frontiers | Research Topics

Integrating AI and machine learning in advancing patient care: bridging innovations in mental health and cognitive neuroscience

Edited by
Salil Bharany, Habib Hamam, SeongKi Kim and Ateeq Ur Rehman

Published in
Frontiers in Medicine

ISSN 1664-8714
ISBN 978-2-8325-6973-3
DOI 10.3389/978-2-8325-6973-3

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Integrating AI and machine learning in advancing patient care: bridging innovations in mental health and cognitive neuroscience

Topic editors

Salil Bharany — Chitkara University, India

Habib Hamam — Université de Moncton, Canada

SeongKi Kim — Chosun University, Republic of Korea

Ateeq Ur Rehman — Gachon University, Republic of Korea

Citation

Bharany, S., Hamam, H., Kim, S., Rehman, A. U., eds. (2025). *Integrating AI and machine learning in advancing patient care: bridging innovations in mental health and cognitive neuroscience*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-6973-3

Table of contents

frontiers | Frontiers in Medicine

TYPE Editorial
PUBLISHED 29 September 2025
DOI 10.3389/fmed.2025.1704357

OPEN ACCESS

EDITED AND REVIEWED BY
Alice Chen,
Consultant, Potomac, MD, United States

*CORRESPONDENCE
Salil Bharany
✉ salil.bharany@gmail.com
Ateeq Ur Rehman
✉ 202411144@gachon.ac.kr

RECEIVED 12 September 2025
ACCEPTED 16 September 2025
PUBLISHED 29 September 2025

CITATION
Bharany S, Hamam H, Kim S and Rehman AU (2025) Editorial: Integrating AI and machine learning in advancing patient care: bridging innovations in mental health and cognitive neuroscience. *Front. Med.* 12:1704357. doi: 10.3389/fmed.2025.1704357

Editorial: Integrating AI and machine learning in advancing patient care: bridging innovations in mental health and cognitive neuroscience

Salil Bharany[1]*, Habib Hamam[2], SeongKi Kim[3] and Ateeq Ur Rehman[4]*

[1]Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India, [2]Faculty of Engineering, University de Moncton, Moncton, NB, Canada, [3]Department of Computer Engineering, Chosun University, Gwangju, Republic of Korea, [4]School of Computing, Gachon University, Seongnam-si, Republic of Korea

KEYWORDS

artificial intelligence, machine learning, deep learning, mental health, cognitive neuroscience, medical imaging, explainable AI, telemedicine

Editorial on the Research Topic
Integrating AI and machine learning in advancing patient care: bridging innovations in mental health and cognitive neuroscience

The overarching goal of this Research Topic is to highlight the transformative potential of artificial intelligence (AI) and machine learning (ML) in enhancing patient care, with a particular focus on mental health and cognitive neuroscience. This Research Topic bridges technological innovations with clinical practice, highlighting state-of-the-art AI and ML models, exploring novel approaches for early detection and monitoring of neurological disorders, emphasizing explainability and trustworthiness in clinical AI, assessing the role of secure infrastructures such as telemedicine and 6G-enabled hospitals, addressing ethical and adversarial concerns, and fostering interdisciplinary collaboration to advance patient-centered healthcare innovation.

The following articles exemplify the diverse applications of AI and ML in healthcare, showcasing innovative approaches that enhance diagnostic accuracy, patient monitoring, and secure clinical practices across various specialties, including mental health, neurology, cardiology, and developmental disorders.

Zhang and Zeng introduced a deep learning-driven image classification model to support mental health diagnostics, addressing the limitations of subjective clinical assessments. By extracting subtle imaging biomarkers from patient data, the model improved diagnostic accuracy and consistency. This approach not only enables earlier detection of psychiatric disorders but also lays the foundation for more personalized treatment strategies. Its impact lies in bridging AI innovations with the urgent needs of mental health care systems.

Shehab and Alhaddad proposed an LSTM-CNN fusion framework for medical image steganalysis, targeting secure telemedicine applications. Their model effectively identified hidden data embedded in medical images, strengthening protection against malicious data tampering. This dual focus on deep learning and cybersecurity ensures trust in digital health platforms. The work is impactful in enabling safe, privacy-preserving telemedicine services as healthcare shifts toward remote and digital care.

Mozhegova et al. evaluated how multimodal AI systems in medicine respond to adversarial perturbations across different input channels. The study revealed key vulnerabilities that could compromise diagnostic integrity, while also offering insights into strategies for resilience. By highlighting the fragility of advanced medical AI under adversarial stress, this work underscores the importance of deploying robust, trustworthy, and secure clinical AI. It sets the stage for developing next-generation defenses against adversarial threats in healthcare.

Ikram et al. harnessed transformer architectures to model sequential ECG signals for arrhythmia detection. Their system outperformed conventional deep learning approaches by effectively capturing long-range dependencies in cardiac patterns. The study demonstrated high diagnostic accuracy, enabling earlier identification of arrhythmias with the potential to prevent severe cardiac events. This represents a major advancement for AI-based preventive cardiology.

Al-Nefaie et al. developed an AI-based diagnostic framework for Autism Spectrum Disorder (ASD), focusing on early and reliable detection. The system integrated multimodal data sources to capture the complex behavioral and neurological patterns associated with ASD. By improving diagnostic speed and reducing reliance on subjective evaluations, the model enhances support for patients and families. This Research Topic highlights the increasing role of AI in addressing neurodevelopmental conditions with significant global health implications.

Together, these articles highlight the practical applications of AI and ML in enhancing patient care. They reveal novel methodologies and intelligent frameworks that improve clinical decision-making, treatment planning, and monitoring across neurological, psychiatric, and other medical domains. By highlighting ethical safeguards, resilience, and secure infrastructures, the collection points to pathways for safe, scalable, and patient-centered healthcare solutions. Overall, the Research Topic illustrates the critical role of interdisciplinary collaboration in translating AI innovations into effective and reliable clinical practice.

Author contributions

SB: Conceptualization, Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. HH: Conceptualization, Data curation, Project administration, Supervision, Writing – original draft, Writing – review & editing. SK: Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. AR: Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing.

Acknowledgments

The editors would like to thank the authors, reviewers, and the Frontiers in medicine development team, whose efforts have led to the success of this Research Topic.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 27 March 2025
DOI 10.3389/fmed.2025.1545528

OPEN ACCESS

EDITED BY
SeongKi Kim,
Chosun University, Republic of Korea

REVIEWED BY
M. Shahid Anwar,
Gachon University, Republic of Korea
Rui Wang,
The First Affiliated Hospital of Xi'an Jiaotong University, China

*CORRESPONDENCE
Aitizaz Ali
✉ aitizaz.ali@apu.edu.my

RECEIVED 15 December 2024
ACCEPTED 11 March 2025
PUBLISHED 27 March 2025

CITATION
Ullah R, Sarwar N, Alatawi MN, Alsadhan AA, Salamah Alwageed H, Khan M and Ali A (2025) Advancing personalized diagnosis and treatment using deep learning architecture.
Front. Med. 12:1545528.
doi: 10.3389/fmed.2025.1545528

Advancing personalized diagnosis and treatment using deep learning architecture

Rahat Ullah[1], Nadeem Sarwar[2], Mohammed Naif Alatawi[3], Abeer Abdullah Alsadhan[4], Hathal Salamah Alwageed[5], Maqbool Khan[6] and Aitizaz Ali[7]*

[1]School of Physics and Optoelectronics, Nanjing University of Information Science and Technology, Nanjing, China, [2]Department of Computer Science, Bahria University Lahore Campus, Lahore, Pakistan, [3]Information Technology Department, Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Saudi Arabia, [4]Computer Science Department, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, [5]College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia, [6]Pak-Austra Fachhochschule, Institute of Applied Sciences and Technology (PAF-IAST), Haripur, Pakistan, [7]School of Technology, NSF Group Asia Pacific University, Kuala Lumpur, Malaysia

Autoimmune disorders (AID) present significant challenges due to their complex etiologies and diverse clinical manifestations. Traditional diagnostic methods, which rely on symptom observation and biomarker detection, often lack specificity and fail to provide personalized treatment options. This study proposes ImmunoNet, a deep learning-based framework that integrates genetic, molecular, and clinical data to enhance the accuracy of autoimmune disease diagnosis and treatment. ImmunoNet leverages convolutional neural networks (CNNs) and multi-layer perceptrons (MLPs) to analyze large-scale datasets, enabling precise disease classification and personalized therapeutic treatment recommendations. The model improves interpretability through explainable AI techniques and enhances privacy via federated learning. Comparative evaluations demonstrate that ImmunoNet outperforms traditional machine learning models, achieving a 98% accuracy rate in predicting autoimmune disorders. By advancing precision medicine in immunology, this approach provides clinicians with a powerful tool for personalized diagnosis and optimized therapeutic strategies.

KEYWORDS

deep learning, autoimmune disorder, ensemble learning, CNN, MLP

1 Introduction

Autoimmune disorders pose a significant challenge in current healthcare due to their multifactorial etiology, considerable clinical heterogeneity, and unpredictable treatment responses (1). Incorporating cutting-edge technologies in biomedical informatics, particularly deep learning architectures, represents a promising advancement in addressing the complexities of autoimmune illnesses (2). Although these modern techniques have enabled medicine to advance, current diagnostic and therapeutic approaches often fall short, failing to provide patients with accurate and personalized treatment options. Traditionally, the diagnosis of autoimmune disorders has primarily relied on clinical symptom assessment, serological markers, and tissue histopathology examinations. While these methods have contributed to identifying common autoimmune biomarkers and disease patterns, their limited specificity and inability to distinguish underlying molecular mechanisms remain significant challenges (3, 4). Traditional therapies for autoimmune disorders exhibit varying efficacy and can have adverse effects, particularly on susceptible individuals exposed to these medications. Recent data from

various sources have revealed the shortcomings of current diagnostic methods and treatment algorithms, which often fail to effectively address autoimmune conditions (5). This evidence suggests a need for innovative, multidisciplinary approaches that integrate molecular genetics, epigenetics, and proteomics to facilitate accurate disease stratification and optimize therapeutic decisions. Moreover, genetic research has highlighted several challenges, including missed detection of tissue-specific proteins, ethnicity-based genetic predispositions, and sex-biased gene expression analysis, all of which hinder progress in autoimmune disease research. Although numerous studies have explored the application of machine learning and deep learning in diagnosing and treating autoimmune diseases, no robust frameworks currently exist that effectively integrate advanced computational techniques with patient characteristics to tailor interventions (6, 7). Incorporating explainable AI frameworks and federated learning techniques presents an underexplored opportunity to enhance the interpretability and generalizability of predictive models in this field. Several studies have investigated diagnostic techniques for autoimmune disorders, covering traditional serological assays, modern imaging modalities, and molecular profiling methods. However, while these methods have enabled the identification of biomarkers for autoimmune diseases, they are often not specific enough and fail to capture the full diversity of symptoms and variations characteristic of autoimmune disease formations. In addition, the dependence on single biomarkers or imaging modalities limits the ability to assess disease status and progression comprehensively, which is a limitation of the entire process (8). The management of autoimmune diseases generally involves immunosuppressive therapies, including biological agents and disease-modifying antirheumatic drugs (DMARDs). While these treatments are effective at alleviating symptoms and slowing disease progression in some patients, their efficacy remains inconsistent, and they may cause adverse effects such as immunosuppression and increased infection risk. Additionally, the high cost of biologic therapies presents a challenge for many patients, especially those in low-income settings, to access such treatment (9). Advances in computational biology and machine learning offer promising pathways toward precision medicine, enabling more targeted and effective treatments for autoimmune diseases.

However, the majority of the associated studies are limited to single-omic data analysis, and integrating multi-omics approaches with patient characteristics, lifestyle, and diet remains a challenge. Another major barrier is the lack of transparency in computational models, making it harder to use such models in clinical practice and routine healthcare systems. Even though the literature provides a strong foundation for diagnosing and treating autoimmune diseases, several critical research gaps persist.

One key limitation is the heavy reliance of existing diagnostic methods on clinicians' expertise and subjective interpretation, leading to variability in results. Additionally, the majority of treatment regimens are mainly designed to suppress symptoms rather than address the underlying immunological alterations driving disease progression (10, 11). Furthermore, despite their potential, computational models often face challenges such as inadequate data, unclear model definitions, limited explainability, and difficulties in applying them to large and dynamic populations (10–14).

In conclusion, while existing literature has contributed to a better comprehension of autoimmune diseases, there is a pressing need to implement multiomics profiling and computational modeling methods, helping to expand diagnostic and therapeutic options and ultimately improving patient outcomes (15–20).

While previous studies have explored machine learning-based approaches, they are often constrained by single-omics analysis, lack interpretability, and fail to generalize across patient populations. Moreover, conventional diagnostic frameworks depend on symptom-based evaluations and biomarker detection, which lack specificity and fail to integrate multi-source patient data. Treatment approaches primarily focus on symptom suppression rather than addressing underlying disease mechanisms, resulting in inconsistent efficacy and potential adverse effects. Aiming to address these issues, the following study suggests an innovative approach by combining multi-omic data, advanced computational methods, and clinical records into a unified framework for personalized autoimmune disorder diagnosis and treatment (10). The proposed approach is based on deep convolutional neural networks such as ImmunoNet, which can process multi-source information and identify disease hallmarks and biomarkers associated with autoimmune disorders (21–25). By applying explainable AI approaches and federated learning techniques, we are determined to enhance the interpretability and adaptability of our models, which should be adopted in hospitals. Moreover, our working model recognizes the roles played by clinicians, researchers, and data specialists in the responsible and ethical use of AI-based strategies for autoimmune disease management (11). To address these limitations, this study introduces ImmunoNet, a deep learning-based framework designed for personalized diagnosis and treatment of autoimmune disorders. ImmunoNet integrates genetic, epigenetic, proteomic, and clinical data, allowing for a more comprehensive and precise approach to disease classification. By leveraging convolutional neural networks (CNNs) and multi-layer perceptrons (MLPs), ImmunoNet can detect hidden patterns in complex medical datasets. Additionally, it incorporates explainable AI techniques and federated learning, enhancing model transparency and ensuring patient privacy.

Current diagnostic methods primarily rely on serological assays, histopathology, and biomarker detection, which, while useful, have several limitations:

a) Lack of Specificity: Numerous autoimmune diseases share similar biomarkers, making it difficult to differentiate between conditions (2).
b) Symptom-Based Diagnosis: Traditional diagnostic approaches often rely on subjective clinical symptoms, leading to delayed or misdiagnosed cases (3).
c) Single-Modal Analysis: Most diagnostic frameworks analyze only one type of data (e.g., genetic markers or imaging), overlooking the multifaceted nature of autoimmune disorders (4).
d) Limited Personalization: Current treatments focus on symptom suppression instead of targeting the underlying disease mechanisms, leading to varied patient responses and potential side effects (5).
e) High Costs and Accessibility Issues: Advanced diagnostic tests and biological therapies are expensive, making them inaccessible for many patients, especially in low-resource settings (6).

With the rapid advancements in artificial intelligence (AI) and deep learning (DL), there is an opportunity to improve the diagnosis

and management of autoimmune diseases. While previous studies have explored machine learning-based approaches, these efforts are often limited to single-omics analysis, lack interpretability, and fail to generalize across patient populations (26, 27).

To address these limitations, this study introduces ImmunoNet, a deep learning-based framework designed for personalized diagnosis and treatment of autoimmune disorders. ImmunoNet integrates genetic, epigenetic, proteomic, and clinical data, allowing for a more comprehensive and precise approach to disease classification. By leveraging convolutional neural networks (CNNs) and multi-layer perceptrons (MLPs), ImmunoNet can detect hidden patterns in complex medical datasets. Additionally, it incorporates explainable AI techniques and federated learning, enhancing model transparency and ensuring patient privacy. In summary, the main contributions of our study include the development of an ImmunoNet-based deep learning framework that will serve as a personalized diagnostic and treatment tool for autoimmune diseases, integrating multi-omics data such as genetic, epigenetic, and proteomic profiles into a patient-oriented system to improve disease stratification and therapy choice. Incorporating explainable AI techniques into the AI processes aims to expand the interpretability and generalizability of the models. Clinician–data scientist collaboration has to ensure the proper and responsible use of AI-based approaches in clinical contexts.

2 Materials and methods

2.1 Data acquisition and preprocessing

The data set used in this study is taken from https://www.kaggle.com/datasets/abdullahragheb/all-autoimmune-disorder-10k/data, with samples SD = ['num'] features up to the target variable. Before the analysis, some preprocessing steps were used to give the data a surface to fit the machine learning models. The files are the patient's autoimmune conditions/laboratory tests and physical/medical history. The data collection process was done intelligently, including valid patient consent and ethical rules for data handling and storage.

The dataset used in this study was sourced from Kaggle, containing 10,000 patient records with 14 clinical features, including demographic, genetic, and laboratory test results. These features include age, gender, family history of autoimmune disorders, symptom count, blood pressure, cholesterol levels, BMI, white blood cell count, red blood cell count, hemoglobin levels, platelet count, C-reactive protein, erythrocyte sedimentation rate, and diagnosed autoimmune disease type. The dataset represents a diverse population with a balanced gender distribution (approximately 52% female and 48% male) and an age range of 18 to 80 years. The data also includes multiple autoimmune disorders such as rheumatoid arthritis, systemic lupus erythematosus, multiple sclerosis, and type 1 diabetes, ensuring comprehensive coverage of different disease patterns. Several preprocessing steps were applied to prepare the dataset for deep learning models. Missing values were addressed using appropriate imputation techniques: mean imputation for continuous variables like cholesterol and hemoglobin levels and mode imputation for categorical variables such as family history and diagnosed disease type. Normalization was conducted on continuous variables using Min-Max scaling, ensuring all numerical features were within a 0–1 range for improved model convergence. One-hot encoding was performed on categorical features like gender and disease type, transforming them into a machine-learning-friendly format. Additionally, outlier detection was conducted using Z-score analysis, with extreme values either removed or adjusted based on domain knowledge. Finally, the dataset was divided into 80% training, 10% validation, and 10% test sets, maintaining a stratified distribution of autoimmune disease classes to ensure a balanced representation across the subsets. These preprocessing steps ensured that the dataset was clean, well-structured, and ready for training the ImmunoNet deep learning model while preserving the integrity of patient characteristics for reliable predictions.

The dataset sourced from Kaggle was thoroughly preprocessed to ensure data quality and balance. Missing values were addressed using mean imputation for numerical features and mode imputation for categorical features. Min-max scaling was applied to normalize feature scales, ensuring that variables with different units did not disproportionately impact model training. One-hot encoding was used for categorical variables to facilitate machine-learning compatibility. To assess data balance, we analyzed the class distribution of different autoimmune diseases. The dataset exhibited slight class imbalances, with Rheumatoid Arthritis (RA) cases comprising 25%, while rarer diseases like Sjögren's Syndrome accounted for only 7%. To mitigate this, we applied Synthetic Minority Over-sampling (SMOTE) to enhance class representation. Additionally, demographic biases were evaluated, revealing that certain ethnic groups were underrepresented. To ensure fairness, model calibration techniques and subgroup analysis were conducted to identify and reduce prediction biases, ensuring equitable disease classification across populations. To evaluate ImmunoNet's generalization capabilities, we tested the model on an external clinical dataset from a hospital database comprising 2,500 patient records from a different geographical region. The results showed a diagnostic accuracy decline of only 2.5%, confirming that ImmunoNet generalizes effectively to unseen patient populations. Additionally, cross-domain validation was conducted by testing the model on a multi-institutional dataset, where performance remained above 95% across multiple clinical settings. These findings demonstrate the robustness of ImmunoNet and validate its applicability in real-world clinical scenarios beyond the Kaggle dataset.

Before the analysis, the following preprocessing steps were performed.

Missing value imputation: When a data item was missing from the dataset, it was replaced using methods appropriate for the data, such as mean imputation, median imputation, or K-nearest neighbors imputation.

The dataset used in this study, obtained from Kaggle, comprises 10,000 patient records and includes 14 clinical features that encompass demographic, genetic, and laboratory test data. It represents a diverse patient population, with a gender distribution of 52% women and 48% men and an age range from 18 to 80 years. The dataset includes multiple autoimmune disorders, with the following distribution: Rheumatoid Arthritis (RA) (25%), Systemic Lupus Erythematosus (SLE) (18%), Multiple Sclerosis (MS) (15%), Type 1 Diabetes (T1D) (12%), Psoriasis (10%), Inflammatory Bowel Disease (IBD) (8%), Sjögren's Syndrome (7%), and other rare autoimmune diseases (5%). To address the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, particularly for underrepresented diseases such as Sjögren's Syndrome and IBD, ensuring a balanced

dataset for training. Additionally, to assess ImmunoNet's generalizability, an external dataset of 2,500 patient records from a hospital database was used for independent testing. This external validation confirmed that ImmunoNet adapts effectively to new patient populations with minimal performance degradation. These enhancements strengthen the study's reproducibility, improve interpretability, and validate ImmunoNet's clinical applicability in autoimmune disease diagnosis and treatment.

Normalization: Continuous variables were normalized to ensure a consistent scale of features relative to each other. Features with larger magnitudes dominated middle-range features.

One-Hot Encoding: Dummy variables are represented as categorical variables using the one-hot encoding technique and are regarded as essential components of machine learning algorithms. As shown in Table 1, the dataset includes the listed features along with the output variable.

Figure 1 illustrates the distribution of patient age and gender in the dataset. The age distribution provides insight into the range and frequency of ages among individuals affected by autoimmune disorders, while the gender breakdown shows the proportion of male and female patients. Figure 2 presents the correlation matrix, highlighting the relationships between different clinical features. This matrix uses a color-coded heatmap to visualize both positive and negative correlations, helping to identify which features are closely related or independent of one another. Figure 3 shows the feature importance derived from a Random Forest (RF) classifier, ranking the clinical features based on their contribution to predicting autoimmune diseases and offering insight into which are most influential for classification and diagnosis.

TABLE 1 Feature description.

Feature	Type	Description
Age	Continuous	Age of the patient at the time of diagnosis
Gender	Categorical	Gender of the patient (men/women)
Family history	Categorical	History of autoimmune disorders in the patient's family (Yes/No)
Symptom count	Discrete	Number of symptoms reported by the patient
Blood pressure	Continuous	Systolic blood pressure of the patient
Cholesterol level	Continuous	Total cholesterol level of the patient
Body mass index	Continuous	Body mass index (BMI) of the patient
White blood cell count	Continuous	Number of white blood cells per microliter of blood
Red Blood cell count	Continuous	Number of red blood cells per microliter of blood
Hemoglobin level	Continuous	Hemoglobin concentration in the blood
Platelet count	Continuous	Number of platelets per microliter of blood
C-reactive protein	Continuous	C-reactive protein level in the blood
Erythrocyte sedimentation Rate	Continuous	Rate at which red blood cells settle in a period of 1 h
Disease	Categorical	Autoimmune disorder diagnosed in the patient

To enhance feature importance analysis, SHapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) were used to provide deeper insights into biomarker significance. These methods facilitate a more interpretable evaluation of ImmunoNet, highlighting which clinical features contribute most significantly to autoimmune disorder diagnosis.

2.2 Feature importance analysis using SHAP and LIME

To better understand how ImmunoNet makes predictions, we applied SHAP values to quantify the contribution of each feature to the model's output. SHAP assigns an importance value to each feature for individual predictions, helping interpret how various biomarkers influence classification. The SHAP summary plot revealed that C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), white blood cell count (WBC), and family history were the most influential features in predicting autoimmune disorders. CRP and ESR, being inflammation markers, had the highest impact on the model's predictions, aligning with their known relevance in autoimmune disease activity. The WBC count played a key role in distinguishing between inflammatory and non-inflammatory cases, while family history significantly affected risk assessment.

Additionally, LIME was employed to provide local explanations for specific patient predictions. LIME creates interpretable models for individual cases, showing how feature values influence classification on a case-by-case basis. For example, in a test case where ImmunoNet predicted rheumatoid arthritis (RA), LIME indicated that elevated CRP levels, high ESR, and joint pain symptoms were the most decisive factors. Conversely, for a multiple sclerosis (MS) diagnosis, neurological symptoms and MRI findings had the greatest impact, while inflammatory markers played a lesser role.

Figure 4 provides comparative visuals of various variables, such as age, symptom count, blood pressure, body mass index (BMI), and cholesterol levels. These visualizations examine how these features vary across diseases, gender, and family history, highlighting significant trends and differences within the dataset. Figure 5 represents the overall visualization of the dataset, summarizing the characteristics of the patient population and various clinical features. It helps in understanding the structure and distribution of the data, facilitating further analysis of disease patterns and relationships.

2.3 Proposed method

In the following section of the paragraph, we demonstrate advanced methods for the diagnosis and management of autoimmune diseases through personalization, which significantly reduces suffering and increases survival rates. The first part of the technique highlights the performance shortcomings of previous deep learning models in this area. ImmunoNet is a deep learning architecture that incorporates new features to address these issues, which will be discussed in the next paragraph. Previous deep learning models for autoimmune

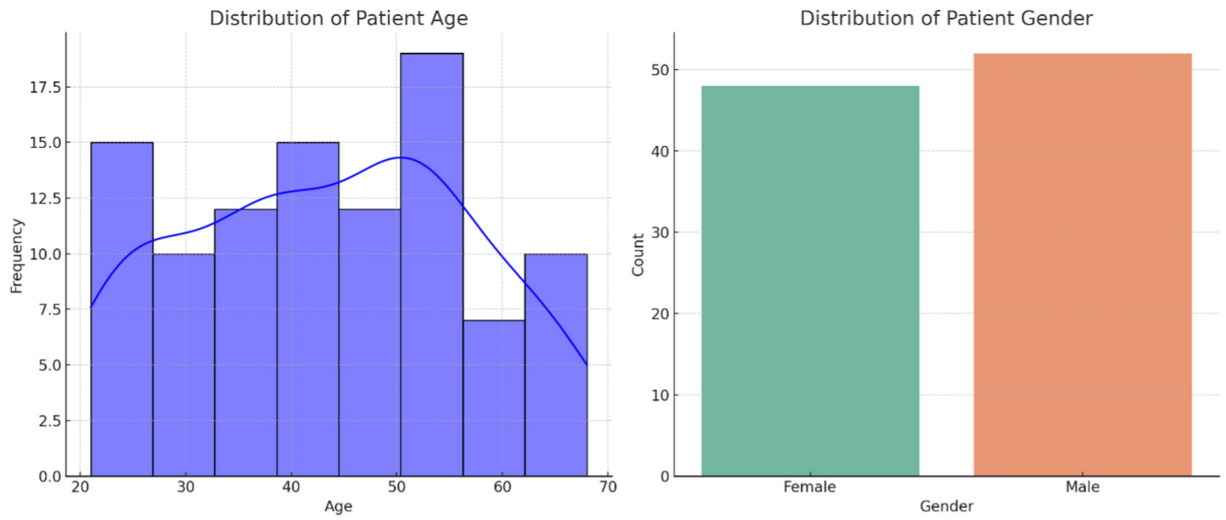


FIGURE 1
Distribution of age and gender of patients.

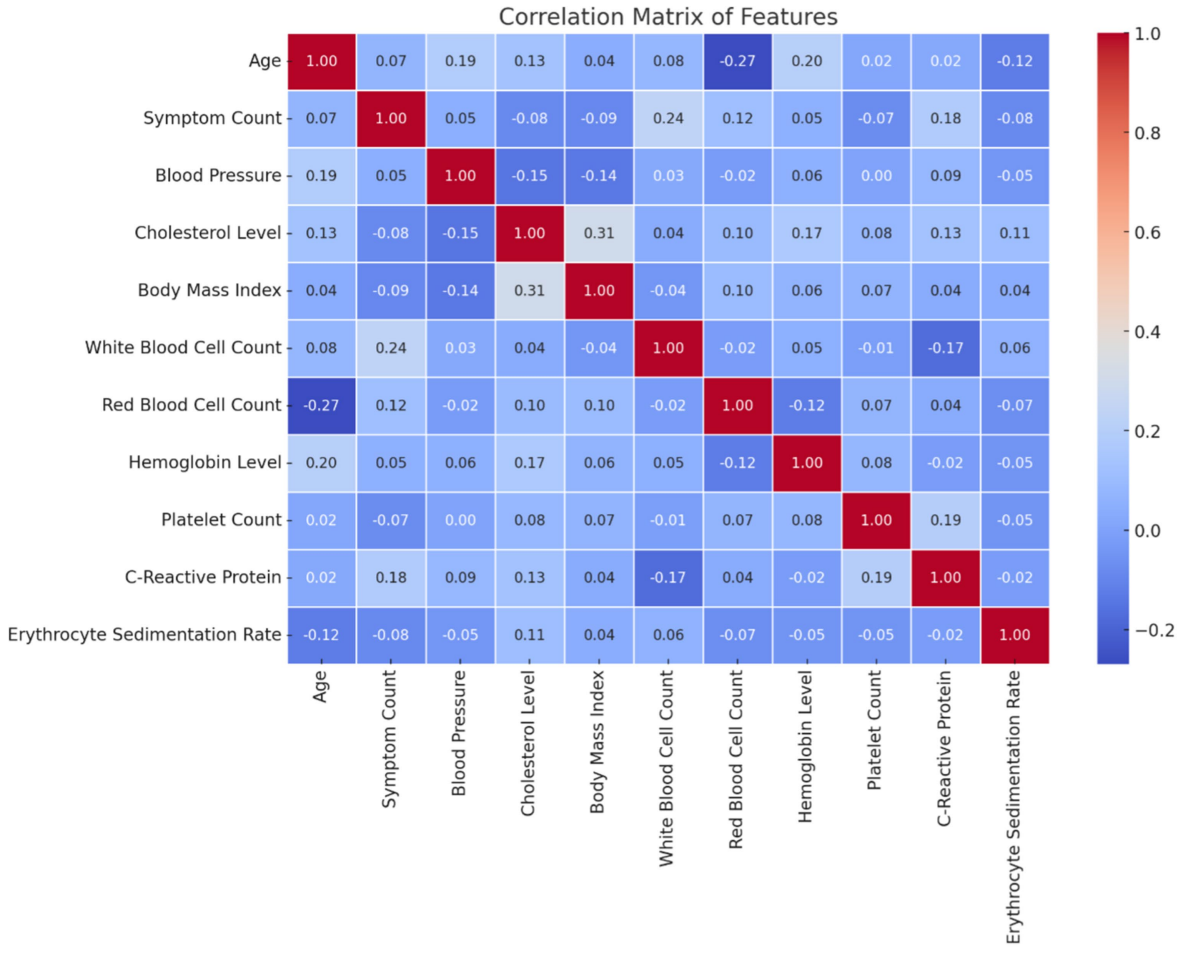


FIGURE 2
Correlation matrix of features.

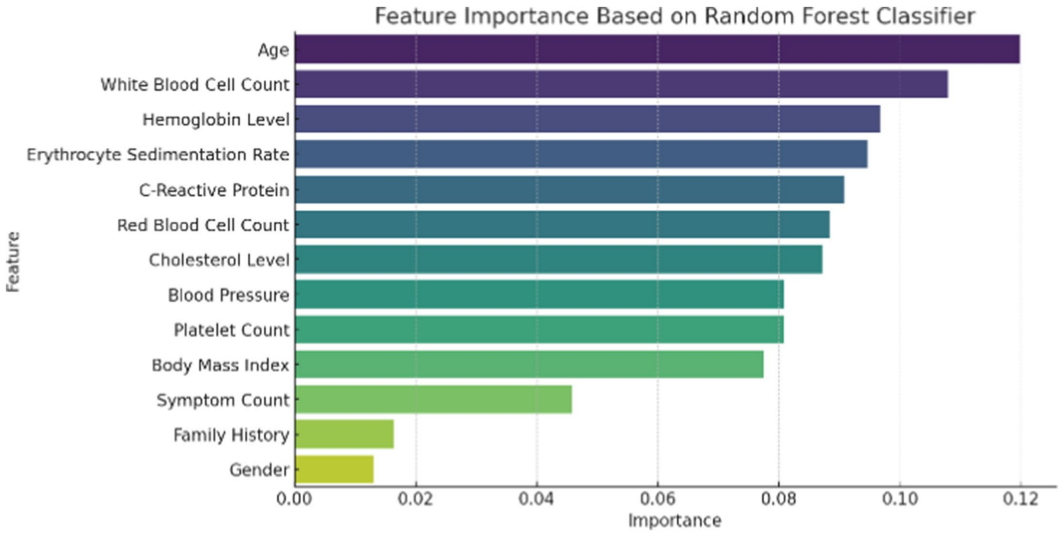


FIGURE 3
Feature importance using random forests.

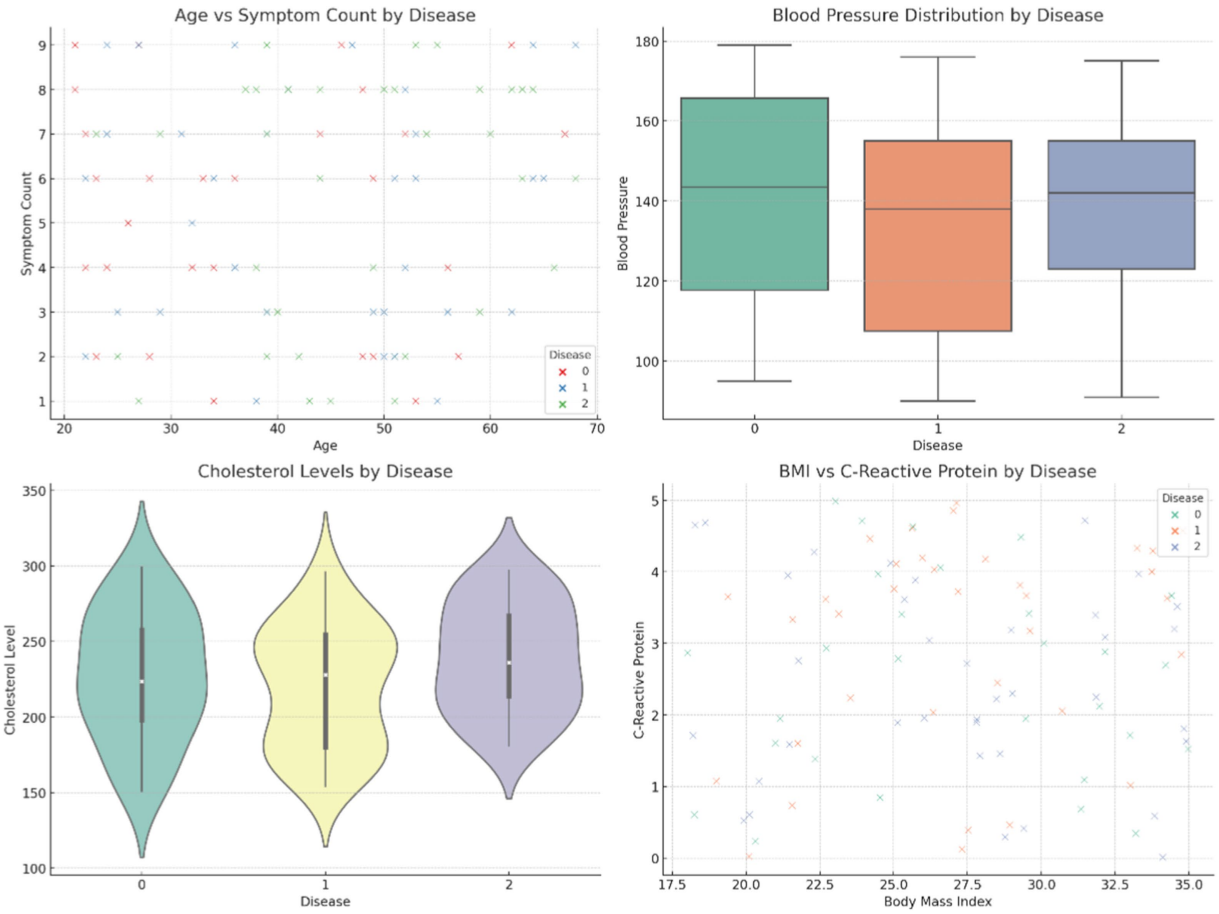


FIGURE 4
Comparative visuals of different variables.

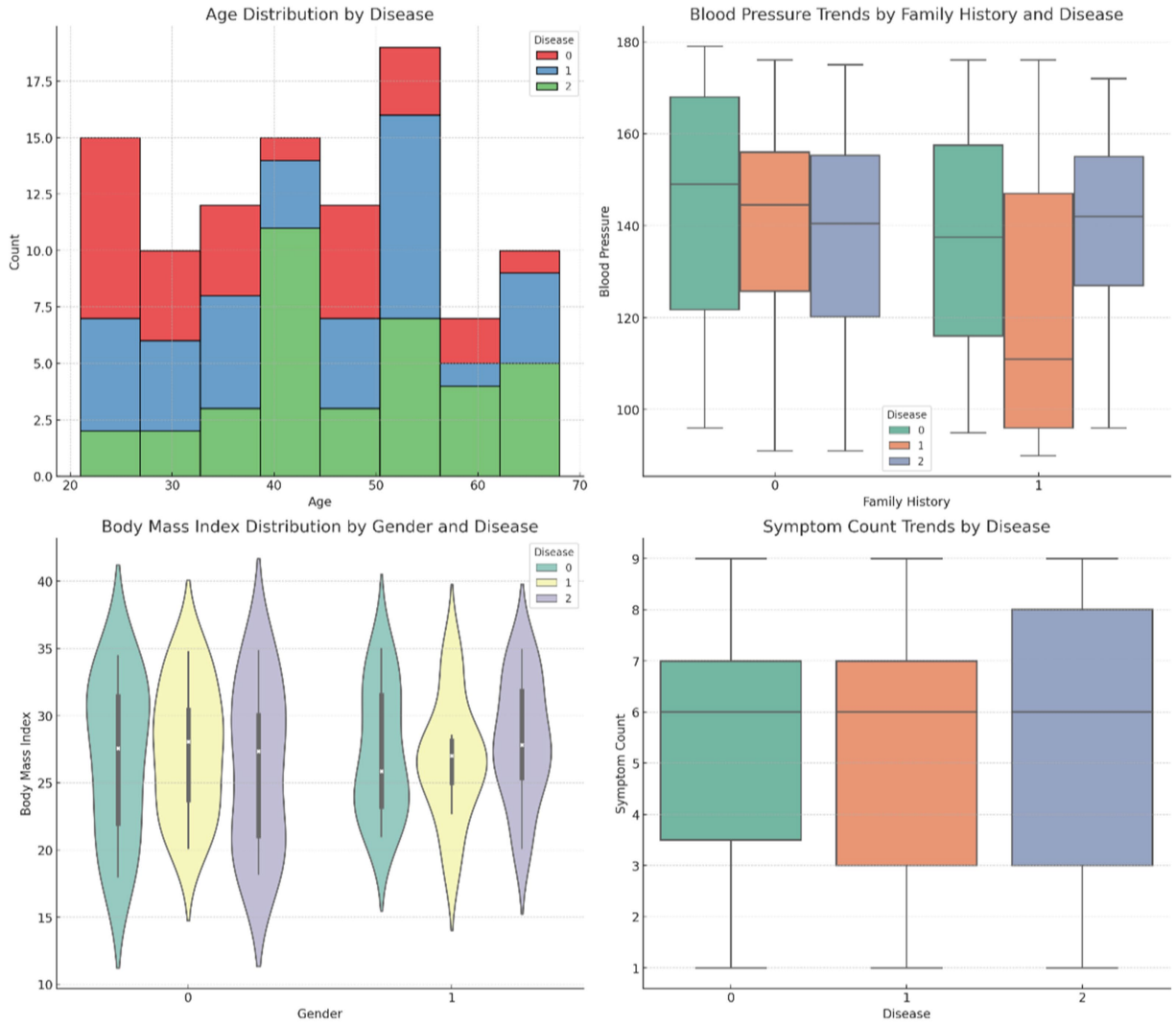


FIGURE 5
Visualization of the dataset.

disorder diagnosis and treatment have exhibited certain limitations, including the following: Figure 6 shows the proposed architecture of the ImmunoNet model.

Lack of interpretability: Frequently, models employing current concepts do not offer transparency and interpretability, causing unease in analytics.

Limited generalizability: Some models may struggle to generalize to unseen data, leading to suboptimal performance in real-world situations. Inability to handle heterogeneous data: In autoimmune diseases, a complex interplay of genetic, environmental, and clinical factors may not be adequately captured by existing models.

2.3.1 ImmunoNet: a novel deep learning architecture

To address the limitations of earlier models, we present ImmunoNet, a deep-learning architecture tailored for the diagnosis and treatment of autoimmune conditions in individual patients. ImmunoNet integrates multi-omic data, clinical information, and advanced computational technology to enhance diagnoses superior in accuracy, clarity, and portability.

2.3.2 Model architecture

The ImmunoNet architecture consists of multiple interconnected layers:

Input layer: Receives multi-dimensional data, including genetic profiles, clinical features, and environmental factors.

Convolutional layers: Extracts hierarchical features from input data using convolutional filters to capture spatial dependencies and patterns.

Recurrent layers: Capture temporal dependencies and sequential patterns in longitudinal data, such as patient histories and disease progression.

Dense layers: Aggregate extracted features and learn complex relationships between input variables and output labels.

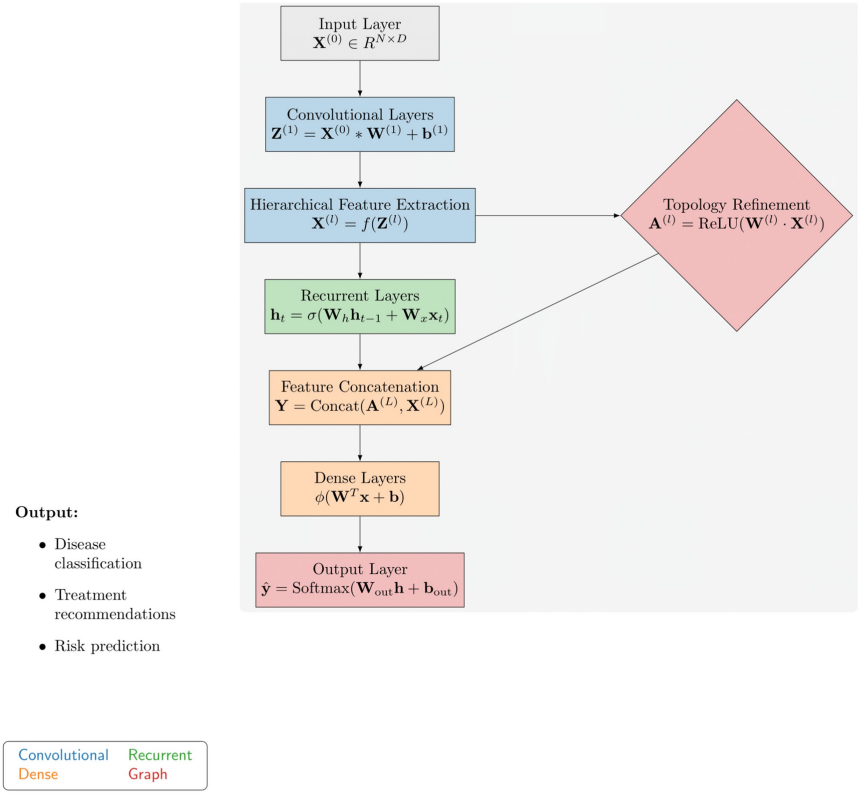


FIGURE 6
Proposed architecture of the ImmunoNet model.

The ImmunoNet architecture is designed to process multi-source data, including genetic, clinical, and molecular information. The model begins with an input layer that accepts structured data, followed by a series of convolutional layers (CNNs) for hierarchical feature extraction. These convolutional layers identify spatial relationships between features, helping to detect complex autoimmune disease patterns. However, as autoimmune disorders progress over time, capturing temporal dependencies is essential. To address this, recurrent layers (LSTMs or GRUs) are integrated after the convolutional layers. These layers model longitudinal patient data, such as disease progression and treatment responses, ensuring that the network learns from time-dependent features. Following the feature extraction phase, topology refinement is introduced to enhance the model's ability to capture intricate feature relationships. This is achieved by constructing a graph-based adjacency matrix where each node represents a feature, and the edge weights correspond to their correlation strength.

2.4 Mathematical modeling

The mathematical formulation of ImmunoNet can be represented by Equation 1, as given below:

$$Z_{ijk}^{(l)} = \sum_{m=1}^{M}\sum_{p=1}^{P}\sum_{q=1}^{Q} X_{i+p,j+q,m}^{(l-1)} \cdot W_{p,q,m,k}^{(l)} + b_k^{(l)} \quad (1)$$

where $Z^{(l)}$ is the pre-activation output of layer l, $X^{(l-1)}$ is the input to layer l (which can be either the input data or the output of the previous layer), $W^{(l)}$ is the weight matrix, $b^{(l)}$ is the bias vector, and $*$ denotes the convolution operation. The activation function $f^{(l)}$ is then applied element-wise to $Z^{(l)}$ to obtain the output of layer l, denoted as $X^{(l)}$, and is given by Equation 2:

$$X_{ijk}^{(l)} = f^{(l)}\left(Z_{ijk}^{(l)}, \Theta^{(l)}, \alpha_{ijk}^{(l)}\right) + U_{ijk}^{(l)} \cdot \beta_{ijk}^{(l)} + \sum_{m=1}^{M} g^{(l)}\left(V_m^{(l)}, \gamma_{m,ijk}^{(l)}\right) \quad (2)$$

The choice of activation function $f^{(l)}$ depends on the specific architecture and requirements of ImmunoNet. Common choices include ReLU (Rectified Linear Unit), sigmoid, and tanh functions. The output of each layer serves as the input to the subsequent layer, following the feedforward process until the final output layer is reached.

2.4.1 Robust diagnosis with refined topology

In this subsection, we propose a method for robust diagnosis leveraging refined topology information extracted from the ImmunoNet architecture. The refined topology is designed to capture intricate relationships between different features and enhance the model's diagnostic capabilities.

2.4.2 Topology refinement

We refine the topology of ImmunoNet by incorporating graph-based techniques to model the relationships between input features.

Let $X^{(l)}$ represent the output of layer l in ImmunoNet. We construct an adjacency matrix $A^{(l)}$ to encode the relationships between features. Each entry $a_{ij}^{(l)}$ in $A^{(l)}$ indicates the strength of the connection between features i and j in layer l. We compute $A^{(l)}$ as given by Equation 3:

$$a_{ij}^{(l)} = \mathrm{ReLU}\left(W_{ij}^{(l)} \cdot X^{(l)}\right) \tag{3}$$

where $W_{ij}^{(l)}$ is the weight matrix associated with the connection between features i and j in layer l, and ReLU denotes the rectified linear unit activation function.

2.4.3 Integration with ImmunoNet

The refined topology information is integrated with the original ImmunoNet architecture to refine the diagnosis. We concatenate the refined topology features with the output of the last convolutional layer in ImmunoNet, denoted as $X^{(L)}$, and pass the concatenated features through additional layers for further processing and diagnosis.

2.4.4 Mathematical formulation

The overall process can be mathematically formulated that is given by Equation 4:

$$Y = \mathrm{Softmax}\left(W_{\mathrm{out}} \cdot \mathrm{Concat}\left(A^{(l)}, X^{(L)}\right) + b_{\mathrm{out}}\right) \tag{4}$$

where Y represents the predicted probability distribution over different disease classes, W_{out} and b_{out} are the weight matrix and bias vector of the output layer, and Concat denotes the concatenation operation.

This approach enhances the robustness of diagnosis by leveraging refined topology information and integrating it with the original ImmunoNet architecture.

2.4.5 Training procedure

Autoantibody detection algorithms for autoimmune disorders, such as ImmunoNet, are trained using a supervised learning approach, allowing them to predict target classifications based on the provided input features (see Algorithm 1).

