Ul Amin, S., Kim, Y., Sami, I., Park, S., & Seo, S. (2023). An efficient attention-based strategy for anomaly detection in surveillance video. *Computer Systems Science and Engineering, 46*(3), 3939-3958. https://doi.org/10.32604/csse.2023.034805

Ullah, F. U. M., Khan, N., Hussain, T., Lee, M. Y., & Baik, S. W. (2021). Diving deep into short-term electricity load forecasting: Comparative analysis and a novel framework. *Mathematics, 9*(6), Article 611. https://doi.org/10.3390/math9060611

Ullah, W., Ullah, A., Hussain, T., Muhammad, K., Heidari, A. A., Del Ser, J., Baik, S. W., & De Albuquerque, V. H. C. (2022). Artificial intelligence of things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data. *Future Generation Computer Systems, 129*, 286-297. https://doi.org/10.1016/j.future.2021.10.033

Vosta, S., & Yow, K.-C. (2022). A CNN-RNN combined structure for real-world violence detection in surveillance cameras. *Applied Sciences, 12*(3), Article 1021. https://doi.org/10.3390/app12031021

Vosta, S., & Yow, K.-C. (2024). KianNet: A violence detection model using an attention-based CNN-LSTM structure. *IEEE Access, 12*, 2198-2209. https://doi.org/10.1109/access.2023.3339379

Wu, P., Pan, C., Yan, Y., Pang, G., Yan, Q., Wang, P., & Zhang, Y. (2026). Deep learning for video anomaly detection: A review. *IEEE Transactions on Neural Networks and Learning Systems, 37*(7), 3010-3030. https://doi.org/10.1109/tnnls.2025.3647892

Yadav, R., Halder, R., & Banda, G. (2024). Masked autoencoders for spatial–temporal relationship in video-based group activity recognition. *IEEE Access, 12*, 132084-132095. https://doi.org/10.1109/access.2024.3457024

You, K., Long, M., Wang, J., & Jordan, M. I. (2019). *How does learning rate decay help modern neural networks?* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1908.01878

REFEREES FOR THE PERTANIKA

JOURNAL OF SCIENCE & TECHNOLOGY

Vol. 34 (S3) 2026

The Editorial Board of the Pertanika Journal of Science and Technology wishes to thank the following:

Komal Kumar N
(SIMATS, India)

Radha Krishna Reddy
(MAHE, India)

Sathiyaraj R.
(MAHE, India)

Sivaram Rajeyyagari
(SU, Saudi Arabia)

Sreenivasa Chakravarthi Sangapu
(AMRITA, India)

K. Suneetha
(JDU, India)

Thulasi Bikku
(AMRITA, India)

Viswanadha Raju
(JNTUH, India)

AMRITA - *Amrita Vishwa Vidyapeetham*
JDU - *Jain (Deemed-to-be University)*
JNTUH - *Jawaharlal Nehru Technological University Hyderabad*
MAHE - *Manipal Academy of Higher Education*
SIMATS - *Saveetha Institute of Medical and Technical Sciences*
SU - *Shaqra University*

While every effort has been made to include a complete list of referees for the period stated above, however if any name(s) have been omitted unintentionally or spelt incorrectly, please notify the Chief Executive Editor, *Pertanika* Journals at executive_editor.pertanika@upm.edu.my

Any inclusion or exclusion of name(s) on this page does not commit the *Pertanika* Editorial Office, nor the UPM Press or the university to provide any liability for whatsoever reason.

An Optimised Deep Learning Framework for Skin Lesion Classification Using ResNet150 and M-GSO 165
Gurram Sunitha, Ahmed Elngar, Pravin R. Kshirsagar, and Subba Rao Polamuri

Enhancing Automotive Safety and Operational Efficiency Through Controller Area Network (CAN) and Internet of Things (IoT) Integration 193
Arava Ramprakash Reddy and Suresh Kallam

SentryNet: Fusion of CNN-LSTM Architecture for Robust Suspicious Activity Detection in Video Surveillance Using Novel Hybrid Optimisation 209
Roopa R and Humera Khanam M

Pertanika Journal of Science & Technology
Vol. 34 (S3) 2026

Content

Preface xi
I. Siva, K. Reddy Madhavi, and Nagaraj Ramrao

Swin Transfer-Based Auxiliary Classifier Generative Adversarial Network for Colorectal Cancer Analysis 1
G. Kanishka and K. Reddy Madhavi

The Sky Emotional Chatbot Companion: Proactive AI Intervention Fostering Emotional Connectivity to the Mind 27
Baby Vadlana, Karnam Akhil, Tamma Bhuwan Vamsi, and Tejasree Lokireddy

Detecting Fake Reviews on Online Products with Machine Learning 55
P. Dhanalakshmi, Ummadi Janardhan Reddy, Ravikanth Garladinne, S. Ramachandra, E. R. Aruna, and Sharath Chandra Parashara

Deep Learning and Signal Processing Synergy: Multitask Deep Neural Networks (MLT-DNNs) and Second-Order Blind Identification (SOBI) for Cybersickness Detection 71
Bisani Leela Padmavathi and J. Avanija

GenCKD-Net: A Lightweight Multilayer Perceptron Pipeline for Chronic Kidney Disease Classification Using the UCI CKD Dataset 101
Karimulla Sha Shaik, Konda Hari Krishna, and K. K. Baseer

Leveraging Business Analytics for Operational Efficiency in Small-Scale Manufacturing Enterprises in Tamil Nadu 125
Geetha N and U. M. Gopal Krishna

Comparison of Deep Learning Models for Classification of Acute Lymphoblastic Leukaemia 149
Beebi Naseeba, Pranavesh Kumar Talupuri, Nitin Bhimineni, V. Jyothsna, Sharath Chandra Parashara, and Puvi Rajan

Pertanika Editorial Office, Journal Division,
Putra Science Park,
1st Floor, IDEA Tower II,
UPM-MTDC Center,
Universiti Putra Malaysia,
43400 UPM Serdang,
Selangor Darul Ehsan
Malaysia

http://www.pertanika.upm.edu.my
Email: executive_editor.pertanika@upm.edu.my
Tel. No.: +603- 9769 1622

http://penerbit.upm.edu.my
Email: penerbit@upm.edu.my
Tel. No.: +603- 9769 3606

e-ISSN 2231-8526