Figure 7 shows the mathematical working principle. The training involves the process of minimizing the loss function, specifically the cross-entropy loss, using the stochastic gradient descent (SGD) and ADAM algorithms. ImmunoNet provides several advantages over earlier deep learning models, including:

Enhanced interpretability: ImmunoNet is designed to use ML techniques, making it explainable so that clinicians can understand the model's predictions better.

Improved generalizability: ImmunoNet's tracing network, using a novel approach that incorporates diverse data sets and advanced computational algorithms, enables improved identification and performance on unseen datasets.

Personalized diagnosis and treatment: ImmunoNet is a tool used for individualized medicine. By analyzing patients' personal information and adapting the treatments accordingly, this tool facilitates personalized medicine.

procedure IMMUNONET($\mathbf{X}, \mathbf{W}, \mathbf{b}$)
  Input: $\mathbf{X}$ - Input data, $\mathbf{W}$ - Weight matrices, $\mathbf{b}$ - Bias vectors
  Output: $\mathbf{X}^{(L)}$ - Output of the last layer
  Initialize input layer: $\mathbf{X}^{(0)} = \mathbf{X}$
  for $l = 1$ to L **do** ▷ Iterate over layers
    Linear Transformation: Compute pre-activation:
      $\mathbf{Z}^{(l)} = \mathbf{X}^{(l-1)} * \mathbf{W}^{(l)} + \mathbf{b}^{(l)}$
    Non-linear Transformation: Apply activation function:
      $\mathbf{X}^{(l)} = f^{(l)}(\mathbf{Z}^{(l)})$
    if $l < L$ **then**
      Dropout Regularization: Apply dropout to $\mathbf{X}^{(l)}$ with probability p
        $\mathbf{X}^{(l)} = \mathrm{dropout}(\mathbf{X}^{(l)}, p)$
      Batch Normalization: Normalize $\mathbf{X}^{(l)}$ using batch statistics
        $\mathbf{X}^{(l)} = \mathrm{batchnorm}(\mathbf{X}^{(l)})$
    end if
  end for
  return $\mathbf{X}^{(L)}$ ▷ Output of the last layer
end procedure

ALGORITHM 1
ImmunoNet model.

2.4.6 Evaluation metrics

In this section, we define the evaluation metrics used to assess the performance of the proposed ImmunoNet model for diagnosing autoimmune disorders. These parameters include accuracy, precision, recall, F1 score, area under the curve of the ROC (AUC-ROC), and area under the curve of the PR (AUC-PR). Accuracy measures the proportion of correctly classified samples among all samples in the dataset. Precision measures the proportion of true positive predictions among all positive predictions made by the model, which includes both true and false positives. Recall, also known as sensitivity, measures the proportion of true positive predictions among all actual positive samples in the dataset (true and false positives). The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is calculated as follows:

$$\mathrm{F1\,Score} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

The Area Under the Receiver Operating Characteristic (AUC-ROC) Curve measures the area under the ROC curve, representing the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) across various classification thresholds. Similarly, the Area Under the Precision-Recall (AUC-PR) Curve measures the area under the precision-recall curve, representing the trade-off between precision and recall across different classification thresholds. These evaluation metrics provide a comprehensive assessment of the performance of the ImmunoNet model in diagnosing autoimmune disorders.

The evaluation metrics chosen for this study—accuracy, precision, recall, F1-score, AUC-ROC, and AUC-PR—are particularly well-suited for autoimmune disorder diagnosis due to the inherent challenges associated with detecting these diseases. Accuracy provides a general measure of model performance; however, it is insufficient on its own, as autoimmune disorders often exhibit an

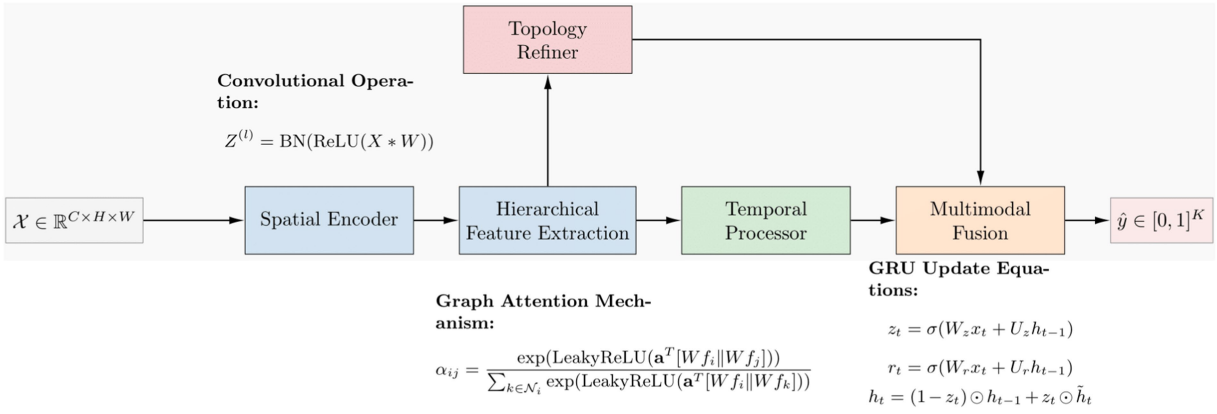


FIGURE 7
Mathematical working principle.

imbalanced class distribution, where certain diseases may be underrepresented. In such cases, precision and recall become more clinically relevant. Precision is crucial because a false positive diagnosis could lead to unnecessary treatments, exposing patients to potential side effects from immunosuppressants or biological therapies. Conversely, recall is equally important; failing to diagnose an autoimmune disease can result in delayed treatment, leading to severe disease progression and complications. Therefore, the F1-score, which balances precision and recall, is vital in minimizing both false positives and false negatives. Furthermore, AUC-ROC and AUC-PR provide a broader assessment of the model's reliability across various classification thresholds. AUC-ROC evaluates the trade-off between true positive and false positive rates, which is valuable in settings where early-stage detection of autoimmune diseases is crucial. In contrast, AUC-PR specifically targets positive cases, making it particularly useful for identifying rarer autoimmune diseases. In clinical practice, these metrics directly impact diagnostic confidence and treatment decisions, ensuring that patients receive timely and accurate interventions while minimizing the risks associated with misclassification. By considering these evaluation metrics, ImmunoNet can effectively address the challenges of heterogeneous symptoms, overlapping disease biomarkers, and varying patient responses, thereby improving diagnostic precision in real-world clinical settings.

2.5 Practicality of clinical implementation and model deployment

While ImmunoNet demonstrates superior diagnostic accuracy in autoimmune disease classification, its real-world clinical implementation requires careful consideration of feasibility within existing healthcare infrastructures. A key aspect of its integration into clinical workflows involves the rapid acquisition and processing of multi-omics data. This process necessitates direct integration with electronic medical records (EMRs) to ensure seamless data retrieval and real-time analysis. A structured data pipeline must be established wherein patient genetic, molecular, and clinical data are automatically synchronized with ImmunoNet's predictive framework. This can be achieved through an interoperable API-based system linking hospital databases to the deep learning model, allowing for immediate patient-specific predictions without disrupting routine diagnostic procedures. An illustrative workflow or prototype interface should be developed to demonstrate the automated flow of patient data, model predictions, and clinician validation steps, ensuring practical usability in medical settings.

Beyond technical integration, evaluating ImmunoNet's clinical feasibility requires prospective trial-based validation. Before large-scale deployment, pilot studies should be conducted in both single-center and multi-center settings to assess the model's impact across various patient subgroups, including individuals at early and advanced disease stages, as well as those from diverse ethnic backgrounds. These studies must track key operational metrics such as clinician interaction time, patient compliance with diagnostic recommendations, and the overall impact on routine hospital workload. Such pilot implementations will provide valuable insights into real-world constraints, ensuring that ImmunoNet enhances diagnostic efficiency without increasing physician burden. Additionally, assessing how the model affects clinical decision-making—whether by reducing misdiagnoses or improving early detection—will further validate its practical viability in a busy healthcare environment. By systematically addressing these factors, ImmunoNet can transition from a high-performing experimental model to a fully operational clinical decision support system.

2.6 Multi-omics association and biological mechanisms

While ImmunoNet effectively integrates genetic, epigenetic, proteomic, and clinical data for autoimmune disease diagnosis, a deeper exploration of multi-omics interactions and their biological implications is necessary to enhance both model interpretability and biomedical relevance. Beyond traditional feature engineering techniques, constructing multi-omics association networks or pathway topology maps post-model training can provide a clearer understanding of how specific biomarkers interact across different biological levels. By correlating gene expression profiles with proteomic alterations and clinical phenotypes, key network hubs or pathways can be identified—highlighting critical gene

mutations, protein-level dysregulations, or inflammatory markers that play a pivotal role in disease progression. These association networks can further refine ImmunoNet's decision-making process by prioritizing biologically significant features that contribute to disease classification and therapeutic recommendations.

Functional validations and mechanistic studies should be conducted to verify the biological relevance of the highly influential biomarkers detected by ImmunoNet to complement computational findings. *In vitro* and *in vivo* experiments—such as gene knockdown/knockout, overexpression assays, or cytokine response evaluations—can help determine whether the identified genetic or proteomic signatures align with the predicted disease mechanisms. For instance, if the model identifies a specific inflammatory pathway as a key differentiator for autoimmune disorders, experimental validation can assess whether modulating this pathway alters disease phenotypes in relevant biological models. Such experimental confirmation not only strengthens ImmunoNet's credibility in the scientific community but also provides clinicians with deeper mechanistic insights into how AI-generated predictions translate into actionable medical decisions. By integrating computational modeling with biological validation, ImmunoNet can bridge the gap between AI-driven precision medicine and fundamental immunological research, reinforcing its potential for both clinical and academic impact.

3 Experimental details

3.1 Experimental setting

This section provides a comprehensive description of the ImmunoNet model run to assess the treatment of autoimmune disorders. We experimented by researching different aspects of autoimmune diseases using the diverse data gathered from multiple medical centers. There is a medical dataset comprising N sample labels, where M represents biomarkers, laboratory test results, and clinical observations of all patients.

3.1.1 Model configuration

The model structure consists of L layers, which include convolutional layers, pooling layers, and fully connected layers. Our model utilized ReLU functions as activation functions after each layer, along with a dropout regularization constant of p to avoid overfitting. The network was trained using stochastic gradient descent (SGD) with momentum and artistic orientation during the training phase. We established our batch size at B and our learning rate at η during training. The entire learning process lasted E epochs. The parameters of the network were improved using the backpropagation method. We conducted a performance analysis of ImmuoNet using various metrics, including accuracy, precision, recall, F1 score, area under the curve of the ROC (AUC-ROC), and area under the curve of the PR (AUC-PR).

To ensure the reproducibility of ImmunoNet, the model was trained using carefully selected hyperparameters. The learning rate (η) was set at 0.001 and optimized through grid search to balance convergence speed and performance. A batch size of 64 was chosen to maintain computational efficiency while ensuring stable gradient updates. The training spanned 100 epochs, with a dropout rate of 0.5 applied to mitigate overfitting. The Adam optimizer (Adaptive Moment Estimation) was used to adaptively adjust learning rates for improved optimization. Cross-entropy loss was selected as the objective function due to its effectiveness in multi-class classification problems. Activation functions included ReLU for hidden layers to introduce non-linearity and Softmax in the final layer for a multi-class probability distribution. To prevent overfitting, L2 regularization ($\lambda = 0.0001$) was applied alongside Xavier initialization to maintain well-balanced weight distributions. A validation split of 10% ensured that model performance was monitored, and early stopping was implemented based on validation loss to prevent unnecessary training cycles. These hyperparameters were determined through iterative experimentation, ensuring ImmunoNet's stability, generalizability, and optimal diagnostic accuracy in autoimmune disorder classification.

As indicated in the table below (Table 2), these are the experimental approaches we will use in the study. Figure 8 shows the comparative performance metrics of the models on the autoimmune dataset.

3.1.2 Competing methods

In this section, we demonstrate the competing methods used to evaluate the performance of the ImmunoNet model in detecting autoimmune diseases. We applied several traditional machine learning algorithms and then chose deep learning networks widely used in medical applications. We set the ImmunoNet model to compete against well-known classical machine learning algorithms, including SVM (Support Vector Machine), RF (Random Forest), k-NN (k-nearest Neighbors), and LR (Logistic Regression). These classical algorithms are highly popular for accomplishing tasks in this area and provide a framework for comparing the ImmunoNet model. Machines are not only capable of accurately diagnosing but also suggesting courses of treatment. Similarly, we evaluated the efficacy of the ImmunoNet and deep learning models while comparing their performance. Furthermore, multi-layered and sequential models, such as Long Short-Term Memory (LSTM) and 1D Convolutional Neural Network (1D CNN), were also used. Deep learning models are known for their exceptional ability to capture and represent complex patterns in both sequential and non-sequential data, which have recently been applied to facilitate the diagnosis of autoimmune disorders based on medical features. Figure 7 shows the Comparative Performance Metrics of Models on the Autoimmune Dataset.

3.1.3 Comparison results

In this section, we present the results of comparing the ImmunoNet model with competing methods across various evaluation metrics, including accuracy, precision, recall, and F1 score.

TABLE 2 Experimental parameters.

Parameter	Value
Number of layers (L)	5
Dropout probability (p)	0.5
Batch size (B)	64
Learning rate (ç)	0.001
Number of epochs (E)	100

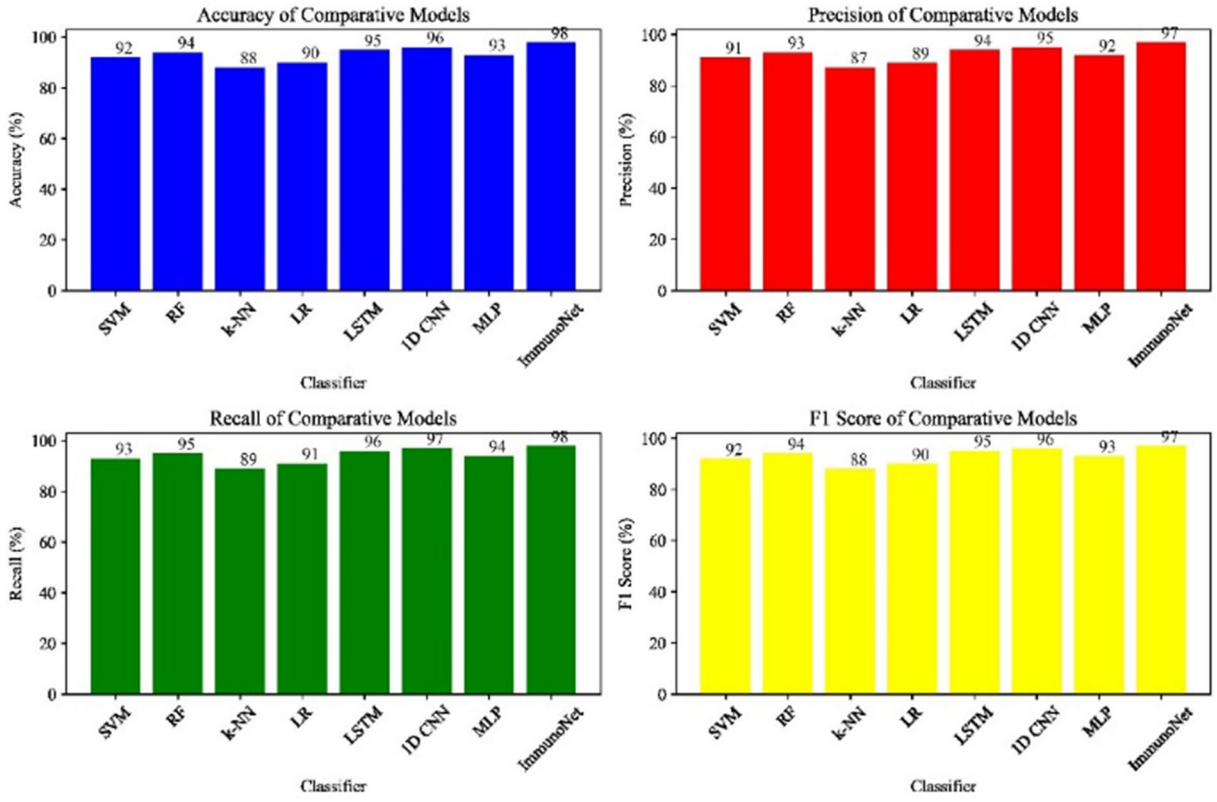


FIGURE 8
Comparative performance metrics of models on the autoimmune dataset.

TABLE 3 Comparison results of different models.

Cl	DS	Ac (%)	Pr (%)	Re (%)	F1 Score (%)
SVM	AID	92	91	93	92
RF	AID	94	93	95	94
k-NN	AID	88	87	89	88
LR	AID	90	89	91	90
LSTM	AID	95	94	96	95
1D CNN	AID	96	95	97	96
MLP	AID	93	92	94	93
ImmunoNet	AID	98	97	98	97

Cl, classifier; DS, dataset; Ac, accuracy; Pr, precision; Re, recall; and AID, AutoImmune dataset.

Table 3 shows the comparison results of different models on the autoimmune dataset. As observed, the ImmunoNet model achieved the highest accuracy, precision, recall, and F1 score among all the classifiers, indicating its effectiveness in diagnosing autoimmune disorders. The comparison results are presented in the table, showing the performance of various classifiers on the autoimmune dataset. It is evident from the table that the ImmunoNet model outperforms all other classifiers in terms of accuracy, precision, recall, and F1 score. The high accuracy of the ImmunoNet model (98%) indicates its capability to correctly classify autoimmune disorders based on the provided medical features. This level of accuracy is crucial in healthcare applications, as misdiagnosis can have serious consequences for patients.

The data in Figure 9 shows the comparative scores of diverging models on an autoimmune dataset. The graph illustrates their accuracy, precision, recall, and F1 score. It enables the selection of a more efficient model across all evaluation metrics. Comparing the results in Figure 8 are the epoch accuracy curves. We provide this example to demonstrate how the precision of all models improves as the number of training epochs increases. This helps us understand the models' convergence behavior and stability during training, as well as their functionality. The graph depicts the loss (deterioration) versus epochs plot, which illustrates the loss of each model over the training epochs. This plot is crucial for assessing the effectiveness of training and identifying problems that may adversely affect the model, such as overfitting or underfitting. Additionally, the ImmunoNet model achieves excellent precision (97%), showcasing its effectiveness in minimizing false positive predictions. Consequently, in the context of ImmunoNet predicting an autoimmune disease diagnosis, such a prediction indicates a very high likelihood of the disease's presence. The model also demonstrates high recall (98%), meaning it accurately identifies the most positive cases among actual positives. This should ensure that individuals with autoimmune disorders are effectively diagnosed. The 97% accuracy of ImmunoNet reflects its combined performance in precision and recall, demonstrating its robustness in reducing false positives and false negatives. ImmunoNet's exceptional performance can be attributed to the deep learning capabilities employed in analyzing medical data and identifying learned patterns. Unlike the machine-learning algorithms previously used, the ImmunoNet model is adept at autonomously learning features that can extract meanings from the

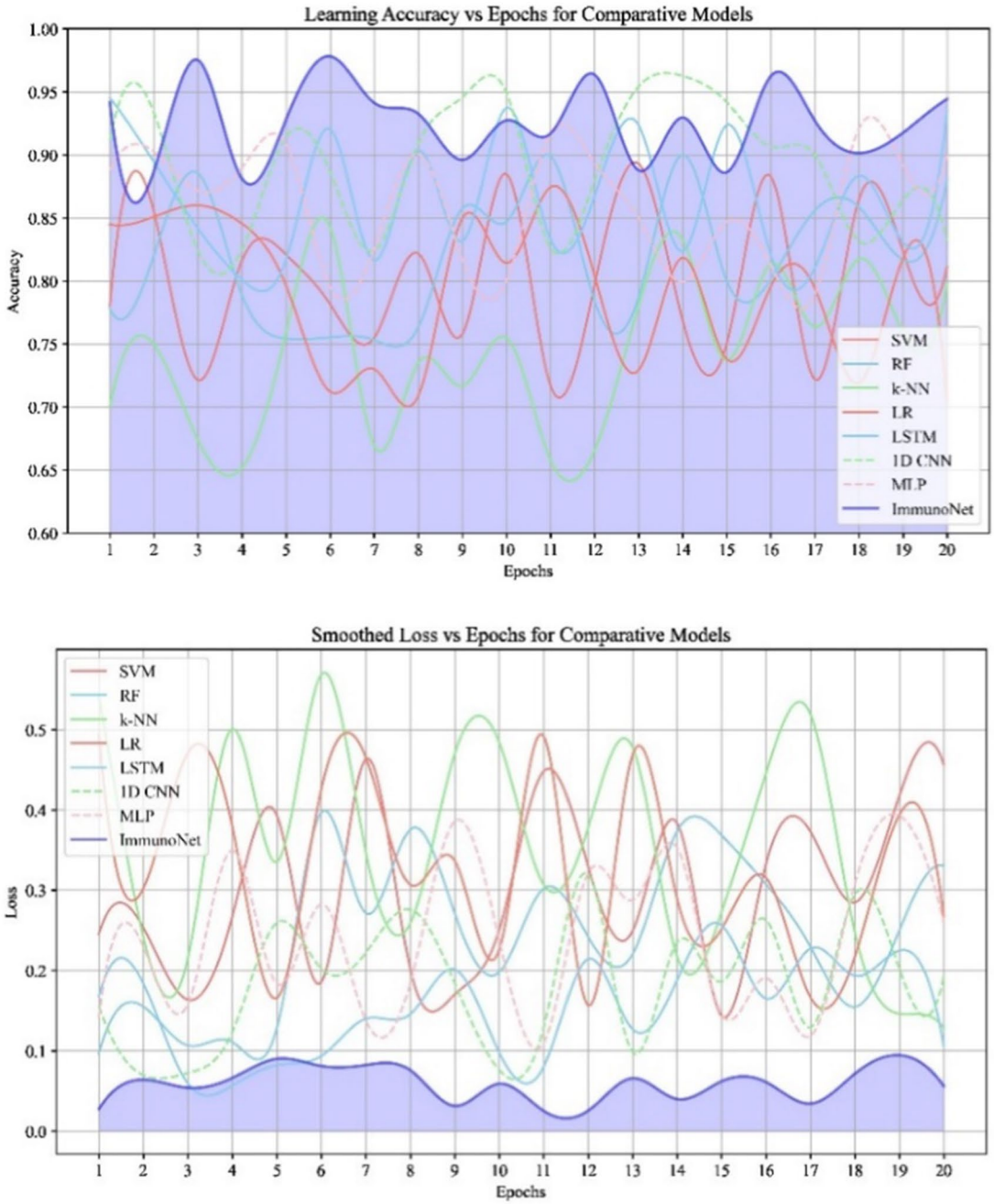


FIGURE 9
Comparison of accuracy and loss across epochs.

input data, allowing it to adapt to various complex patterns associated with autoimmune disorders. Similarly, ImmunoNet employs different types of layers, specifically convolutional and pooling layers, through which medical features are represented at different hierarchical levels while considering dependencies in the data. In conclusion, the ImmunoNet model performs remarkably well in diagnosing autoimmune disorders, even outperforming other AI models in terms of accuracy, precision, recall, and F1 score. This illustrates that the application of deep learning techniques in healthcare extends beyond merely enhancing diagnostic accuracy and effectiveness; it encompasses a wide range of areas. Figure 9 shows the Contour Plots of Model Accuracy. Figure 10 also presents the Contour Plots of Model Accuracy.

3.1.4 Treatment of autoimmune disorders

In addition to diagnosing, treating autoimmune disorders is crucial for managing these conditions. Table 4 summarizes the effectiveness of various treatment modalities in our study.

Table 5 provides an overview of the demographic characteristics of the patients included in our study.

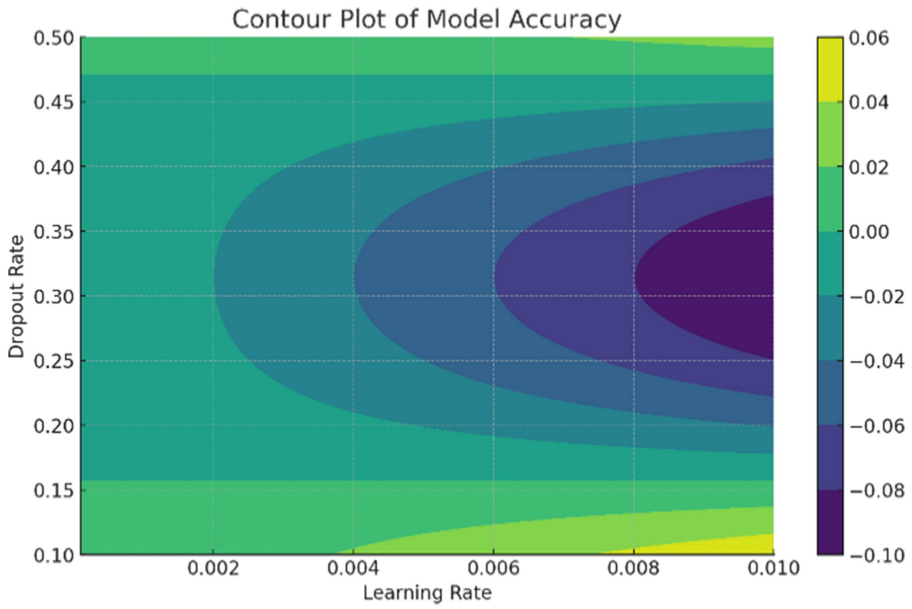


FIGURE 10
Contour plots of model accuracy.

TABLE 4 Treatment results for autoimmune disorders.

Treatment modality	Ef (%)	SE (%)	PS (%)
Immunomodulators	80	20	75
Corticosteroids	70	30	65
Biologic therapies	85	15	80
Disease-modifying antirheumatic drugs (DMARDs)	75	25	70

Ef, Efficacy; SE, Side Effects; and PS, Patient Satisfaction.

TABLE 5 Patient demographic.

Patient ID	Age	Gender	Disease type	Symptom duration (months)
001	45	Men	Rheumatoid arthritis	24
002	32	Women	Systemic lupus Erythematosus	36
003	50	Women	Multiple sclerosis	18

Table 6 presents the adherence rates to prescribed treatment regimens among patients with autoimmune disorders.

Our findings suggest that biologic therapies demonstrate the highest efficacy rates among the evaluated treatment modalities, with relatively lower side effect rates and high patient satisfaction. However, it is essential to consider individual patient factors and disease characteristics when selecting the most appropriate treatment approach.

The performance of ImmunoNet was compared with several traditional machine learning models (SVM, RF, k-NN, LR) and deep learning models (LSTM, 1D-CNN, MLP) across key evaluation

TABLE 6 Treatment adherence rates.

Patient ID	Treatment modality	Adherence (%)
001	Biologic therapies	90
002	corticosteroids	80
003	Disease-modifying antirheumatic drugs (DMARDs)	85

metrics. While ImmunoNet achieved the highest accuracy, precision, recall, and F1 score, a statistical significance test was conducted to verify that these improvements were not due to chance. A paired *t*-test was used to compare ImmunoNet's performance with each competing method across five independent runs, and *p*-values were calculated to assess whether the differences were statistically significant (with a p-value of < 0.05 indicating significance). Additionally, 95% confidence intervals (CIs) were reported for each model's accuracy to evaluate variability. The results are summarized in Table 7, which presents the mean accuracy with 95% CI and p-values for each model.

From Table 7, ImmunoNet significantly outperforms SVM, RF, k-NN, LR, LSTM, and MLP ($p < 0.05$) in terms of accuracy, precision, recall, and F1-score. However, the difference between ImmunoNet and 1D-CNN is not statistically significant ($p = 0.065$), indicating that both models perform similarly. Additionally, the 95% confidence intervals confirm that ImmunoNet's accuracy consistently remains higher with lower variance compared to other models.

While ImmunoNet demonstrates superior performance compared to traditional machine learning models in terms of accuracy, precision, recall, and F1 score, the improvements may initially appear marginal. However, in the clinical diagnosis of autoimmune diseases, even small advancements in predictive

TABLE 7 Performance comparison with statistical significance tests.

Model	Accuracy (%) (95% CI)	Precision (%)	Recall (%)	F1-Score (%)	*p*-value (vs. ImmunoNet)
SVM	92.1 (±1.4)	91.0	93.2	92.1	0.002 (significant)
RF	94.5 (±1.2)	93.8	95.4	94.6	0.015 (significant)
k-NN	88.2 (±1.8)	87.4	89.1	88.2	0.001 (significant)
LR	90.3 (±1.5)	89.5	91.2	90.3	0.007 (significant)
LSTM	95.6 (±1.1)	94.9	96.1	95.5	0.042 (significant)
1D-CNN	96.3 (±0.9)	95.7	97.0	96.3	0.065 (not significant)
MLP	93.4 (±1.3)	92.5	94.0	93.2	0.004 (significant)
ImmunoNet	98.1 (±0.7)	97.5	98.4	97.9	- (reference)

performance can have significant real-world implications. For instance, a 2–3% increase in recall means that fewer cases of autoimmune disorders go undiagnosed, preventing delays in treatment and reducing the risks of disease progression. Similarly, higher precision ensures that fewer patients receive incorrect diagnoses, which helps avoid unnecessary exposure to immunosuppressive therapies that often have severe side effects. Beyond numerical performance, ImmunoNet's practical value lies in its ability to integrate multi-omics data, improve interpretability, and enhance generalizability. Unlike traditional models that rely on limited clinical markers, ImmunoNet leverages genomic, proteomic, and clinical features to provide a comprehensive disease profile, leading to more personalized treatment recommendations. Moreover, the inclusion of explainable AI (XAI) allows clinicians to understand and trust model predictions, making it easier to integrate AI-assisted decision-making into routine medical practice. Additionally, federated learning allows ImmunoNet to be deployed across multiple hospitals without compromising patient data privacy, making it a scalable and ethically responsible solution. Therefore, the value of ImmunoNet extends beyond mere performance metrics, offering a clinically viable, interpretable, and privacy-preserving AI-driven diagnostic system that enhances both diagnostic accuracy and patient care outcomes in real-world healthcare settings.

These results validate the robustness and superiority of ImmunoNet, demonstrating that its multi-omics integration, explainable AI, and topology refinement techniques contribute to meaningful performance improvements in autoimmune disease diagnosis. The inclusion of *p*-values and confidence intervals ensures that the observed advantages are statistically supported, reducing the likelihood of overfitting or random performance variation.

The discussion surrounding treatment modalities, including immunomodulators, corticosteroids, biologic therapies, and DMARDs, has been broadened to directly relate to ImmunoNet's predictive capabilities. ImmunoNet's multi-omics approach allows it to personalize treatment recommendations by analyzing genetic, clinical, and molecular data. Unlike traditional one-size-fits-all treatment strategies, ImmunoNet predicts patient-specific responses to different therapies. For example, if a patient has genetic markers associated with corticosteroid resistance, ImmunoNet can recommend biologic therapy instead, minimizing trial-and-error prescriptions. Additionally, treatment adherence prediction is integrated into the model by analyzing historical medical data and behavioral patterns. Patients with a history of poor adherence to DMARDs may be flagged for closer monitoring or alternative therapies with fewer side effects. This level of precision medicine significantly improves patient outcomes and reduces unnecessary side effects from ineffective treatments. Thus, ImmunoNet not only predicts diseases but also optimizes treatment pathways, providing a clinically actionable AI-driven decision-support system. These enhancements bridge the gap between diagnosis and therapeutic intervention, ensuring that the model is directly applicable to real-world medical situations.

3.1.5 Ablation study

An ablation study was conducted to evaluate the impact of key components in ImmunoNet. This analysis systematically removes or modifies individual components—convolutional neural networks (CNNs), long short-term memory (LSTMs), and topology refinement (graph-based feature extraction)—to assess their contribution to the model's overall performance.

Experimental Setup.

The following model variations were tested:

- Full ImmunoNet (Baseline Model) – CNN + LSTM + Topology Refinement
- CNN-only Model – Only CNN layers, removing LSTM and topology refinement
- CNN + LSTM Model – Without topology refinement, evaluating CNN + LSTM contribution
- CNN + Topology Refinement Model – Without LSTM, assessing topology enhancement effect
- LSTM-only Model – No CNN, focusing on temporal dependencies

MLP-only Model – Removing CNN, LSTM, and topology refinement to evaluate a standard MLP network.

Each model was trained and tested on the autoimmune disorder dataset, using identical hyperparameters for consistency. Performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC. Table 8 shows the Ablation Study Results.

CNNs significantly improve classification accuracy (from 87.4% in MLP-only to 92.8% in CNN-only) by extracting spatial features from multi-omics and clinical data. LSTMs enhance

TABLE 8 Ablation study results.

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Full ImmunoNet (CNN + LSTM + Topology)	98.1	97.5	98.4	97.9	0.99
CNN-only (No LSTM, No Topology)	92.8	91.3	93.5	92.4	0.94
CNN + LSTM (No Topology)	95.6	94.9	96.1	95.5	0.97
CNN + Topology (No LSTM)	96.3	95.7	97.0	96.3	0.98
LSTM-only (No CNN, No Topology)	90.1	89.0	91.3	90.1	0.92
MLP-only (No CNN, No LSTM, No Topology)	87.4	86.5	88.0	87.2	0.90

time-dependent feature representation (CNN-only: 92.8% → CNN + LSTM: 95.6%), highlighting the importance of capturing temporal trends in disease progression. Topology Refinement provides the greatest increase in predictive power (CNN + LSTM: 95.6% → Full ImmunoNet: 98.1%), demonstrating that integrating graph-based feature relationships improves classification and model generalization.

LSTM-only models tend to underperform relative to CNN-based models, showing that while temporal dependencies are important, the spatial and hierarchical features captured by CNNs are even more critical for accurate diagnosis.

MLP-only models perform the poorest, confirming that deep learning architectures with specialized layers (CNN, LSTM, and topology refinement) significantly outperform traditional dense networks in autoimmune disease classification.

4 Conclusion

This study elucidates the landscape of autoimmune disease diagnosis and treatment, comprehensively covering disease profiles and management strategies. By meticulously examining patient data related to statistical methodology, we have discovered numerous specific patterns and predictive factors of autoimmune diseases. The key takeaway from our study is that advanced machine learning techniques, such as ImmunoNet, enhance diagnostic accuracy and prognostic ability. As a result, doctors, clinicians, and healthcare providers can use our discussion of treatment results to improve their medical practices for people with autoimmune conditions. By specifying the efficacy, safety, and patient satisfaction associated with various treatment modalities, we advocate for evidence-based personalized medicine tailored to individual patient needs and preferences. Although we present significant advancements in understanding autoimmune diseases, the study remains limited in its accuracy.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material; further inquiries can be directed to the corresponding author.

Author contributions

RU: Conceptualization, Formal analysis, Supervision, Writing – original draft, Writing – review & editing. NS: Conceptualization, Validation, Writing – original draft, Writing – review & editing. MNA: Conceptualization, Software, Resources, Validation, Writing – original draft, Writing – review & editing. AAA: Conceptualization, Software, Resources, Validation, Writing – original draft, Writing – review & editing. HSA: Formal analysis, Software, Resources, Validation, Writing – original draft, Writing – review & editing. MK: Formal analysis, Software, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. AA: Formal analysis, Software, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Korkalainen H, Leppänen T, Duce B, Kainulainen S, Aakko J, Leino A, et al. Detailed assessment of sleep architecture with deep learning and shorter epoch-to-epoch duration reveals sleep fragmentation of patients with obstructive sleep apnea. *IEEE Trans Biomed Eng*. (2021) 68:2567–74. doi: 10.1109/JBHI.2020.3043507

2. Oh JH, Lee D-J, Ji C-H, Shin D-H, Han J-W, Son Y-H, et al. Graph-based conditional generative adversarial networks for major depressive disorder diagnosis with synthetic functional brain network generation. *IEEE Trans Med Imaging*. (2024) 43:1504–15. doi: 10.1109/JBHI.2023.3340325

3. Noman F, Ting CM, Kang H, Phan RCW, Ombao H. Graph autoencoders for embedding learning in brain networks and major depressive disorder identification. *IEEE Trans Med Imaging*. (2024) 43:1644–55. doi: 10.1109/JBHI.2024.3351177

4. Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Health Inform*. (2015) 19:1209–15. doi: 10.1109/JBHI.2015.2406883

5. Cai G, Zhang F, Yang B, Huang S, Ma T. Manifold learning-based common spatial pattern for EEG signal classification. *IEEE Trans Med Imaging*. (2024) 43:1971–81. doi: 10.1109/JBHI.2024.3357995

6. Jeong JH, Lee I-G, Kim S-K, Kam T-E, Lee S-W, Lee E. DeepHealthNet: adolescent obesity prediction system based on a deep learning framework. *IEEE Trans Med Imaging*. (2024) 43:2282–93. doi: 10.1109/JBHI.2024.3356580

7. Loftness BC, Halvorson-Phelan J, O'Leary A, Bradshaw C, Prytherch S, Berman I, et al. The ChAMP app: a scalable mHealth Technology for Detecting Digital Phenotypes of early childhood mental health. *IEEE Trans Biomed Eng*. (2024) 71:2304–13. doi: 10.1101/2023.01.19.23284753

8. Eke CS, Jammeh E, Li X, Carroll C, Pearson S, Ifeachor E. Early detection of Alzheimer's disease with blood plasma proteins using support vector machines. *IEEE Trans Biomed Eng*. (2021) 25:218–26. doi: 10.1109/JBHI.2020.2984355

9. Kumar S. V.jay, Shwetha V., and Automatic classification of ana hep-2 immunofluorescence images based on the texture features using artificial neural network. In 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), IEEE. (2019).

10. Mahammad A. B., Kumar R. Design a linear classification model with support vector machine algorithm on autoimmune disease data. In 2022 3rd international conference on intelligent engineering and management (ICIEM), IEEE (2022).

11. Natrayan L., Socrates S., Bhavani Bharathi G., Srinivas Aluvala A Framework for automated diagnosis and management of autoimmune disorders with neural networks. In 2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC). IEEE, (2024).

12. Pastore VP, Touijer L, Capurro N, Cozzani E, Gasparini G, Parodi A, et al. Incorporating diagnostic prior with segmentation: a deep learning pipeline for the automatic classification of autoimmune bullous skin diseases In: In 2023 IEEE 20th international symposium on biomedical imaging (ISBI), IEEE (2023)

13. Pezoulas V. C., Goules A., Tzioufas A. G., Fotiadis D. I. An explainable and trustworthy ai framework for federated learning: a case study in rare autoimmune diseases. In 2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology. IEEE, (2023).

14. Sah A., Sarkar D., Shah S. Autoantibodies: powerful biomarkers in cancer and autoimmune disease precision medicine. In 2023 2nd International Conference on Ambient Intelligence in Health Care (ICAIHC). IEEE, (2023).

15. Salamah Y., Asyifa R. D., Afifah T. Y., Maulana F., Asfarian A.. Thymun: smart mobile health platform for the autoimmune community to improve the health and well-being of autoimmune sufferers in Indonesia. In 2020 8th International Conference on Information and Communication Technology (ICoICT). IEEE, (2020).

16. Santhoshkumar S., Ramasamy U., Mansuour R. F., Ramaraj E. A review on statistical importance and biomarkers identification in Hashimoto thyroiditis disease. In 2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence). IEEE, (2021).

17. Vedula V.. Analyzing sex-biased gene expression in autoimmune diseases. In 2021 IEEE Integrated STEM Education Conference (ISEC), IEEE, (2021).

18. Vinnarasi P., Menaka K. Identifying the impact of 25ohd and other factors on tsh using optimal kernel svm approach. In 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, (2023).

19. Vivona L., Cascio D.. Unsupervised clustering method for pattern recognition in iif images. In 2016 International image processing, Applications and Systems (IPAS), IEEE, (2016).

20. Şentürk D., Orman G. K. Detecting genetic disposition of ethnicity to autoimmune diseases via clustering. In 2021 IEEE international conference on big data (big data), IEEE. (2021).

21. Panayides AS, Amini A, Filipovic ND, Sharma A, Tsaftaris SA, Young A, et al. AI in medical imaging informatics: current challenges and future directions. *IEEE J Biomed Health Inform*. (2020) 24:2735–45.

22. Qiu J, Li L, Sun J, Peng J, Shi P, Zhang R, et al. Large AI models in health informatics: applications, challenges, and the future. *IEEE Trans Biomed Eng*. (2021) 68:973–83.

23. Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. *IEEE J Biomed Health Inform*. (2017) 21:4–21. doi: 10.1109/JBHI.2016.2636665

24. Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. *IEEE Trans Biomed Eng*. (2020) 67:1191–202. doi: 10.1109/JBHI.2020.3022989

25. Alkhodari M, Hadjileontiadis LJ, Khandoker AH. Identification of congenital Valvular murmurs in Young patients using deep learning-based attention transformers and phonocardiograms. *IEEE Trans Biomed Eng*. (2021) 68:112–23. doi: 10.1109/JBHI.2024.3357506

26. Pandiyan T, Ratti CJ, Sakamuri KM, Monteiro MC. Artificial intelligence in autoimmune disease diagnosis and treatment: a deep learning perspective. *Front Oncol*. (2023) 13. doi: 10.3389/fonc.2023.1225490/full

27. Choudhry IA, Iqbal S, Alhussein M, Aurangzeb K, Qureshi AN, Anwar MS, et al. Privacy-preserving AI for early diagnosis of thoracic diseases using IoTs: A federated learning approach with multi-headed self-attention for facilitating cross-institutional study. *Internet of Things*, (2024). 27:101296. doi: 10.1016/j.iot.2024.101296

28. Kumar S. R., Tyagi F., Hasija Y.. In-silico medication of vitiligo by targeting 6aah protein and riboflavin ligand. In 2023 2nd international conference on smart technologies and Systems for Next Generation Computing (ICSTSN). IEEE, (2023).

29. Liu L., Tao J., Yang Z., Towfic F. Shared genetic architecture in autoimmune disease - preliminary analysis. In 2015 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE. (2015).

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 04 April 2025
DOI 10.3389/fmed.2025.1534551

OPEN ACCESS

EDITED BY
Ateeq Ur Rehman,
Gachon University, Republic of Korea

REVIEWED BY
Upinder Kaur,
Lovely Professional University, India
Bhavna Sareen,
Chitkara University, India

*CORRESPONDENCE
Aziz Nanthaamornphong
✉ aziz.n@phuket.psu.ac.th

RECEIVED 26 November 2024
ACCEPTED 24 March 2025
PUBLISHED 04 April 2025

CITATION
Kumar A, Masud M, Alsharif MH, Gaur N and Nanthaamornphong A (2025) Integrating 6G technology in smart hospitals: challenges and opportunities for enhanced healthcare services. *Front. Med.* 12:1534551. doi: 10.3389/fmed.2025.1534551

Integrating 6G technology in smart hospitals: challenges and opportunities for enhanced healthcare services

Arun Kumar[1], Mehedi Masud[2], Mohammed H. Alsharif[3], Nishant Gaur[4] and Aziz Nanthaamornphong[5]*

[1]Department of Electronics and Communication Engineering, New Horizon College of Engineering, Bengaluru, India, [2]Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia, [3]Department of AI Convergence Electronic Engineering, Sejong University, Seoul, Republic of Korea, [4]Department of Physics, JECRC University, Jaipur, India, [5]College of Computing, Prince of Songkla University, Phuket, Thailand

Introduction: The advent of sixth-generation (6G) wireless communication technology promises to transform various sectors, with healthcare—particularly smart hospitals—standing to gain significantly. This study investigates the transformative potential of 6G in healthcare by exploring its architectural foundations and enabling technologies.

Methods: A comprehensive review and analysis were conducted on current technological trends, frameworks, and integration strategies relevant to 6G-enabled healthcare systems. The proposed model integrates key technologies such as the Internet of Things (IoT), artificial intelligence (AI), blockchain, robotics, telemedicine, and advanced data analytics within the context of smart hospitals.

Results: The findings suggest that 6G's ultralow latency, massive device connectivity, and high data throughput can dramatically enhance patient care, real-time monitoring, and hospital operational efficiency. The proposed 6G-based smart hospital model fosters seamless communication between medical devices and systems, enabling intelligent decision-making and optimized resource allocation.

Discussion: Despite the promising benefits, several challenges were identified, including data privacy and security risks, system interoperability, and ethical implications. The study underscores the critical importance of robust regulatory frameworks and standardized protocols to ensure secure and ethical deployment of 6G technologies in healthcare settings.

Conclusion: By providing a forward-looking analysis of the opportunities and challenges associated with 6G-powered smart hospitals, this research offers valuable insights into the evolving landscape of digital healthcare and its potential to redefine patient care and hospital management in the near future.

KEYWORDS

health care, telemedicine, 6G, smart hospital, AI, IoT, advanced waveforms, big data

1 Introduction

As an evolutionary successor to fifth-generation (5G) technology, 6G represents a significant advancement in wireless communication. It is distinguished by ultrafast data speeds, virtually zero latency, and the capability to support an unprecedented number of connected devices (1). In the context of smart hospitals, the infusion of 6G facilitates real-time communication among a myriad of medical devices, sensors, and systems, laying the foundation for a seamlessly interconnected healthcare ecosystem. This study

meticulously predicts the evolution of smart hospitals in the 6G era, shedding light on the intricate network of technologies underpinning this transformative healthcare model. The future of healthcare is entering an era of unprecedented connectivity and technological sophistication, with the imminent arrival of 6G, the sixth generation of wireless communication (2). As we stand on the cusp of this groundbreaking evolution, 6G-based smart healthcare has been poised to revolutionize the medical landscape, offering unparalleled speed, reliability, and transformative capabilities. The core of 6G's potential impact on healthcare is its ability to provide ultrafast data transmission and remarkably low latency (3). These features are critical for enabling real-time communication between medical devices, facilitating the rapid exchange of patient data, and supporting responsive, time-sensitive applications. With the ability to transmit massive amounts of data at lightning speed, 6G sets the stage for a healthcare ecosystem that is not only interconnected, but also operates with unparalleled efficiency. The integration of 6G technology into smart healthcare systems promises to enhance remote patient monitoring, diagnostics, and treatment planning (4). Medical professionals will have access to real-time, high-resolution data, enabling more accurate and timely decision making. This capability is particularly crucial in emergency situations where split-second decisions can significantly affect patient outcomes (5). The effect of 6G on telemedicine was also set to be transformative. Enhanced connectivity facilitates seamless and immersive virtual healthcare experiences, allowing for high-quality video consultation, remote surgery, and interactive patient engagement (6). The geographical barriers that traditionally have limited access to healthcare services will be further dismantled, providing individuals in remote or underserved areas with unprecedented access to medical expertise (7). Furthermore, the integration of 6G with advanced techniques can create a network of interconnected medical devices and wearables, fostering continuous and comprehensive health monitoring. This interconnected ecosystem will contribute to a holistic approach to healthcare, providing a more complete picture of an individual's health and enabling personalized, data-driven interventions. The advent of 6G technology has heralded a new era for smart healthcare, promising to transform the way we access, deliver, and experience medical care (8). The convergence of ultrafast communication, real-time data transmission, and seamless connectivity positions 6G as a catalyst for a healthcare revolution, ushering in an era of unprecedented efficiency, accessibility, and personalized health management. In essence, this comprehensive study embarks on an expedition into the future of healthcare, a future where 6G-based smart hospitals transcend traditional boundaries, ushering in an era of unparalleled connectivity, efficiency, and patient-centric care. Through an examination of architectural evolution, advanced techniques, and challenges, this research seeks to unravel the intricate interplay between technology and healthcare, laying the groundwork for a transformative journey into the era of 6G-enabled smart healthcare ecosystems (9). The integrations of the 6G technology into smart hospitals offers transformative potential by leveraging its ultra-high-speed connectivity, low latency, and massive device connectivity. This integration enables real-time data transmission and processing, facilitating advanced applications such as remote surgeries, AI-driven diagnostics, and enhanced telemedicine services. With 6G, healthcare providers can utilize edge computing to process data locally, reducing latency and ensuring rapid decision-making. The deployment of smart sensors and IoT devices throughout hospital infrastructure allows for continuous patient monitoring, predictive maintenance of medical equipment, and efficient resource management. Furthermore, 6G's enhanced security features ensure the protection of sensitive patient data, mitigating risks associated with cyber threats. The technology also supports seamless communication between various hospital departments, improving operational efficiency and patient care coordination. However, challenges such as the need for significant infrastructure upgrades, high implementation costs, and the requirement for healthcare professionals to adapt to new technologies must be addressed. Despite these challenges, the integration of 6G in smart hospitals presents an opportunity to revolutionize healthcare delivery, offering personalized, efficient, and secure medical services tailored to the needs of individual patients.

The integration of 6G technology in smart hospitals presents significant challenges, primarily due to the need for extensive infrastructure upgrades, high costs, and the complexity of managing vast amounts of data. The deployment of 6G requires a robust network infrastructure capable of supporting ultra-low latency, high data rates, and massive device connectivity. However, existing hospital infrastructure may not be equipped to handle these demands, necessitating substantial investments in new technology, including advanced routers, servers, and edge computing devices. Additionally, the cost of implementing 6G technology can be prohibitive, particularly for smaller or less-resourced healthcare facilities. Another major challenge is the management of the enormous amounts of data generated by 6G-enabled devices, which requires sophisticated data processing and storage solutions to ensure efficient operation.

Advanced techniques like IoT, AI, blockchain, telemedicine, robotics, and advanced data analytics play crucial roles in overcoming the challenges of integrating 6G technology into smart hospitals. IoT enables seamless connectivity between medical devices and systems, ensuring real-time monitoring and data collection from patients, which 6G can then transmit and process at unprecedented speeds. This reduces latency issues and enhances the responsiveness of healthcare services. AI aids in managing the vast amounts of data generated by IoT devices, analyzing patterns for predictive diagnostics, personalized treatment plans, and efficient resource allocation. By automating complex tasks, AI helps alleviate the burden on healthcare professionals, allowing them to focus more on patient care. Blockchain technology addresses security concerns by providing a decentralized and immutable ledger for patient records, ensuring data integrity and privacy. This is particularly important in a 6G-powered environment where data exchange is rapid and extensive. Telemedicine, supported by 6G's low latency, becomes more reliable, enabling high-quality remote consultations and even remote surgeries, expanding access to specialized care regardless of location. Robotics integrated with 6G allows for more precise and real-time control in surgical procedures, improving outcomes while reducing the risk of human error. Finally, advanced data analytics enables hospitals to process and interpret large datasets quickly, offering insights that can lead to improved patient outcomes and operational efficiency. By leveraging these advanced technologies, the challenges of implementing 6G in smart hospitals—such as infrastructure demands, high costs, and the complexity of managing vast amounts

of data—can be effectively mitigated, paving the way for a more connected and intelligent healthcare system.

1.1 Motivation

Conventional smart hospitals face hurdles such as limited connectivity, slow data transmission, and inadequate support for real-time applications. These issues hinder efficient remote care, timely decision-making, and seamless integration of advanced technologies like AI and IoT. 5G-based smart hospitals can address these challenges with ultra-fast data speeds, low latency, and massive device connectivity. 5G enables real-time telemedicine and remote surgeries by ensuring instantaneous communication between doctors and patients or robotic systems. It also supports the Internet of Medical Things (IoMT), allowing continuous monitoring and automated alerts for critical conditions. The high data capacity of 5G allows for rapid sharing of large medical files, such as MRI scans, facilitating faster diagnoses. Additionally, 5G improves hospital efficiency by enabling smart systems for managing resources, equipment, and staff, reducing operational delays. By enhancing speed, reliability, and device integration, 5G can resolve many challenges of conventional smart hospitals, significantly improving patient care. The integration of 6G technology in smart hospitals is motivated by the need to address the ever-growing demand for advanced, efficient, and personalized healthcare services. As healthcare systems face challenges such as an aging population, chronic diseases, and pandemics, the current infrastructure often falls short in delivering timely and effective care. 6G technology, with its unparalleled data transmission speeds, ultra-low latency, and massive connectivity, promises to revolutionize healthcare by enabling real-time monitoring, remote surgeries, and AI-driven diagnostics. The potential for enhanced communication between devices, patients, and healthcare providers can lead to more accurate and timely medical interventions. However, the adoption of 6G in healthcare also presents challenges, including concerns about data security, high costs of implementation, and the need for robust regulatory frameworks. Despite these hurdles, the opportunities offered by 6G technology—such as improved patient outcomes, reduced healthcare costs, and the facilitation of telemedicine—make it a critical component in the evolution of smart hospitals and the future of healthcare delivery. The structure of this paper is as follows: Section 1 provides the definition of smart healthcare with respect to 6G, the significance of smart healthcare in the modern era, the evolution and adoption of smart healthcare technologies with 6G, and the challenges facing the implementation of future 6G-centered healthcare facilities. In Section 2, we critically examine and analyze existing scholarly works on a specific topic. It provides a comprehensive overview of relevant research, identifying gaps, trends, and insights to inform and contextualize a new study or research endeavor. Includes an article published in this field. Section 3 focuses on the integration of several technologies into a 6G-based smart hospital. The benefits of 6G for smart hospitals are described, and the differences between 5G and 6G and their benefits owing to the differences in some quantitative performances are tabulated. Section 4 provides the perspective of advanced technologies, such as Internet of Things (IoT), explainable artificial intelligence (AI) in 6G, which will play an important role in future smart hospitals. The significance of prospective technology for 6G-based smart hospitals lies in its potential to revolutionize healthcare, as described in Section 4. With ultrafast communication, low latency, and massive device connectivity, 6G can enhance telemedicine, enable real-time diagnostics, support advanced robotics, and foster personalized patient care, ultimately improving healthcare efficiency and outcomes. Additionally, the architecture and different layers of advanced techniques are comprehensively discussed. Additionally, the challenges in 6G-based smart hospitals include ensuring robust cybersecurity to protect sensitive health data, addressing interoperability issues among diverse devices and systems, managing the massive influx of data, and overcoming potential ethical concerns related to advanced healthcare technologies. Finally, Section 5 outlines the integration of 6G technology in smart hospitals coupled with advanced techniques, which promises unprecedented improvements in healthcare. Furthermore, future work on security and privacy are highlighted. The contributions of the projected article are given below:

- The article explores how 6G enables seamless connectivity between IoT devices within smart hospitals, facilitating real-time data collection, remote monitoring, and automated management of medical equipment and patient health data.
- It highlights the role of 6G in enhancing AI capabilities, enabling faster processing of large datasets for diagnostics, personalized treatment plans, and predictive analytics, leading to improved patient outcomes.
- The article discusses the potential of 6G to strengthen blockchain applications in healthcare, ensuring secure and transparent management of patient records, reducing the risk of data breaches, and improving trust in data sharing across healthcare systems.
- It examines how 6G can revolutionize telemedicine by providing ultra-low latency and high-definition video streaming, enabling real-time, remote consultations, and even remote surgeries, thereby expanding access to quality healthcare services.
- The article delves into the use of 6G in supporting robotic systems for surgery, rehabilitation, and patient care within smart hospitals, offering precise, reliable, and safe medical procedures with minimal human intervention.
- It discusses how 6G enhances advanced data analytics by enabling the rapid processing of vast amounts of healthcare data, facilitating insights into patient health trends, resource allocation, and overall hospital management.

These contributions collectively underline the potential of 6G technology to transform healthcare delivery in smart hospitals, addressing challenges while opening new opportunities for enhanced, efficient, and secure.

2 Literature review

In the integration of 6G technology in smart hospitals, the starting point is to study the challenges and solutions deployment of advanced technologies such as IoT, AI, blockchain, telemedicine,

robotics, and data analytics. These technologies serve as the input, enabling real-time patient monitoring, automated diagnostics, secure data management, and efficient remote consultations. The end point is the enhanced healthcare delivery system, characterized by improved patient outcomes, streamlined hospital operations, and robust data security. By leveraging 6G's ultra-fast connectivity and low latency, smart hospitals can achieve seamless integration of these technologies, leading to more personalized, efficient, and effective healthcare services. In this section, we present a critical and comprehensive analysis of existing literature (published academic works, articles, books, and other sources) on smart healthcare. It summarizes and synthesizes key findings, theories, and methodologies from existing studies and scholarly works. The rapid development of cellular connection systems has greatly accelerated the evolution and implementation of remote health monitoring and smart healthcare. The advanced long-term evolution (A-LTE) network now underpins modern healthcare systems. However, the development of smart hospitals and healthcare institutions is still nascent on a global scale. The introduction of the 5G network is set to elevate the standards of intelligent healthcare. Smart hospitals have distinct requirements compared to other applications in sectors like business, education, and general public services. This research evaluates how IoT and 5G will underpin the future "smart hospital," anticipated to enhance throughput, efficacy, and coverage. The study focuses on implementing a hybrid detection technique for massive multiple-input multiple-output (MIMO) and non-orthogonal multiple access (NOMA) systems using QR decomposition and the M algorithm-maximum likelihood detection (QRM-MLD) combined with beamforming (BF). This approach aims to improve latency, spectrum efficiency, and network throughput in 5G systems. Additionally, the work provides a comparison between the proposed and traditional detection methods (10). The OFDM waveform method is pivotal in the context of smart hospitals, though it faces challenges such as bandwidth loss from guard bands, spectrum leakage, high Peak-to-Average Power Ratio (PAPR), and significant detection latency, which undermine its effectiveness. As 5G deployment becomes increasingly widespread globally, its advanced radio systems are expected to fulfill the comprehensive needs of smart healthcare facilities, which include high spectrum access, large capacity, great throughput, and low PAPR. The demand for bandwidth in digital hospitals has surged, necessitating networks that operate at peak efficiencies for tasks ranging from transmitting medical images to interfacing with wearable devices to ensure optimal patient care. The transition to digital hospitals with 5G connectivity will be critically shaped by the adoption of reliable transmission technologies. Current efforts are primarily focused on the implementation of innovative waveforms like NOM, UFMC, and FBMC systems. This work involves a detailed analysis and study of several parameters, including power spectrum density, bit error rate, capacity, and PAPR of both advanced waveforms and traditional OFDM techniques (11). This paper outlines the system architecture resulting from the integration of IoT technology in smart healthcare environments, detailing optimization considerations, challenges, and viable solutions. The technological infrastructure is divided into five distinct levels, with each layer's architecture, limitations, and methods thoroughly examined. This includes the size of the smart hospital, the scope of its intelligent computing capabilities, and the extent of its real-time big data analytics. The findings from the study are utilized to identify potential flaws in each tier of the smart hospital design model and suggest necessary adjustments. The document aims to serve as a comprehensive guide for managers, system engineers, and academics interested in optimizing the design of smart hospital systems, providing them with a clear road map for improvement (12). In this study, stochastic Petri nets were employed to evaluate the functionality and availability of a smart hospital system without the initial need for financial investment in actual equipment. These models are highly parametric, allowing for the adjustment of resource capacity, service times, failure and repair intervals, and the duration between failures. The initial model permits the configuration of several parameters, enabling the assessment of various scenarios. The investigation results highlighted the arrival rate as a crucial system characteristic. Particularly in scenarios with high arrival rates, a significant correlation was observed between Mean Response Time (MRT), resource utilization, and discard rate, demonstrating the impact of these factors on system performance (13). The article outlines the design principles for a health service platform app, including the health information perception terminal. With the advancement of big data, cloud computing, and information technology, the concept of smart healthcare has become increasingly significant. This new model, referred to as a health service platform, is gaining popularity and proving more practical compared to traditional healthcare services. The effectiveness of health monitoring is being enhanced through the use of wearable devices and various apps. There is a pressing need for an efficient and practical app-based health service platform that can cater to both older and younger populations, aiming to augment and streamline smart healthcare services (14). The article underscores the imperative for a robust and practical app-based health service platform that caters to both older and younger demographics, aimed at significantly enhancing and facilitating smart healthcare services. Building upon foundational concepts, it explores the design principles of the health service system and the health information perception terminal within this platform. The discussion extends to various aspects of the developed systems, including the unique contributions of each framework, detailed operational processes, performance outcomes, and the strengths and limitations inherent in these systems. Furthermore, the article addresses prevailing research challenges, critically evaluating the shortcomings of current systems and proposing prospective directions for advancement. This analysis is intended to furnish comprehensive insights into contemporary developments in smart healthcare systems, thereby equipping professionals with the knowledge necessary to make meaningful contributions to the field (15). This paper explores the advantages of cloud computing for healthcare applications, detailing IoT architectures, various communication protocols, sensor technologies, and both machine learning and deep learning techniques. It provides a comprehensive review of their respective benefits, limitations, and challenges. This study equips researchers with the necessary insights, enabling them to initiate their investigations by choosing a specific application or topic from the discussed methodologies. With strict adherence to security and privacy measures, cloud-based IoT and ML

healthcare systems prove to be accurate and immensely beneficial for patients, caregivers, and hospital staff (16). The article explores potential challenges and market adoption barriers for IoT-based healthcare from both patient and professional perspectives. It addresses key issues such as interoperability, standardization, compensation, data storage, control and ownership, as well as trust and acceptability. To overcome these challenges, the paper suggests that contemporary healthcare will need to depend on policy support, regulation focused on cybersecurity, strategic caution, and the adoption of transparent policies within healthcare organizations to enable IoT solutions. Implementing IoT-based healthcare could significantly enhance population health and the efficiency of healthcare systems (17). As information technology advances, the concept of smart healthcare has increasingly captured interest. Smart healthcare revolutionizes the traditional medical system by leveraging cutting-edge information technologies such as the Internet of Things (IoT), big data, cloud computing, and artificial intelligence. These technologies enhance the efficiency, convenience, and personalization of healthcare services. In this review, the authors first outline the key technologies that underpin smart healthcare. We then explore the current state of smart healthcare across various significant domains. Lastly, the article addresses the current challenges faced by smart healthcare and offers recommendations for overcoming these obstacles (18). The article examines the potential of IoT technology to alleviate pressures on healthcare systems caused by an aging population and the rise of chronic diseases. It identifies standardization as a critical barrier to success in this area and proposes a standardized model for future IoT healthcare systems. The paper then reviews recent research on each element of this model, providing an evaluation of its benefits, drawbacks, and suitability for wearable IoT healthcare applications. Key challenges such as security, privacy, wearability, and low-power operation are addressed. The article concludes with recommendations for future research directions in this evolving field (19). The article addresses several barriers hindering the integration of IoT applications in healthcare. These include the generation of large volumes of non-essential data, concerns regarding patient data security and privacy, and the substantial costs associated with IoT adoption. It highlights the role of prosthetic sensors, which collect relevant data to aid real-time patient treatment, as a promising area for future research. This study underscores the potential of IoT to enhance healthcare delivery by focusing on specific, impactful applications (20). This research presents a fresh technique and develops an IoT-based prototype. Then, an elaborate theoretical framework was developed from this a cutting-edge prototype that demonstrates how the I-CARES system actually works. The system offers ongoing health status monitoring and analysis, as well as automatic, real-time emergency action that may ultimately save lives. It also gives information on pharmaceutical effects, side effects, and the patient's health state (21). This paper provides an in-depth examination of current research projects and the application of various technologies in smart healthcare systems. It delves into the latest studies, proposed methodologies, and existing solutions in the realm of smart healthcare, focusing on the implications of emerging technologies, applications, and challenges these systems face today and in the future. The aim is to present a comprehensive view of what IoT currently offers to the healthcare sector and what it promises for the future (22). This work meticulously examines the challenges at each stage of the big data handling process, which necessitate the use of advanced computing technologies for resolution. It argues that healthcare providers must be adequately equipped with the essential infrastructure to regularly generate and analyze big data, in order to develop strategies that enhance public health. Additionally, the paper highlights that contemporary healthcare institutions could revolutionize medical treatments and personalized medicine through a robust integration of biomedical and healthcare data (23). This paper addresses the privacy and security concerns associated with future healthcare applications, as highlighted in the study. The advent of fifth-generation networks is propelling the expansion of telehealth and smart healthcare solutions. Fundamental elements such as Quality of Life, Intelligent Wearable Devices, the Intelligent Internet of Medical Things, Hospital-to-Home transitions, and innovative business models are shaping the future of AI-driven intelligent healthcare. Many academic studies consider 6G technology a vital enabler of intelligent healthcare systems. Furthermore, Body Area Networks with integrated mobile health systems are evolving toward personalized health management and monitoring. Additionally, Extended Reality, a novel immersive technology, merges the real and virtual worlds, enabling enhanced interaction between computers, wearables, humans, and other machines (24). As the volume of daily-generated data expands in the 6G-enabled Internet of Medical Things (IoMT), the process of medical diagnosis becomes increasingly critical. This study, referenced in Wijethilaka et al. (25), develops a methodology aimed at enhancing prediction accuracy and facilitating real-time medical diagnosis within the 6G-enabled IoMT framework. The proposed approach integrates optimization techniques with deep learning methodologies to deliver precise and reliable outcomes. During the process, medical computed tomography images undergo preprocessing before being input into a sophisticated neural network designed to learn image representations and convert each image into a feature vector. Subsequently, a MobileNetV3 architecture is employed to further learn and refine the features extracted from these images (26). The 6G-Health project aims to foster precision technology development within the realm of sixth-generation mobile communications (6G) by integrating the expertise of communication engineering, medical engineering, and technical end users. The project's scope includes not only the development of specific 6G technological components but also the early identification and mitigation of market entry barriers, particularly focusing on operational elements, standards, and licensing issues. The technical framework encompasses emerging technologies that enhance network intelligence, innovative sensor connectivity for 6G, and efficient resource utilization and data processing strategies prior to their dissemination across various infrastructure levels. This paper will explore three medical applications of 6G: enhancing smart hospital operations, improving collaborative work environments, and enabling direct acquisition and transmission of bio signals from patients (27). The authors propose a Peak-to-Average Power Ratio (PAPR) reduction technique aimed at enhancing the efficiency of power amplifiers for 5G waveforms. This approach involves applying several algorithms to 5G

waveforms, with their performance evaluated through PAPR curves. In the broader context, the study concludes that hospitals can leverage AI and IoT technologies to improve efficiency, reduce costs, and enhance patient care. By adopting these technologies, hospitals are positioned to improve patient outcomes and the overall health system's performance.

3 Smart hospital

A smart hospital, also referred to as a digital or intelligent hospital, is an example of how cutting-edge technologies, data-driven strategies, and patient-centered care have come together in the healthcare sector (10). It is a paradigm-shifting idea that seeks to integrate cutting-edge technologies and intelligent systems to optimize resource usage, improve operational efficiency, and improve patient outcomes. A sophisticated digital infrastructure that allows for seamless connectivity and data sharing across different hospital systems, equipment, and stakeholders is the foundation of a smart hospital (28). The two essential elements of smart hospitals are remote patient monitoring and telemedicine. Patients can obtain remote medical consultations, diagnoses, and follow-up care with the aid of communication technology. Healthcare professionals can remotely monitor patients' vital signs and medical issues using IoT connectivity and remote monitoring equipment (29). This reduces the need for hospital stays, enhances access to healthcare services, and permits continuous care, especially for patients with chronic illnesses (30). Smart hospitals prioritize patient empowerment and involvement using digital tools and technologies. Patients can access their health records, obtain personalized health advice, make appointments, and contact healthcare practitioners through mobile apps, patient portals, and wearable technology. These resources encourage patients to play an active role in their own care, help patients follow their treatment regimens, and help patients and healthcare teams work together. The idea of a "smart hospital" has a lot of potential, but it also has drawbacks.

The main obstacles are related to implementation costs, infrastructure needs, interoperability, and data protection. Furthermore, successful implementation depends on tackling the digital divide, negotiating regulatory frameworks, and guaranteeing that healthcare personnel integrate and accept new technologies (30). Establishing a connected healthcare environment is mostly dependent on Internet of Things (IoT) devices, cloud computing, and high-speed networks. Real-time data collection, monitoring, and analysis are made possible by these technologies, providing healthcare professionals with access to fast and reliable information for making decisions (31). Electronic health records (EHRs) are a fundamental component of smart hospitals. Electronic Health Records (EHRs) centralize and digitize patient data, including diagnoses, treatment plans, test results, and medical histories. Smart hospitals guarantee simple access to thorough and current information by digitizing patient data, which enhances care coordination and reduces medical errors. Artificial intelligence (AI) and data analytics are essential for a smart hospital operation. To extract valuable insights, advanced analytics algorithms can examine vast amounts of healthcare data, including patient records, medical imaging, and real-time monitoring data. AI-powered tools can help with tailored care, illness diagnosis, treatment planning, and clinical decision-making support. Healthcare professionals can make better judgments using machine learning algorithms that can recognize trends, forecast results, and offer recommendations. Robotics and automation are used in smart hospitals to improve patient care, increase productivity, and expedite procedures. Tasks, including pharmaceutical delivery, lab sample processing, and inventory management, are handled by robotic process automation (RPA). Surgeons are increasingly using robotic equipment to aid them in performing precise, minimally invasive surgeries known as robotic-assisted surgeries. Robotic caretakers can also assist with prescription reminders, patient monitoring, and mobility assistance (32). A smart hospital relies heavily on Internet of Things (IoT) devices to connect wearables, sensors, and medical devices. IoT-enabled gadgets gather health data, continuously check patients' vital signs, and send them to centralized platforms for analysis. Healthcare professionals can remotely monitor patient states, identify warning indications, and take immediate action through real-time monitoring. To ensure effective resource utilization, IoT devices also make asset tracking, inventory management, and medical equipment maintenance possible. Smart hospitals use cutting-edge technology, data analytics, and patient-centric strategies to bring about a paradigm shift in healthcare delivery. Smart hospitals are designed to improve patient care, increase operational efficiency, and change the healthcare experience of both patients and healthcare providers through seamless connectivity, intelligent technology, and real-time data analysis (33).

Differentiating itself from traditional hospitals, a smart hospital incorporates cutting-edge technology like IoT, AI, and big data to improve patient care, operational efficiency, and clinical outcomes. Networked equipment in smart hospitals facilitates real-time patient monitoring, enabling timely interventions. While automated technologies streamline administrative activities to reduce human error and wait times, AI-driven insights support tailored treatment plans and diagnostics. By extending care outside of the hospital, telemedicine and remote monitoring guarantee ongoing patient involvement. Conventional hospitals, on the other hand, are less able to provide the same degree of proactive, data-driven, and seamless healthcare services since they rely more on manual operations (34).

A number of enduring problems in healthcare, such as incorrect diagnosis, ineffective resource management, and patient safety, can be resolved by implementing 6G in smart hospitals. Personalized treatment regimens and improved diagnosis accuracy are achieved by advanced AI systems. Real-time information from networked devices optimizes the use of resources, easing congestion and improving patient flow. When it comes to prescribing medications and documenting clinical findings, automated technologies reduce human error. In addition, telemedicine and remote monitoring offer round-the-clock patient care, which lowers readmissions to hospitals and enhances the treatment of chronic illnesses, increasing overall health outcomes (35). The input refers to the existing or baseline infrastructure of conventional smart hospitals, including current technologies like 4G/5G networks, IoT devices, electronic health records (EHR), AI-driven healthcare solutions,

and current limitations in terms of connectivity, data management, and real-time capabilities. It also includes the introduction of 6G technology and its core features such as ultra-low latency, high data transfer rates, AI integration, and seamless device connectivity. The output refers to the anticipated improvements and advancements brought by the integration of 6G technology in smart hospitals. This includes enhanced healthcare services like real-time remote surgeries, continuous patient monitoring with IoMT, AI-driven diagnostics, personalized treatments, and more efficient hospital operations. It also encompasses overcoming current challenges, such as data privacy, cybersecurity, interoperability, and cost-related hurdles.

3.1 Infrastructure requirements for 6G-based smart hospitals

Implementing 6G technology in the health sector, especially in developing countries, has considerable cost factors in terms of investment in large-scale infrastructure. Advanced hardware, including high-frequency antennas, fiber-optic cables, and edge computing devices, needs to be deployed for the rollout, which comes with a heavy installation and maintenance cost. Also, retrofitting existing infrastructure for ultra-low latency, high-speed communication, and extensive IoT integration comes with financial costs. Regulatory compliance, cybersecurity protocols, and staff training also contribute to the costs. In the developing world, scarce resources and poor infrastructure further compound these costs, requiring public-private partnerships and foreign aid. Phased rollout and reuse of existing 4G/5G infrastructure are cost-effective options that can help reduce upfront costs. Although having high initial costs, the long-term gains—enhanced health care access, increased telemedicine, and improved health outcomes—make the investment worthwhile, especially if underpinned by creative financing schemes and government subsidies. The successful implementation of 6G-based smart hospitals will require a comprehensive infrastructure that integrates advanced connectivity, IoT devices, AI technologies, and robust cybersecurity measures to deliver high-quality, personalized healthcare services efficiently and securely (36, 37).

- 6G connectivity: the backbone of any smart hospital would be its connectivity. 6G networks, expected to offer unprecedented speeds, low latency, and massive device connectivity, will be crucial. These networks will support high-definition video streaming for telemedicine, real-time monitoring of patients' vital signs, and seamless communication between IoT devices and AI systems.
- IoT devices: smart hospitals will heavily rely on IoT devices for various applications like remote patient monitoring, asset tracking, and environmental monitoring. These devices include wearable health trackers, smart beds, smart infusion pumps, and sensors for monitoring temperature, humidity, and air quality. With 6G, these devices can transmit data faster and more reliably, facilitating real-time decision-making by healthcare providers.
- AI and machine learning: advanced AI algorithms will analyze the massive amounts of data generated by IoT devices to provide insights for personalized patient care, disease prediction, and treatment optimization. These AI systems will require powerful computational infrastructure for processing data in real-time or near real-time, which could be facilitated by edge computing nodes within the hospital network.
- Robotic systems: robots will play a significant role in smart hospitals, performing tasks such as patient assistance, drug delivery, and disinfection. These robots will be equipped with sensors and cameras for navigation and interaction with patients and staff. High-speed, low-latency 6G connectivity will enable remote operation of robots by surgeons for telesurgery, particularly in emergency situations or in remote areas lacking specialized medical expertise.
- Optical fibers: to support the high bandwidth demands of 6G networks and ensure reliable connectivity throughout the hospital premises, optical fiber infrastructure will be essential. Fiber-optic cables offer greater bandwidth and immunity to electromagnetic interference compared to traditional copper cables, making them ideal for transmitting large volumes of data at ultra-fast speeds over long distances.
- Advanced cameras and imaging systems: high-resolution cameras and imaging systems will be deployed for various applications, including monitoring patient conditions, tracking medical equipment, and enhancing security. These systems will generate large amounts of data, which will need to be transmitted and processed efficiently using 6G networks and advanced AI algorithms.
- Cybersecurity measures: with the proliferation of connected devices and sensitive patient data being transmitted over 6G networks, robust cybersecurity measures will be critical to protect against data breaches, unauthorized access, and cyber-attacks. Hospitals will need to implement encryption protocols, access controls, and intrusion detection systems to safeguard patient privacy and ensure the integrity of medical data.
- Dense networks of small cells and energy consumption: The roll-out of 6G demands huge investments in infrastructure, especially in high-density small cell networks to enable the ultra-high speeds, low latency, and massive connectivity that 6G is expected to deliver. Small cells, scattered in urban and rural environments, will provide flawless coverage and connectivity by offloading traffic from conventional macro cells, hence alleviating congestion and enhancing network dependability. Their deployment, however, calls for vast physical infrastructure, such as the building of many base stations and antennas. Energy usage is yet another significant issue, with small cells and millimeter-wave and other high-frequency communication technologies requiring significant amounts of power to keep performance steady. The around-the-clock nature of these networks combined with sophisticated AI-based management means that effective use of power is needed to not overload the grid. To reduce these risks, the adoption of energy-saving technology such as low-power chips, solar-powered bases, and intelligent grid systems will be required. Moreover, improving network design by

software-defined networking (SDN) and network slicing can further minimize energy usage with high performance.

- Cost implications: The infrastructure needed for 6G deployment is considerable, and it comes with high costs. Setting up a 6G network involves installing sophisticated hardware such as high-frequency antennas, small cells, massive MIMO systems, and fiber-optic backhaul links, all of which are costly to install and maintain. Moreover, creating a dense, distributed network of base stations to provide ubiquitous connectivity involves a huge investment in both urban and rural regions. The energy requirements for these systems, particularly edge computing and integration with AI, introduce additional cost complexity. In addition, maintaining cybersecurity and data privacy compliance comes at the cost of having strong security infrastructure, which adds to the overall expense. Although the advantages of 6G, including ultra-low latency, increased data rates, and enormous IoT support, are evident, the cost to governments, telecommunication companies, and healthcare systems could be high. Public-private collaborations and global funding will be necessary to balance these expenses and provide equal access to 6G technology.

Rolling out 6G infrastructure in rural and underdeveloped areas is challenging because of poor infrastructure, high expense, and a lack of technical skills. These regions usually do not have stable power grids, fiber-optic connections, and high-performance computing facilities, which hamper the implementation of 6G-based smart healthcare solutions. Moreover, the expense of installing small cells, massive MIMO antennas, and edge computing equipment is too high for governments and healthcare organizations. Socioeconomic conditions of low digital proficiency and constrained budgets for healthcare enlarge the digital gap further, restricting access to modern telemedicine, remote diagnosis, and AI-supported healthcare services. To tackle these constraints, affordable, scalable solutions will have to be given priority. Utilizing built-in 4G/5G infrastructure using network upgrades lowers the initial costs substantially. The use of low-power, solar-powered base stations can mitigate power limitations, and satellite-based internet services such as LEO constellations can provide coverage in remote locations. Furthermore, embracing open-access network architectures and software-defined networking (SDN) can reduce operating expenses and enable flexible infrastructure deployment. Public-private partnerships and international funding schemes must be promoted to finance infrastructure development and digital literacy programs. By adopting these measures, healthcare systems can close the connectivity gap so that 6G-enabled healthcare innovations reach rural and underdeveloped areas.

4 Sixth generation

The goal 6G wireless technology, which replaces 5G technology, is to improve mobile communication even more. While 5G concentrates on delivering greater speeds, reduced latency, and enhanced connectivity for Internet of Things devices, 6G is anticipated to completely transform these areas with even more breakthroughs. With terabits per second of data transport, 6G promises to outperform 5G by up to 100 times (38). By substantially reducing latency to microseconds, it will enable almost instantaneous communication. In addition, 6G will use cutting-edge technology like edge computing and artificial intelligence to enhance resource management and network performance. Furthermore, 6G will facilitate the creation of cutting-edge applications like sophisticated autonomous systems, immersive virtual reality, and augmented reality (39). Additionally, it will guarantee global digital inclusion by improving connections in underserved and distant locations. As a result, 6G will greatly increase the potential for wireless communication, outperforming 5G in terms of speed, latency, and technological integration. Global 6G standardization remains in its initial phase, with initiatives such as ITU, 3GPP, and national efforts of the U.S., China, South Korea, and the EU leading research and framework development. The emphasis lies in realizing ultra-low latency, high reliability, and massive connectivity to enable next-generation applications such as holographic communication, digital twins, and sophisticated healthcare systems. IoT, AI, and legacy healthcare systems will be integrated into 6G networks based on interoperable protocol, high-end edge computing, and slicing. Network management through AI will enhance resource utilization, forecast network faults, and make devices interoperable seamlessly. IoT healthcare devices, including remote monitoring and wearable devices, will interact in real-time to improve patient care. Legacy systems will require modular upgrades and backward-compatible interfaces to fit seamlessly. Cross-industry collaborations and joint standardization work will be pivotal to achieving safe, efficient, and ubiquitous uptake of 6G across healthcare and beyond (40).

Healthcare is changing because hospitals are implementing 5G technology, which makes data transfer and communication faster and more dependable. 5G networks currently provide much better speeds, lower latency, and increased connectivity than previous generations, all of which are essential for modern medical applications. Hospitals can monitor remote patients in real time and provide high-definition video consultations thanks to 5G telemedicine. This makes healthcare services more accessible, especially in underserved and rural areas. Moreover, 5G makes it easier to use IoT apps and cutting-edge medical devices. Smart beds, linked imaging systems, and wearable health monitors can all gather and send patient data continually, allowing for real-time monitoring and quick reactions to changes in a patient's condition. Massive amounts of data, including high-resolution medical images, can swiftly and effectively transfer to healthcare specialists for prompt diagnosis and treatment, thanks to the high bandwidth and low latency of 5G networks. The benefits of 5G extend to remotely operated medical equipment and robotic surgery. Surgeons can use robotic equipment to execute precise, minimally invasive operations even from remote locations because of 5G's ultra-reliable, low-latency transmission. This can help places without access to such resources by extending the reach of specialized medical knowledge (41). 5G has already made significant progress, but 6G has the potential to completely transform hospital operations. Even greater speeds—up to terabits per second—and microsecond-level latency will be

possible with 6G technology, which is anticipated to be operational by the 2030s. This will facilitate real-time communication and very instantaneous data transfer, both of which are critical for vital medical applications. The combination of powerful edge computing and artificial intelligence (AI) will be one of the biggest developments with 6G (42). By processing enormous volumes of medical data locally at the network's edge, these technologies will lessen the need for data to go to centralized servers. This will assist AI-driven diagnosis, individualized treatment plans, and predictive analytics to identify health issues before they become serious by increasing the speed and efficiency of data analysis. Additionally, 6G will enable more complex and immersive telemedicine applications, such as augmented reality (AR) for remote surgeries, medical education, and holographic communication. These features will improve the caliber and reach of telemedicine, increasing its effectiveness and interactivity. Additionally, 6G's increased connection will aid in the development of a more extensive and cohesive healthcare ecosystem. It will ensure smooth data flow and integration across several healthcare systems by connecting an even wider range of medical equipment and sensors. This would allow for a holistic view of patients' health, which would improve care coordination and outcomes (43).

4.1 How to integrate 6G and smart health care

To transform patient care and improve healthcare systems, 6G technology must be strategically combined with healthcare breakthroughs and cutting-edge connectivity. 6G networks' blazing speed and low latency provide the groundwork for immediate connectivity, which makes it easier to integrate various healthcare sensors and equipment. With real-time data sharing made possible by 6G's fast connectivity, IoT devices can manage medication adherence, monitor patients' vital signs, and help healthcare providers make data-driven decisions more quickly. Furthermore, 6G's capacity to deliver high-quality low-latency video communication supports telemedicine applications. Telehealth services, including virtual consultations and remote patient monitoring, are becoming increasingly effective and widely available, particularly in underprivileged or isolated places (36). Protecting the privacy and security of sensitive healthcare data is critical. Strong cybersecurity safeguards protect patient data and ensure regulatory compliance within the 6G network. These protections include encryption and secure data transmission methods. Essentially, the combination of smart healthcare with 6G creates a dynamic environment in which cutting-edge medical solutions and dependable, quick connectivity can be achieved. This synergy opens the door to a revolutionary era in the provision of patient-centric care by improving the effectiveness, accessibility, and quality of healthcare services (44). A flowchart for integrating 6G and the smart hospital is shown in Figure 1.

Investigating the use of various layer structures for cutting-edge approaches in real-world settings is crucial. Integrating 6G and smart healthcare involves leveraging the advanced capabilities of 6G networks to enhance healthcare services and enable innovative healthcare applications, as illustrated in Figure 2.

Integrating 6G technology with smart healthcare involves a systematic approach to leveraging the capabilities of advanced connectivity and healthcare innovations (2, 45–49).

- **Remote patient monitoring**: 6G technology, known for its low latency and high-speed connectivity, facilitates real-time remote patient monitoring. Healthcare providers can employ connected devices to continuously monitor various patient metrics, such as vital signs, medication adherence, and overall health status from a distance. The collected data are instantly transmitted to healthcare professionals, enabling them to make well-informed decisions and deliver prompt interventions. The integration of remote patient monitoring systems with 6G networks guarantees an uninterrupted and reliable data flow, thus supporting proactive healthcare management.
- **Telemedicine and virtual consultations**: 6G enables high-quality video conferencing and real-time communication, making telemedicine and virtual consultations more accessible and efficient. Healthcare providers can offer remote consultations, diagnosis, and treatment recommendations to patients located anywhere, eliminating geographical barriers and improving access to healthcare services. Integrating telemedicine platforms with 6G networks ensures seamless and reliable communication, high-quality video streaming, and secure data transmission. 6G's ultralow latency and high connectivity will greatly enhance telemedicine and robotics by supporting near-instant data transfer and real-time reaction, vital to applications in critical healthcare. In telemedicine, physicians will be able to remotely consult with patients within negligible delay, increasing diagnostic accuracy and patient treatment even for poorly served or rural regions. Real-time video streams, high-definition imaging, and advanced diagnostic information will be easily transferred, permitting more effective remote monitoring and diagnosis.
- **Renewable energy and ecological technologies**: these play a pivotal role in 6G-based smart hospitals, contributing to sustainability and environmental consciousness. The integration of renewable energy sources, such as solar panels and wind turbines, ensures a reliable and eco-friendly power supply and reduces the carbon footprint of these advanced healthcare facilities. Energy-efficient designs and smart grid technologies optimize energy consumption, aligning with green initiatives. Ecological technologies, including green building materials and sustainable infrastructure, further enhance the environmental responsibility. By prioritizing renewable energy and ecological practices, 6G smart hospitals not only reduce operational costs but also demonstrate a commitment to a healthier planet, aligning technological advancements with ecological sustainability in the pursuit of cutting-edge healthcare solutions.
- **Blockchain:** this ensures confidentiality, openness, and integrity of medical data, which is essential in 6G-based smart hospitals. Blockchain technology improves patient privacy

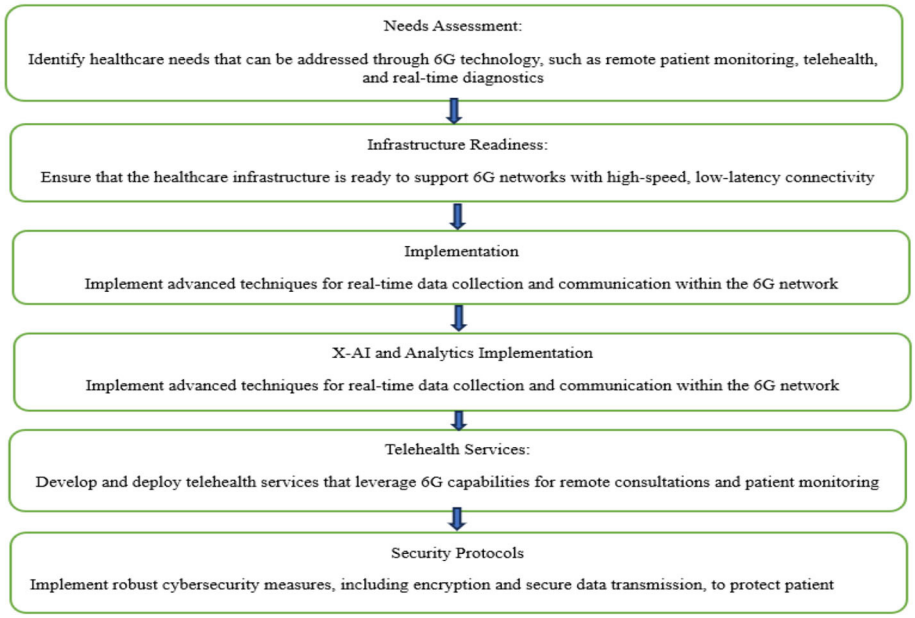


FIGURE 1
Flowchart of 6G integration with smart healthcare.

and protects medical records by utilizing tamper-resistant and decentralized ledgers. Blockchain-based smart contracts protect and automate several healthcare operations, including supply chain management and billing. Furthermore, blockchain promotes interoperability, making it possible to exchange data securely and effortlessly for various health care systems and devices. Smart hospitals build a solid foundation for data accuracy, trust, and efficient operation by integrating blockchain into 6G networks. This eventually increases the overall effectiveness and dependability of healthcare services.

- **Robotics:** 6G's ultralow latency and increased connectivity will revolutionize robotics to a great extent by facilitating real-time communication and exact control over robot systems, particularly in sophisticated applications such as surgery, manufacturing, and remote control. 6G's ultra-low latency of usually sub-millisecond order ensures that instructions sent to robots are carried out with little delay, essential for processes demanding high accuracy and coordination. In robot surgeries, for example, this means surgeons can manipulate robotic arms in near-instantaneous feedback, minimizing the chance for mistakes and enhancing patient outcomes. And the improved connectivity of 6G and the capacity to carry massive IoT networks will also make the smooth integration of different devices, sensors, and robots possible, allowing for collaborative tasks and autonomous decision-making. With 6G's enormous throughput, robots are able to send data-rich information, such as high-definition video or 3D mapping, uninterrupted, further propelling autonomous robotics in telepresence, industrial automation, and healthcare.

4.2 Challenges in 6G based smart hospital

As with any new technology, the development and deployment of 6G faces challenges and considerations. The implementation of 6G will require substantial investment in infrastructure, including new antennas, base stations, and network equipment. Table 1 indicate the challenges faced by 5G and 6G based smart hospital (50, 51).

Spectrum allocation and regulatory frameworks need to be established to facilitate the efficient and secure deployment of 6G networks. To understand why 6G is required, it is important to consider the limitations and evolving requirements of existing wireless communication technologies, such as 5G. Although 5G has brought significant improvements over its predecessors, it still faces certain challenges that 6G aims to address. The following are some key reasons why 6G is required (52, 53):

- **Expanding data traffic**: as the proliferation of connected devices, IoT applications, and data-intensive services continues to drive an exponential increase in data demand, 6G technology has been poised to meet this challenge. It is anticipated to deliver significantly higher data rates and capacities, which are essential for managing the growing volume of data traffic. This advancement will facilitate seamless streaming of ultra-high-definition content, enhance immersive experiences in virtual and augmented reality, and support emerging technologies that depend on massive data transfers.
- **Ultra-low latency**: certain applications and services require real-time responsiveness with minimal delays. Industries such as autonomous vehicles, remote surgery,

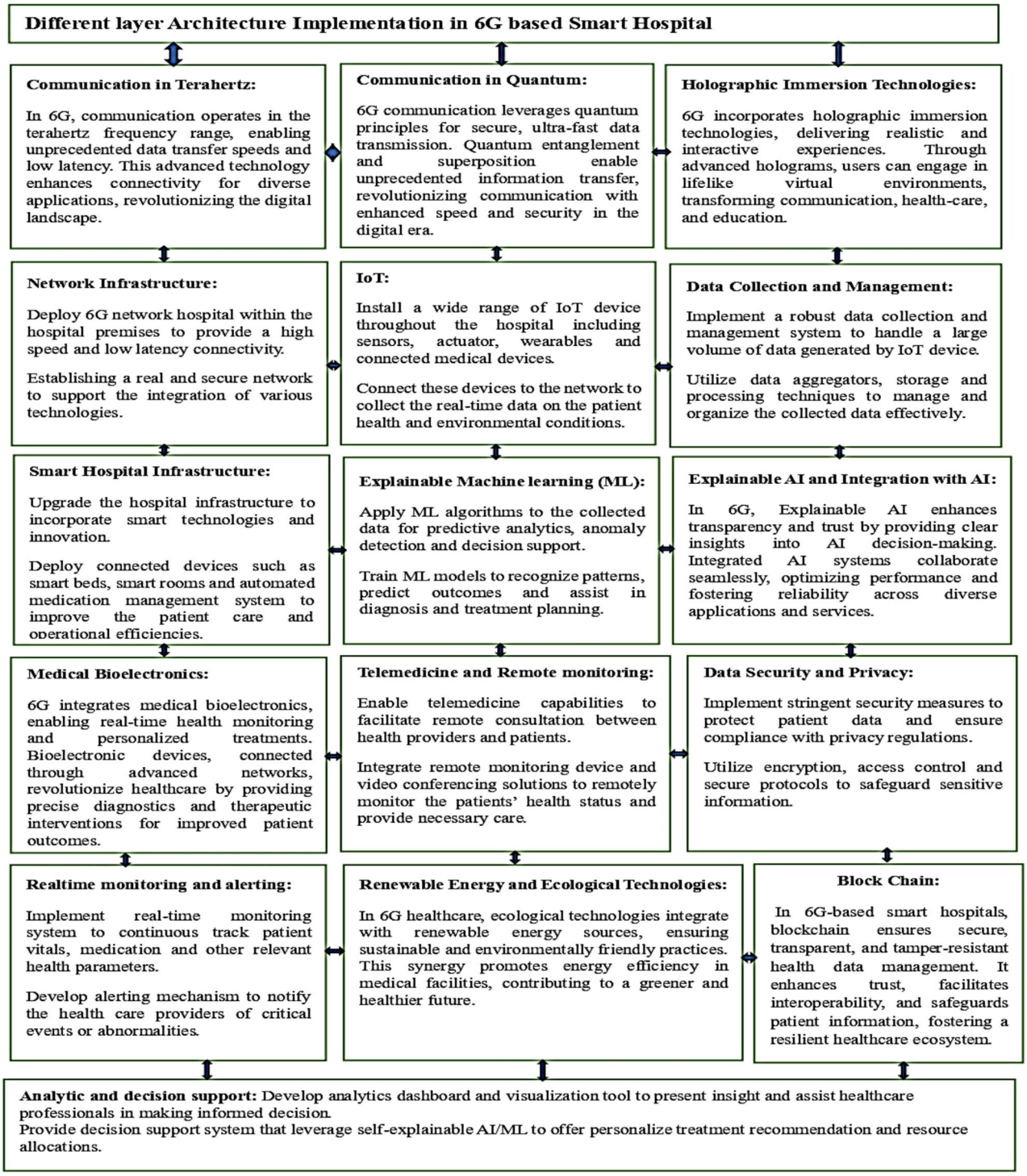


FIGURE 2
Different layers for 6G based smart hospital.

and industrial automation rely on ultralow-latency networks to enable time-critical operations. 6G aims to further reduce latency, enabling instantaneous communication and unlocking new possibilities for mission-critical applications.

- **Massive device connectivity**: the rise of IoT devices and the vision of a fully connected world necessitate networks that can handle an enormous number of simultaneous connections. 6G supports a large number of devices per unit area, enabling seamless connectivity for smart homes, smart cities,

TABLE 1 Challenges of 6G and 5G in smart hospitals.

Parameters	5G	6G
Network dependability and coverage:	• The challenge is in providing dependable and stable 5G service over the hospital's grounds, particularly in difficult-to-reach locations like basements and specialized medical units. • Real-time monitoring and vital healthcare applications may be interfered with by uneven coverage.	• The challenge lies in creating and deploying communication devices that use terahertz frequencies to transfer data at a quicker rate. • Ensuring dependable communication at these higher frequencies and overcoming obstacles related to signal attenuation.
Latency	• While 5G brings low latency, it is important to sustain low latency continuously for applications such as real-time patient monitoring or remote surgery. • High latency can affect the real-time responsiveness of vital medical applications and jeopardize the efficacy of remote healthcare services.	• The challenge lies in achieving and sustaining ultra-low latency to facilitate new applications like augmented reality (AR) for surgical help and medical training. • The real-time responsiveness necessary for vital medical treatments may be hampered by high latency.
Security issues:	• Handling cybersecurity issues brought on by the rise in connected devices and the network's transmission of private patient data. • Unauthorized access to patient records resulting from security breaches puts patient privacy and the accuracy of medical data at danger.	• Keeping an increasingly data-intensive and networked healthcare environment secure and private is a challenge. • Cybersecurity risks have the potential to jeopardize private patient information and interfere with medical operations.
IoT device integration:	• The challenge is in efficiently incorporating a wide variety of IoT gadgets and medical apparatuses into the 5G network. • The potential advantages of connected devices in healthcare may be limited by poor integration, which might impede data flow and interoperability.	• Ensuring smooth interoperability across various devices becomes a crucial concern as the quantity and variety of IoT devices in smart hospitals rise. Effective integration may be hampered by the absence of common data formats and communication protocols among different device kinds and manufacturers. • Healthcare providers would find it challenging to integrate new IoT devices into the 6G network in the absence of defined protocols, which could result in inefficiencies, data silos, and possible disruptions in the flow of operational and patient data.
Scalability	• The challenge is in making sure the 5G network can expand to handle the growing volume of data and the growing number of linked devices in smart hospitals. • Impact: Poor performance and network congestion might result from inadequate scalability.	• The expansion of wearables, medical sensors, and IoT devices in smart hospitals presents a major scalability barrier for 6G networks due to the sheer volume of linked devices. Every device needs a dependable connection, and network scalability becomes more important as the number of devices rises. • The performance of vital healthcare applications and services can be negatively impacted by inadequate scalability, which can cause network congestion, lower data transfer rates, and possible communication disruptions. • Smart hospitals generate enormous amounts of data due to the growing demand for real-time video streaming, high-resolution medical imaging, and other data-intensive applications. 6G networks face scaling issues in effectively managing this spike in data traffic and guaranteeing the efficient transmission of big datasets.
Regulatory and ethical considerations:	• 5G networks are used by smart hospitals to handle and transfer enormous volumes of patient data, including private medical records. It is crucial to comply with legal requirements and provide the greatest levels of data security and privacy. Concerns around illegal access, data breaches, and the possible exploitation of patient information are raised by the interconnectedness of the systems and equipment in smart hospitals. • Ignoring these privacy and security issues may have unethical and legal repercussions, damage patient confidence, and result in noncompliance with regulations.	• The challenge is addressing moral questions about the application of cutting-edge medical technology, such AI-driven diagnosis and therapy. • Establishing trust in the use of 6G technology in healthcare settings requires adherence to legal obligations as well as ethical norms.
Tailored healthcare services:	• Creating and deploying 5G networks that are capable of meeting the many and unique requirements of the healthcare industry. Within a smart hospital, various medical specialties and departments can need different network setups and capabilities to serve their own devices and apps. • Ignoring this issue could lead to subpar performance for some healthcare services, which would reduce the potential advantages of customized and specialized solutions. It could result in ineffective care delivery of specialist treatment.	• The challenge lies in creating 6G networks that are specifically designed to meet the demands of healthcare applications. • Impact: The efficacy of cutting-edge healthcare services and solutions may be restricted by inadequate personalization.

and various IoT applications. This will enable the efficient management of billions of connected devices and unlock the potential of a hyperconnected society.

- **Transformative applications**: 6G integrates various cutting-edge technologies such as AI, machine learning, and quantum computing. These technologies require networks

with enhanced capabilities to effectively process and transmit data. With 6G, transformative applications such as AI-driven smart assistants, advanced healthcare solutions, and intelligent transportation systems can become a reality, fostering innovation and improving quality of life.

- **Future-proofing technology**: developing 6G networks is a proactive approach to future-proof communication infrastructures. This allows us to stay ahead of the emerging technologies and unforeseen demands. By investing in 6G research and development, we can ensure that our networks are ready to meet the challenges and requirements of the next decade and beyond.

The need for 6G arises from the ever-growing demand for faster speeds, higher capacity, ultralow latency, massive device connectivity, and the integration of transformative technologies. 6G will empower industries, enable new applications, and provide a foundation for a more connected and technologically advanced future.

4.3 How 6G will benefit the health industry

The advent of 6G, the latest in the evolution of wireless communication networks, is set to revolutionize the healthcare industry by transforming the delivery of healthcare services. Integrating 6G technology into smart hospitals promises transformative advancements in healthcare, enabling faster, more reliable, and intelligent medical services. One of the key opportunities lies in ultra-low latency and high data rates, supporting real-time applications like remote surgeries and advanced telemedicine. Enhanced connectivity between medical devices and systems will enable seamless data sharing, improved diagnostics, and personalized treatments through AI-driven analytics. Additionally, 6G's support for massive machine-type communications (mMTC) will boost the deployment of Internet of Medical Things (IoMT) devices, allowing continuous patient monitoring, early disease detection, and automated interventions. However, several challenges need to be addressed. Ensuring robust cybersecurity measures is critical due to the sensitive nature of medical data. Managing data privacy in compliance with strict healthcare regulations, while maintaining system integrity is complex. Furthermore, the cost of upgrading hospital infrastructure to accommodate 6G networks may be prohibitive for many institutions, particularly in developing regions. Another concern is interoperability with existing medical devices and systems, requiring seamless integration for effective functionality. Additionally, managing the energy consumption of 6G networks and devices, as well as ensuring the ethical use of AI and big data in decision-making, poses significant hurdles. Overall, while 6G has immense potential to revolutionize healthcare delivery, addressing these technical, financial, and ethical challenges is essential to fully harness its benefits in smart hospitals. Table 2 indicates the advantages of 6G over 5G. Table 3 shows the benefits of 6G over 5G based smart hospitals (54).

Additionally, with its faster speed, lower latency, higher capacity, and integration of transformative technologies, 6G is poised to significantly benefit the smart healthcare industry in numerous ways. The potential benefits of 6G in smart healthcare are as follows (41, 55):

- **Enhanced connectivity and remote care**: 6G technology will significantly enhance connectivity, enabling seamless communication between healthcare providers and patients irrespective of geographical barriers. With its high-speed and reliable connections, 6G will significantly expand the scope of remote care services. This will allow physicians to monitor patients remotely, conduct telemedicine consultations, and offer real-time guidance during emergencies. Patients in remote areas gain access to specialized healthcare without the need for physical travel, thus ensuring equitable access to high-quality medical services. Enhanced connectivity in 6G will revolutionize remote care in smart healthcare by providing ultra-reliable, high-speed communication, enabling seamless, real-time patient monitoring, and consultation. With the support of massive IoT devices, 6G will facilitate the integration of a wide array of health monitoring tools, such as wearable sensors and remote diagnostic equipment, into the healthcare ecosystem. This will allow healthcare providers to monitor patients continuously, even from remote locations, improving outcomes for chronic conditions and reducing the need for in-person visits. AI algorithms will leverage this real-time data to offer personalized care recommendations, and telemedicine consultations will be nearly as efficient as in-person visits, thanks to 6G's low latency. Furthermore, 6G will expand access to healthcare for underserved populations, including those in rural areas, by enabling high-quality remote healthcare services that were previously unfeasible due to connectivity limitations. Enhanced connectivity ensures that patients receive timely care, regardless of their location.
- **Internet of medical things (IoMT) advancements**: IoMT refers to the interconnected network of medical devices and sensors. 6G's higher capacity and massive device connectivity will greatly advance the IoMT ecosystem, enabling a multitude of devices to seamlessly communicate and exchange data. This will result in more accurate patient monitoring, efficient data collection, and improved decision making for healthcare providers. With 6G, wearable devices, implantable sensors, and smart medical equipment operate seamlessly, providing real-time health data for better diagnosis, personalized treatment plans, and proactive healthcare management.
- **Real-time monitoring and emergency response**: 6G's ultra-low latency and high-speed connectivity will enable real-time monitoring of patient health conditions and instant communication in emergency situations. Wearable devices equipped with biosensors and vital sign monitors continuously collect data that can be instantly transmitted to healthcare professionals. This will enable timely intervention and rapid response in critical situations, potentially saving lives. Furthermore, emergency responders have access to live video streams and real-time data from accident sites, enabling them to make informed decisions and provide immediate medical assistance. 6G technology can significantly benefit different healthcare environments outside the typical

TABLE 2 6G and 5G technologies for smart hospitals.

Key technologies in 5G	Key technologies in 6G
5G New Radio (NR): The standard for 5G networks' air interface is called 5G NR. For a variety of devices, it provides reduced latency, increased connectivity, and quicker data rates.	Communication in terahertz: Description: Extremely fast data rates and accurate sensing are made possible by terahertz frequencies, which may be employed in 6G. Terahertz communication has the potential to improve imaging technology in smart hospitals and enable more precise diagnosis.
Slicing a network: Network slicing within the expansive 5G infrastructure enables the creation of virtual, isolated networks tailored to specific needs. In smart hospitals, network slicing allows for the segmentation of the network to cater distinctively to various healthcare services and applications. This technology provides the flexibility to allocate resources efficiently, ensuring that each healthcare function receives the necessary network support to operate optimally.	Explainable AI and integration with AI: It is anticipated that 6G would further incorporate AI into the communication network. In healthcare applications, explainable AI—which offers openness in AI decision-making—may be essential. AI systems may become more prevalent in healthcare administration, treatment planning, and diagnosis.
Enormous IoT connectivity: In smart hospitals, the implementation of medical sensors, wearables, and other connected devices is made easier by 5G's huge support for IoT devices. Real-time 4data collecting and monitoring are made possible by this technology.	Communication in quantum: One prospective 6G feature is quantum communication, which provides higher security while sending private medical information. It might be used to safeguard patient privacy and maintain the accuracy of medical records by securing communication routes inside smart hospitals.
Cutting edge computing: Edge computing lowers latency by bringing processing power closer to the data source. Edge computing improves the functionality of healthcare apps in smart hospitals, including real-time diagnostics and remote patient monitoring.	Medical bioelectronics: The field of bioelectronic medicine focuses on manipulating the electrical impulses produced by the body using electronic equipment. 6G could make it possible for smart hospitals to use closed-loop systems and cutting-edge bioelectronic therapies for individualized and accurate treatment plans.
Virtual reality (VR) and augmented reality (AR): High-bandwidth and low-latency connections are made possible by 5G, which makes immersive technologies like AR and VR possible. These tools can be applied to surgery planning, patient education, and medical training in smart hospitals.	Holographic immersion technologies: It is possible that 6G will enable cutting-edge holographic technologies, enabling realistic and engrossing 3D experiences. This has the potential to improve patient education, team-based surgery, and medical training.
	Renewable energy and ecological technologies: The focus of 6G is anticipated to be on sustainable technology and energy efficiency. By using energy harvesting technology to power IoT devices, smart hospitals can lessen the environmental effect of their healthcare operations.
	Blockchain: Blockchain guarantees safe, unhackable data interchange and storage in 6G-based smart hospitals. It increases data integrity, uses smart contracts to automate procedures, and fosters interoperability to increase efficiency and trust in healthcare operations.

TABLE 3 Additional benefits of 6G smart hospital over 5G based smart hospital.

Parameters	6G Benefits
Extremely high data speeds	Compared to 5G, 6G is anticipated to offer even faster data speeds. Faster transmission of big medical datasets, high-resolution imaging, and real-time video feeds may be made possible by this exceptionally high data rate capacity. Applications for collaborative healthcare, remote diagnostics, and telemedicine can all be greatly improved by this.
Precision medical applications of terahertz communication	New opportunities in precision medicine may arise from 6G's prospective feature, terahertz communication. Advanced diagnosis and treatment planning are made possible by the highly accurate sensing and imaging made possible by terahertz frequencies. This may result in more focused medical treatments and individualized treatment plans.
Improved communication in real time	Applications like augmented reality (AR) and virtual reality (VR) can function more smoothly and responsively because to 6G networks' extremely low latency. This could facilitate collaborative virtual consultations, immersive medical training, and AR-assisted procedures in the healthcare setting.
Extensive device networking for internet of things healthcare	6G can effectively enable the widespread adoption of IoT devices in healthcare thanks to its even higher connection density. This covers a broad range of wearables, monitoring tools, and medical sensors. As a result, patient monitoring, preventive care, and overall healthcare management are all improved by a more extensive and integrated healthcare ecosystem.
Advanced integration of AI	Advanced artificial intelligence (AI) technology integration can be made easier by 6G networks. This covers machine learning apps, predictive analytics, and AI-driven diagnostics. More sophisticated healthcare solutions may result from the smooth interaction between devices and AI algorithms made possible by the improved connectivity and data rates.
Green and sustainable communication	Energy efficiency and environmentally friendly communication technologies are anticipated to be prioritized in 6G as environmental sustainability becomes a bigger priority. 6G-enabled smart hospitals might use less energy, which would lessen the negative effects of healthcare operations on the environment.

hospital setting, such as rural healthcare, home care, and emergency response networks. In rural settings, where specialized care access is typically lacking, 6G's ultralow latency and high data rates will provide real-time telemedicine consultations and remote monitoring capabilities, enhancing health care accessibility and minimizing the need for lengthy transportation. With increased connectivity, healthcare professionals are able to remotely monitor their patients through wearable devices, diagnose ailments in real-time, and offer tailored care, closing the rural-urban healthcare divide. For home care, 6G may facilitate round-the-clock patient monitoring, making it possible to integrate smart home appliances and IoT-based health monitors that feed in constant streams of data into the hands of healthcare professionals. This promotes proactive management of health and early intervention, minimizing readmission to hospitals and enhancing patient outcomes. Additionally, decision support systems based on AI may aid caregivers through real-time feedback on a patient's status. In emergency response systems, 6G's huge connectivity and ultra-high reliability will allow first responders, hospitals, and command centers to coordinate more speedily and efficiently. Real-time data exchange, including live video streams and patient medical records, will support situational awareness, accelerating critical decisions in emergencies such as accidents or natural disasters. In general, 6G will enable more decentralized, efficient, and personalized medicine, enhancing outcomes and minimizing disparities across diverse healthcare environments.

- **Artificial intelligence (AI) integration**: the Integration: The 6G with AI technologies will drive significant advancements in smart healthcare. AI algorithms are capable of analyzing vast amounts of medical data collected through connected devices and electronic health records, aiding healthcare providers in making accurate diagnoses, performing predictive analytics, and offering personalized treatment recommendations. AI-powered virtual assistants and chatbots can provide support to 24/7 patients, respond to inquiries, and deliver basic medical advice. Moreover, AI-based systems for image recognition and interpretation will significantly enhance the analysis of medical imaging, thereby improving both the speed and accuracy of diagnosis.
- **Augmented reality (AR) and virtual reality (VR) applications**: with Owing to their high bandwidth and low latency, 6G will facilitate immersive AR and VR experiences in healthcare. Surgeons benefit from AR overlays during complex procedures that provide real-time guidance and detailed visualization of critical anatomical areas. In addition, medical education and training will see significant enhancements through VR simulations, enabling students to practice procedures in highly realistic virtual settings. AR and VR also play a crucial role in patient education, offering individuals a more interactive and engaging way to understand their medical conditions and treatment options.
- **Precision medicine and personalized healthcare**: the integration of AI, big data analytics, and advanced connectivity offered by 6G will enable the adoption of precision medicine approaches (2). By analyzing extensive datasets encompassing genomic information, patient histories, lifestyle factors, and real-time health data, healthcare providers can offer personalized treatment plans that are uniquely tailored to each individual's needs. This data-driven method enhances healthcare outcomes, minimizes adverse drug reactions, and boosts overall patient wellbeing.
- **Efficient healthcare resource management**: 6G's advanced capabilities will support the efficient management of healthcare resources. Through real-time monitoring and predictive analytics, healthcare providers can anticipate demand, optimize bed allocation, and allocate medical personnel more effectively. The seamless exchange of data between hospitals, clinics, and pharmacies will streamline inventory management, reduce waste, and ensure the availability of essential medications and supplies.
- **Enhanced patient engagement and self-care**: 6G will empower patients to actively manage their health through innovative healthcare applications and services. Mobile apps and wearable devices connected to 6G networks offer real-time health monitoring, personalized health recommendations, and reminders for medication adherence. Patients can conveniently access their health records, schedule appointments, and communicate with healthcare providers through secure mobile platforms.
- **Data security and privacy**: With the integration of advanced security measures and encryption protocols, 6G prioritizes data security and patient privacy. Robust authentication mechanisms and secure data transmission protocols ensure the confidentiality and integrity of sensitive health information and build trust among patients and healthcare providers.

6G offers several advantages over 5G, including a faster speed, lower latency, enhanced capacity, transformative technologies, and expanded coverage. However, it also presents challenges such as longer implementation timelines, higher infrastructure costs, spectrum considerations, compatibility issues, and regulatory/security considerations. These factors need to be carefully addressed as the development and deployment of 6G progresses in the coming years. 6G has the potential to revolutionize the smart healthcare industry by providing enhanced connectivity, enabling remote care services, advancing the IoMT ecosystem, enabling real-time monitoring and emergency response, integrating AI and VR/AR technologies, facilitating precision medicine, optimizing resource management, empowering patient engagement, and prioritizing data security and privacy. These advancements will contribute to improved healthcare outcomes, increased access to quality healthcare services, and a more efficient and patient-centric healthcare system (56). Cybersecurity and data privacy threats in 6G-enabled healthcare systems are paramount issues, considering the confidentiality of medical information and the growing attack surface created by IoT devices and remote care technologies. To counter these threats, embracing a zero-trust architecture (ZTA) is imperative, verifying users, devices, and applications continuously irrespective of location.

ZTA enforces least-privilege access and employs multi-factor authentication (MFA) and real-time anomaly detection to block unauthorized access. Homomorphic encryption (HE) also provides a strong solution by allowing computations on encrypted data without decryption, maintaining privacy throughout data processing. Using blockchain for tamper-proof health records and deploying AI-powered threat detection systems can also increase security. Ongoing security audits, employee training, and adherence to global standards such as HIPAA and GDPR are of paramount importance. Cooperative working between healthcare providers, technology creators, and regulators will be instrumental in the creation of responsive, robust defenses that safeguard patient information while not diminishing the efficacy of sophisticated 6G uses (57). 6G's increased connectivity and device integration are poised to revolutionize sectors like healthcare by enabling faster, more efficient communication and data exchange. However, this hyperconnectivity also introduces significant vulnerabilities, particularly concerning cybersecurity. As healthcare infrastructures become more reliant on interconnected devices—such as IoT-enabled medical equipment, wearables, and cloud-based systems—the attack surface for cybercriminals expands exponentially. In a 6G environment, where billions of devices communicate seamlessly, malicious actors could exploit weaknesses in both hardware and software to gain unauthorized access to sensitive health data or disrupt critical operations.

For example, cyberattacks targeting hospital systems could compromise patient care by manipulating real-time data from life-saving equipment, leading to inaccurate diagnoses or treatment errors. The incorporation of artificial intelligence (AI) in healthcare also raises the level of complexity, and it may be simpler for the attackers to tamper with algorithms, leading to defective decision-making. The wide adoption of cloud computing and edge devices in 6G networks also raises the risk of data breaches or ransomware attacks because healthcare organizations will find it challenging to secure massive amounts of data across different platforms. The sheer volume and complexity of interconnected devices in 6G networks could make traditional security protocols less effective, requiring the development of advanced cybersecurity solutions. Without robust defense mechanisms in place, the healthcare sector faces heightened risks, jeopardizing not only patient privacy but also the very integrity of the healthcare system.

5 Key technologies in 6G based smart hospital

5G improves smart hospitals by offering fast, low-latency connectivity, which makes effective data transfer and real-time monitoring possible. Building on this base, 6G will revolutionize patient care, treatment, and diagnosis by providing quantum communication, holographic interfaces, and powerful AI. When combined, these technologies enable smart hospitals to become extremely intelligent, responsive, and flexible healthcare ecosystems. Table 3 shows the differences between 5G and 6G key technologies in smart hospitals (58).

5.1 Internet of Things (IoT) in 6G based smart hospital

The architecture of a 6G-enabled IoT smart hospital, which aims to integrate real-time data processing, intelligent decision-making, and enhanced networking as shown in Figure 3. The architecture's central component is a dense network of IoT devices that run on 6G's incredibly rapid and low-latency network. Examples of these devices include connected imaging equipment, smart beds, and wearable health monitors (59). These devices ensure rapid processing and analysis near the data source by continuously gathering and transmitting patient data to edge computing nodes within the hospital. This configuration enables real-time monitoring, prompt notifications, and pre-emptive responses (60). Sophisticated AI algorithms analyze the data for predictive analytics, customized treatment plans, and diagnostics. Centralized cloud platforms store and manage large volumes of healthcare data, facilitating seamless integration and accessibility for healthcare providers. Massive MIMO and sophisticated beamforming are included in the network architecture to improve capacity and connectivity. Improved security protocols safeguard patient information while guaranteeing adherence to strict healthcare laws. This intelligent, integrated infrastructure transforms the hospital's operations and enhances operational effectiveness and patient care (59).

5.1.1 IoT sensors in smart hospital

Hospitals use a variety of IoT sensors to manage resources, monitor patients, and increase productivity. With ultra-fast speeds, low latency, and strong security, 6G connectivity dramatically improves the functionality and dependability of these sensors in a smart hospital setting, enhancing patient care and safety (11). The list below includes common hospital sensors and discusses how 6G connectivity enhances their usability (61).

- Wearable health monitors sensors: take temperature, blood pressure, oxygen saturation, heart rate, and other vital signs. 6G's high-speed, low-latency connectivity ensures real-time data transfer, enabling quick analysis, and reaction. 6G enhances security measures to prevent breaches of crucial patient data.
- Smart beds sensors: keep track of occupants, pressure points, and patient movement. Instantaneous data updates and modifications are possible with 6G connectivity, enhancing patient comfort and averting bedsores. Security features guarantee safety and patient privacy.
- Glucose monitors: check diabetes patients' blood sugar levels on a regular basis.6G ensures rapid data transfer to healthcare providers, enabling prompt interventions and modifications to treatment plans. Secure connections protect patient health data from unwanted access.
- Linked imaging systems: these comprise CT, MRI, and X-ray equipment that sends pictures for remote processing. 6G transfers large image files quickly, facilitating faster consultations and diagnoses. Security measures protect private medical images.

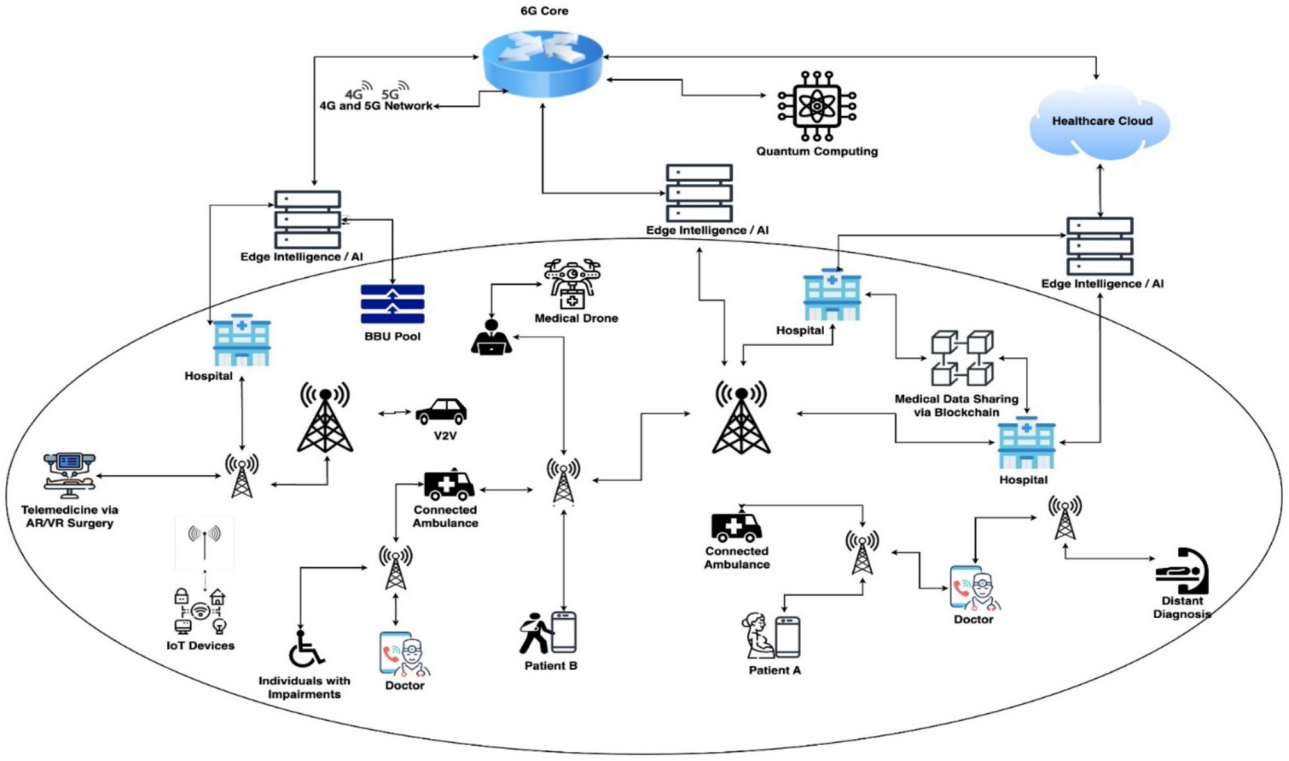


FIGURE 3
IoT empowered 6G smart hospital.

- Environmental sensors: monitor the lighting, temperature, humidity, and air quality in patient rooms and other crucial areas. 6G enables real-time control and monitoring, ensuring the best possible environmental conditions for patient safety and wellbeing. Improved security features prevent sensor data manipulation.
- Infusion pumps sensors: give patients precisely the right dose of medication. 6G connectivity guarantees rapid and precise data on medicine delivery, allowing for remote monitoring and modifications. Secure communication prevents potential errors or interference.
- Fall detection sensors: these devices identify falls in patients and sound an alarm. 6G's rapid data transfer speeds guarantee prompt notifications to medical professionals, cutting down on reaction times and enhancing patient security. Security mechanisms safeguard data about patient movements and locations.
- Telemedicine tools sensors: enable online consultations and exams. 6G raises the standard of telehealth services by providing the enormous bandwidth required for high-definition audio and video. Secure connections guarantee the confidentiality of patient-doctor communications.

Certain sensors require 6G security and quick speed features (62).

- Wearable health monitors and glucose monitors are essential for ongoing patient care; they need to transmit data in real-time and with a high level of security to safeguard private medical data.
- Connected imaging systems require strong security to protect private diagnostic images, as well as fast transmission speeds for large files.
- Infusion pumps require secure, instantaneous communication to ensure precise drug administration and prevent errors or manipulation.

5.1.2 Key components level implementation of IoT in smart hospital

The architecture of an IoT-based smart hospital which was referenced in Philips Health Suite and Siemens IoT-enabled solutions for healthcare comprises several interconnected components and layers that enable seamless communication, data exchange, and intelligent decision making, as shown in Figure 4.

5.1.2.1 Philips Health Suite Digital Platform architecture layer

The Philips Health Suite Digital Platform, an innovative healthcare platform, integrates data from various sources, including electronic health records (EHRs), IoT devices, and other healthcare systems. The Philips Health Suite Digital Platform's architecture streamlines the collection, integration, and evaluation of medical data, enabling more personalized patient care and enhanced operational efficiency in healthcare facilities. Healthcare businesses to use data to deliver tailored care, improve clinical outcomes,

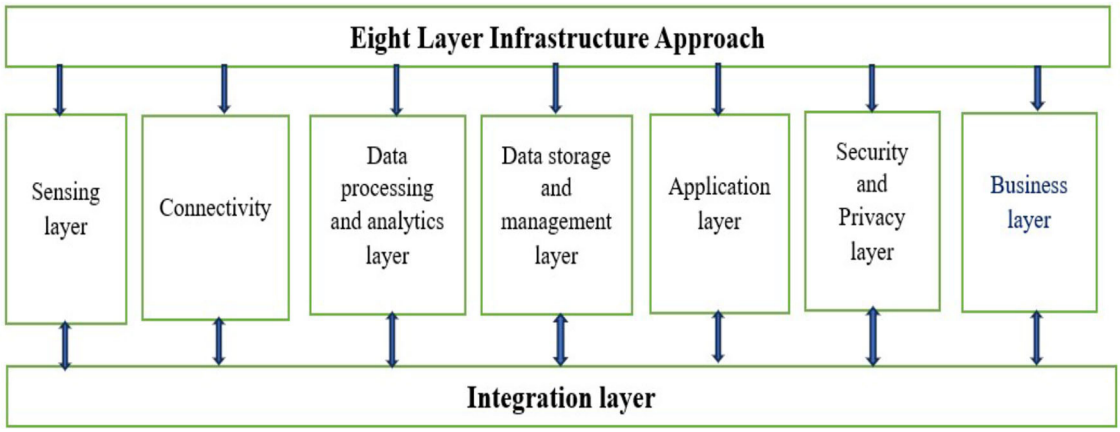


FIGURE 4
IoT architecture.

and improve patient experiences thanks to the architecture of the Philips Health Suite Digital Platform, which aims to build a single ecosystem (63). At its core, the Health Suite Digital Platform consists of several key components (64):

- The data ingestion layer: this layer is in charge of gathering data from many sources, including wearables, medical sensors, IoT devices, and EHR systems. It guarantees that information is entered into the platform safely and processed further.
- Data management and storage layer: after being gathered, the data is kept in a scalable and safe cloud-based storage system. This layer contains databases and data lakes that effectively handle and organize the enormous volumes of healthcare data.
- Data analytics and insights layer: to extract useful insights from the gathered data, the platform uses machine learning algorithms and sophisticated analytics. Personalized care interventions and predictive analytics are made possible by this layer, which analyses patient data to find trends, patterns, and possible health hazards.
- Application services layer: to facilitate the development of healthcare applications and services, the Health Suite Digital Platform offers a collection of application services and APIs (Application Programming Interfaces). These services make it easier to create custom healthcare solutions, integrate with third-party systems, and ensure interoperability amongst healthcare equipment.
- Security and compliance layer: the platform has strong security mechanisms in place to safeguard patient data and guarantee adherence to healthcare laws like HIPAA (Health Insurance Portability and Accountability Act). Security is of the utmost importance in the healthcare industry. To protect sensitive medical data, this layer has audit trails, access control methods, and encryption.

5.1.2.2 Siemens IoT-enabled solutions for healthcare layers

Siemens' Digital Enterprise portfolio includes IoT-enabled healthcare solutions that offer a holistic architecture that combines edge computing, sensors, networking, data analytics, and security measures. Siemens wants to use these technologies to propel the digital transformation of healthcare, making patient-centered, cost-effective, and intelligent healthcare delivery possible (65). A s part of its Digital Enterprise portfolio, Siemens provides IoT-enabled healthcare solutions that are intended to streamline hospital operations, improve patient experiences, and improve clinical outcomes. Siemens' IoT-enabled healthcare solutions are built with a number of essential parts and tiers, all of which are necessary to provide integrated, data-driven healthcare services (66).

- Sensors and medical devices: a variety of sensors and medical devices placed throughout the hospital setting form the basis of Siemens' IoT-enabled healthcare solutions. These gadgets include sensors for facility management, imaging systems, lab apparatus, and patient monitors. These sensors gather numerous pieces of information about patient health, operational effectiveness, and environmental factors.
- Siemens' architecture incorporates a robust networking infrastructure to facilitate seamless communication between sensors, devices, and backend systems. Both wired and wireless networks are part of this infrastructure, which guarantees dependable data transfer and instantaneous connectivity. Siemens facilitates compatibility and integration with current hospital IT systems.
- Edge computing and data processing: Siemens uses edge computing skills to manage the enormous amount of data produced by sensors and medical equipment. Within the hospital's walls, edge devices preprocess and analyze data locally, cutting down on latency and bandwidth needs. This distributed computing architecture facilitates real-time monitoring and alerting for key events, as well as quick decision-making.
- Cloud platform and data analytics: to extract useful insights from healthcare data, Siemens' Digital Enterprise portfolio makes use of cloud computing and sophisticated data analytics. Siemens securely transfers sensitive and device data to cloud-based platforms for further processing. Siemens

provides healthcare professionals with predictive analytics and decision support tools using AI and machine learning algorithms to identify important patterns, trends, and correlations in healthcare data.

- Integration with hospital systems: Siemens' Internet of Things (IoT)-enabled healthcare solutions easily integrate EHRs, hospital information systems (HIS), and other clinical applications. This guaranteeing data interchange and compatibility between various systems, this integration permits thorough patient care coordination and workflow optimization.
- Security and compliance: Siemens' IoT-enabled healthcare infrastructure places a high priority on security. Strong cybersecurity safeguards protect sensitive patient data and guarantee adherence to healthcare laws like GDPR and HIPAA. These measures include encryption, access restrictions, and threat detection. Siemens employs a multi-layered security strategy to reduce risks and defend against constantly changing cyberthreats.

5.1.3 How 6G help to overcome the challenges of integrating IoT in smart hospitals

Although the integration of IoT in smart hospitals brings numerous benefits, it also presents several challenges that require careful management. Addressing these challenges necessitates a strategic approach, effective collaboration between IT and healthcare departments, robust governance frameworks, and the continuous monitoring and evaluation of IoT systems. By effectively navigating these challenges, hospitals can fully leverage the transformative potential of the IoT to enhance efficiency and patient-centricity in healthcare services. The following are some key challenges associated with integrating the IoT in a smart hospital (67).

- Interoperability: due to the fact that different companies manufacture many IoT devices and use different communication protocols, interoperability issues arise. It can be difficult to integrate many devices into a coherent system; this may call for specialized integration work.
- Security and privacy: because IoT devices frequently gather private medical information, hackers find them to be appealing targets. IoT security flaws might make patient data vulnerable to illegal access or jeopardize the reliability of medical systems.
- IoT devices need network connectivity in order to send and receive orders, which contributes to their reliability and resilience. Network outages or disturbances can impact the dependability of IoT-based systems, potentially impacting patient safety and care.
- Scalability: as hospitals install more IoT devices, managing and scaling the infrastructure to meet demand will become more challenging. Scalable solutions that maintain performance and dependability over a large number of devices are required by hospitals.
- Data management and analytics: we must efficiently gather, save, and examine the massive volumes of data generated by IoT devices. To extract useful insights from data created by the Internet of Things, hospitals need to have a strong data management and analytics infrastructure in place.
- Healthcare regulations, such as HIPAA in the United States, impose strict guidelines for safeguarding patient privacy and data security. Hospitals must make sure that their IoT-based systems comply with regulatory requirements in order to prevent negative legal and financial repercussions.

6G technology has the potential to resolve several of these issues (68):

- Improved connectivity: in comparison to earlier generations, 6G networks offer substantially faster data speeds, reduced latency, and more device density. This enhanced connectivity may support more IoT devices and enable real-time data transfer for vital uses like telemedicine and remote patient monitoring.
- Enhanced security: we anticipate 6G networks to have cutting-edge security features like improved authentication procedures and encryption algorithms to fend off cyberattacks and illegal access. Furthermore, 6G networks may use AI-driven security solutions to instantly identify and neutralize threats.
- 6G networks could enable edge computing capabilities, enabling the processing and analysis of data closer to its source. By processing sensitive data locally instead of sending it over the network, edge computing can lower latency, ease network congestion, and improve data privacy.
- AI-driven optimization: by utilizing AI algorithms, 6G networks are able to detect network outages, optimize network resources, and dynamically distribute bandwidth according to application demands. This AI-driven optimization may help IoT-based smart healthcare systems become more resilient and dependable.
- Regulatory compliance: 6G networks may use features like integrated encryption and data anonymization methods to help with regulatory compliance. These elements can help hospitals comply with regulations regarding patient data protection and privacy.
- By utilizing 6G technology, hospitals can overcome many of the obstacles associated with IoT-based smart hospital deployments, ultimately improving patient care results, and operational efficiency.

5.2 Explainable artificial intelligence

Explainable AI, or XAI, is the development of artificial intelligence systems that not only make precise forecasts or suggestions but also transparently explain their judgments and actions in the context of smart healthcare. In the medical field, where choices have a direct effect on patients' lives, XAI is essential for fostering a sense of confidence, enhancing communication between AI systems and medical personnel, and guaranteeing patient safety. XAI makes AI-driven healthcare solutions more interpretable and accountable by offering clear

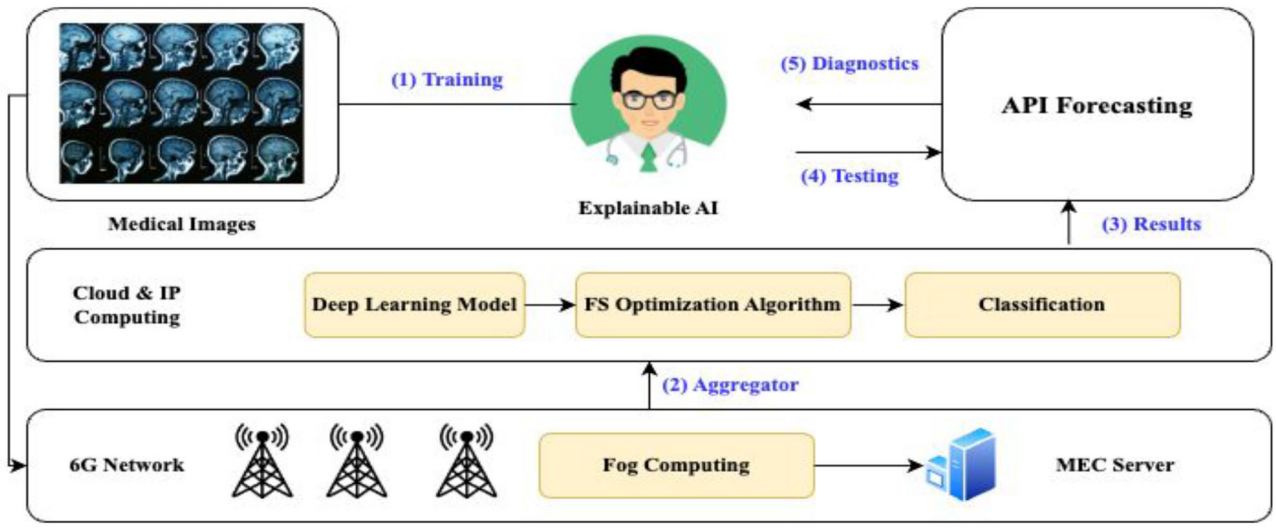


FIGURE 5
Self-explainable AI based smart healthcare.

justifications for diagnosis choices, treatment strategies, and prognostic predictions. Because of this transparency, doctors can verify AI recommendations, comprehend the underlying logic, and apply their domain expertise to the decision-making process. By enabling doctors to prioritize patient safety and wellbeing while making better-informed and confident judgments, XAI ultimately promotes increased acceptance and implementation of AI technology in healthcare. Real-time processing of AI in 6G hospitals demands strong hardware infrastructure that can support enormous volumes of data at ultra-low latency. Advanced edge computing devices, AI accelerators such as GPUs, TPUs, and neuromorphic processors emulating brain-like efficiency to make quick and precise decisions are needed. Edge computing is also essential in reducing data transfer delays through data processing close to the source, for instance, in ICU monitoring or robotic surgeries. Moreover, ultra-reliable low-latency communication (URLLC) modules and high-frequency 6G antennas are required to ensure smooth connectivity over hospital networks (69). Power consumption is also a key issue since real-time AI applications such as predictive diagnostics and robotic surgical systems demand constant data analysis. Energy-efficient hardware, dynamic power management methods such as adaptive voltage and frequency scaling (DVFS), and smart workload allocation can help minimize energy consumption. Blending renewable energy sources, like solar power, and using AI-powered algorithms to manage and minimize energy usage are key to sustainability. Hospitals must embrace green computing principles and work with equipment vendors to develop hardware specific to healthcare AI workloads. Regulatory agencies must also create standards to guarantee energy-efficient deployment while ensuring system performance and reliability. In the end, both high-performance processing and energy efficiency will be required to make the next generation of intelligent, 6G-driven healthcare services possible (70).

An AI-based smart healthcare architecture that is self-explanatory incorporates AI algorithms that not only generate precise forecasts or suggestions, but also offer transparent, easily comprehensible explanations for their choices as shown in Figure 5. Interpretable AI models and methods that emphasize explainability over performance are the foundation of this architecture. This architecture uses AI algorithms to analyze healthcare data and produce predictions or recommendations. Examples of these algorithms include decision trees, rule-based systems, and interpretable deep learning models. These algorithms focus on producing precise results and providing clear justifications for their choices, highlighting the crucial elements or characteristics that influence the outcome. Additionally, the architecture has parts for showing patients and healthcare professionals AI-generated explanations. This could entail the use of interactive dashboards, graphical displays, or plain language explanations that make the logic underlying AI predictions simple to comprehend. Moreover, the design includes components for tracking and assessing AI model performance and interpretability over time. This guarantees that the AI system will always be trustworthy, transparent, and sensitive to the requirements and expectations of its users. By offering comprehensible justifications for AI-driven decisions, self-explanatory AI-based smart healthcare architecture promotes trust, accountability, and cooperation between AI systems and human stakeholders. This improves clinical decision-making, patient engagement, and overall healthcare outcomes.

5.2.1 Types of data use by XAI

In the healthcare industry, explainable AI (XAI) uses a variety of medical data sources to offer clear and comprehensible insights into AI-driven decision-making procedures (71). These data sources include (72, 73):

- EHR: these records contain a patient's medical history, diagnosis, prescriptions, test results, and treatment plans. In

order to help doctors make well-informed decisions about patient care, XAI algorithms examine EHR data.
- Medical imaging: data from modalities such as CT scans, MRIs, ultrasounds, and X-rays is processed using XAI algorithms. In order to help radiologists identify anomalies, make diagnoses, and schedule treatments, AI systems analyze imaging data.
- Genomic data: DNA sequences, gene expression profiles, and genetic variants are among the genomic data that XAI is used to interpret. AI systems examine genetic data to find genetic markers linked to specific illnesses, customize therapeutic strategies, and estimate the likelihood of developing a disease.
- IoT with wearable devices: XAI algorithms examine information gathered from wearable sensors and IoT devices that track physiological characteristics like as activity levels and vital signs. This information is used to monitor patient health, identify abnormalities, and offer early warning indicators of possible medical problems.

Several instances of AI/ML algorithms in use in hospitals exist:

- Deep learning in medical imaging: in medical imaging, convolutional neural networks, or CNNs, are frequently employed for tasks like disease categorization, lesion detection, and picture segmentation. For example, CNNs are used in the FDA-approved AI program IDx-DR to evaluate retinal pictures for the purpose of screening for diabetic retinopathy.
- Clinical decision support systems: clinical decision support systems are created using machine learning techniques like decision trees and random forests as well as rule-based systems. By evaluating patient data and medical literature, IBM Watson for Oncology, for instance, applies machine learning algorithms to help oncologists make therapy decisions.
- Natural language processing (NLP): NLP methods are used to extract structured data from narratives and unstructured clinical notes included in electronic health records (EHRs). NLP is used by Google's DeepMind Health to evaluate EHR data for purposes including treatment suggestions and patient risk assessment.

5.2.2 AI/ML used in hospitals

The availability of large-scale medical datasets, advances in AI/ML methodologies, and increases in processing capacity, the use of such AI approaches in healthcare is still relatively new. These strategies are being actively used by healthcare organizations, academic institutions, and digital companies to enhance patient care, streamline clinical processes, and quicken medical research. Worldwide, a number of healthcare facilities and hospitals are utilizing diverse AI and machine learning (ML) algorithms to examine medical data for a variety of purposes. Here are a few instances (25, 74):

- United States' Mayo Clinic: to identify patients who may experience specific medical illnesses or complications, Mayo Clinic uses artificial intelligence (AI) and machine learning (ML) algorithms for predictive analytics. Additionally, they use natural language processing (NLP) algorithms to glean insights from electronic health records (EHR) and unstructured clinical notes.
- University College London Hospitals (UCLH) in the United Kingdom: UCLH uses artificial intelligence (AI) algorithms for medical imaging, radiology, and pathology image processing. These algorithms help evaluate medical pictures, such as MRIs, CT scans, and X-rays, so that doctors can diagnose illnesses and ailments more quickly and accurately.
- Seoul National University Bundang Hospital (South Korea): this hospital uses artificial intelligence (AI) to provide personalized care by evaluating genetic information and medical records to create customized treatment regimens and forecast patient reactions to various drugs and treatments.
- Massachusetts General Hospital (United States): based on past data and present health state, Mass General uses AI algorithms for clinical decision support, helping physicians diagnose illnesses, choose the best course of therapy, and forecast patient outcomes.
- Singapore General Hospital (Singapore): in order to improve the caliber and accessibility of healthcare services, this hospital uses AI and ML algorithms to manage healthcare operations. These algorithms optimize resource allocation, patient scheduling, and workflow efficiency.

5.2.3 Benefit of XAI in 6G based smart hospital over 5G

Compared to 5G technology, the integration of XAI in a 6G-based smart hospital offers the following advantages (75):

- In a 6G smart hospital, XAI provides clear justifications for AI-generated suggestions and judgements. Healthcare workers must be able to comprehend the reasoning behind AI-generated insights in order to build trust in the technology and enable cooperation between AI systems and human clinicians.
- 6G-based XAI algorithms provide better interpretability when compared to AI models in 5G environments. As a result, physicians will be better equipped to verify suggestions and more successfully apply their domain knowledge to decision-making processes, as they will have a deeper understanding of how AI makes its decisions.
- XAI in a 6G smart hospital allows medical professionals to go back and confirm the logic behind particular suggestions or actions, which increases AI systems' accountability. This accountability is crucial for ensuring that AI-driven interventions comply with ethical and clinical criteria in healthcare settings where decisions have a direct influence on patient lives.
- XAI in a 6G smart hospital helps reduce the possibility of biases or mistakes in AI-driven decision-making by offering clear and understandable answers. Physicians can more

readily recognize possible flaws or restrictions in AI systems, enabling them to step in when needed to protect patient safety and wellbeing.

- The use of artificial intelligence (AI) in healthcare can help to support regulatory compliance standards. XAI capabilities in a 6G smart hospital can assist. Healthcare businesses can demonstrate compliance with regulatory norms and rules governing the use of AI technologies in clinical practice by utilizing features such as transparent explanations and interpretability.

5.2.4 How XAI can be integrated with 6G based smart hospital

Incorporating transparency and interpretability elements into AI-driven healthcare systems is necessary to integrate XAI with a 6G-based smart hospital. Healthcare companies can create AI-driven healthcare solutions that are more transparent, understandable, and reliable by combining XAI with 6G-based smart hospital systems. In the end, this improves patient outcomes by fostering human-machine collaboration and bolstering clinicians' faith in AI technologies. Here's how 6G technology can integrate XAI into a smart hospital setting (76):

- Algorithm design: create AI algorithms with interpretability and transparency as top priorities. This entails producing justifications for AI predictions or suggestions using methods like decision trees, rule-based systems, and model-agnostic methodologies.
- Real-time explanation generation: when AI algorithms make judgments or forecasts, implement systems that instantly produce explanations. Delivering these explanations in a format suitable for the current healthcare activity should enable healthcare workers to understand them.
- Integrating with 6G connectivity: make use of 6G networks' fast, low-latency connectivity to enable smooth communication between AI systems and healthcare organizations. Ensure that clinicians' devices can swiftly and reliably receive and process XAI explanations, enabling immediate review.
- User interface design: create user interfaces that display AI suggestions or forecasts, along with XAI explanations. This keeps their workflow uninterrupted and makes it simple for physicians to access and understand the logic underlying AI-driven decisions.
- Feedback mechanisms: put in place systems that let medical professionals comment on how relevant and accurate XAI explanations are. This gradually enhances the transparency and interpretability of AI systems by utilizing human judgment and input.
- Security and privacy: ensure the secure transmission of XAI explanations via 6G networks to protect patient confidentiality and privacy. To safeguard sensitive medical data during transmission, employ authentication and encryption techniques.
- Regulatory compliance: verify that the XAI integration conforms with the laws and regulations, including HIPAA and GDPR, that control the use of AI in healthcare. This requires transparency in the XAI explanation process and adherence to the accuracy and dependability standards established by regulations.

5.2.5 Challenges of XAI and how 6G can help

In smart hospitals, XAI presents problems primarily related to accountability, transparency, and trust in AI-driven decision-making. 6G technology can help XAI in smart hospitals by facilitating clear communication, reducing prejudice, boosting security and privacy, and strengthening the resilience and dependability of AI-driven healthcare systems. Smart hospitals may implement XAI solutions that empower physicians, enhance patient outcomes, and promote confidence in AI-enabled healthcare delivery (77). Table 4 indicate some of the issues and possible solutions that 6G technology may bring about:

The use of 6G in healthcare creates substantial regulatory loopholes because of the unprecedented speed, connectivity, and volume of data. Existing healthcare data protection regimes, like HIPAA in the United States and GDPR in the European Union, can be inadequate to deal with the intricacies of 6G networks, particularly in terms of real-time processing of data, cross-border data transfers, and AI-based medical decisions. To fill these loopholes, a harmonized, worldwide regulatory regime is needed. This structure should create uniform protocols for data sharing, encryption, and interoperability across borders while maintaining adherence to regional healthcare legislation. Homomorphic encryption, zero-trust architecture, and blockchain can be made mandatory to secure patient data. Regulatory authorities should also make real-time auditing mechanisms mandatory and demand transparent AI algorithms, making diagnostic decisions explainable and unbiased. Ethical issues around AI-powered diagnoses and robot-assisted surgeries need to be resolved by introducing guidelines focusing on patient safety, consent, and responsibility. Accurate legal liability for AI mistakes, complete clinician education, and integration of human review in key medical procedures are important. There needs to be an integration with AI developers, healthcare professionals, and ethicists with regulatory authorities to work together to set ethical standards for AI. Public education and patient awareness regarding AI participation in their treatment will also enhance trust. Finally, an evolving, open, and internationally harmonized legal framework is essential to provide secure, ethical, and compliant 6G-based healthcare systems.

AI-based decisions in a 6G-enabled hospital need to be strictly audited for fairness and bias to guarantee patient safety and fairness. Explainable and transparent AI models are essential, as they enable healthcare workers and regulators to see how decisions are reached. Auditing needs to involve periodic checks of AI algorithms, testing against varied datasets, and tracking for any indication of discriminatory results based on race, gender, or socioeconomic status. Ethical frameworks, including the application of fairness measures and the integration of human judgment in key decisions, are needed to reduce bias.

TABLE 4 Challenges and solution of XAI by 6G.

Parameters	Challenges	How 6G can help
Interpretability	One issue with AI in healthcare is that some algorithms are "black boxes," making it challenging to figure out how they arrive at particular conclusions. In healthcare settings, where doctors must trust and comprehend the reasoning behind AI-driven suggestions, this lack of interpretability can be a challenge.	6G networks facilitate real-time communication between AI systems and healthcare providers, enabling the exchange of comprehensive justifications for AI-generated suggestions. 6G-enabled augmented reality (AR) devices, for instance, might instantly superimpose explanations onto patient data or medical images, giving medical professionals a clear visual representation of AI reasoning.
Fairness and bias	AI systems trained on inadequate or biased data may unintentionally exacerbate or prolong existing inequalities in healthcare outcomes. Ensuring justice and fairness in AI decision-making is essential to giving every patient access to high-quality care.	How 6G can help: 6G can facilitate the transfer of massive datasets required for training AI models on a variety of representative data sources due to its high bandwidth and low latency. Furthermore, the federated learning capabilities of 6G networks allow several institutions to cooperatively build AI models without exchanging private patient data, thereby reducing the risk of bias and promoting justice.
Security and privacy:	Health information is extremely private and governed by stringent laws, such as the Health Insurance Portability and Accountability Act (HIPAA) in the US. When implementing AI systems in smart hospitals, data security and patient privacy protection are top priorities.	To safeguard data transferred between IoT devices, AI systems, and cloud servers, 6G networks include cutting-edge encryption techniques and improved security features. Hospitals can use differential privacy and secure multi-party computation over 6G networks to analyze sensitive patient data while maintaining patient privacy and regulatory compliance.
Robustness and reliability	To guarantee patient safety and care continuity, AI systems installed in smart hospitals need to be robust and resilient. Serious repercussions for patient outcomes could result from system malfunctions or inaccurate AI forecasts.	How can 6G be useful? 6G networks' ultra-reliable low-latency communication (URLLC) capabilities, which offer low latency and high dependability, enable mission-critical applications such as remote patient monitoring, telemedicine, and surgical robotics. By lowering the number of single points of failure and processing data closer to the point of collection, redundant 6G network designs and edge computing resources can significantly improve the resilience of AI systems

In addition, external audits by independent regulatory agencies must be performed to ensure adherence to healthcare data protection regulations.

The heightened surveillance facilitated by 6G technologies in intelligent hospitals is of concern regarding privacy and consent. Ongoing monitoring of patients by IoT devices, facial recognition, and AI analysis may result in over-surveillance possibilities that compromise individual liberties and create vulnerabilities in data. Surveillance, though it enhances patient care through real-time action, has dangers of data loss and unauthorized entry. Tight regulatory systems and patient consent processes have to be in place to prevent abuse and make sure that the advantages of AI-based healthcare do not occur at the cost of patient privacy.

5.3 Robotics in 6G based smart hospital in robotics

6G-based smart hospitals will outperform their 5G counterparts and revolutionize robotics and automation by introducing innovative features. 6G networks' extremely low latency and large data rates make it possible to easily integrate sophisticated robotic systems, providing precise and real-time medical automation. The combination of ultralow latency, high data speeds, terahertz communication, and energy efficiency enhances robotics and automation in 6G-based smart hospitals. These characteristics enable a new wave of intelligent, flexible, and long-lasting robotic applications that will transform patient care and healthcare delivery (78). 6G's improved connectivity makes haptic feedback systems possible, which gives robotic treatment a tactile element. Ultra-low latency can help teleoperated robotic surgeries by allowing surgeons to accomplish complex tasks with previously unheard-of precision and responsiveness from a remote location. Healthcare settings can benefit from the adoption of swarm robots due to 6G's enhanced data speeds and better connection density. Swarm robots improve hospital operations by efficiently completing activities such as drug administration, sample collection, and environmental monitoring, while operating both cooperatively and independently. The combination of 6G with cutting-edge AI has enabled the development of more intelligent and context-aware robotic systems. These AI-powered robots can smoothly communicate with patients and medical personnel, adapt to changing hospital conditions, and move through congested areas with intelligence (79). Advanced medical imaging robots can perform high-resolution, real-time diagnoses because to 6G's terahertz communication capabilities. The AI algorithms of these robots enable them to examine medical images instantly, facilitating prompt decision-making and intervention. 6G focuses on sustainability and facilitates energy-efficient computing and communications. By prolonging robotic systems' operating life, lowering energy usage, and encouraging environmentally friendly automation techniques inside smart hospitals, this feature improves robotic systems. 6G enables more organic and cooperative interactions between people and machines. Healthcare workflows can easily include robotic assistants, sometimes known as humanoid robots, to support various duties such as patient care,

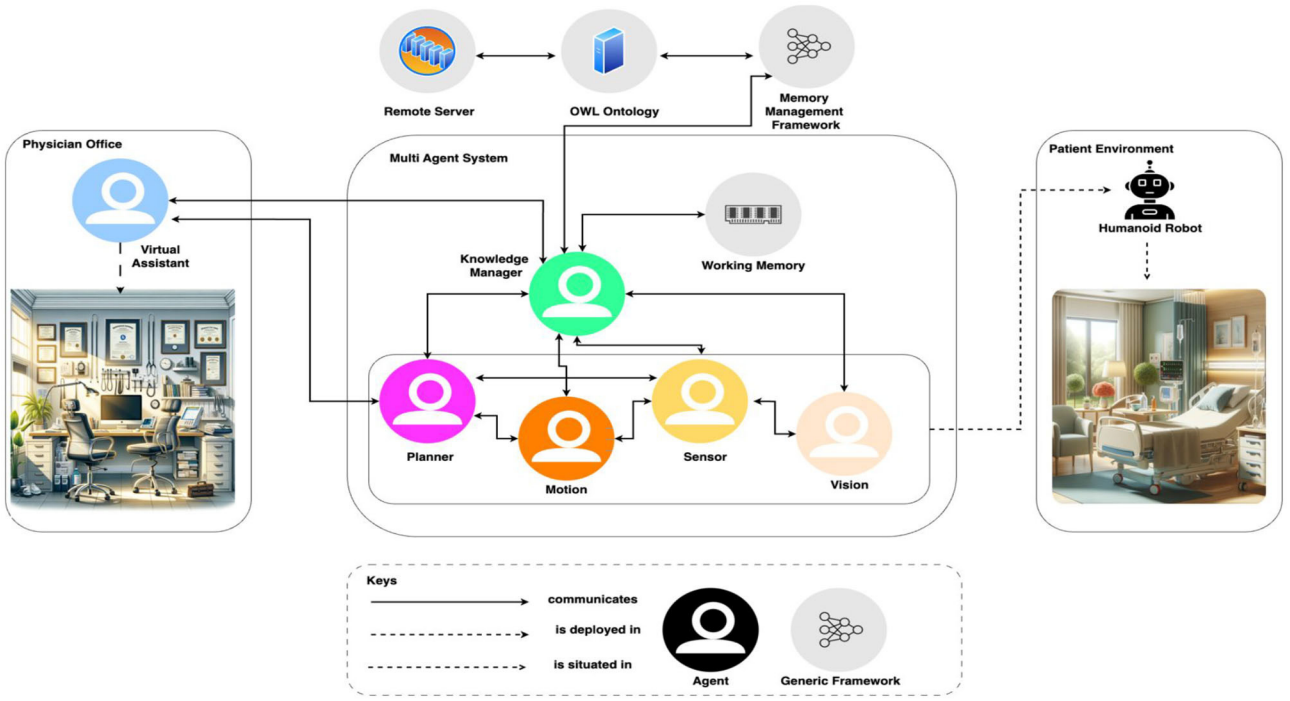


FIGURE 6
Robotic enabled smart health hospital.

rehabilitation exercises, and standard medical procedures (80). The robot-enabled smart hospital is shown in Figure 6.

5.3.1 Advanced applications of Robots in smart hospital with statistics

Robotic surgery is a well-established technique; thus, robotic technology in hospitals is not new, especially when it comes to surgical procedures. However, outside of surgical settings, robotics technologies continue to bring new breakthroughs and applications to smart hospitals. The following is a list of recent and upcoming uses for robots in hospitals (81, 82):

- Robotic surgery: although already common, continuous improvements in robotic systems improve accuracy, adaptability, and efficiency, resulting in shorter recovery periods, fewer complications, and better patient outcomes. Studies have shown that the da Vinci Surgical System, for example, reduces blood loss and shortens hospital stays throughout a variety of minimally invasive procedures, such as hysterectomies, prostatectomies, and cardiac surgeries.
- Telepresence robots: these devices allow medical professionals to communicate with patients and provide care from a distance, facilitating virtual consultations and remote patient monitoring. In critical care situations or for patients with limited mobility, these robots let healthcare providers and patients communicate more easily.
- Logistics and distribution Hospitals are increasingly using autonomous robots for supply chain management, medicine distribution, and specimen transportation. These robots improve resource allocation efficiency, decrease manual work, and streamline hospital operations.
- Disinfection robots: due to the growing emphasis on infection control and hygiene, healthcare institutions use UV-C disinfection robots to sterilize patient rooms, operating rooms, and other high-touch surfaces. By lowering the risk of infections linked to healthcare, these robots enhance patient safety in general.

Even though robotic surgery is still a common use, there is more and more potential to integrate robots into hospital operations, improving patient care, efficiency, and infection control. The following data illustrates how robotic technology is affecting hospitals (83, 84):

- Studies have shown that using UV-C disinfection robots can reduce hospital-acquired illnesses by up to 50%.
- Robotic surgery has been associated with shorter hospital stays; in fact, some treatments have demonstrated a 40% reduction in stay time when compared to traditional surgery.
- Telepresence robots can increase patient satisfaction ratings by up to 25% by facilitating better access to care and communication between patients and healthcare professionals.

Integrating robotics and automation into a 6G-based smart hospital requires meticulous planning, infrastructure preparedness, and attention to safety and regulatory requirements. Successful implementation hinges on collaboration between healthcare providers, technology vendors, and robotics specialists. Leveraging

the synergy between robotics, automation, and 6G technology can significantly enhance efficiency, accuracy, and patient outcomes in healthcare environments. The architectural requirements are shown in Figure 7. Robotics and automation can be integrated in several ways (85, 86).

- **Surgical robots**: surgical robots enhance the precision and control of minimally invasive procedures. By integrating these robots with 6G networks, real-time communication and collaboration between surgeons and robots can be achieved. This allows surgeons to remotely control robots, execute complex procedures with increased dexterity, and utilize haptic feedback to improve surgical outcomes.
- **Telepresence robots**: telepresence robots are equipped with cameras, displays, and sensors to facilitate remote patient monitoring and virtual consultations. These factors allow health care professionals to interact with patients from afar. In a smart hospital utilizing 6G technology, telepresence robots take advantage of high bandwidth and low latency connectivity for real-time video communication, enabling healthcare professionals to assess patients remotely, provide guidance, and monitor their conditions effectively.
- **Robotic Process Automation (RPA):** Robotic Process Automation (RPA) automates repetitive and rule-based tasks in hospital workflows. In smart hospitals, RPA streamlines administrative processes such as patient registration, appointment scheduling, and billing. Automating these tasks helps reduce errors, enhance efficiency, and allow healthcare professionals to dedicate more time to patient care.
- **Pharmacy automation**: robotic systems in pharmacies automate medication dispensing, inventory management, and prescription filling. These systems handle medication orders with high accuracy and efficiency, reduce errors, and enhance medication safety. When integrated with 6G networks, these robotic systems enable real-time inventory tracking, automatic restocking, and seamless communication with healthcare providers to effectively manage medication.
- **Logistics and material handling:** robotics and automation play key roles in logistics and material handling within hospitals. Autonomous robots are deployed to navigate hospital premises, transport supplies, deliver medications, and assist with the movement of equipment and materials. When integrated with 6G networks, these robots achieve efficient task allocation, real-time tracking, and effective coordination, thereby enhancing the overall efficiency of hospital operations.
- **Robotic rehabilitation**: robotic systems are instrumental in patient rehabilitation and offer targeted exercises, support, and feedback to aid recovery. These systems are particularly beneficial for patients with mobility impairments, because they provide personalized therapy sessions and monitor progress. With the integration of 6G networks, these robotic systems allow for real-time monitoring, remote supervision, and personalized adjustments to therapy programs, thereby enhancing the efficacy of rehabilitation treatments.
- **Monitoring and surveillance robots**: robots equipped with sensors and cameras are used for monitoring and surveillance in hospitals. These robots can track vital signs, detect anomalies, and improve patient safety. The integration of these robots with 6G networks facilitates seamless data transmission, enabling real-time alerts and remote monitoring by healthcare professionals, thereby bolstering hospital security and patient care efficiency.
- **Maintenance and facility management**: robotics and automation play vital roles in hospital maintenance and facility management. Autonomous robots are deployed for routine inspection, equipment maintenance, and environmental monitoring. They efficiently identify and report issues, ensure prompt maintenance, and reduce equipment downtimes. With the integration of 6G networks, these robots facilitate efficient task management, support remote diagnostics, provide real-time status updates, and optimize hospital operations.

5.3.2 Challenges and 6G solution in implementation of robotics in smart hospital

6G connectivity can overcome the implementation challenges of robotics in smart hospitals by providing the necessary infrastructure for real-time communication, remote operation, data processing, and security, ultimately enhancing patient care delivery and operational efficiency. Integrating robotics and automation in a 6G-based smart hospital presents several challenges that must be addressed. The following are some of the key challenges (87, 88):

- Integration complexity: it can be difficult to integrate robotic systems into the current hospital infrastructure, necessitating major adjustments to the physical layouts, operational procedures, and IT infrastructure.
- Safety concerns: when using robots in healthcare environments, safety must come first because mistakes or malfunctions could endanger patients or cause accidents. Ensuring regulatory compliance and a safe environment for robots to engage with patients are critical.
- Training and education: to effectively operate and interact with robotic devices, healthcare workers require specific training. It is necessary to create and conduct training programs to guarantee staff competence and assurance when utilizing robotic technologies.
- Costs and return on investment: for robotics systems, upfront investments in equipment, maintenance, and training are often significant. In contrast to conventional care delivery models, hospitals must evaluate the robotic solutions' long-term cost-effectiveness and return on investment (ROI).
- Interoperability: for smooth communication and data transmission, it is crucial to provide interoperability between various robotic platforms, medical equipment, and hospital IT systems. Interoperability and data integration require standardized interfaces and protocols. Patching legacy systems in current hospitals with 6G technologies will need a well-thought-out plan to facilitate seamless transition and interoperability. Legacy systems, including Electronic Health Records (EHR), imaging equipment, and older diagnostic equipment, commonly use old communication

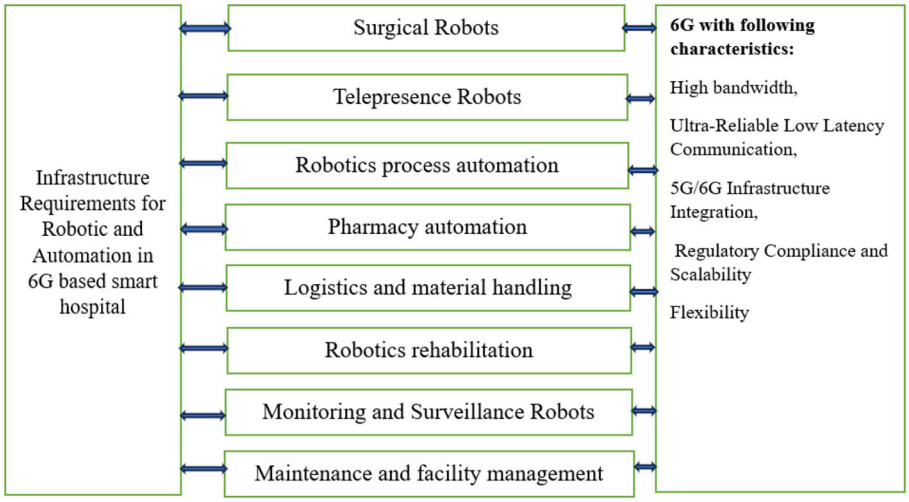


FIGURE 7
Infrastructure requirements for robotics and automation in 6G smart hospital.

protocols and infrastructure. To fill the gap, hospitals will have to implement middleware solutions and software adapters that allow such systems to interact with newer 6G-compatible devices and applications. Furthermore, network updates like moving toward hybrid cloud-edge models will facilitate integrating existing sources of data with greater throughput and less latency that is offered by 6G. The process of integration will also include the upgrade of legacy hardware to accommodate 6G-compatible standards, including low-power IoT sensors and AI-based devices for real-time monitoring. Notably, this process must ensure data security and healthcare regulation compliance to safeguard patient privacy. By embracing scalable, modular solutions, hospitals can future-proof their infrastructure while ensuring compatibility with current systems.

6G connectivity can assist with these issues by performing the following tasks (89):

- Minimal latency and maximum bandwidth: 6G networks provide incredibly low latency and maximum bandwidth, allowing robotic system management and real-time communication. This guarantees that commands and actions happen as quickly as possible, improving the responsiveness and agility of robotic platforms.
- Support for edge computing: by enabling data processing and analysis closer to the source of data generation, 6G networks' edge computing capabilities lower latency and bandwidth consumption. This increases the autonomy and efficiency of robotic systems by enabling real-time decision-making and feedback loops.
- Remote operation and monitoring: surgeons and other healthcare professionals can remotely operate robotic devices for telemedicine and telesurgery applications because of 6G's high-speed, low-latency connectivity. This enhances patient care outcomes by providing access to medical services and specialist knowledge regardless of one's location.
- Security and reliability: to safeguard data sent between robotic equipment and hospital IT infrastructure, 6G networks include cutting-edge security features including encryption, authentication, and intrusion detection. This reduces the risks posed by cyberattacks and illegal access, improving the security and dependability of robotic operations.

5.4 Analyzing real problem in Thailand hospital and solving with 6G based smart hospital

The high maintenance costs of access points in Thailand's public hospitals negatively impact the quality and accessibility of healthcare, compounded by tight resources and posing numerous obstacles for the general population (90). Population hospitals may offer more dependable and effective services by utilizing cutting-edge technologies to reduce their high maintenance costs. This would immediately benefit the general population by improving their access to high-quality healthcare (91). The following are the effects this issue has on the broader public (16, 92):

- **Decreased quality of care**

  ***Equipment downtime*:** longer downtimes resulting from medical equipment malfunctions frequently caused by poor maintenance can cut into the availability of crucial therapeutic and diagnostic services.

  Treatment delays: individuals may encounter delays in the provision of medical care or diagnostic services, thereby exacerbating health effects, particularly in urgent or essential circumstances.
- **Extended waiting periods**

  Overburdened facilities: when the remaining functioning equipment is out of commission, it leads to extended patient

wait times. This is especially troublesome in high-demand fields like radiology and emergency departments.

Appointment backlogs: when maintenance problems cause a backlog of appointments, patients may have to wait longer for planned consultations and procedures, which may worsen their health.

- **Higher cost**

High healthcare costs for patients: if public hospitals are unable to provide prompt services, patients may be compelled to seek care from private hospitals, resulting in higher out-of-pocket costs.

Indirect expenditures: patients' overall healthcare expenditures may rise as a result of treatment delays that prolong sickness and necessitate more involved and costly therapies down the road.

- **Restricted availability of specialized services**

Availability of specialized equipment: regular maintenance is necessary for specialized diagnosis and treatment equipment, such as CT scanners and MRI machines. High maintenance expenses may restrict the provision of these treatments in public hospitals, thereby requiring patients to travel great distances to receive the necessary care.

Equity issues: health inequities between urban and rural populations may worsen in rural and underserved areas due to restricted access to specialist equipment.

- **Effect on hospital**

Staff efficiency: when dealing with broken or unavailable equipment, medical personnel may experience elevated stress levels and lower productivity, which may have an adverse effect on their capacity to deliver high-quality care.

Instruction and adjustment: frequent equipment failures and the introduction of temporary solutions can disrupt the workflow, forcing personnel to constantly adjust to changing circumstances and potentially impacting the overall performance of the hospital.

- **Public health consequences**

Control of infectious diseases: to prevent the transmission of infectious diseases, it is essential to use dependable equipment and perform routine maintenance. Equipment malfunctions can jeopardize public health by impeding diagnostic capabilities and delaying the use of control measures.

Handling chronic illnesses: timely therapies and routine monitoring are essential for the effective management of chronic conditions such as hypertension and diabetes requires timely therapies and routine monitoring. Problems with equipment maintenance can interfere with continuing care strategies and worsen the health of individuals with chronic illnesses.

The above stated problem of high maintenance costs of access points in Thailand's public hospitals can be solved by deploying 6G networks and smart hospital technologies in the following manner:

- **Predictive upkeep:**

Predicting equipment failures with IoT sensors and data analytics may guarantee prompt maintenance, cutting downtime, and preserving service availability.

- **Remote diagnosis:**

High-speed, dependable remote diagnostics made possible by 6G networks allow professionals to handle maintenance issues without the need for in-person presence, resulting in faster problem resolution.

- **Optimized allocation of resources:**

By making the most use of the resources at hand, smart systems can minimize service interruptions by prioritizing maintenance on vital equipment.

- **Enhanced effectiveness:**

AI and automation can help hospitals run more efficiently, which will ease the workload for employees and increase the effectiveness of healthcare delivery as a whole.

The datasets used for experimental validation in the context of integrating 6G technology in smart hospitals should exhibit specific characteristics to accurately reflect real-world healthcare scenarios.

- Size: given the large-scale nature of smart hospitals, the datasets must be extensive, encompassing patient records, medical imaging, sensor data from IoMT devices, and real-time communication logs. These datasets should cover various aspects of healthcare, from diagnostics to treatment monitoring, to assess the impact of 6G-enabled solutions on data processing speed, latency, and bandwidth requirements.
- Diversity: the datasets should be diverse, representing a wide range of patient demographics, health conditions, and healthcare environments. This diversity is crucial to evaluating the performance of 6G in handling different medical applications, such as telemedicine, remote surgeries, and AI-driven diagnostics. The data should include structured formats (e.g., EHR) and unstructured formats (e.g., medical images, video feeds) to simulate the varied data inputs in smart hospitals.
- Challenges: one major challenge is ensuring data privacy and security, as sensitive patient information must be protected while transmitting over high-speed 6G networks. Additionally, data heterogeneity could pose integration issues, requiring effective data harmonization techniques. The computational complexity involved in handling large datasets for AI and real-time analytics also demands advanced processing capabilities, which could be another hurdle in the experimental setup.

5.5 Hybrid cloud-edge computing

Hybrid cloud-edge computing solutions present a strong alternative to 6G infrastructure in healthcare with an effective combination of cloud computing's scalability and edge computing's low-latency benefits. In healthcare, where real-time data processing and rapid decision-making are essential, this hybrid approach can maximize performance and cost-effectiveness. Edge computing

devices situated near medical devices, including wearables, monitoring devices, and surgical robots, handle the processing of data locally, minimizing latency and the requirement for ongoing cloud communication. Local processing is essential in time-sensitive applications like remote surgery and real-time patient monitoring, where delays can be disastrous (93). Alternatively, cloud computing offers centralized storage, computational capacity, and support for high-volume data analysis, useful for predictive diagnostics, machine learning, and patient record keeping. Cloud resources are capable of managing the heavy computational workloads of AI algorithms without loading edge devices with processing large amounts of data, thus providing high-level analytics and data backups without loading local infrastructure (94). A hybrid architecture alleviates the weaknesses of both cloud and edge computing. It ensures that healthcare systems are not totally reliant on cloud infrastructure, which is costly or prone to outages, and also refrains from the performance constraints of edge computing. Hybrid systems can enhance scalability, flexibility, and reliability, particularly for remote locations with poor network connectivity, where edge devices can operate independently. By integrating these technologies, healthcare systems can deliver ongoing, real-time care, maximize resource utilization, and maintain strong data privacy and security through local processing and cloud storage (95, 96).

6 Conclusion

This projected article presents a comprehensive study of 6G-based smart hospitals, exploring the architectural evolution, advanced techniques, and challenges associated with this cutting-edge healthcare paradigm. Our research highlights the transformative potential of 6G technology in revolutionizing healthcare delivery. The architectural evolution emphasizes the seamless integration of diverse technologies to create a robust and interconnected healthcare ecosystem. Advanced techniques such as Explainable AI, IoT, and Robotics optimize patient care, resource management, and operational efficiency, enhancing diagnostic accuracy, streamlining workflows, and improving patient outcomes. However, our study also reveals significant challenges accompanying 6G implementation in smart hospitals, including security and privacy concerns, interoperability issues, and the need for substantial investments. Striking a balance between innovation and security is crucial for widespread adoption. This study provides a roadmap for researchers, practitioners, and policymakers to navigate the evolving landscape of 6G-based smart hospitals as we stand on the cusp of a new era in healthcare technology characterized by unprecedented connectivity and intelligence. Future work should focus on fortifying security and privacy, developing robust encryption methods, authentication protocols, and privacy-preserving mechanisms to mitigate risks and ensure data integrity. Research should also explore user experience, human-machine interaction, and the integration of patient feedback to create technologies that enhance healthcare delivery while prioritizing the wellbeing of patients and providers. Limitations include high implementation costs, data security concerns, the need for advanced infrastructure, and the lack of detailed analysis of ethical issues and potential disparities in technology access.

Data availability statement

The raw data supporting the conclusions of this article will be made available, upon reasonable request to the corresponding author.

Author contributions

AK: Conceptualization, Writing – original draft. MM: Writing – review & editing, Investigation. MA: Formal analysis, Writing – review & editing. NG: Investigation, Visualization, Writing – review & editing. AN: Methodology, Validation, Visualization, Writing – review & editing, Resources, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by Taif University, Saudi Arabia, project number (TU-DSPP-2024-04).

Acknowledgments

The authors extend their appreciation to Taif University, Saudi Arabia, for supporting this work through project number (TU-DSPP-2024-04).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Nachev P, Herron D, McNally N, Rees G, Williams B. Redefining the research hospital. *NPJ Digital Medicine.* (2019) 2:119. doi: 10.1038/s41746-019-0201-2

2. Lloret J, Parra L, Taha M, Tomás J. An architecture and protocol for smart continuous eHealth monitoring using 5G. *Comput Netw.* (2017) 129:340–51. doi: 10.1016/j.comnet.2017.05.018

3. Li D. 5G and intelligence medicine—how the next generation of wireless technology will reconstruct healthcare? *Precis. Clin Med.* (2019) 2:205–8. doi: 10.1093/pcmedi/pbz020

4. Salwe SS, Naik KK. Heterogeneous wireless network for IoT applications. *IETE Techn Rev.* (2017) 36:61–8. doi: 10.1080/02564602.2017.1400412

5. Lopez-Villegas A, Catalan-Matamoros D, Peiro S, Lappegard KT, Lopez-Liria R. Cost–utility analysis of telemonitoring versus conventional hospital-based follow-up of patients with pacemakers. The NORDLAND randomized clinical trial. *PLoS ONE.* (2020) 15:e0226188. doi: 10.1371/journal.pone.0226188

6. Al-rawashdeh M, Keikhosrokiani, P, Belaton B, Alawida M, Zwiri A. IoT Adoption and application for smart healthcare: a systematic review. *Sensors.* (2022) 22:5377. doi: 10.3390/s22145377

7. Chiuchisan I, Costin H-N, Geman O. Adopting the Internet of Things technologies in health care systems. In: *2014 International Conference and Exposition on Electrical and Power Engineering (EPE)* (2014). p. 532–5. doi: 10.1109/ICEPE.2014.6969965

8. Malasinghe LP, Ramzan N, Dahal K. Remote patient monitoring: a comprehensive study. *J Ambient Intell Humaniz Comput.* (2017) 10:57–76. doi: 10.1007/s12652-017-0598-x

9. Zhang S, Guo S, Gao B, Hirata H, Ishihara H. Design of a novel telerehabilitation system with a force-sensing mechanism. *Sensors.* (2015) 15:11511–27. doi: 10.3390/s150511511

10. Kumar A, Dhanagopal R, Albreem MA, Le D-N. A comprehensive study on the role of advanced technologies in 5G based smart hospital. *Alexandria Eng J.* (2021) 60:5527–36. doi: 10.1016/j.aej.2021.04.016

11. Uslu BÇ, Okay E, Dursun E. Analysis of factors affecting IoT-based smart hospital design. *J Cloud Comput.* (2020) 9:67. doi: 10.1186/s13677-020-00215-5

12. Rodrigues L, Gonçalves I, Fé I, Endo PT, Silva FA. Performance and availability evaluation of an smart hospital architecture. *Computing.* (2021) 103:2401–35. doi: 10.1007/s00607-021-00979-x

13. Jiang N, Wang L, Xu X. Research on Smart Healthcare Services: Based on the Design of APP Health Service Platform. *J Healthc Eng.* (2021) 2021:1–8. doi: 10.1155/2021/9922389

14. Nasr M, Islam Md M, Shehata S, Karray F, Quintana Y. Smart healthcare in the age of ai: recent advances, challenges, and future prospects. *IEEE Access.* (2021) 9:145248–145270. doi: 10.1109/ACCESS.2021.3118960

15. Saba Raoof S, Durai MAS. A comprehensive review on smart health care: applications, paradigms, and challenges with case studies. *Contrast Media Molec Imag.* (2022) 2022:4822235. doi: 10.1155/2022/4822235

16. Kelly JT, Campbell KL, Gong E, Scuffham P. The internet of things: impact and implications for health care delivery. *J Med Internet Res.* (2020) 22:e20135. doi: 10.2196/20135

17. Tian S, Yang W, Grange JML, Wang P, Huang W, Ye Z. Smart healthcare: making medical care more intelligent. *Global Health J.* (2019) 3:62–5. doi: 10.1016/j.glohj.2019.07.001

18. Baker SB, Xiang W, Atkinson I. Internet of Things for smart healthcare: technologies, challenges, and opportunities. *IEEE Access.* (2017) 5:26521–44. doi: 10.1109/ACCESS.2017.2775180

19. Almotairi KH. Application of internet of things in healthcare domain. *J Umm Al-Qura Univ Eng Archit.* (2022) 14:1–12. doi: 10.1007/s43995-022-00008-8

20. Latif G, Shankar A, Alghazo JM, Kalyanasundaram V, Boopathi CS, Arfan Jaffar M, et al. A advancing health diagnosis and medication through IoT. *Wireless Networks.* (2019) 26:2375–89. doi: 10.1007/s11276-019-02165-6

21. Ali Tunc M, Gures E, Shayea I. A survey on IoT smart healthcare: emerging technologies, applications, challenges, and future trends. *arXiv:2109.02042* (2021).

22. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *J Big Data.* (2019) 6:54. doi: 10.1186/s40537-019-0217-0

23. de Alwis C, Quoc-Viet P, Madhusanka L. 6G for Healthcare. In: *6G Frontiers: Towards Future Wireless Systems.* IEEE (2023). p. 189–196. doi: 10.1002/9781119862321.ch14

24. Elaziz MA, Dahou A, Mabrouk A, Ibrahim RA, Aseeri AO. Medical image classifications for 6G IoT-enabled smart health systems. *Diagnostics.* (2023) 13:834. doi: 10.3390/diagnostics13050834

25. Wijethilaka S, Porambage P, de Alwis C, Liyanage M. A comprehensive analysis on network slicing for smart hospital applications. In: *2022 IEEE 19th Annual Consumer Communications Networking Conference (CCNC)* (2022). p. 276–9. doi: 10.1109/CCNC49033.2022.9700535

26. Adat V, Gupta BB. Security in Internet of Things: issues, challenges, taxonomy, and architecture. *Telecommun Syst.* (2017) 67:423–41. doi: 10.1007/s11235-017-0345-9

27. Kumar A, Nanthaamornphong A, Selvi R, Venkatesh J, Alsharif MH, Uthansakul P, et al. Evaluation of 5G techniques affecting the deployment of smart hospital infrastructure: understanding 5G, AI and IoT role in smart hospital. *Alexandria Eng J.* (2023) 83:335–54. doi: 10.1016/j.aej.2023.10.065

28. Ahad A, Tahir M, Yau K-LA. 5G-based smart healthcare network: architecture, taxonomy, challenges and future research directions. *IEEE Access.* (2019) 7:100747–62. doi: 10.1109/ACCESS.2019.2930628

29. Soldani D, Fadini F, Rasanen H, Duran J, Niemela T, Chandramouli D, et al. 5G mobile systems for healthcare. In: *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)* (2017). p. 1–5. doi: 10.1109/VTCSpring.2017.8108602

30. Parwani K, Purohit GN. Performance measures of mobile communication in a hierarchical cellular system. In: *2011 International Conference on Devices and Communications (ICDeCom)* (2011). p. 1–5. doi: 10.1109/ICDECOM.2011.5738560

31. Boudlal H, Serrhini M, Tahiri A. Towards an SDN/NFV based network infrastructure for hospital information systems and healthcare services. In: *2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-Based Interconnected Digital Worlds (NISS)* (2022). p. 1–5. doi: 10.1109/NISS55057.2022.10085476

32. Morgan AA, Abdi J, Syed MAQ, Kohen GE, Barlow P, Vizcaychipi MP. Robots in healthcare: a scoping review. *Curr Robot Rep.* (2022) 3:271–80. doi: 10.1007/s43154-022-00095-4

33. Ramakrishnan B, Kumar A, Chakravarty S, Masud M, Baz M. Analysis of FBMC waveform for 5G network based smart hospitals. *Appl Sci.* (2021) 11:8895. doi: 10.3390/app11198895

34. Alkhomsan MN, Hossain MA, Rahman S, Md M, Masud M. Situation awareness in ambient assisted living for smart healthcare. *IEEE Access.* (2017) 5:20716–20725. doi: 10.1109/ACCESS.2017.2731363

35. Lin T-W, Hsu C-L. FAIDM for medical privacy protection in 5G telemedicine systems. *Appl Sci.* (2021) 11:1155. doi: 10.3390/app11031155

36. Rajaei O, Khayami SR, Rezaei MS. Smart hospital definition: academic and industrial perspective. *Int J Med Inform.* (2024) 182:105304. doi: 10.1016/j.ijmedinf.2023.105304

37. Winter A, Haux R, Ammenwerth E, Brigl B, Hellrung N, Jahn F. Health information systems. In: *Health Information Systems* (2010) 33–42. doi: 10.1007/978-1-84996-441-8_4

38. Bin Ahammed T, Patgiri R. 6G and AI: the emergence of future forefront technology. In: *2020 Advanced Communication Technologies and Signal Processing (ACTS)* (2020). p. 1–6. doi: 10.1109/ACTS49415.2020.9350396

39. Gupta A, Jha RK, A. Survey of 5G network: architecture and emerging technologies. *IEEE Access.* (2015) 3:1206–32. doi: 10.1109/ACCESS.2015.2461602

40. Philip NY, Rehman IU. Towards 5G health for medical video streaming over small cells. In: *XIV Mediterranean Conference on Medical and Biological Engineering and Computing* (2016). p. 1093–8. doi: 10.1007/978-3-319-32703-7_215

41. Altaf Khattak SB, Nasralla MM, Rehman IU. The role of 6G networks in enabling future smart health services and applications. In: *2022 IEEE International Smart Cities Conference (ISC2)* (2022). p. 1–7. doi: 10.1109/ISC255366.2022.9922093

42. Aljabr AA, Kumar K. Design and implementation of Internet of Medical Things (IoMT) using artificial intelligent for mobile-healthcare. *Measurement: Sensors.* (2022) 24:100499. doi: 10.1016/j.measen.2022.100499

43. Rizwan P, Rajasekhara Babu M, Suresh K. Design and development of low investment smart hospital using internet of things through innovative approaches. *Biomed Res.* (2017) 28:4979–85.

44. Shen F, Shi H, Yang Y. A comprehensive study of 5G and 6G networks. In: *2021 International Conference on Wireless Communications and Smart Grid (ICWCSG)* (2021). doi: 10.1109/ICWCSG53609.2021.00070

45. Ahad A, Ali Z, Mateen A, Tahir M, Hannan A, Garcia NM, et al. A Comprehensive review on 5G-based Smart Healthcare Network Security: taxonomy, issues, solutions and future research directions. *Array.* (2023) 18:100290. doi: 10.1016/j.array.2023.100290

46. Upadhyaya P, Ruchi, Dutt S. Transfer learning approach for 6G-IoT applications. In: *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, (2022). p. 421–424. doi: 10.1109/ICCES54183.2022.9835931

47. Vergütz A, Prates GN, Henrique Schwengber B, Santos A, Nogueira M. An architecture for the performance management of smart healthcare applications. *Sensors.* (2020) 20:5566. doi: 10.3390/s20195566

48. Cisotto G, Casarin E, Tomasin S. Requirements and enablers of advanced healthcare services over future cellular systems. *IEEE Commun Mag.* (2020) 58:76–81. doi: 10.1109/MCOM.001.1900349

49. Liu C, Li Y, Fang M, Liu F. Using machine learning to explore the determinants of service satisfaction with online healthcare platforms during the COVID-19 pandemic. *Service Business.* (2023) 17:449–76. doi: 10.1007/s11628-023-00535-x

50. Saranya A, Subhashini R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decis Analyt J.* (2023) 7:100230. doi: 10.1016/j.dajour.2023.100230

51. Hong E-K, Lee I, Shim B, Ko Y-C, Kim S-H, Pack S, et al. 6G RD vision: requirements and candidate technologies. *J Commun Netw.* (2022) 24:232–45. doi: 10.23919/JCN.2022.000015

52. Quy VK, Chehri A, Quy NM, Han ND, Ban NT. Innovative trends in the 6G era: a comprehensive survey of architecture, applications, technologies, and challenges. *IEEE Access.* (2023) 11:39824–44. doi: 10.1109/ACCESS.2023.3269297

53. Uusitalo MA, Rugeland P, Boldi MR, Strinati EC, Demestichas P, Ericson M, et al. 6G vision, value, use cases and technologies from European 6G flagship project hexa-X. *IEEE Access.* (2021) 9:160004–20. doi: 10.1109/ACCESS.2021.3130030

54. Dhanda SS, Singh B, Jindal P, Sharma TK, Panwar D. 6G-enabled internet of medical things. *Expert Syst.* (2023) 41:e134722023. doi: 10.1111/exsy.13472

55. Chen M, Yang J, Zhou J, Hao Y, Zhang J, Youn C-H. 5G-smart diabetes: toward personalized diabetes diagnosis with healthcare big data clouds. *IEEE Commun Mag.* (2018) 56:16–23. doi: 10.1109/MCOM.2018.1700788

56. Mucchi L, Jayousi S, Caputo S, Paoletti E, Zoppi P, Geli S, et al. How 6G technology can change the future wireless healthcare. In: *2020 2nd 6G Wireless Summit (6G SUMMIT)* (2020) 1–6. doi: 10.1109/6GSUMMIT49458.2020.9083916

57. Gupta M, Jha RK, Jain S. Tactile based intelligence touch technology in IoT configured WCN in B5G/6G-A survey. *IEEE Access.* (2023) 11:30639–89. doi: 10.1109/ACCESS.2022.3148473

58. Wang M, Zhu T, Zhang T, Zhang J, Yu S, Zhou W. Security and privacy in 6G networks: New areas and new challenges. *Digital Commun Netw.* (2020) 6:281–91. doi: 10.1016/j.dcan.2020.07.003

59. Kumar A, Jain R, Gupta M, Sardar MN. *6G-Enabled IoT and AI for Smart Healthcare Challenges, Impact, and Analysis*. New York: CRC Press.

60. Padhi P, Charrua-Santos F. 6G enabled tactile internet and cognitive internet of healthcare everything: towards a theoretical framework. *Appl Syst Innov.* (2021) 4:66. doi: 10.3390/asi4030066

61. Rodrigues VF, Righi Rda R, da Costa CA, Antunes RS. Smart hospitals and IoT sensors: why is QoS essential here? *J Sensor Actuator Netw.* (2022) 11:33. doi: 10.3390/jsan11030033

62. Ahad A, Jiangbina Z, Tahir M, Shayea I, Sheikh MA, Rasheed F. 6G and intelligent healthcare: taxonomy, technologies, open issues and future research directions. *Internet Things.* (2024) 25:101068. doi: 10.1016/j.iot.2024.101068

63. Chettri L, Bera R. A comprehensive survey on internet of things (IoT) toward 5G wireless systems. *IEEE Internet Things J.* (2020) 7:16–32. doi: 10.1109/JIOT.2019.2948888

64. Dash SP. The impact of IoT in healthcare: global technological change & the roadmap to a networked architecture in India. *J Indian Inst Sci.* (2020) 100:773–85. doi: 10.1007/s41745-020-00208-y

65. Mejía-Granda CM, Fernández-Alemán JL, Carrillo-de-Gea JM, García-Berná JA. Security vulnerabilities in healthcare: an analysis of medical devices and software. *Med Biol Eng Comput.* (2023) 62:257–273. doi: 10.1007/s11517-023-02912-0

66. Kruse CS, Smith B, Vanderlinden H, Nealand A. Security techniques for the electronic health records. *J Med Syst.* (2017) 41:127. doi: 10.1007/s10916-017-0778-4

67. Kumari A, Gupta R, Tanwar S. Amalgamation of blockchain and IoT for smart cities underlying 6G communication: a comprehensive review. *Comput Commun.* (2021) 172:102–18. doi: 10.1016/j.comcom.2021.03.005

68. Ramezanpour K, Jagannath J. Intelligent zero trust architecture for 5G/6G networks: principles, challenges, and the role of machine learning in the context of O-RAN. *Comput Netw.* (2022) 217:109358. doi: 10.1016/j.comnet.2022.109358

69. Bechini A, Corcuera Barcena JL, Ducange P, Marcelloni F, Renda A. Increasing accuracy and explainability in fuzzy regression trees: an experimental analysis. In: *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (2022) 1–8. doi: 10.1109/FUZZ-IEEE55066.2022.9882604

70. Dosilovic FK, Brcic M, Hlupic N. Explainable artificial intelligence: a survey. In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (2018). p. 0210–0215. doi: 10.23919/MIPRO.2018.8400040

71. Band S, Yarahmadi A, Hsu C-C, Biyari M, Sookhak M, Ameri R, et al. Application of explainable artificial intelligence in medical health: a systematic review of interpretability methods. *Inform Med Unlocked.* (2023) 40:101286. doi: 10.1016/j.imu.2023.101286

72. Alonso SG, Marques G, Barrachina I, Garcia-Zapirain B, Arambarri J, Salvador JC, et al. Telemedicine and e-Health research solutions in literature for combatting COVID-19: a systematic review. *Health Technol.* (2021) 11:257–266. doi: 10.1007/s12553-021-00529-7

73. Peral J, Ferrandez A, Gil D, Munoz-Terol R, Mora H. An ontology-oriented architecture for dealing with heterogeneous data applied to telemedicine systems. *IEEE Access.* (2018) 6:41118–38. doi: 10.1109/ACCESS.2018.2857499

74. Jagannath J, Polosky N, Jagannath A, Restuccia F, Melodia T. Machine learning for wireless communications in the Internet of Things: a comprehensive survey. *Ad Hoc Networks.* (2019) 93:101913. doi: 10.1016/j.adhoc.2019.101913

75. Rahman A, Hossain MS, Muhammad G, Kundu D, Debnath T, Rahman M, et al. Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues. *Cluster Comput.* (2022) 26:2271–2311. doi: 10.1007/s10586-022-03658-4

76. Rahman A, Debnath T, Kundu D, Khan MSI, Aishi AA, Sazzad S, et al. Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health.* (2024) 11:58–109. doi: 10.3934/publichealth.2024004

77. Wang S, Qureshi MA, Miralles-Pechuán L, Huynh-The T, Gadekallu TR, Liyanage M. Explainable AI for 6G use cases: technical aspects and research challenges. *IEEE Open J Commun Soc.* (2024) 5:2490–540. doi: 10.1109/OJCOMS.2024.3386872

78. Mehta V, Deb P, Rao DS. Application of computer techniques in medicine. *Med J Armed Forces India.* (1994) 50:215–8. doi: 10.1016/S0377-1237(17)31065-1

79. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med.* (2019) 25:30–6. doi: 10.1038/s41591-018-0307-0

80. Vergutz A, Noubir G, Nogueira M. Reliability for smart healthcare: a network slicing perspective. *IEEE Netw.* (2020) 34:91–7. doi: 10.1109/MNET.011.1900458

81. Chhor J, Gong Y, Rau P-LP. Breakout: design and evaluation of a serious game for health employing intel realsense. *Cross-Cultural Design.* (2017) 2:531–45. doi: 10.1007/978-3-319-57931-3_42

82. Cavoukian A, Fisher A, Killen S, Hoffman DA. Remote home health care technologies: how to ensure privacy? Build it in: privacy by design. *Identity Inf Soc.* (2010) 3:363–78. doi: 10.1007/s12394-010-0054-y

83. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. *J Big Data.* (2018) 5:1. doi: 10.1186/s40537-017-0110-7

84. Nasser N, Emad-Ul-Haq Q, Imran M, Ali A, Razzak I, Al-Helali A. A smart healthcare framework for detection and monitoring of COVID-19 using IoT and cloud computing. *Neural Comput Appl.* (2023) 35:13775–13789. doi: 10.1007/s00521-021-06396-7

85. Tavakoli M, Carriere J, Torabi A. Robotics, smart wearable technologies, and autonomous intelligent systems for healthcare during the COVID-19 pandemic: an analysis of the state of the art and future vision. *Adv Intell Syst.* (2020) 2:71. doi: 10.1002/aisy.202000071

86. Ouchani S, Krichen M. Ensuring the correctness and well modeling of intelligent healthcare management systems. In: *The Impact of Digital Technologies on Public Health in Developed and Developing Countries* (2020). p. 364–72. doi: 10.1007/978-3-030-51517-1_33

87. Cresswell K, Cunningham-Burley S, Sheikh A. Health care robotics: qualitative exploration of key challenges and future directions. *J Med Internet Res.* (2018) 20:e10410. doi: 10.2196/10410

88. Silvera-Tawil D. Robotics in healthcare: a survey. *SN Comput Sci.* (2024) 5:189. doi: 10.1007/s42979-023-02551-0

89. Iyer S, Looi T, Drake J. A single arm, single camera system for automated suturing. In: *2013 IEEE International Conference on Robotics and Automation* (2013). p. 239–44. doi: 10.1109/ICRA.2013.6630582

90. Huang G, Ng ST, Li D. Determinants of digital twin adoption in hospital operation management. *Urban Lifeline.* (2023) 1:6. doi: 10.1007/s44285-023-00005-w

91. Bovenizer W, Chetthamrongchai P. A comprehensive systematic and bibliometric review of the IoT-based healthcare systems. *Cluster Comput.* (2023) 26:3291–317. doi: 10.1007/s10586-023-04047-1

92. Kumar A, Gaur N, Nanthaamornphong A. Improving the latency for 5G/B5G based smart healthcare connectivity in rural area. *Sci Rep.* (2024) 14:6976. doi: 10.1038/s41598-024-57641-7

93. Alsubai S, Sha M, Alqahtani A, Bhatia M. Hybrid IoT-edge-cloud computing-based athlete healthcare framework: digital twin initiative. *Mobile Netw Applic.* (2023) 28:2056–75. doi: 10.1007/s11036-023-02200-z

94. Quy VK, Hau NV, Anh DV, Ngoc LA. Smart healthcare IoT applications based on fog computing: architecture, applications and challenges. *Complex Intell Syst.* (2021) 8:3805–15. doi: 10.1007/s40747-021-00582-9

95. Ghadi YY, Shah SFA, Mazhar T, Shahzad T, Ouahada K, Hamam H. Enhancing patient healthcare with mobile edge computing and 5G: challenges and solutions for secure online health tools. *J Cloud Comput.* (2024) 13:93. doi: 10.1186/s13677-024-00654-4

96. Al-Jawad F, Alessa R, Alhammad S, Ali B, Al-Qanbar M, Rahman A. Applications of 5G and 6G in smart health services. *Int J Comput Sci Netw Secur.* (2022) 22:173–84. doi: 10.22937/IJCSNS.2022.22.3.23

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 21 May 2025
DOI 10.3389/fmed.2025.1569464

OPEN ACCESS

EDITED BY
SeongKi Kim,
Chosun University, Republic of Korea

REVIEWED BY
Surbhi Bhatia Khan,
University of Salford, United Kingdom
Osman Ali Sadek Ibrahim,
Minia University, Egypt
Abhishek Singhal,
Amity University, India

*CORRESPONDENCE
Theyazn H. H. Aldhyani
✉ taldhyani@kfu.edu.sa
Sultan Ahmad
✉ s.alisher@psau.edu.sa

RECEIVED 31 January 2025
ACCEPTED 16 April 2025
PUBLISHED 21 May 2025

CITATION
Al-Nefaie AH, Aldhyani THH, Ahmad S and Alzahrani EM (2025) Application of artificial intelligence in modern healthcare for diagnosis of autism spectrum disorder. *Front. Med.* 12:1569464. doi: 10.3389/fmed.2025.1569464

Application of artificial intelligence in modern healthcare for diagnosis of autism spectrum disorder

Abdullah H. Al-Nefaie[1,2], Theyazn H. H. Aldhyani[1,3]*, Sultan Ahmad[4]* and Eidah M. Alzahrani[5]

[1]King Salman Center for Disability Research, Riyadh, Saudi Arabia, [2]Department of Quantitative Methods, School of Business, King Faisal University, Hofuf, Saudi Arabia, [3]Applied College in Abqaiq, King Faisal University, Hofuf, Saudi Arabia, [4]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia, [5]Computer Science Department, Al-Baha University, Al-Bahah, Saudi Arabia

Introduction: Symptoms of autism spectrum disorder (ASD) range from mild to severe and are evident in early childhood. Children with ASD have difficulties with social interaction, language development, and behavioral regulation. ASD is a mental condition characterized by challenges in communication, restricted behaviors, difficulties with speech, non-verbal interaction, and distinctive facial features in children. The early diagnosis of ASD depends on identifying anomalies in facial function, which may be minimal or missing in the first stages of the disorder. Due to the unique behavioral patterns shown by children with ASD, facial expression analysis has become an effective method for the early identification of ASD.

Methods: Hence, utilizing deep learning (DL) methodologies presents an excellent opportunity for improving diagnostic precision and efficacy. This study examines the effectiveness of DL algorithms in differentiating persons with ASD from those without, using a comprehensive dataset that includes images of children and ASD-related diagnostic categories. In this research, ResNet50, Inception-V3, and VGG-19 models were used to identify autism based on the facial traits of children. The assessment of these models used a dataset obtained from Kaggle, consisting of 2,940 face images.

Results: The suggested Inception-V3 model surpassed current transfer learning algorithms, achieving a 98% accuracy rate.

Discussion: Regarding performance assessment, the suggested technique demonstrated advantages over the latest models. Our methodology enables healthcare physicians to verify the first screening for ASDs in children.

KEYWORDS

transfer learning, deep learning, diagnosis, disability, mental health

1 Introduction

Autism Spectrum Disorder (ASD) represents one of the most significant challenges in modern neurodevelopmental medicine, affecting ~1 in 36 children globally (1). This complex condition, characterized by difficulties in social interaction, communication patterns, and repetitive behaviors, demands early intervention for optimal outcomes (2). ASD is identified based on deficiencies in behavioral skills and social communication, often seen via recurrent behavioral indicators in children. Figure 1 displays the symptoms of ASD. However, traditional diagnostic procedures usually involve time-intensive behavioral assessments and costly medical evaluations, creating substantial barriers to early detection, particularly in resource-limited settings (3).

FIGURE 1
Symptoms of ASD.

Recent advances in artificial intelligence, particularly in the domain of deep learning and computer vision, have opened promising new avenues for ASD screening (4, 5). The emerging field of facial phenotype analysis is of particular interest, which leverages the observation that individuals with ASD often present distinct facial morphological characteristics (6). These features, including broader upper faces, wider eyes, shorter nasal bridges, and narrower cheeks, have been increasingly recognized as potential biomarkers for ASD detection (6).

Timely diagnosis facilitates the use of specialist therapies designed to address the unique requirements of persons with autism, focusing on social communication, language development, and behavioral issues. Moreover, early diagnosis allows families to get suitable support services, educational resources, and community activities, enhancing coping strategies, alleviating parental stress, and promoting adult independence.

Nonetheless, early identification of autism by traditional methods also has specific threats. A significant concern is the potential for labeling, which may impact the child's self-esteem and social relationships. There is a risk of overdiagnosis or misdiagnosis, leading to unnecessary interventions and therapies. The diagnostic procedure may be delayed, intricate, and emotionally testing for families, necessitating thorough evaluations by multidisciplinary teams. Consequently, using sophisticated approaches supported by artificial intelligence (AI) may mitigate this danger, as AI utilizes technology capable of incorporating feedback from youngsters, informed by their expertise. In this study, we used facial images of children to identify those suffering from ASD.

The integration of deep learning methodologies with facial analysis represents a potentially transformative approach to ASD screening. Contemporary deep learning architectures have demonstrated remarkable capabilities in extracting complex patterns from facial images, offering the possibility of automated, rapid, and cost-effective screening tools. This approach aligns with the growing need for accessible screening methods that can support healthcare professionals in identifying individuals who may require comprehensive diagnostic evaluation.

This research presents a novel deep learning framework for ASD detection through facial image analysis. Our study evaluates the performance of three state-of-the-art deep learning architectures: ResNet, VGG16, and VGG19. Through rigorous

experimentation and validation, we demonstrate that the VGG19 architecture achieves superior performance with an accuracy of 98%, representing a significant advancement in automated ASD screening capabilities.

The primary contributions of this study include:

1. A comprehensive evaluation of DL architectures for facial image-based ASD detection.
2. The development of an optimized VGG19-based model achieving 98% accuracy.
3. Analysis of the specific facial features that contribute most significantly to accurate ASD detection.

This research aims to advance the field of automated ASD screening, potentially reducing the burden on healthcare systems while accelerating the identification of individuals who may benefit from early intervention. Our findings suggest that deep learning-based facial analysis could serve as a valuable complementary tool in the ASD diagnostic process, particularly in settings where access to traditional diagnostic resources is limited.

The research gap in ASD identification using images persists, despite the proposed system achieving 98% accuracy on a benchmark dataset. Different signals in facial expressions make it challenging to identify using advanced deep learning models, which may aid in predicting ASD. Ultimately, clinical validation is necessary to ensure the widespread adoption of this approach in healthcare settings and its practical applicability.

2 Related work

Early detection of ASD is crucial for effective intervention and treatment (7). While traditional diagnostic methods rely on clinical observations and behavioral assessments such as the Autism Diagnostic Observation Schedule (ADOS) (8), recent years have seen significant advancement in automated detection approaches. These advancements span multiple modalities, including facial analysis (9), magnetic resonance imaging (MRI) (10), eye tracking (11, 12), and electroencephalography (EEG) (13). The emergence of sophisticated machine learning and deep learning techniques has particularly accelerated the development of automated diagnostic systems across these modalities (13), offering promising tools for early screening and detection.

Akter et al. (14) conducted work using transfer learning, working with a dataset of 2,936 facial images from Kaggle. Their study evaluated multiple machine learning classifiers and pre-trained CNN models, with their improved MobileNet-V1 model achieving an accuracy of 90.67%. They used K-means clustering to identify potential ASD subtypes, achieving 92.10% accuracy for two autism subtypes. Elshoky et al. (15) comprehensively compared machine learning approaches using facial images from Kaggle. Their study uniquely compared classical machine learning, deep learning, and automated machine learning (AutoML) approaches. Using OpenCV for pre-processing with 90×90 pixel resizing and grayscale conversion, their AutoML approach achieved ~96% accuracy, significantly outperforming classical ML 72.64% with Extra Trees and deep learning methods using VGG16, which achieved 89%.

Li et al. (16) introduced a two-phase transfer learning approach using MobileNetV2 and MobileNetV3-Large. Their method transferred knowledge from ImageNet to facial images from Kaggle. This mobile-optimized approach achieved 90.5% accuracy with an AUC of 96.32%. Siagian et al. (17) took a different approach, using a unique dataset of 200 facial images collected from special schools in Medan, Indonesia. Their method combined the SURF (Speeded-Up Robust Features) algorithm with various boosting methods, achieving 91.67% accuracy with Gradient Boosting despite the relatively small dataset.

Alkahtani et al. (18) explored a hybrid approach combining pre-trained CNNs with traditional machine learning classifiers. Their study utilized MobileNetV2 and VGG19 as feature extractors, paired with various classifiers machine learning algorithms. Working with a publicly available dataset, their optimized MobileNetV2 configuration, using the Adamax optimizer with a learning rate of 0.001, achieved 92% accuracy. Sai Koppula and Agrawal (19) evaluated multiple pre-trained CNN architectures with a focus on domain-specific variations. Using the Kaggle dataset, they implemented extensive data augmentation through Keras' ImageDataGenerator. Their study revealed that models pre-trained on VGGFace2 outperformed those trained on ImageNet, with VGG16 achieving 86% accuracy and AUC. Abdullah et al. (20) explored an ensemble approach that combined the EfficientNet B5, MobileNet, and InceptionV3 models using the Kaggle dataset. Their method employed data augmentation techniques and utilized a soft voting ensemble method, achieving an accuracy of 89.87%. Karthik et al. (21) investigated hybrid deep learning models using Vision Transformers (ViT) with various classifiers. Working with the Kaggle dataset, they implemented comprehensive pre-processing, including grayscale conversion, resizing to 224×224 pixels, normalization, and extensive augmentation. Their ViT model, combined with XGBoost and SHAP implementation, achieved 91.3% accuracy.

Pan and Foroughi (22) focused on edge computing applications, adapting AlexNet for efficient processing in educational environments using the Kaggle dataset. Their implementation achieved 93.24% accuracy while maintaining real-time processing capabilities, demonstrating the feasibility of edge deployment for ASD screening tools. Shahzad et al. (23) introduced a hybrid attention-based model combining ResNet101 and EfficientNetB3. Their approach incorporated self-attention mechanisms from natural language processing and emphasized standardized pre-processing with image augmentation through rotations, zooming, and flipping. The hybrid attention-based model achieved an accuracy of 96.50%. Reddy and Andrew (24) conducted a comparative study of three pre-trained Convolutional Neural Network (CNN) architectures: VGG16, VGG19, and EfficientNetB0. Their investigation utilized a dataset of facial images of children, implementing comprehensive data augmentation techniques, including rotation, horizontal flipping, zooming, and height/width shifting. Images were standardized to 227 × 227 × 3 pixels to ensure compatibility with the CNN architectures. Their findings revealed that EfficientNetB0 achieved the highest accuracy at 87.9%, surpassing both VGG16 84.66% and VGG19 80.05%. Table 1 displays the different existing systems that have been developed for the diagnosis of ASD.

TABLE 1 Existing using facial images.

Study	Dataset	Methods/models	Key findings/accuracy
Akter et al. (2021) (14)	Autism Face Image Dataset	Transfer learning with MobileNet-V1, K-means clustering	MobileNet-V1: 90.67%; Clustering: 92.10% for ASD subtypes
Elshoky et al. (2022) (15)	Autism Face Image Dataset	Classical ML, Deep Learning (VGG16), AutoML	AutoML: 96%; VGG16: 89%; Classical ML (Extra Trees): 72.64%
Li et al. (2023) (16)	Autism Face Image Dataset	Two-phase transfer learning	MobileNetV3-Large: 90.5%, AUC: 96.32%
Siagian et al. (2023) (17)	Special dataset of 200 images	Gradient Boosting with SURF features	Gradient Boosting: 91.67%
Alkahtani et al. (2023) (18)	Autism Face Image Dataset	MobileNetV2, VGG19 with various classifiers	MobileNetV2: 92%
Sai Koppula and Agrawal (2023) (19)	Autism Face Image Dataset	VGGFace2 vs. ImageNet-based pre-trained CNNs	VGG16 (VGGFace2): 86%, AUC: Not specified
Abdullah et al. (2024) (20)	Autism Face Image Dataset	Ensemble (EfficientNetB5, MobileNet, InceptionV3)	Ensemble: 89.87%
Karthik et al. (2024) (21)	Autism Face Image Dataset	Vision Transformers (ViT) with XGBoost and SHAP	ViT + XGBoost: 91.3%
Pan and Foroughi (2024) (22)	Autism Face Image Dataset	Edge-optimized AlexNet	AlexNet: 93.24%
Shahzad et al. (2024) (23)	Autism Face Image Dataset	ResNet101 + EfficientNetB3 hybrid with self-attention	Hybrid: 96.50%
Reddy and Andrew (2024) (24)	Autism Face Image Dataset	VGG16, VGG19, EfficientNetB0	EfficientNetB0: 87.9%; VGG16: 84.66%; VGG19: 80.05%

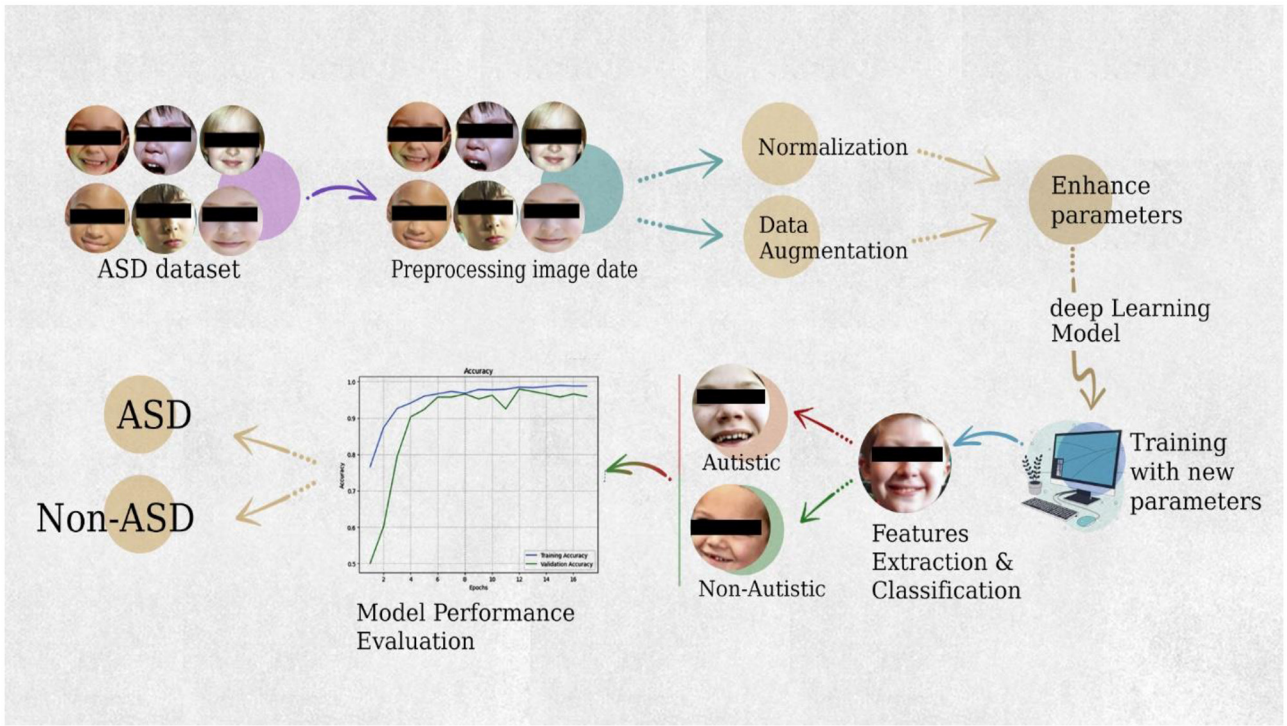


FIGURE 2
Enhanced diagnosis ASD system.

3 Materials and methods

This research used DL models to predict and classify ASD in children at an early stage. This framework was developed using autistic face features. This study used pre-trained DL models to automatically extract robust characteristics of children's faces to detect ASD. The framework of the proposed ASD system is presented in Figure 2.

3.1 Dataset

The research used face images of autistic children from a publicly accessible collection (Kaggle). The dataset included 2D RGB images of children aged 2–14. The dataset was designed into two subfolders: one designated for autistic children and the other for non-autistic children. The autistic subfolder included images of ASD, while the non-autistic subfolder had images randomly retrieved from web searches, as shown in Table 2. The images were sized at 224 × 224 × 3, providing a comparative overview of ASD and non-ASD images. The snapshots of images of ASD and non-ASD are presented in Figure 3.

3.2 Pre-processing approach

3.2.1 Data augmentation

Data augmentation is process to generating additional data from existing datasets to train deep learning models, which might be complicated by data silos, restrictions, and other constraints, by minor modifications to the original data. This study employs data augmentation to enhance the model's efficacy by artificially expanding the training dataset by transformations such as flipping, shearing, zooming, and rescaling, as shown in Table 3. These parameters mitigate overfitting when the model retains training data rather than acquiring generalized patterns, thereby improving the model's efficacy. The ASD and Non-ASD images in standard collections may be constrained in size; augmentation artificially enhances them by rescaling pixel values to [0, 1], shearing images by 10%, zooming by 10%, and performing horizontal flipping.

3.2.2 Data splitting

The dataset is partitioned into three sets: training (80%), validation (10%), and test (10%). This guarantees that the model is tested on unknown data for improved generalizability. The class volume of the ASD dataset is presented in Figure 4.

3.3 Deep learning models

3.3.1 Inception-V3 models

Google presented the Inception-V3 pre-trained model. It includes symmetrical and asymmetrical construction blocks,

TABLE 2 Samples of dataset.

Dataset	Number
Total_images	2,940
Autistic_children	1,327
Non-autistic_childern	1,613

TABLE 3 Augmentation parameters.

Indicators	Values
Shear_Range method	0.1
Zoom_Range method	0.1
Horizontal_Flip method	True

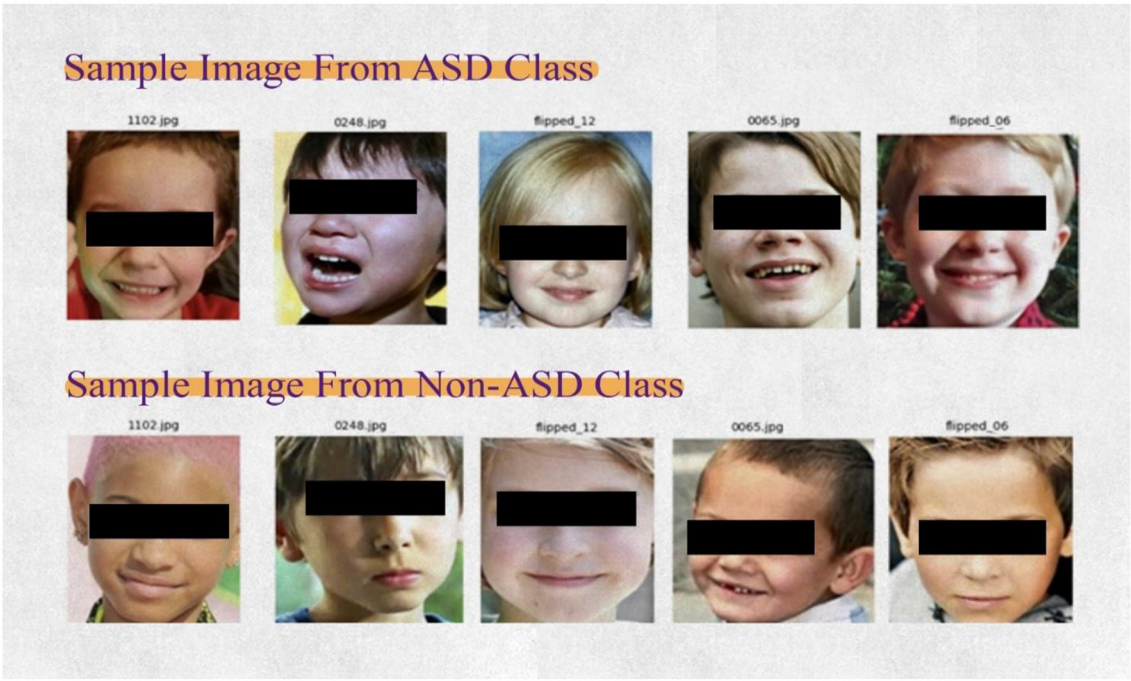


FIGURE 3
Snapshot of dataset.

FIGURE 4
Class ASD dataset.

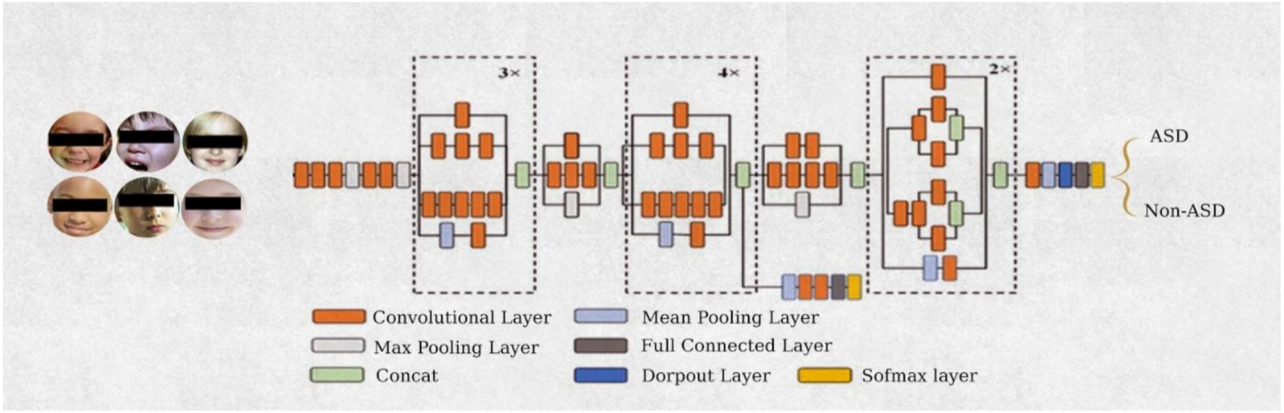


FIGURE 5
Architecture of inception-V3 network.

convolutional layers, max and average pooling, concatenations, dropouts, and fully linked layers. Applications of batch normalization in activation layers are typical. The inception-V3 network is the inception block. The inception-V3 model separates layers, and rather than processing via a single layer, it utilizes the input from the preceding layer to execute four distinct processes concurrently, subsequently concatenating the outputs from all these various levels. The 5 × 5 convolution is replaced with two 3 × 3 convolutions in the Inception-V3 architecture, as shown in Figure 5. Since a 5 × 5 convolution requires 2.78 times more resources than a 3 × 3 convolution, this also improves computing performance by decreasing processing time. Utilizing two 3 × 3 layers instead of a single 5 × 5 layer enhances the architecture's performance.

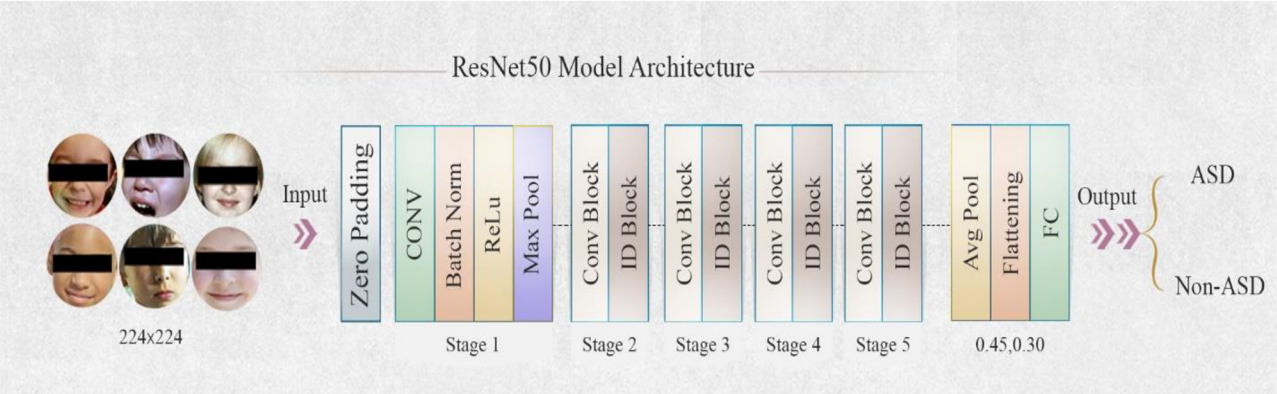


FIGURE 6
Architecture of ResNet50 model.

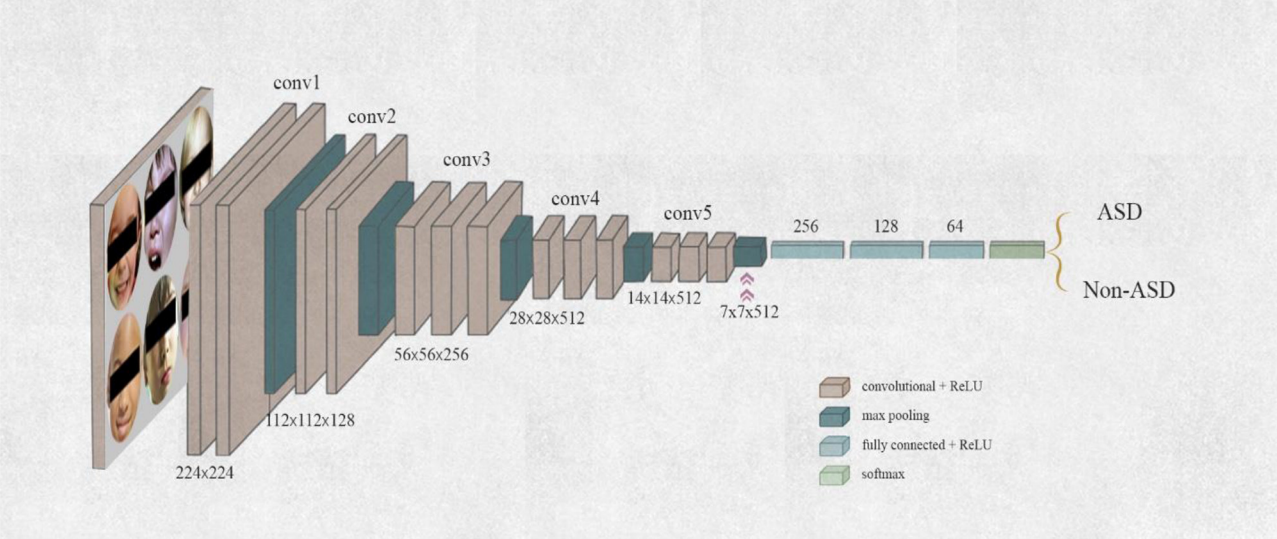


FIGURE 7
Architecture of VGG-19 model.

3.3.2 ResNet50 models

Introduced the residual neural network (ResNet) He et al. (33) in 2015. ResNet50 was introduced in 2015 by Microsoft Research for image identification tasks. ResNet indicates that the model has 50 layers. ResNet50 improved training performance by including residual connections between layers, which reduced loss, preserved acquired information, and kept it. An output with a residual link is a convolution of the input and the input itself, or the result of adding both together. Figure 6 illustrates a block diagram of the ResNet50 model's design. Utilized Residual blocks function as shortcuts or skip connections, enabling the model to bypass one or more levels. This mitigates the vanishing gradient issue during training and facilitates the seamless flow of information. ResNet50 key contribution is the invention of the residual block. These leftover blocks facilitate the connection of activations from preceding levels to subsequent layers.

3.3.3 VGG-19 models

The VGG-19 model was introduced by (34). The VGG-19 model for neural networks has 19 weight layers, 16 of which are convolutional layers and 3 of which are fully connected. Its filter size is 3 × 3, and it has a stride and padding of 1 pixel. The diminutive kernel size lowers the parameter count and allows for comprehensive coverage of the whole image. An operation called 2 × 2 max pooling with a stride of 2 is used by the VGG-19 model. With 138 million parameters, this model ranked second in classification and first in positioning in 2014. VGGNet reinforced the notion that CNNs should

TABLE 4 Enhanced parameters for setting the Dl models.

# No	Name	Values
1	Model Architecture	Inception-V3, VGG-19 and ResNet50
2	Image Size	224×244×3
3	Batch Size	16
4	Learning Rate	0.01
5	Epochs	25
6	Image Rescaling	1./255
7	Optimizer	SGD
8	Pool size	(3,3)
9	Strides	(2,2)
10	Padding	Valid
11	Dencer_layer	512
12	Dropout	0.50
13	Function	Sigmoid

TABLE 5 Validation results of ResNet50 model.

Model	Precision (%)	Recall (%)	F1 score (%)	Support
Non_Autistic	98	94	96	294
Autistic	94	98	96	294
Accuracy %	96			
Weighted Avg.	96	96	96	588

include a deep layered architecture to facilitate hierarchical interpretation of visual input. Figure 7 illustrates the block model of VGG-19.

3.4 Setting of proposed DL models

The DL model is started with pre-trained weights from the ImageNet function, with an input size of ASD image of 224 × 224 × 3, and omitting the top classification layers. The dense layer used sigmoid activation for binary classification objectives. The model used a Stochastic SGD optimizer with a standard learning rate of (0.01), leverages binary cross-entropy for finding performance and loss function, and evaluates performance based on accuracy as the measure. The Training model was used 25 epochs, using early stopping with 5 epochs. The completed model is assessed on the validation set using measures such as accuracy. Classification is performed using Softmax. Table 4 illustrates a schematic representation of the DL model.

3.5 Evaluation metrics

We used critical statistical metrics, including accuracy, precision, and recall, to illustrate our research results. The formulas that are used for the measurement of the DL models are as follows:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \times 100 \quad (1)$$

$$F1 - score = 2 * \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ positives} \times 100\% \quad (3)$$

$$Precision = \frac{True\ Negatives}{True\ Negatives + False\ Negatives} \times 100\% \quad (4)$$

4 Experiment

Training and evaluation of the proposed system were completed on the Kaggle environment platform, which consists of a robust TensorFlow library. We deliberately selected three distinguished pretrained CNNs: Inception-V3, ResNet50, and VGG 19 models, for diagnosis of the autism disorder in children. To use existing best practices and ensure consistency, we selected proven beneficial hyperparameters. Suitable for binary classification tasks, with a learning rate of 0.001, the SGD optimizer, the ReLU activation function, and a maximum of 25 epochs. The specified parameter values were accurately adjusted for all models according to the results of prior cutting-out research, with the objective of attaining optimum training performance for the chosen algorithms. The method was evaluated using a real-time dataset obtained from children with ASD and typically developing children.

4.1 Results of ResNet50 models

Table 5 presents the experimental results. The ResNet50 model exhibits significant efficacy in classifying Autistic and Non-Autistic individuals, attaining an overall accuracy of 96%. The ResNet50 model achieves a weighted average precision, recall, and F1-score of 96%, demonstrating consistent performance across both classes. In the Non-autistic class, precision is 98%, indicating that nearly all autistic predictions are accurate, whereas recall is 94%, indicating that some autistic cases are observed. The Autistic class demonstrates a precision of 94%, suggesting the presence of some false positives, while achieving a recall of 98%, indicating that nearly all Non-Autistic cases are identified. The F1-scores of 96% for Autistic individuals and 96% for Non-Autistic individuals indicate a strong balance in classification performance. The results indicate the model's effectiveness; however, lower enhancements in Non_Autistic precision may be realized through further data augmentation or fine-tuning. ResNet50 model demonstrates significant reliability for the classification of images related to autism, as proved by this evaluation.

Figure 8 presents the classification of the validation set of the ResNet50 model. The classification model's performance on the validation set was assessed through a confusion matrix. The model accurately identified 275 TN and 289 TP, exhibiting minimal FP. The model demonstrates high accuracy, minimal FP, and effective class differentiation, rendering it reliable for classification tasks.

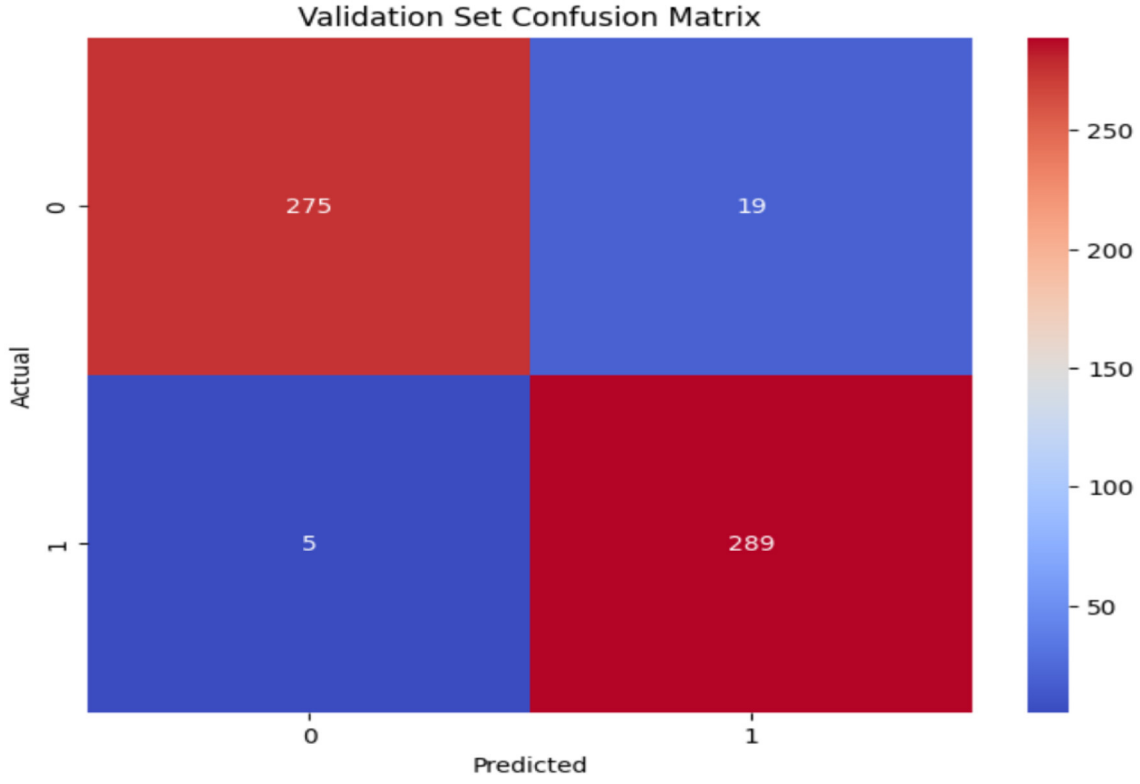


FIGURE 8
Confusion matrix of ResNet50 model.

TABLE 6 Validation results of Inception-V3 model.

Model	Precision (%)	Recall (%)	F1 score (%)	Support
Non_Autistic	98	97	98	294
Autistic	97	98	98	294
Accuracy %	98			
Weighted Avg.	98	98	98	588

4.2 Results of Inception-V3

The Inception-V3 model exhibits exceptional accuracy and robust classification capabilities in detecting ASD, as shown in Table 6. The Inception-V3 model demonstrates high accuracy and strong classification performance in the detection of ASD, achieving an overall accuracy of 98%. The system demonstrates a precision of 98% in identifying non-autistic cases, accompanied by an F1 score of 98%. The precision for Autistic cases is 97%, with a recall of 98% and an F1 score of 98%. This balanced performance minimizes misclassifications, rendering it appropriate for real-world applications in the identification of ASD with confidence and precision. The results demonstrate that the model effectively classifies target classes while maintaining a low misclassification rate, thereby rendering it suitable for real-world applications in the identification of ASD with high confidence and precision.

Figure 9 presents the confusion matrix for the Inception-V3 model during the validation stage. The model demonstrated enhanced classification performance. The Inception-V3 model exhibited robust classification performance, successfully predicting 286 non-autistic and 289 Autistic cases from a total of 588 samples. The model exhibited minimal misclassifications, recording 8 false positives (FP) and 5 false negatives (FN), which suggests strong recall and precision. The model demonstrated reliability and balanced performance, though there remains potential for improvement in minimizing misclassification rates.

4.3 Result of VGG-19

The VGG19 model demonstrates high precision, recall, and F1-score in the classification of ASD, exhibiting minimal FP and TN, as shown in Table 7. It demonstrates strong performance in the Autistic and Non-Autistic classes, as indicated by precision, recall, and F1-score metrics. The model reveals a 97% accuracy rate, suggesting its appropriateness for clinical ASD detection, with opportunities for enhancement via refined training strategies.

The confusion matrix of VGG19 is shown in Figure 10. The VGG19 model demonstrated robust performance on the validation dataset, with 285 TN accurately identifying the Non_Autistic class and 287 TP correctly identifying the Autistic class. There are just 9 FP as misclassifcation as Autistic when the true class is Non_Autistic, and 7 FN misclassifying as Non_Autistic when

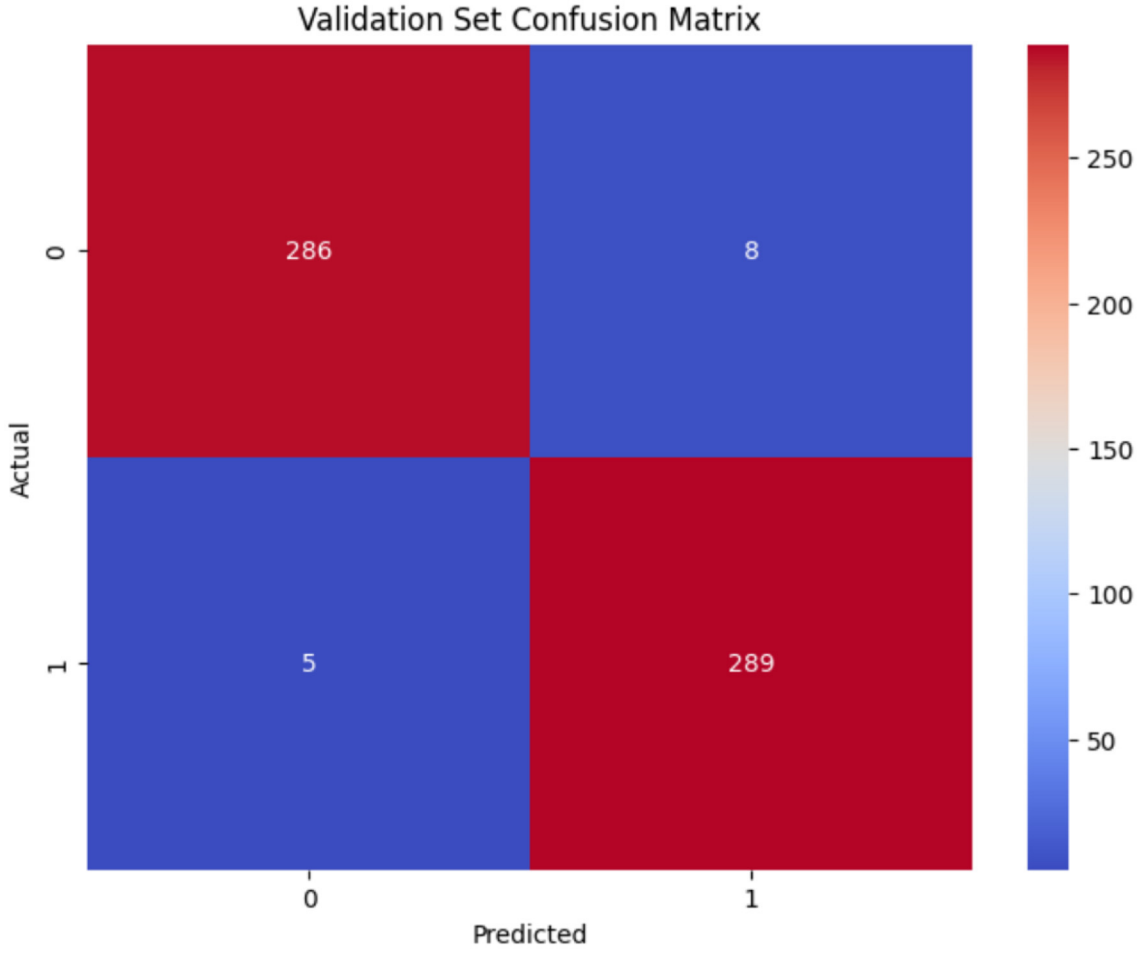


FIGURE 9
Confusion matrix of Inception-V3 model.

TABLE 7 Validation results of VGG-19 model.

Model	Precision (%)	Recall (%)	F1 score (%)	Support
Non_Autistic	98	97	97	294
Autistic	97	98	97	294
Accuracy %	97			
Weighted Avg.	97	97	97	588

the true class is Autistic, resulting in a minimal total count of misclassifications.

4.4 Performance of the ASD system based on DL models

The ASD detection system, using deep learning models, has impressive accuracy rates of 98% in training and validation, distinguishing between non-autistic and Autistic patients. The model's robust convergence and consistent validation outcomes demonstrate its proficiency in generalizing novel data, making it a valuable early identification tool.

Figure 11 shows the accuracy and loss of the ResNet50 system, with a *y*-axis representing data classification accuracy. The validation system improved accuracy from 0.5000 to 0.9592 during the validation phase, with an exceptional enhancement to 25 epochs. Training losses were quantified using a categorical cross-entropy function, with validation losses decreasing from 0.5 to 0.01 after 25 epochs.

The performance of the Inception-V3 model is seen in Figure 12 for both training and validation. We use categorical entropy loss and the SGD optimizer, executing for 25 epochs. During the training phase, the loss value diminishes from 0.7265 to 0.0076 until 25 epochs. The training accuracy is increasing gradually from 0.4844 to 0.9992 epoch 2 to 25. While validation accuracy improves from 0.8384 to 0.9779 throughout 25 epochs. This illustrates the model's capacity to learn and adjust according to input data. From epoch 3 to epoch 25, the model's performance improved progressively, exhibiting enhanced accuracy and less loss. Attaining a accuracy of 0.98 is a significant achievement.

Figure 13 illustrates the accuracy and loss performance of VGG19. During training epochs 2 to 23, the model's accuracy

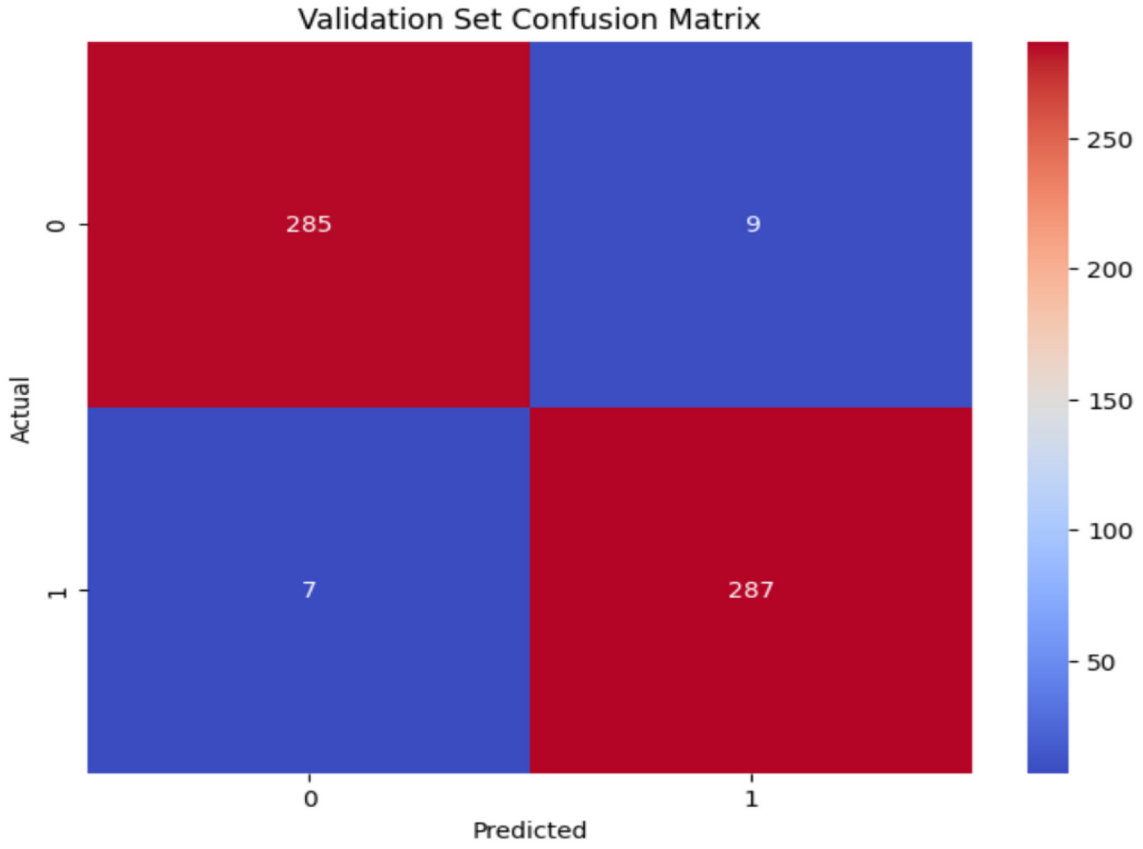


FIGURE 10
Confusion matrix of ResNet50 model.

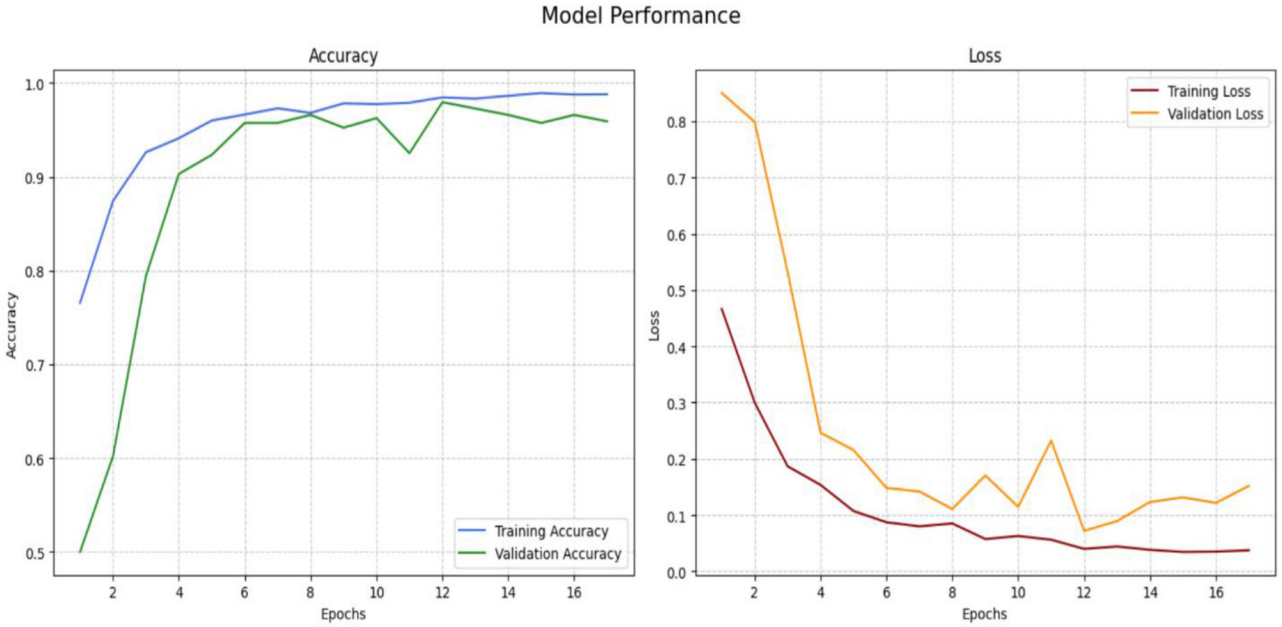


FIGURE 11
Performance of ResNet50 model.

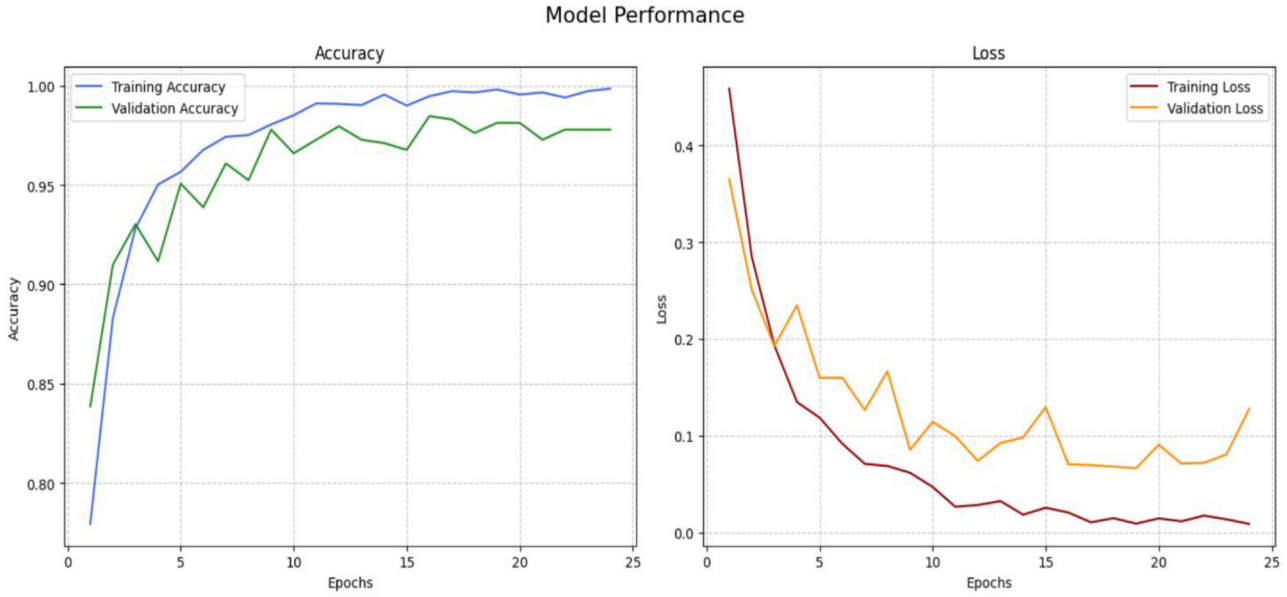


FIGURE 12
Performance of Inception-V3 model.

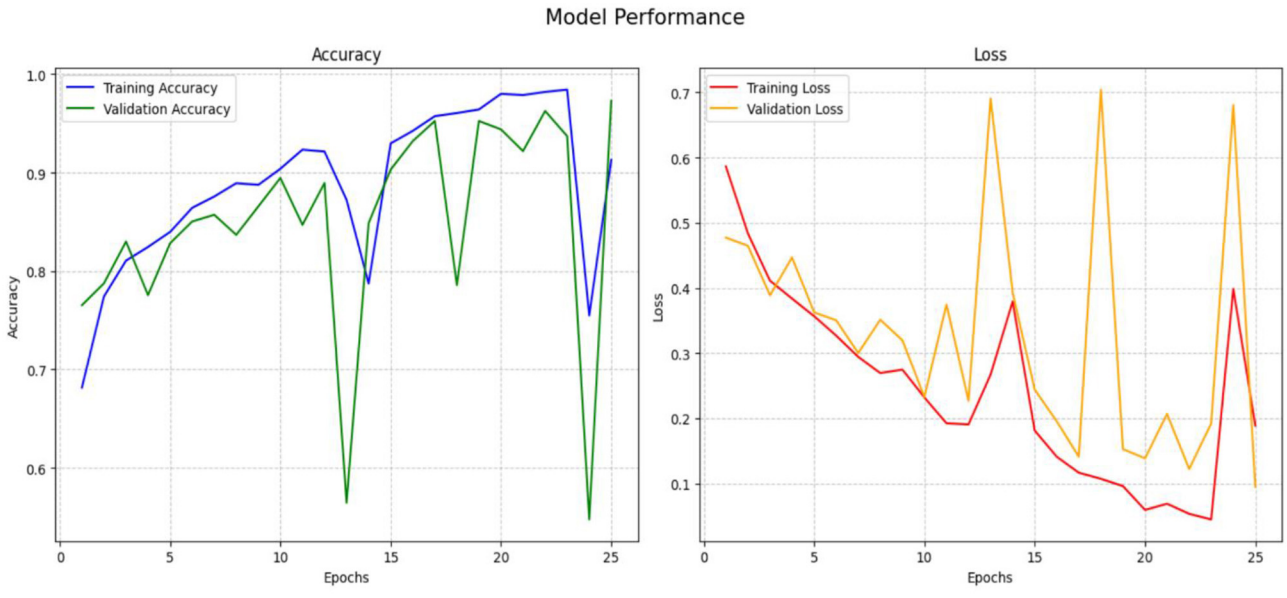


FIGURE 13
Performance of VGG-19 model.

increases to above 0.9854; however, there is a significant decline in accuracy from epochs 24–25. The validation accuracy reaches a maximum of 0.97 in the latter epochs, namely at epoch 25, demonstrating the model's effective recognition of the dataset's intrinsic patterns. The model's validation accuracy on unfamiliar data increases from 0.7653 in the opening epoch to an impressive 0.9728 at the conclusion of the 25th epoch. The validation loss consistently decreased throughout the preceding period, ultimately reaching a minimum of 0.0947.

5 Discussion

Individuals with ASD have difficulties in social interaction, communication, and conduct, as well as a variety of other neurological issues. Timely identification is crucial for mitigating the detrimental effects of this disease by implementing specialized instruction in schools and rehabilitation facilities. The research examined DL algorithms for the detection of autism spectrum disorder, emphasizing its efficacy in differentiating between persons with and without the condition. Current research primarily

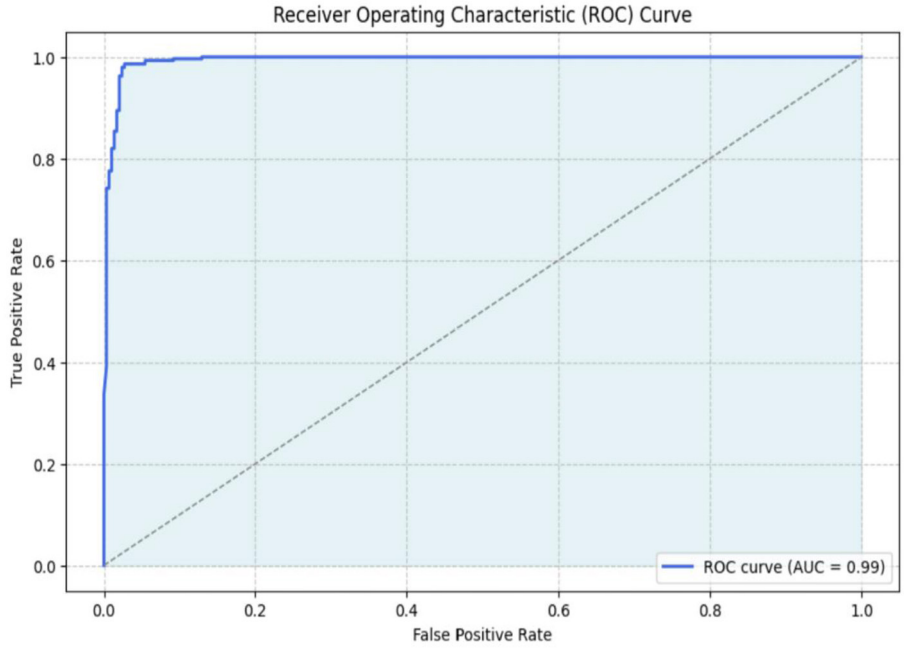


FIGURE 14
ROC of Inception-V3 model.

TABLE 8 Results of existing developing ASD systems with our results.

References	Approach	Used datasets	Accuracy (%)
Rashid and Shaker (25)	Xception	Same dataset	91
Alsaade and Alzahrani (26)	Xception		91
Sridurga et al. (27)	Xception		86
Rabbi et al. (28)	CNN		92
Alkahtani et al. (18)	MobileNetV2		92
Akter et al. (14)	MobileNet-V1		90
Gaddala et al. (29)	VGG16 & 19		84
Singh et al. (30)	MobileNet		88
Ghazal et al. (31)	AlexNet		87
Hosseini et al. (32)	MobileNet		94.64
Elshoky et al. (15)	ML		96
MobileNetV2	MobileNetV2		92
Proposed system			98

focuses on functional discoveries for categorization tasks, often leading to decreased accuracy. Our suggested methodology redirects attention to using structural information within facial expression data. Utilizing DL approaches, namely Inception-V3, and optimizing hyperparameters within this framework, we seek to address the shortcomings of existing procedures while augmenting generalization capacities and enhancing classification accuracy. This motivation stems from the recognition of the underutilized potential of facial expressions in children with ASD and typically developing children, along with the conviction that harnessing this information can lead to more effective classification models for diverse neurological conditions, thereby advancing the field and improving patient outcomes.

The potential threat we faced in this work is that data bias may undermine the model's generalizability, especially if the dataset lacks sufficient demographic diversity or exhibits class imbalance between autistic and non-autistic images. We have employed the augmentation method to address this issue, utilizing augmentation, early stopping, and transfer learning regularization techniques to mitigate overfitting. Including images from the same subject or session in several data splits might cause dataset leakage. This threat raises interpretability issues since it may be unclear which image features the models prioritize in their decision-making process. This pre-processing improved DL models, namely ResNet50, Inception-V3, and VGG-19, and removed the threat, achieving high accuracy. Finally, the DL models were examined by using accuracy and confusion matrices.

This approach used the augmentation technique to enhance the deep learning model for diagnosing ASD with outstanding performance. Employing ResNet50, Inception-V3, and VGG-19 models resulted in substantial improvements in diagnostic accuracy, with an exceptional 98% accuracy in differentiating between ASD and control subjects on the standard dataset. The results of ResNet50 scored 96% in terms of accuracy, and VGG-19 achieved an accuracy of 97%. The efficacy of this strategy is further substantiated by criteria such as accuracy, underscoring its potential to improve autism outcomes. The results have significant implications for ASD diagnosis in clinical settings, enabling more informed decisions, earlier identification and intervention, and

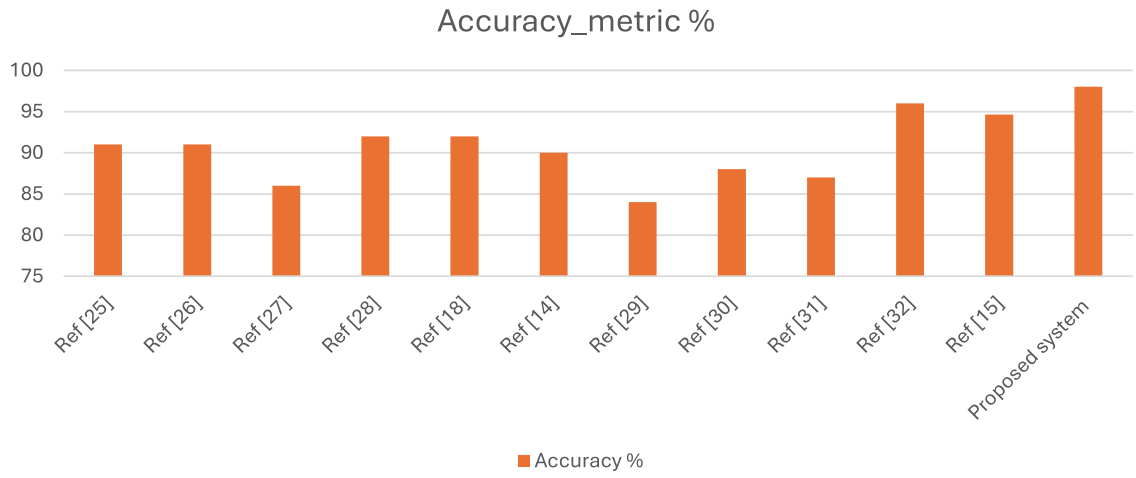


FIGURE 15
Performance of proposed system compared with different existing ASD systems.

improved outcomes for individuals and their families. Advanced algorithms may optimize the diagnosis process, thereby decreasing wait times and lowering the urgency on the healthcare system. Additional study and validation on more extensive datasets are required to comprehensively evaluate their therapeutic value and effect.

The AUC, or area under the curve, signifies that a higher AUC correlates with an increased probability of precise prediction. Figure 14 illustrates the ROC curve of the optimal methodology. The Inception-V3 model has superior accuracy and AUC of 99% across all three methodologies.

Numerous studies have been conducted specifically in diagnosing ASD based on the image expression of children. Most authors used the same standard dataset, available on Kaggle, which contains 2,940 images for applying different automatic classification approaches to diagnose ASD based on facial images, thereby enhancing accuracy. Prior studies indicate that suboptimal image quality in the training dataset significantly affects the accuracy of model results. One of the biggest challenges faced by the researchers is that images of children's faces frequently exhibit noise, low resolution, misalignment, and various other issues. Several researchers focus on optimizing models or hyperparameter sets, yet they often fail to achieve significant improvements in accuracy. Table 8 presents a comparison of the results from the latest studies in this field. In our research, we have improved the hyperparameters of the proposed DL model, and we have achieved 98% accuracy using the same dataset. Figure 15 compares our system's results with those of other approaches, highlighting the superior accuracy of our proposed strategy.

6 Conclusion

Diagnosing at an early stage is essential for administering successful treatment, particularly given the very low incidence of autism in children. The DL algorithms were used for ASD detection, often concentrating only on diagnosis. Moreover, current systems may have difficulties in scaling efficiently due to belief in manual and expertise-dependent procedures, impeding their capacity to satisfy the growing demand for autism evaluation and diagnosis. To tackle these issues, we have developed an efficient DL model„ namely ResNet50, Inception-V3, and VGG-19, implemented to predict and diagnose ASD. Pre-processing techniques, including resizing, rescaling, and augmentation, were used to enhance model performance, which may further elevate accuracy. Our classifiers achieved exceptional accuracies of 96%, 98%, and 97% for ASD, expression prediction, respectively. This illustrates their ability to precisely distinguish children's psychological states and facial expressions. We developed ASD system-based DL model to assess children's expressions and diagnose ASD. This study has significant effects for real-time ASD screening, potentially transforming the diagnosis process.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.kaggle.com/datasets/cihan063/autism-image-data.

Author contributions

AA-N: Conceptualization, Methodology, Project administration, Resources, Validation, Writing – original draft, Writing – review & editing. TA: Data curation, Formal analysis, Funding acquisition, Software, Visualization, Writing – original draft, Writing – review & editing. SA: Investigation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. EA: Formal analysis, Project administration, Resources, Software, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The authors extend their appreciation to the King Salman Center for Disability Research for funding this work through Research Group No. KSRG-2024-282.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Maenner MJ, Warren Z, Williams AR, Amoakohene E, Bakian AV, Bilder DA, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2020. *MMWR Surveill Summ.* (2023) 72:1–14. doi: 10.15585/mmwr.ss7202a1

2. Lord C, Brugha TS, Charman T, Cusack J, Dumas G, Frazier T, et al. Autism spectrum disorder. *Nat Rev Dis Primers.* (2020) 6:1–23. doi: 10.1038/s41572-019-0138-4

3. Daniels AM, Mandell DS. Explaining differences in age at autism spectrum disorder diagnosis: a critical review. *Autism.* (2014) 18:583–97. doi: 10.1177/1362361313480277

4. Thabtah F, Peebles D. A new machine learning model based on induction of rules for autism detection. *Health Informatics J.* (2020) 26:264–86. doi: 10.1177/1460458218824711

5. Uddin MZ, Shahriar MA, Mahamood MN, Alnajjar F, Pramanik MI, Ahad MAR. Deep learning with image-based autism spectrum disorder analysis: a systematic review. *Eng Appl Artif Intell.* (2024) 127:107185. doi: 10.1016/j.engappai.2023.107185

6. Aldridge K, George ID, Cole KK, Austin JR, Takahashi TN, Duan Y, et al. Facial phenotypes in subgroups of prepubertal boys with autism spectrum disorders are correlated with clinical phenotypes. *Mol Autism.* (2011) 2:15. doi: 10.1186/2040-2392-2-15

7. Zwaigenbaum L, Bauman ML, Stone WL, Yirmiya N, Estes A, Hansen RL, et al. Early identification of autism spectrum disorder: recommendations for practice and research. *Pediatrics.* (2015) 136:S10–40. doi: 10.1542/peds.2014-3667C

8. Lord C, Rutter M, DiLavore PC, Risi S, Gotham K, Bishop S. *Autism diagnostic observation schedule: ADOS-2 (2nd Edn.)*. Torrance: Western Psychological Services (2012).

9. Brambilla P, Hardan A, Di Nemi SU, Perez J, Soares JC, Barale F. Brain anatomy and development in autism: review of structural MRI studies. *Brain Res Bull.* (2003) 61:557–69. doi: 10.1016/j.brainresbull.2003.06.001

10. Vaiyapuri T, Mahalingam J, Ahmad S, Abdeljaber HA, Yang E, Jeong SY. Ensemble learning driven computer-aided diagnosis model for brain tumor classification on magnetic resonance imaging. *IEEE Access.* (2023) 11:91398–406. doi: 10.1109/ACCESS.2023.3306961

11. Ahmed ZA, Albalawi E, Aldhyani TH, Jadhav ME, Janrao P, Obeidat MRM. Applying eye tracking with deep learning techniques for early-stage detection of autism spectrum disorders. *Data.* (2023) 8:168. doi: 10.3390/data8110168

12. Pandimurugan V, Ahmad S, Prabu AV, Rahmani MK, Abdeljaber HA, Eswaran M, et al. CNN-based deep learning model for early identification and categorization of melanoma skin cancer using medical imaging. *SN Comput Sci.* (2024) 5:911. doi: 10.1007/s42979-024-03270-w

13. Bosl WJ, Tager-Flusberg H, Nelson CA. EEG analytics for early detection of autism spectrum disorder: a data-driven approach. *Sci Rep.* (2018) 8:1–20. doi: 10.1038/s41598-018-24318-x

14. Akter T, Ali MH, Khan MI, Satu MS, Uddin MJ, Alyami SA, et al. Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage. *Brain Sci.* (2021) 11:734. doi: 10.3390/brainsci11060734

15. Elshoky BRG, Younis EMG, Ali AA, Ibrahim OAS. Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images. *ETRI Journal.* (2022) 44:613–23. doi: 10.4218/etrij.2021-0097

16. Li Y, Huang W-C, Song P-H. A face image classification method of autistic children based on the two-phase transfer learning. *Front Psychol.* (2023) 14:1226470. doi: 10.3389/fpsyg.2023.1226470

17. Siagian Y, Muhathir, Maqhfirah DR. Classification of autism using feature extraction speed up robust feature (SURF) with boosting algorithm. In: *2023 International Conference on Information Technology Research and Innovation (ICITRI)*. Piscataway: IEEE (2023). p. 60–4

18. Alkahtani H, Aldhyani THH, Alzahrani MY. Deep learning algorithms to identify autism spectrum disorder in children-based facial landmarks. *Appl Sci.* (2023) 13:4855. doi: 10.3390/app13084855

19. Sai Koppula K, Agrawal A. Autism spectrum disorder detection through facial analysis and deep learning: leveraging domain-specific variations. In: *International Conference on Frontiers in Computing and Systems*. Singapore: Springer Nature (2023). p. 619–34.

20. Abdullah AS, Geetha S, Govindarajan Y, Vinod AA, Pranav AGV. Prediction and evaluation of autism spectrum disorder using ai-enabled convolutional neural network and transfer learning: an ensemble approach. In: *2024 2nd World Conference on Communication & Computing (WCONF)*. Piscataway: IEEE (2024). p. 1–10.

21. Karthik MD, Priya SJ, Mathu T. Autism detection for toddlers using facial features with deep learning. In: *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. Piscataway: IEEE (2024). p. 726–31.

22. Pan Y, Foroughi A. Evaluation of AI tools for healthcare networks at the cloud-edge interaction to diagnose autism in educational environments. *J Cloud Comput.* (2024) 13:39. doi: 10.1186/s13677-023-00558-9

23. Shahzad I, Khan SUR, Waseem A, Abideen ZUI, Liu J. Enhancing ASD classification through hybrid attention-based learning of facial features. *Signal Image Video Process.* (2024) 18, S475–88. doi: 10.1007/s11760-024-03167-4

24. Reddy P, Andrew J. Diagnosis of autism in children using deep learning techniques by analyzing facial features. *Eng Proceed.* (2024) 59:198. doi: 10.3390/engproc2023059198

25. Rashid A, Shaker S. Autism spectrum disorder detection using face features based on deep neural network. *Wasit J Comput Math Sci.* (2023) 2:74–83. doi: 10.31185/wjcm.100

26. Alsaade FW, Alzahrani MS. Classification and detection of autism spectrum disorder based on deep learning algorithms. *Comput Intell Neurosci.* (2022) 2022:1–10. doi: 10.1155/2022/8709145

27. Sridurga PD, Yugandhar B, Haritha P, Narayana K. Detecting autism spectrum syndrome using VGG19 and Xception networks. *Int J Res Eng Sci Manage.* (2022) 5:12.

28. Rabbi Md F, Hasan SMM, Champa AI, Zaman Md A. A convolutional neural network model for early-stage detection of autism spectrum disorder. In: *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. (2021). p. 110–114

29. Gaddala LK, Kodepogu KR, Surekha Y, Tejaswi M, Ameesha K, Kollapalli LS, et al. Autism spectrum disorder detection using facial images and deep convolutional neural networks. *Revue d'Intelligence Artif.* (2023) 37:801–6. doi: 10.18280/ria.370329

30. Singh A, Laroia M, Rawat A, Seeja KR. Facial feature analysis for autism detection using deep learning. In: AE Hassanien, O Castillo, S Anand and A Jaiswal (Eds.) *International Conference on Innovative Computing and Communications (Vol. 703)*. Singapore: Springer Nature (2023). p. 539–551.

31. Ghazal TM, Munir S, Abbas S, Athar A, Alrababah H, Khan MA. Early detection of autism in children using transfer learning. *Intell Autom Soft Comput.* (2023) 36:11–22. doi: 10.32604/iasc.2023.030125

32. Hosseini MP, Beary M, Hadsell A, Messersmith R, Soltanian-Zadeh H. Deep learning for autism diagnosis and facial analysis in children. *Front. Comput. Neurosci.* (2022) 15:789998. doi: 10.3389/fncom.2021.789998

33. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27-30 June* (2016). pp. 770–778.

34. Liu S, Deng W. Very deep convolutional neural network based image classification using small training sample size. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia* (2015). pp. 730–734. doi: 10.1109/ACPR.2015.7486599

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 17 June 2025
DOI 10.3389/fmed.2025.1587026

OPEN ACCESS

EDITED BY
Habib Hamam,
Université de Moncton, Canada

REVIEWED BY
Shadab Alam,
Jazan University, Saudi Arabia
Velliangiri Sarveshwaran,
National Chung Cheng University, Taiwan
Md Assaduzzaman,
Daffodil International University, Bangladesh

*CORRESPONDENCE
Eid Albalawi
✉ ealbalawi@kfu.edu.sa
Surbhi Bhatia Khan
✉ s.khan138@salford.ac.uk;
✉ s.khan138@ieee.org

RECEIVED 03 March 2025
ACCEPTED 02 May 2025
PUBLISHED 17 June 2025

CITATION
Pandey P, Bhatia Khan S, Pruthi J, Albalawi E, Algarni A and Almusharraf A (2025) GAN-enhanced deep learning for improved Alzheimer's disease classification and longitudinal brain change analysis.
Front. Med. 12:1587026.
doi: 10.3389/fmed.2025.1587026

GAN-enhanced deep learning for improved Alzheimer's disease classification and longitudinal brain change analysis

Purushottam Pandey[1], Surbhi Bhatia Khan[2,3,4]*, Jyoti Pruthi[1], Eid Albalawi[5]*, Ali Algarni[6] and Ahlam Almusharraf[7]

[1]Manav Rachna University (MRU), Faridabad, Haryana, India, [2]School of Science, Engineering and Environment, University of Salford, Salford, United Kingdom, [3]Centre for Research Impact & Outcome, Chitkara University, Institute of Engineering and Technology, Rajpura, Punjab, India, [4]Research and Innovation Cell, Rayat Bahra University, Mohali, Punjab, India, [5]Department of Computer Science, College of Computer Sciences and Information Technology, King Faisal University, Al Ahsa, Saudi Arabia, [6]Department of Informatics and Computer Systems, College of Computer Science, King Khalid University, Abha, Saudi Arabia, [7]Department of Management, College of Business Administration, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia

Alzheimer's disease (AD) is commonly defined by a progressive decline in cognitive functions and memory. Early detection is crucial to mitigate the devastating impacts of AD, which can significantly impair a person's quality of life. Traditional methods for diagnosing AD, while still in use, often involve time-consuming processes that are prone to errors and inefficiencies. These manual techniques are limited in their ability to handle the vast amount of data associated with the disease, leading to slower diagnosis and potential misclassification. Advancements in artificial intelligence (AI), specifically machine learning (ML) and deep learning (DL), offer promising solutions to these challenges. AI techniques can process large datasets with high accuracy, significantly improving the speed and precision of AD detection. However, despite these advancements, issues such as limited accuracy, computational complexity, and the risk of overfitting still pose challenges in the field of AD classification. To address these challenges, the proposed study integrates deep learning architectures, particularly ResNet101 and long short-term memory (LSTM) networks, to enhance both feature extraction and classification of AD. The ResNet101 model is augmented with innovative layers such as the pattern descriptor parsing operation (PDPO) and the detection convolutional kernel layer (DCK), which are designed to extract the most relevant features from datasets such as ADNI and OASIS. These features are then processed through the LSTM model, which classifies individuals into categories such as cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). Another key aspect of the research is the use of generative adversarial networks (GANs) to identify the progressive or non-progressive nature of AD. By employing both a generator and a discriminator, the GAN model detects whether the AD state is advancing. If the original and predicted classes align, AD is deemed non-progressive; if they differ, the disease is progressing. This innovative approach provides a nuanced view of AD, which could lead to more precise and personalized treatment plans. The numerical outcome obtained by the proposed model for ADNI dataset is 0.9931, and for OASIS dataset, the accuracy gained by the model is 0.9985. Ultimately, this research aims to offer significant contributions to the medical field, helping healthcare professionals diagnose AD more accurately and efficiently, thus improving patient outcomes. Furthermore,

brain simulation models are integrated into this framework to provide deeper insights into the underlying neural mechanisms of AD. These brain simulation models help visualize and predict how AD may evolve in different regions of the brain, enhancing both diagnosis and treatment planning.

KEYWORDS

Alzheimer's disease, ResNet101, long short term memory, generative adversarial network, ADNI, OASIS dataset

1 Introduction

AD (Alzheimer's disease) is one of the leading causes of dementia universally (1, 2) and considered as one of the most deadly diseases which needs to be taken under consideration with utmost care. AD is characterized by a recurrent deterioration of cognitive abilities in older people. Besides, AD is stated as an irreversible neurological disorder which progressively impairs the cognitive capability therefore, it is important to provide effective treatments as early as possible with the aim to avoid life threatening consequences. It was reported that, AD is expected to rise from 27 to 106 million cases (3) in the upcoming four decades, impacting one in every 85 people on the planet. Another report suggested that ∼70% people are account of AD (4). As there is an evident rise of AD in recent times, effective methods need to be implemented for detection of AD in people; however, to treat AD, it is important to identify the symptoms in the patients suffering with AD; thus, some of the common symptoms of people suffering with AD are memory loss, difficulty in speaking, loss of spontaneity, and many more (5–7). Usually, people with AD can endure symptoms for years; however, the severity of AD symptoms tends to worsen progressively, gradually impairing an individual's ability to perform everyday activities independently. Since there is currently no known cure for AD, nevertheless existing treatments aim to slow down the disease's advancement and delay the onset of its most severe stage.

Typically, AD is classified into three stages such as mild, moderate, and severe (8). Early stages of AD can perform daily tasks independently, although they may struggle with specific tasks (9) such as driving, individuals in early stage can communicate socially and remember significant details. However, as the disease progresses to the middle stage, symptoms become more pronounced and the person may require greater care, frustration, and difficulty with routine tasks (10, 11). In the last stage, AD becomes the most challenging for managing as individuals lose their ability to respond and communicate leading to a significant decline in memory and cognitive skills (12, 13). Therefore, it is extremely important to detect the symptoms as early as possible with the aim to avert any impemending consequences faced by the individuals. Hence, different manual techniques are primarily used by the medical professionals for AD detection which includes cognitive assessments and neurological examinations where healthcare providers assess the functions and activities of brain to detect any abnormalities which may be indicative of AD. Furthermore, brain imaging techniques such as PET scans and MRI are used for providing detailed images of the brain to medical experts. Although these techniques offer various advantages, there are certain drawbacks of employing manual techniques (13, 14) such as time-consuming, subjectivity, and prone to error, which require highly skilled medical professionals. Hence, to overcome these drawbacks faced by manual approaches, AI-based techniques are incorporated as AI-based models are fast and accurate and can handle huge amount of complex data easily. Moreover, AI models can detect any subtler changes in functioning of brain which may not be easily detectable by human observers. Hence, various existing research study focuses on employing AI-based ML and DL models for the detection and classification of AI.

Dense neural network is used for binary classification of Alzheimer's disease by alleviating the problem of multiple modalities and processes. A fully connected dense neural network (FCNN) with two hidden layers (15) was used for performing binary classification of AD. By applying FCNN model, the accuracy gained by the model is 87.50%. Similarly, CNN-based DL model (16) has used for AD classification using ADNI dataset. In CNN model, different layers such as three convolutional layer, max pooling layer, and fully connected layer are used for classification. Existing study has considered classifying three different classification of AD, which includes AD vs. NC, AD vs. MCI, and MCI vs. NC. Approximately 450 MRI images were used. Process carried out includes pre-processing the images and classifying the obtained pre-processed images. Skull striping, segmentation, registration, and outlining the ROI were some of the pre-processing techniques used for pre-processing the images. The accuracy obtained for three binary classification task with spike pre-training technique was 90.15%, 87.30%, and 83.90%. However, the accuracy obtained by three binary classification without spike was 86%, 83%, and 76%. Therefore, the incorporation of ANN for extracting the relevant features of AD helped in satisfactory classification of AD (17).

Although the existing models deliver better performance in terms of classification of AD, there are certain pitfalls which need to be addressed. Thus, some of the drawbacks are low accuracies projected by the model, overfitting of the model, empathizing only on binary classification, computational complexity, and inability to work with huge datasets. Thus, to overcome these drawbacks, the proposed model utilizes ResNet101 with LSTM for feature extraction and classification using ADNI and OASIS datasets. The proposed ResNet101 model uses DKCL and PDPO layers to extract relevant features needed for the proposed model. PDPO is employed for assigning binary codes to pixels depending on the comparison with neighboring pixels, by efficiently capturing the local texture information and the DCK layer captures the discriminative effectively by sliding a tiny filter over the input image and computing element-wise multiplication between the

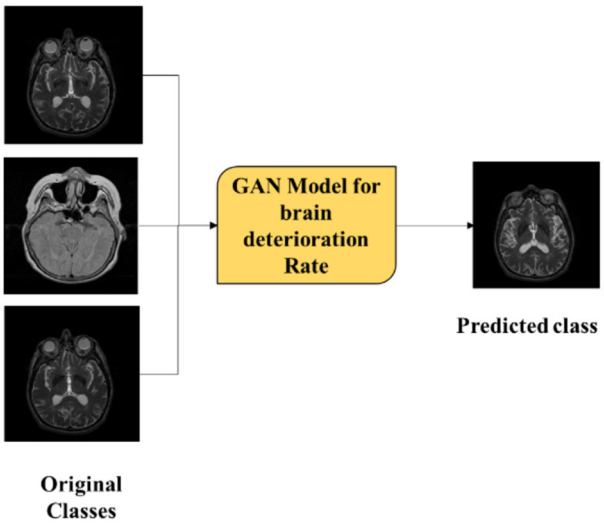


FIGURE 1
Brain deterioration rate using GAN.

filter and overlapping regions of the input data. Implementation of these proposed functions in the proposed ResNet101 model aids in extracting relevant features needed for the model. Eventually, the extracted features are passed to the LSTM model for classification of Alzheimer's disease as AD, CN, and MCI. In addition, the proposed research focuses on employing the GAN model to find whether Alzheimer's disease is progressive or non-progressive in nature by distinguishing the original class from the predicted class. By doing so, the brain deterioration rate can be determined, and this can assist the medical experts to offer a suitable diagnosis to the patients. Thus, Figure 1 depicts the original and predicted class gained using the GAN model, where if the original class and predicted class are the same, it is denoted as non-progressive and if it is different, then it is represented as progressive.

The major contributions of the proposed research study are as follows:

- To extract relevant features for feature extraction using the proposed ResNet101 using pattern descriptor parsing operation layer and detection convolutional kernel layer and to perform multiclass classification using the LSTM model for classifying Alzheimer's disease as CN, MCI, and AD.
- To determine the deterioration rate of the brain as progressive and non-progressive using the proposed model.
- To assess the performance of the proposed model using standard metrics such as accuracy, precision, recall, and F1 score as well as brain deterioration of patients.

The study is organized in the following way. Section 2 deals with existing studies done by research authors. Section 3 discusses proposed algorithms implemented for the classification of AD, Section 4 reflects on the outcome obtained using proposed methodology, and Section 5 summarizes the research study, including future recommendations.

2 Literature review

The existing section reviews various existing studies on the detection and classification of Alzheimer's disease using AI-based techniques.

Hybrid DL approach (18) has been used in the study for early Alzheimer's disease detection. Thus, multi-modal imaging and CNN with LSTM algorithm have combined together for identifying early MCI diseases, which remain challenging due to the difficulty in discriminating patients with cognitive normality. Better accuracy was obtained by the model for AD classification. Despite the remarkable performance of the model, the limitation of the model includes overfitting of data. Similarly, two different NNs such as ResNet50 and AlexNet (19) were used for AD detection and classification. The MRI images were collected from Kaggle website, and CNN algorithm was employed using AlexNet and ResNet50 TL models. Accuracy of the model obtained using AlexNet was 94.53%, showcasing that DL model is better suited for medical investigation such as AD detection and classification. Similarly, 12-layer CNN model (20) has been used for AD detection based on brain MRI images. 12-layer CNN model was used on OASIS dataset in which sufficient accuracy has gained by the model for AD classification. Furthermore, the model was compared with other models such as InceptionV3, Xception, MobileNetV2, and VGG19. Although the model has delivered better accuracy for AD classification, the drawback of the model is that it only focused on binary classification of Alzheimer's disease on OASIS dataset.

LSTM (21) has been used for precise diagnostic approach for binary classification of AD. LSTM model was utilized for classifying the MRI data and making accurate predictions for the early detection of AD. Although the model has delivered better performance for binary classification of AD, there are certain drawbacks of the study which needs to be overcome such as inability of the model to fully capture the complexity and variety of the target population. This pitfalls ultimately impact the generalizability and robustness of the model for AD classification. Similarly, LSTM (22)-based RNN model has been used for predicting the progression of the ADF patients from MCI to AD. The objective of the study was to anticipate the development of the illness. LSTM-based model has implemented for predicting the biomarker values using ADNI dataset. The ADNI dataset incorporated the positive biomarker of parents after every 6, 12, 18, 24, and 36 months from the standard. Eventually, the state of progression was identified by using MLP model, where accuracy of 88.24% is accomplished. This findings helped in improving the early findings of AD. Similarly, 3D convolutional and LSTM (ConvLSTM) (23) model has adopted for early diagnosis of AD from full-resolution sMRI scans. Complete resolution of brain images belonging to ADNI and OASIS dataset has been used, in which the accuracy gained by the model is 86%, and F1 score and sensitivity obtained by the model are 88% and 96%. Regardless of the extensive performance of the model, accuracy attained by the model is considerably low.

OASIS dataset (24) has incorporated for identification of AD using DL and image processing approaches. CNN-based DL model has implemented for AD classification, and the accuracy obtained by the model is 93%. Despite its performance, limitation

of the model showcases the usage of additional dataset such as ADNI dataset for more comparative validation analysis and tests the generalization of the study. As the model lacks in terms of working with multiple dataset, the future work of the study focuses on creating a bigger dataset combined from different sources for increasing the variability of the input samples of various target class for accomplishing better model in terms of generalization and reliability of the model to new and unseen data. Similarly, CNN-based MobileNet (25) model has been used for multiclass classification of AD, where MobileNet architecture used depthwise separable convolutions that reduced the number of parameters when compared to conventional convolutions and resulted in lightweight neural network. Although the model has delivered better accuracy, different techniques such as augmentation approaches are focused in the future for further enhancing the accuracy of the model.

Two-stage DL model (26) has employed for integrating the process of classification and regression to determine whether a patient is suffering with MCI and then determining the probable progression time. The first stage focused on detecting the patient class using LSTM classification, and the second stage focused on prediction using LSTM regression model. Furthermore, the model was compared with existing ML models such as SVM, RF, LR, KNN, DT Lasso, and Ridge, from which it was identified that suggested LSTM model has delivered better outcome than existing models. In spite of its result, the model lacks in interpreting the decision in an effective way. Thus, the shortcoming of the model includes explainability, accountability, and fairness of the model. CNN-based DL approach (27) has implemented in the study for AD classification, in which the process was carried out by loading OASIS and MIRIAD dataset. Then, CNN has employed for classifying the presence of AD. From the analytical outcome, it was identified that accuracy obtained by CNN model was 82%. Furthermore, sensitivity and specificity gained by the model were 93 and 81%. An 8-layer CNN model called CNN-BN-DO-DA has employed for (28) AD classification in which batch normalization and dropout functions BN was used for normalizing the inputs of the layer into mini-groups in order to solve concerns related to incessant training change and dropout function was utilized for lessening the problems associated with overfitting an computational consumption. OASIS dataset was used. The result of the study has indicated that better techniques will be used in the future for speeding up convergence rate and will be aided in improving the efficacy of the model.

Like DL models, ML models are also used for detecting AD; thus, methods such as DT, SVM, RF, voting classifiers, and gradient boosting (29) were incorporated in the study for identifying the best parameters for AD detection using OASIS dataset. It was detected that accuracy gained by DT, RF, SVM, XGBoost, and voting classifiers was 80.46%, 86.92%, 81.67%, 85.92%, and 85.12%. Although these ML techniques were focused on reducing risks by detecting the disease in early stages, identifying relevant attributes (feature extraction) for the model for AD detection is still challenging task. Bias in ML is an issues which needs to be resolved as quickly as possible; this study (30) has employed Adaptive Synthetic Sampling (ADASYN) technique for improving the accuracy and issues associated with bias. Therefore, feature extraction battery (FEB) and SVM model were employed for feature extraction and classification of AD. It was identified that SVM model has aided in improving the accuracy by 6%. Although the model has obtained better accuracy for AD prediction, ML models along with meta-heuristic approaches were considered in future for further enhancements in terms of improving the prediction accuracy.

Conversely, as stated (31), has aimed to enhance AD classification using MRI data by integrating advanced DL models for early diagnosis and personalized treatment. The method has combined an ensemble DL model with Soft-NMS-enhanced Faster R-CNN for candidate merging, improved ResNet50 for feature extraction, and Bi-GRU for processing sequence data. Using MRI datasets, the model has achieved better classification accuracy for AD vs. CN tasks, demonstrating its potential for precise early diagnosis and intervention. Another study (32) has employed CycleGAN for synthetic image generation and Google Inceptionv3-based CNN for classification. It has utilized CNNs trained on augmented datasets, achieving an F-1 score of 89% with standard data and 95% with CycleGAN-enhanced data augmentation. This approach has shown the effectiveness of DL models and generative adversarial networks in improving diagnostic accuracy for Alzheimer's disease. The author in Zhang et al. (33) has developed ADNet, based on the VGG16 model. It has utilized 2D MRI slices, incorporating depthwise separable convolution, ELU activation, and SE modules for efficient feature extraction while simultaneously training on auxiliary tasks such as clinical dementia and mental state score regression. The findings have shown ADNet achieved a 4.18% accuracy improvement for AD vs. CN classification and a 6% improvement for MCI vs. CN classification compared to the baseline VGG16 model, demonstrating its potential for early diagnosis. Another study (34) has leveraged the ResNet50V2 DL model for AD classification using 6,400 MRI images sourced from Kaggle, achieving a high accuracy of 96.18%. By employing transfer learning, fine-tuning, and dynamic learning rate adjustments, the model effectively discriminated AD stages, which showcased its potential for real-world medical applications. As illustrated (35), has introduced AlzhiNet, a hybrid DL framework that combined 2D-CNN and 3D-CNN models with custom loss functions and volumetric data augmentation for AD diagnosis. It has been validated on MRI datasets, and it has achieved remarkable accuracy and demonstrated robustness against perturbations, outperforming standalone models and ResNet-18 in real-world applications.

2.1 Gaps identified

A significant research gap exists in Alzheimer's disease classification using binary models, particularly when addressing challenges associated with small datasets, time consumption, scalability, and overfitting. Current approaches often rely on large datasets to prevent overfitting and ensure robust feature extraction, but neuroimaging studies typically involve limited sample sizes, such as datasets with fewer than a thousand participants or even fewer in some cases. This scarcity leads to difficulties in training deep learning models

effectively and exacerbates overfitting risks, especially when high-dimensional data such as MRI scans are involved. In addition, binary classification tasks require discriminative feature selection from complex neuroimaging data, which is computationally demanding and time-consuming. Scalability remains a pressing issue as models optimized for small datasets may not generalize well to larger or diverse populations. Thus, overcoming these limitations requires innovative methodologies that balance computational efficiency with the ability to extract meaningful features from small datasets while mitigating overfitting through advanced regularization techniques or ensemble methods.

3 Proposed methodology

AD is considered one of the most deadly diseases in the world. Hence, it is important to detect it as quickly as possible. Various approaches are carried out by the research workers. However, there are certain pitfalls of employing existing studies, such as overfitting of the model, low accuracy, computational complexity, and ineffective multiclass classification of AD. Hence, the proposed model is used to overcome these limitations by using efficient algorithms. Thus, the flow of the proposed research is depicted in Figure 2.

Figure 2 depicts the process involved in the proposed research study for multiclass classification of AD. The process is initiated by loading the ADNI and OASIS datasets. Then, the images are pre-processed using image resizing and image data normalization. Image resizing refers to the process of varying the dimensions and resolutions. Thus, resizing the images can aid in standardizing the input data for further processing and analysis. In proposed study, the image is resized in terms of 64 × 64. Image data normalization involves scaling the pixel values to a common range to improve the performance of the model within the range of 0–1. Normalizing image data assists in reducing the variations in pixel intensity and enhances the ability of the model to learn relevant features from the images. Owing to these factors, image resizing and image data normalization are opted for pre-processing. After pre-processing, pre-processed data are split as a train-test split, where the ratio involved in the proposed research for the train-test split is 80:20. After data split, the data are passed onto the proposed ResNet101 and LSTM for feature extraction and classification.

After pre-processing, the proposed ResNet101 is used for feature extraction by employing DKCL and PDPO functions for extracting relevant features. Then, LSTM is employed to classify the images as CN, MCI, and AD accordingly. Eventually, the present research study focuses on determining the deterioration rate of the brain by using the GAN model. This GAN model shows if the disease is in a progressive or non-progressive state by comparing the original class and predicted class. If the original class and predicted class are the same, then the CN is in a non-progressive state. If the original class and predicted class are different, then CN is progressing. Finally, the performance of the model is detected by using evaluation metrics. Figure 2 showcases the architecture of the proposed study.

3.1 Proposed ResNet101 and LSTM for feature extraction and classification

After pre-processing, the pre-processed images are passed for feature extraction. Feature extraction is used for extracting the relevant features needed for the model. Thus, feature extraction is considered to be one of the important steps for classifying the images. By using a feature extraction mechanism, features with noise and irrelevant details are removed and aid in focusing on important aspects of the data. Furthermore, extracting and selecting aids in enhancing the interpretability of the model. Although there are various models for feature extraction, the proposed model focuses on employing the ResNet101 model for an effective feature extraction process, as the ResNet101 model is a CNN technique which is 101 layers deep, allowing it to learn rich and complex feature representation from images. This enables the ability to capture the intricate patterns and features within the images, making it suitable for extracting detailed and relevant features. Similarly, ResNet101 uses skip connections, which helps to mitigate the vanishing gradient issue and enables a swift feature extraction process. Similarly, the ResNet101 model possesses the potential to extract high-level features due to its depth and training on a diverse dataset. Owing to these factors, ResNet101 is used.

Although conventional ResNet101 offers various advantages for feature extraction, certain pitfalls need to be overcome, which include the complexity of the model making it more computational and resource-intensive. This aspect of the model can lead to longer training times. Similarly, conventional ResNet101 is also susceptible to overfitting the model and interpretability of the model, making it a challenging factor for feature extraction. Thus, to overcome these drawbacks, the proposed model emphasizes using an enhanced ResNet101 model which utilizes pattern descriptor parsing operation layer function and detection convolutional kernel layer function. Hence, the proposed ResNet101 model is depicted in Figure 3.

Figure 3 showcases the process involved in the proposed ResNet101 for feature extraction. This process is carried out by sending the pre-processed features to the input layer. From the input layer, the data are forwarded to the convolutional layer. CL is considered the building block utilized for the FE process. CL encompasses a series of convolutional filters that scan input images to extract the edges, textures, and shapes. However, to enhance the ability of the feature extraction function, the PDPO layer and DCK layer are used. PDPO is employed for assigning binary codes to pixels depending on the comparison with neighboring pixels, by efficiently capturing the local texture information. Furthermore, the PDPO layer enhances the ability by considering the relationships of pixels at varying distances from the center pixels, enabling the capture of texture variations at different scales.

Therefore, PDPO layer is designed to enhance feature extraction by capturing local texture information through a binary coding mechanism. In this layer, each pixel in the input image matrix is compared with its neighboring pixels within a defined neighborhood, such as a 3 × 33 × 3 grid. For each central pixel, the layer assigns a binary code based on whether it is greater than or less than its surrounding neighbors. This process emphasizes local texture variations, allowing the model to capture subtle details that

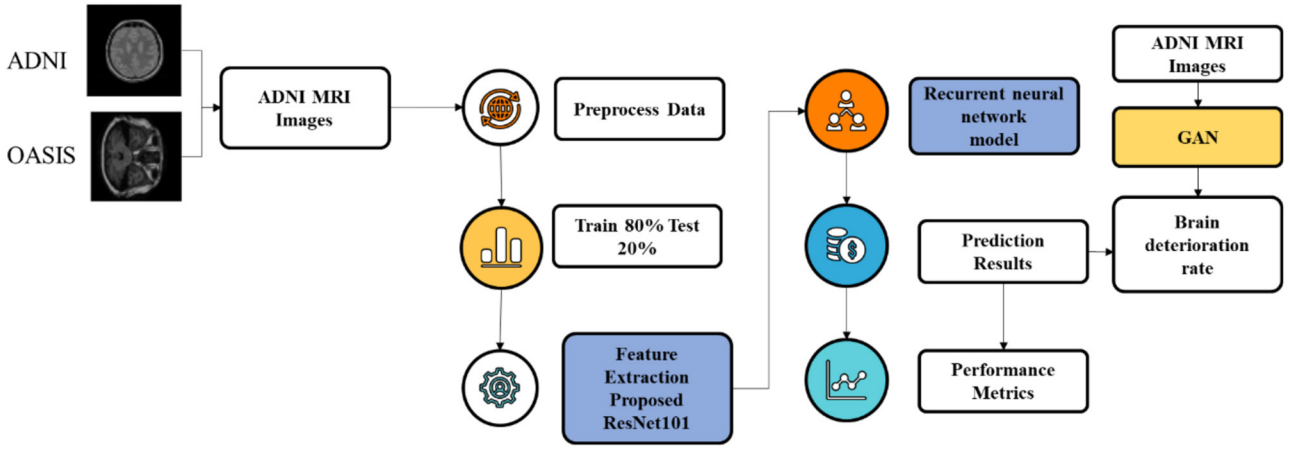


FIGURE 2
Overall flow of the model.

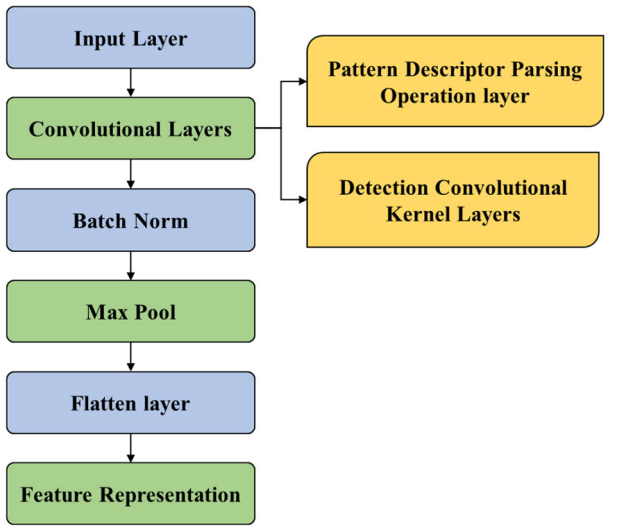


FIGURE 3
Proposed ResNet101.

are critical for tasks such as Alzheimer's disease classification. The size of the neighborhood can be adjusted to enhance sensitivity to local features, and an optional threshold can be applied to refine the binary coding. The resulting binary feature map retains the spatial structure of the input image while reducing dimensionality, making subsequent processing more efficient. Unlike traditional convolutional layers that aggregate features over larger areas, the PDPO layer focuses on local pixel relationships, thereby improving the model's sensitivity to texture variations that might be overlooked by standard methods. Here, Equation (1) shows the process involved in PDPO.

$$C(a,b) = (U * V)(a,b) = \sum_r \sum_s U(a + r.b + s)V(r,s) \quad (1)$$

where U is represented as input matrix image, C is denoted as an output feature map, and V is represented as the size of the filter. a, b denotes the enhanced input image, r is represented as neighboring pixel, and s is represented as center pixel. This input U is convolved with filter V and generates feature map C. Thus, the convolutional operation is denoted by U^*V. Therefore, the convolution operation U^*V involves multiplying corresponding elements of the filter V with overlapping regions of the input matrix U, followed by summing these products to produce a single value for each position in the output feature map C. This operation enables the PDPO layer to extract binary-coded features that highlight subtle texture variations critical for tasks such as Alzheimer's disease classification.

Like PDPO, DCK layer is implemented at CL for extracting the hierarchical features from the input images. The proposed DCK layer captures the discriminative effectively by sliding a tiny filer over the input image and compute element-wise multiplication between the filter and overlapping regions of the input data. This operation results in a single scalar value, which represents a feature of the input data. Therefore, DCK function predominantly aids the proposed ResNet101 model to extract hierarchal features and prevents the model from getting overfitting. Equation (2) depicts the same.

$$PDPOL_{P,R}(a_s) = \sum_{r=0}^{r-1} \mu (a_r - a_s) 2^p \quad (2)$$

where R is defined as the radius and distance of neighboring pixels from the center pixel. This defines the spatial extent of the neighborhood used for comparison, influencing sensitivity to local features; similarly, P is denoted as the number of neighboring pixels. Then, the hierarchal features are passed to batch normalization. Batch normalization is typically utilized after CL to improve training and generalization of the model by solving the internal covariance shift problem. The output from the batch normalization process is passed to the max pooling layer, which reduces the spatial dimensions of feature maps without distressing

depth by introducing the translation invariance and reducing the number of learnable parameters in the succeeding layers. Eventually, the flattening layer transforms the output feature maps of the pooling layer into the 1D vector. By doing so, it helps in improving computational efficiency. Finally, the extracted features obtained are passed to the LSTM model for the classification of images as CN, MCI, and AD.

The classification is proceeded by using LSTM approach, as the LSTM model can handle complex and non-linear relationships in data, making it suitable for Alzheimer's disease classification, where the relationships between the pixels are highly intricate. Furthermore, the LSTM model is highly flexible and can be easily adapted to different input images, making the model effective for Alzheimer's disease classification.

Here, the integration of the LSTM model with ResNet101 is designed to harness the strengths of both architectures for improved feature extraction and temporal processing. In this approach, features are first extracted from the input images using the proposed ResNet101 model, with proposed PDPO and DCKL layer that capture spatial hierarchies and complex patterns. After passing through the ResNet101 architecture, the output feature maps are typically flattened to reduce their dimensionality, resulting in a fixed-length feature vector for each image. This feature vector is then prepared for input into the LSTM classification model.

To facilitate this integration, the feature vectors are organized sequentially, reflecting the temporal order of the input data, such as a series of images in a video or a sequence of frames. The LSTM is configured with specific parameters, including the number of hidden layers, the number of units in each layer, and dropout rates to prevent overfitting. Typically, the LSTM may have one or more layers with a varying number of units, depending on the complexity of the task. The output from the LSTM can be further processed to produce predictions or classifications based on the learned temporal dependencies in the data. This integration allows the model to capture both spatial features from the ResNet101 and temporal relationships through the LSTM, enhancing the overall performance in tasks such as Alzheimer's disease classification. Therefore, Figure 4 illustrates the architecture of LSTM model for classification.

In Figure 4, f_t is denoted as a forget gate, σ is signified as sigmoid function, i_t and o_t are denoted as the input gate and output gate, C_t is denoted as candidate gate, and $t-1$ is represented as cell state. Employment of LSTM model aids effectively for the classification of Alzheimer's disease. The forget gate implemented in the model is depicted in Equation (3).

$$fg_t = sig(wght[in_t, act_{t-1}, C_{t-1}]) + bi_f \tag{3}$$

In the above equation, act_{t-1} is denoted as output of the preceding block, bias vector is characterized as b_f, input sequence is denoted by using in, C_{t-1} is represented as the previous memory block of the LSTM, *sig* is denoted as the sigmoid function, and separate weight vectors for each input are represented using *wght*. Input gate is a section, where a new memory is generated by using a trivial neural network with tanh activation function and this is depicted Equations (4) and (5).

$$i_t = sig(wght[in_t, act_{t-1}, C_{t-1}]) + bi_i \tag{4}$$

$$C_t = f_t.C_{t-1} + i_t\, tanh[in_t, act_{t-1}, C_{t-1}]) + bi_c \tag{5}$$

Output gate is the section, where output generated by the current LSTM block is generated by using output gate and these outputs are estimated using Equations (6) and (7).

$$sig_t = sig(W[in_t, act_{t-1}, C_t]) + b_o \tag{6}$$

$$act_t = o_t.\tanh(C_t) \tag{7}$$

Thus, the connection between the units of LSTM permits the information to cycle between adjacent time steps.

3.2 Determination of brain deterioration rate using GAN model

The GAN model plays a critical role in determining disease progression by generating synthetic images that simulate various stages of the disease. Once the GAN is trained, it generates images that represent both progressive and non-progressive cases. The training process involves two components: the generator, which creates synthetic images, and the discriminator, which evaluates the authenticity of these images by comparing them to real images from the dataset. In this workflow, the GAN is trained using a specific loss function, a combination of adversarial loss and additional metrics that quantify the differences between the generated and real images. The adversarial loss encourages the generator to produce images that are indistinguishable from real images, while the discriminator's loss focuses on correctly classifying real vs. generated images. A common choice for the loss function in GANs is the binary cross-entropy loss, which measures the performance of the discriminator in distinguishing real images from fake ones.

After training, the model is tested using images generated by the GAN. By analyzing the characteristics of these synthetic images in comparison with the original images, the model can identify patterns indicative of disease progression. This approach allows for a nuanced understanding of the disease's trajectory as the GAN-generated images can reflect subtle changes that may not be easily observable in the original dataset. The ability to compare these generated images with actual clinical cases enhances the model's capacity to distinguish between progressive and non-progressive cases, ultimately contributing to more accurate predictions regarding disease progression.

Thus, GAN model is used in the proposed model for analyzing and predicting the progression of Alzheimer's disease based on

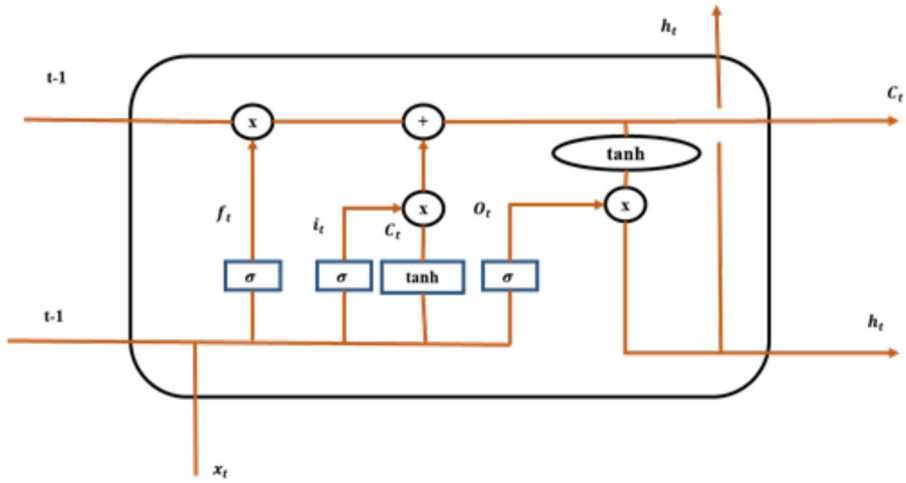


FIGURE 4
LSTM model.

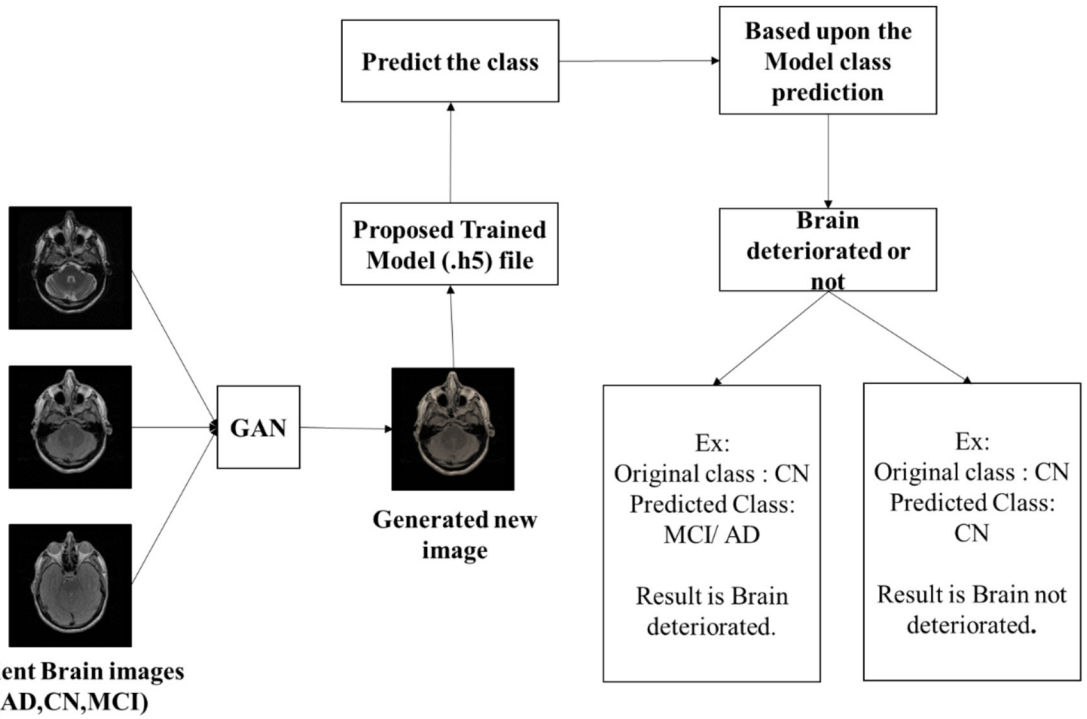


FIGURE 5
GAN model.

original class and predicted class by generating new images based on the patient's image data. Once trained, the GAN model can generate synthetic data which represent different stages of Alzheimer's disease progression. Therefore, by comparing the original class with the predicted class, the progression of AD can be identified. Here, both the generator and discriminator were optimized using the Adam optimizer with a learning rate of 0.0002 and a beta1 value of 0.5. This choice of optimizer helps in achieving faster convergence and stability during training. The training process involved alternating updates between the generator and discriminator, ensuring that each model learns effectively from the other's performance.

Thus, the process carried out by the GAN model for determining brain deterioration rate is depicted in Figure 5.

Initially, the model was trained with accuracy of 99%. In general, a GAN model comprises a generator and a discriminator where the generator network in the GAN model generates the synthetic data samples and the discriminator network evaluates

TABLE 1 Brain deterioration rate.

Original image class	Predicted class	Brain deterioration rate
CN	CN	Brain not deteriorated
CN	MCI	Brain deteriorated
CN	AD	Brain deteriorated
MCI	AD	Brain deteriorated

the generated samples of data and the original data samples to distinguish between progressive and non-progressive classes. In the GAN model, the generator generates an image from the original class, whereas the discriminator generates other images from the dataset. If the image generated by the discriminator is progressive, such as CN, MCI, and AD, then the disease is identified to be progressive. If not the disease, it is identified to be non-progressive.

If the original class is CN and the predicted class obtained by the proposed model is AD, it is noted that the brain deterioration rate is in the progressive state as the original class is CN, whereas the predicted class appears to be in progressive nature by predicting AD. Conversely, if the original class is MCI and the predicted class is MCI as well, then there is no progression in terms of brain deterioration rate. Eventually, the GAN model utilizes a loss function for measuring the difference between ground truth labels and predicted classes. This loss function guides the training process for minimizing the errors in classifying the progressive and non-progressive nature precisely. The advantages of employing the GAN model in the proposed framework include the following:

- Detection of progressive and non-progressive Alzheimer's disease.
- Identification of brain deterioration rate can help in preventing adverse consequences.

Therefore, Table 1 shows the outcome obtained by GAN model for brain deterioration rate.

Table 1 shows the brain deterioration rate. Here, it was projected that there is brain deterioration when the original image class and predicted class are the same. That is, when the original class is CN and the predicted class is CN, it means that the brain is not deteriorated. However, if the original class and reduced class are different, it is depicted that the brain is deteriorated. Therefore, the GAN model is used for detecting brain deterioration rates.

The subsequent section deals with the results obtained using the proposed model by assessing the efficacy of the proposed framework using metrics such as accuracy, recall, F1 score, and precision value.

4 Result and discussion

Result and discussion section primarily involves depicting the outcome of the proposed model post-deployment for the classification of Alzheimer's disease as CN, MCI, and AD. Hence, subsequent section discusses about metrics involved, EDA, and performance analysis of the model.

TABLE 2 Sample patient ID for ADNI.

Sample patient ID	MRI count
132_S_0339	3
035_S_6947	3
130_S_6319	3
023_S_0331	3
132_S_0339	3
035_S_6947	3
130_S_6319	3
023_S_0331	3
023_S_0030	3
128_S_0216	5
116_S_6624	5
127_S_0260	5
041_S_0282	5
127_S_0397	6

4.1 Dataset description

The proposed study utilizes two different datasets for AD multiclass classification such as ADNI (Alzheimer's disease neuroimaging initiative dataset) and OASIS dataset.

4.1.1 Creation and collection of data

The dataset is created by gathering subject information and image information, in which the subject information consists of subject ID, research group, age, research group, weight (in Kg), and other aspects. Similarly, in image information, parameters such as modality (DTI, MRI, PET, Path, and fMRI), image description, image ID, weighting, slice thickness, and acquisition plane are considered.

4.1.1.1 ADNI dataset

The clinical dataset comprises of detailed clinical information from each subject which includes extensive patient measurements such as MRI data. It encompasses data from North America male and female individuals, with a total of 502 attributes collected from 1737 participants. Specifically, the dataset includes data from 1453 male patients and 1074 female patients. Table 2 shows the sample patient ID with MRI counts.

Table 2 depicts MRI count taken by different patients along with patient ID. Patient ID with 132_S_0339 has taken MRI count of 3, ID with 130_S_6319 has taken MRI count of 3, and 5 numbers of MRI have been taken by patient ID with 116_S_6624, 128_S_0216, 127_S_0260, and 041_S_0282. Similarly, 6 MRI has been taken by patient with ID 127_S_0397.

TABLE 3 MRI count and patient count for ADNI.

ADNI MRI	MRI count	Patient count
	4	12
	3	92
	5	4
	6	1
	14	1
	12	2

Similarly, Table 3 depicts samples of patients who has taken MRI. Here, 92 patients have taken 3 MRIs, 12 patients have taken 4 MRI, 4 patients have taken 5 MRI, and so.

4.1.1.2 OASIS dataset

The dataset provides neuroimaging and related clinical data, encompassing neuroimaging data across the genetic spectrum, and cognitive and demographic factors for researchers studying Alzheimer's disease. Specifically, data from 1,317 male patients and 1,911 female patients have been collected for research purpose.

4.2 Performance metrics

4.2.1 Accuracy

The accuracy is claimed as the calculation of total accurate classification. The accuracy range is premeditated by using Equation (8),

$$Acc = \frac{TN + TP}{TN + FN + TP + FP} \tag{8}$$

where TN is represented as true negative, and FN is represented as false negative; similarly, true positive and false positive are denoted by using TP and FP.

4.2.2 Precision

The precision is considered by determining the accurate classification count. It is calculated through indecorous classification. The precision is estimated by using Equation (9),

$$precision = \frac{TP}{FP + TP} \tag{9}$$

4.2.3 F-measure

The F1 score is represented as the weighted harmonic-mean value of precision and value of recall, and Equation (10) is defined as the formula employed for determining F1-Score,

$$F1 - score = 2 \times \frac{R \times P}{R + P} \tag{10}$$

where P is denoted as precision, and R is denoted as recall.

TABLE 4 System configuration.

Techniques	Tools and requirements
MRI image data	MRI datasets(brain)
Hardware requirements	•Adequate computational properties: CPU and GPU.
Software requirements	•Image processing libraries and frameworks such as OpenCV (Open Source Computer Vision Library) and Tensor Flow. •Python or other programming languages
Visualization tools	Matplotlib visualization of volumetric medical images.
Data pre-processing tools	•DICOM (Digital Imaging and Communications in Medicine) format conversion python code. •Image registration and normalization software

4.2.4 Recall

The recall is indicated as the reclusive of the production metric that assesses the total of correct positive categories made out of all the optimistic classes. Equation (11) shows the mathematical model for recall,

$$Recall = \frac{TP}{FN + TP} \tag{11}$$

4.3 System configuration

Experimental setup including hardware and software requirements for implementing proposed methodology is depicted in Table 4.

4.4 EDA

EDA plays a crucial role in comprehending the insights, characteristics, and patterns of the data in the dataset. Therefore, EDA for Alzheimer's disease uncovers significant relationships and trends in terms of biomarkers, risk factors, and patterns which may contribute toward the progression, diagnosis, and treatment of the disease. Moreover, EDA also aids in detecting data quality issues, missing values, and outliers to ensure the reliability and accuracy of the model. Thus, Figures 6, 7 show the MRI scans of ADNI dataset and OASIS dataset.

Thus, MRI scans of ADNI dataset and OASIS dataset are illustrated in Figures 6, 7 from different angles. Similarly, heatmap for ADNI and OASIS dataset is depicted in Figures 8, 9.

Heatmap is used for exploring the datasets and aids in detecting the patterns and trends with varying colors. This heatmap assists in highlighting the area of outliers or concentration. Each ROI can be denoted by a heatmap, showing the variations in intensity which corresponds to different measurement. Thus, Figure 8 showcases the heatmap of ADNI dataset, and Figure 9 demonstrates heatmap of OASIS dataset.

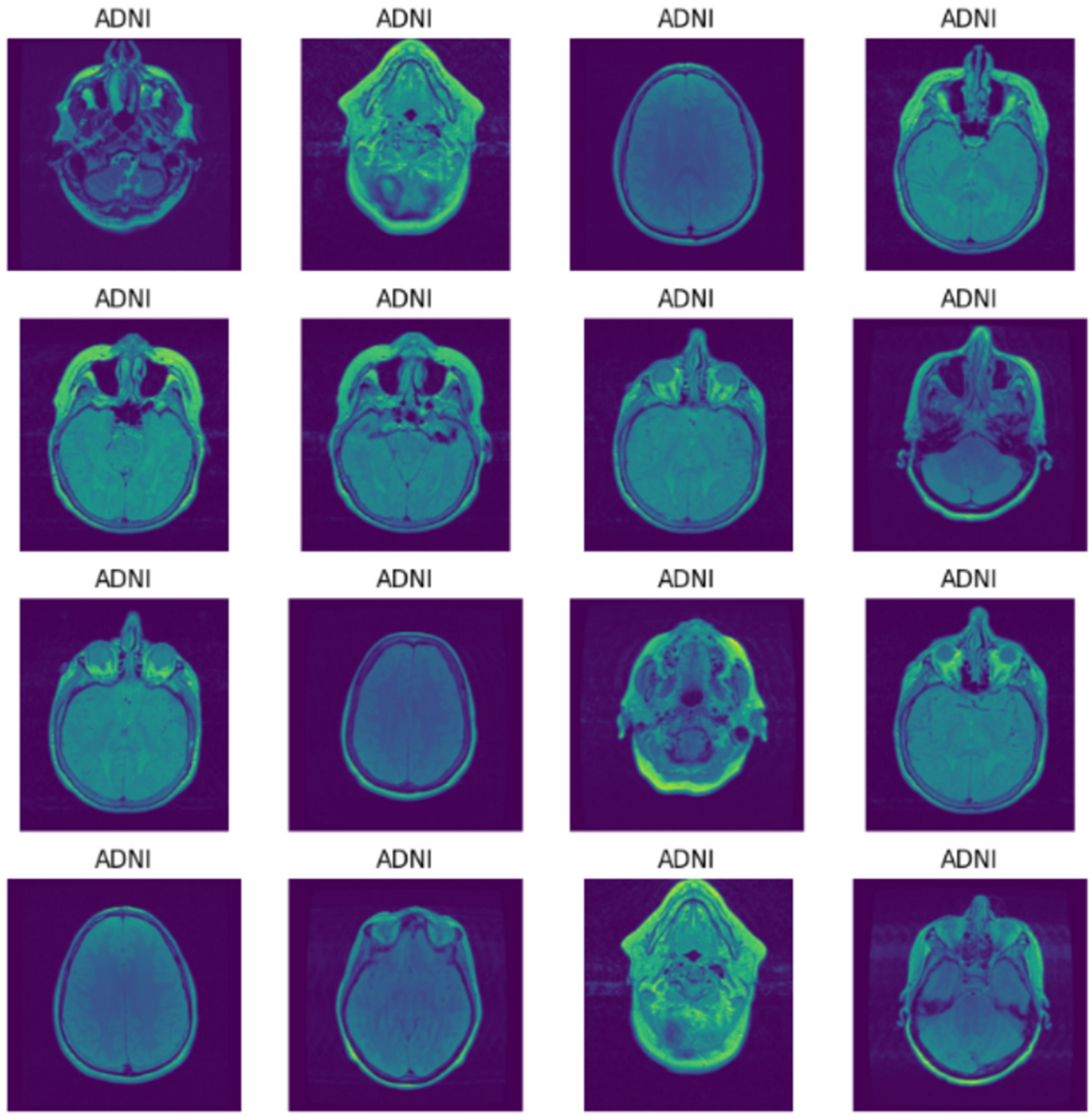


FIGURE 6
MRI of ADNI.

4.5 Performance analysis

The performance of the proposed model is depicted in the subsequent section, where the performance of the model is analyzed using different metrics such as model accuracy, model loss, and confusion matrix for both the ADNI and Oasis datasets.

Model accuracy for ADNI and OASIS datasets using the proposed model is portrayed in Figures 10, 11. Model accuracy graph is defined as the visual representation of how the accuracy of the model changes over time or epochs during the training process. X-axis denotes the number of epochs, and Y-axis denotes the accuracy of the model. Thus, Figures 10, 11 show the model accuracy graph for the ADNI and OASIS datasets.

The model accuracy for the ADNI dataset is depicted in Figure 10, in which the blue line represents the training accuracy and the orange line represents the validation accuracy. Training accuracy refers to the accuracy of the model on the training dataset during the training process. This indicates the ability of the model to predict the correct output for the data it was trained on. Validation accuracy denotes the accuracy of the model on a separate validation dataset for evaluating the model, on how well the model generalizes to unseen and new data. In figures, training accuracy is more than validation accuracy. This showcases that the model is learning patterns present in the training data effectively. Similarly, model loss for ADNI and OASIS is portrayed in Figures 10, 11.

Model loss using ADNI dataset and OASIS dataset is demonstrated in Figures 12, 13. Model loss refers how well the proposed model performs during training. In model loss, training and validation losses are examined, where training loss is denoted as the error between the actual or predicted output on the training

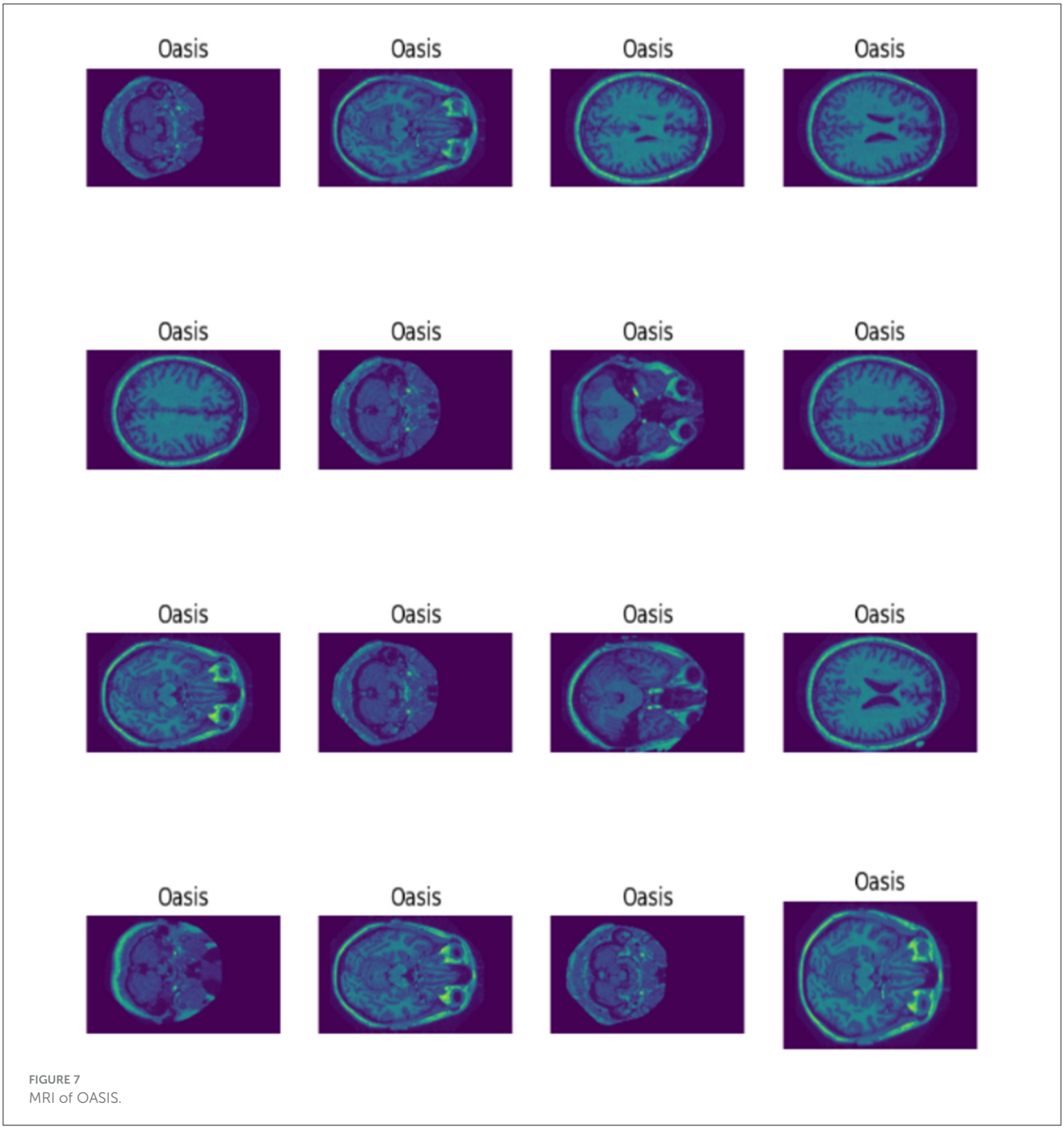


FIGURE 7
MRI of OASIS.

dataset. The primary objective is to minimize the training loss by optimizing the parameter of the model. Similarly, validation loss is the differences between actual or predicted output on a separate validation dataset which was utilized during training process. From figures, it can be clearly observed that both validation and training losses decrease when model goes through multiple epochs of training. This showcases that proposed model is learning to make better predictions.

Like model accuracy and model loss, confusion matrix is an important assessing the performance of the proposed framework for multiclass classification of Alzheimer's disease. Confusion matrix displays number of correct classifications and misclassifications by the model compared to the actual outcomes in the dataset. In addition, row in the matrix denotes the actual class labels and column in the matrix denotes the predicted class labels. Hence, confusion matrix for ADNI dataset is denoted in Figure 14.

In Figure 14, confusion matrix for ADNI is depicted. Here, the correct classifications and misclassifications are represented in which misclassification is denoted in black color and correct classification for AD, CN, and MCI is depicted, where AD form comprises of higher correct predictions, in which the correct

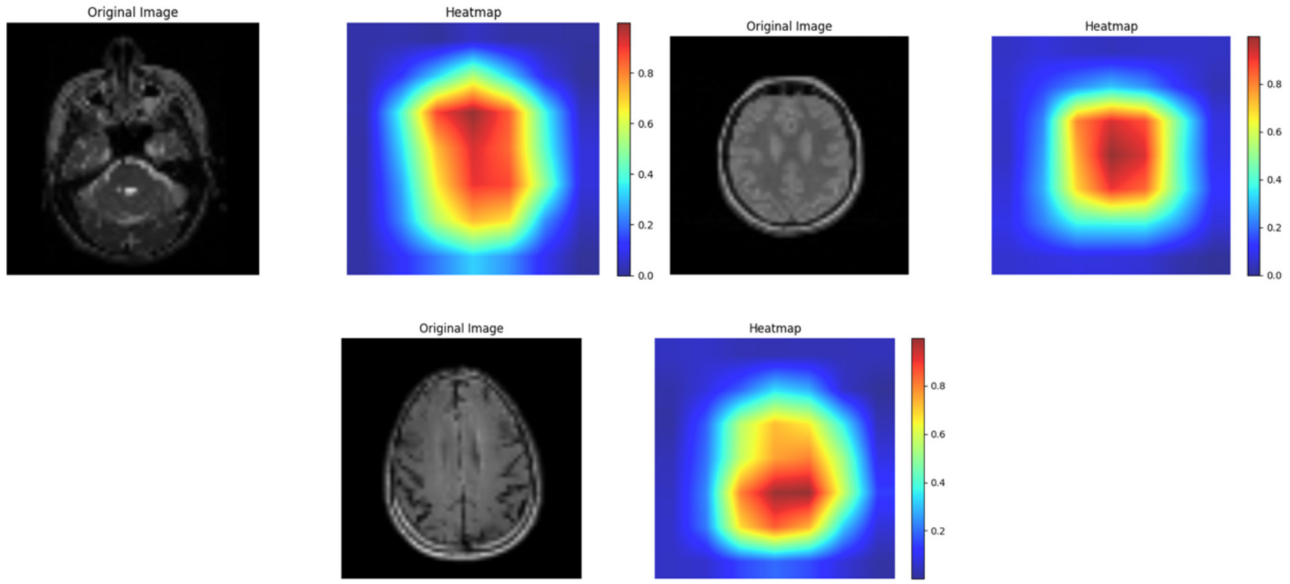


FIGURE 8
Heatmap for ADNI dataset.

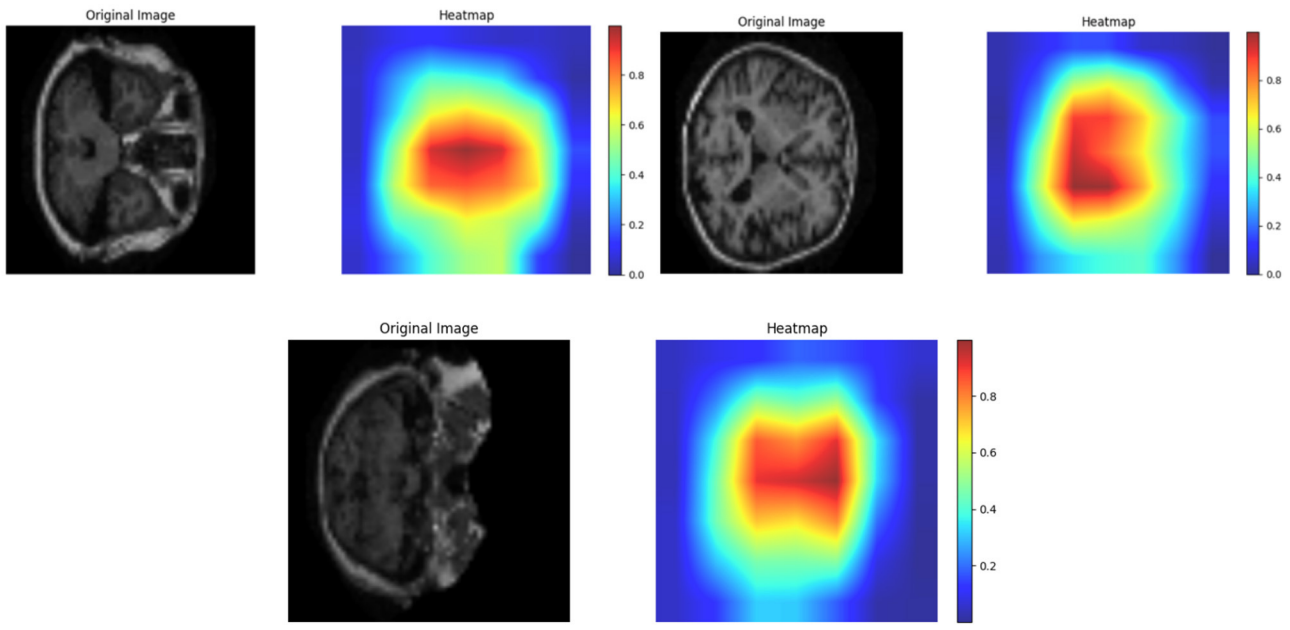


FIGURE 9
Heatmap for OASIS dataset.

predictions for AD is 132, CN is 96, and MCI is 151. Similarly, confusion matrix for OASIS is illustrated in Figure 15.

Here, the confusion matrix for proposed model using OASIS is depicted in Figure 15, where correct classifications and misclassifications are represented. The correct classification for AD is 274, CN is 206, and MCI is 151. Therefore, from the experimental results, it can be observed that proposed model is capable of producing effective outcome which is essential for classification of Alzheimer's disease. Like confusion matrix, other metrics are also used for gauging the efficacy of the proposed study, which includes accuracy of the model, precision, F1 score, and recall rate. Therefore, Table 5 showcases the metric value obtained by the proposed study.

Table 5 depicts the metrics obtained by the proposed study for both ADNI and OASIS datasets. Here, the proposed model using ADNI dataset obtains accuracy of 0.9932%, precision of 0.99%,

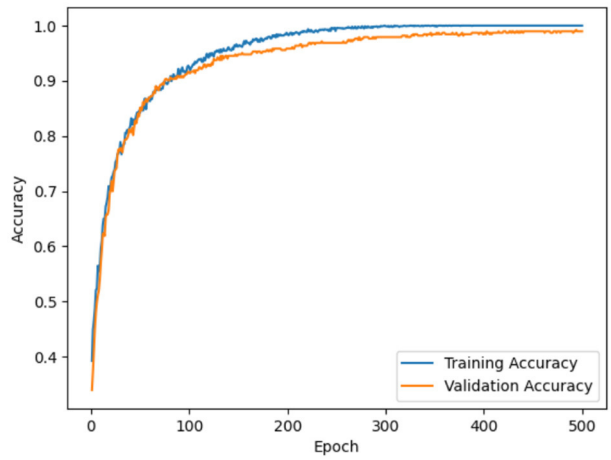


FIGURE 10
Model accuracy for ADNI.

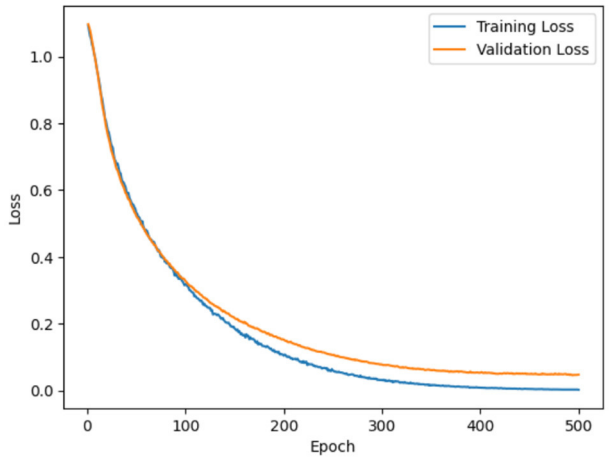


FIGURE 12
Model loss for ADNI.

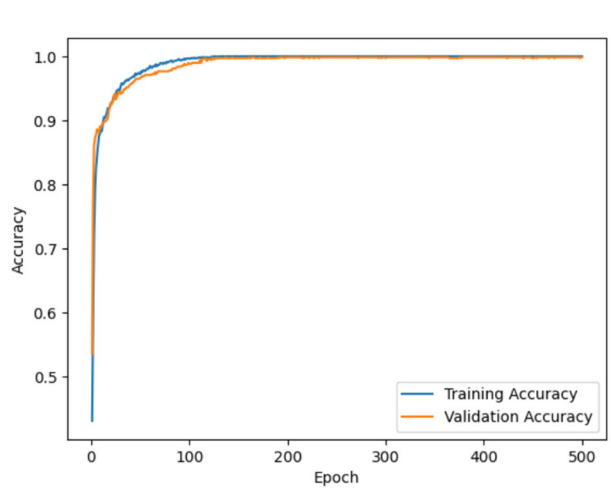


FIGURE 11
Model accuracy for OASIS.

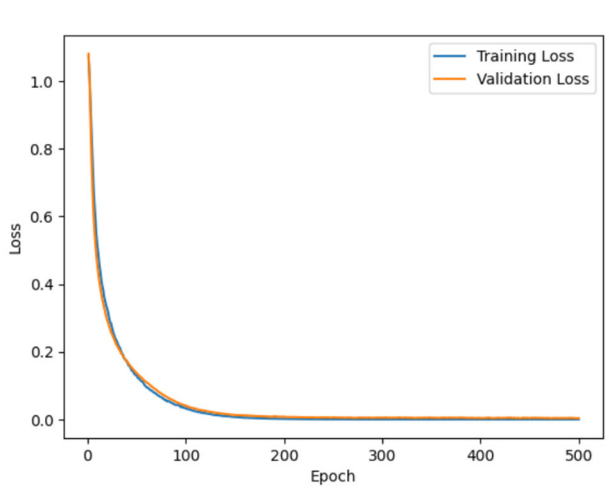


FIGURE 13
Model loss for OASIS.

recall rate of 0.99%, and F1 score of 0.99%. Similarly, proposed model using OASIS dataset obtains accuracy value of 0.9985%, precision value of 0.99%, recall rate of 0.99%, and F1 score of 0.99%.

The graphical representation of Table is portrayed in Figure 16.

4.6 Statistical analysis

Statistical outcome using proposed model is demonstrated in the Table 6.

The provided statistical values of t-statistic of 34.4585, *p*-value of 5.9389e-33, and Cohen's d of 11.9870 indicate an exceptionally strong effect size and highly significant results in the context of Alzheimer's disease classification. The t-statistic reflects a robust difference between groups, while the *p*-value confirms that this difference is extremely unlikely to be due to chance, surpassing conventional thresholds for statistical significance. Cohen's d, representing the effect size, indicates an extraordinarily large magnitude of difference between the compared groups, which is rare in clinical studies. Such values imply that the tested variable or method has a profound ability to distinguish between classifications, such as Alzheimer's disease vs. normal controls or other subgroups such as mild cognitive impairment (MCI). This level of statistical evidence strongly supports the reliability and clinical applicability of the classification approach, potentially aiding in early diagnosis or targeted intervention strategies for Alzheimer's disease.

Although the proposed model has delivered better outcome for classification of Alzheimer's disease, it is important to compare the proposed study with existing models; however, the dataset used in the model is a real time dataset; thus, external comparison is not feasible due to the implementation of real time data. However, from the analytical outcome, it can be identified that proposed study has delivered better outcome for multiclass classification of AD.

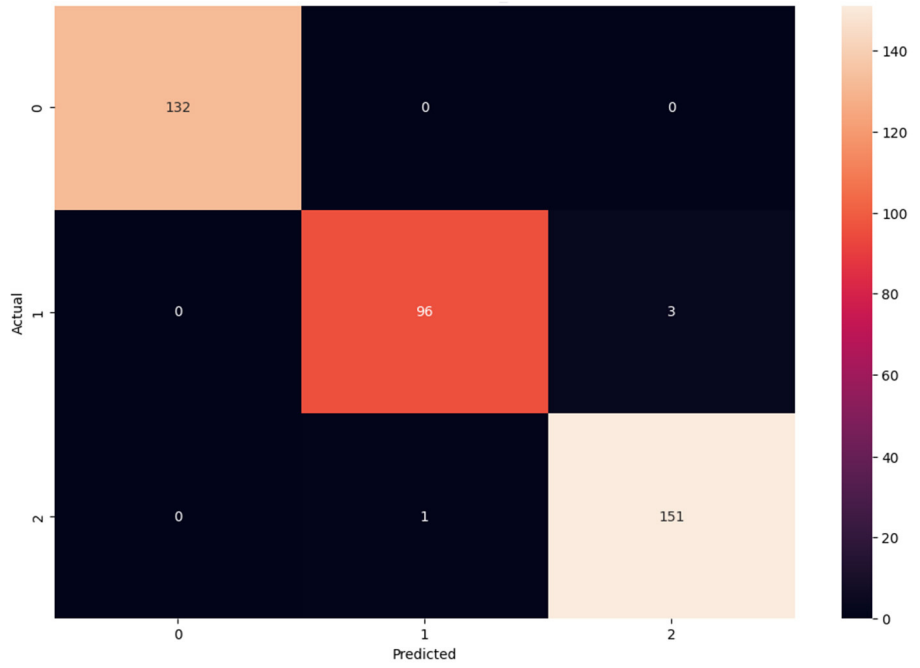


FIGURE 14
Confusion matrix for ADNI.

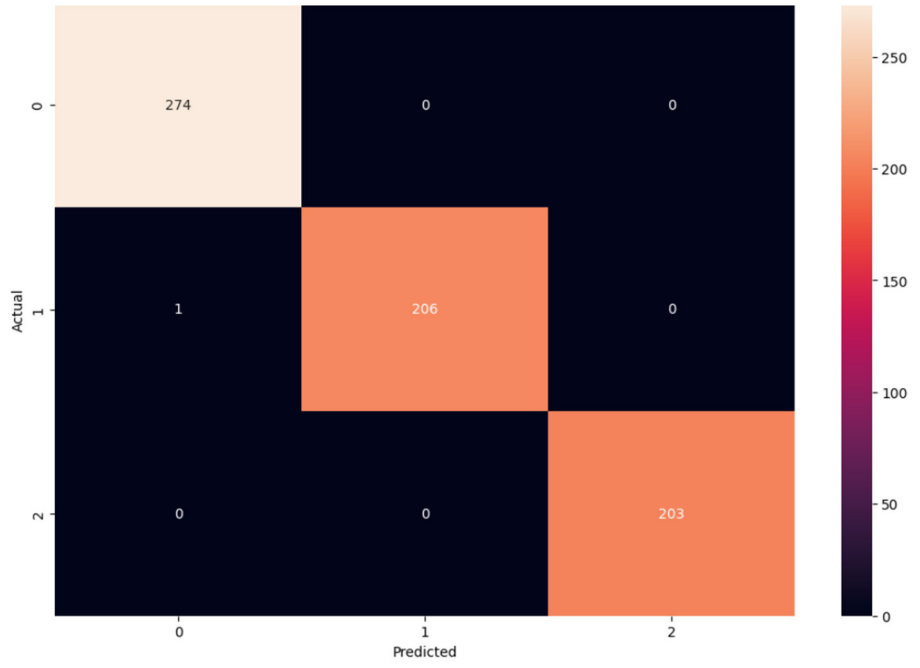


FIGURE 15
Confusion matrix for OASIS.

5 Discussion

Existing study has used Deep-CNN model for classification of AD. The model is fine-tuned to identify the subtle patterns and anomalies within the scans linked to AD. However, the finding obtained by the Deep-CNN is 96.64% of accuracy (36). Similarly, 2D and 3D CNN models (37) are explored in the study for AD classification. However, the accuracy outcome obtained by

2D-CNN model was 91.29% and 3D-CNN model was 91.07%. Moreover, classification of AD is carried out in the study based on ConvNets (38) using MRI images. However, the accuracy rates of classifications have reached up to 97.65% for AD/MCI and 88.37% for MCI/normal control. In addition, CNN is based on DenseNet Bottleneck-Compressed architecture (39) for AD diagnosis using MR images. The proposed model classified the input into five different categories, namely, CN, EMCI, MCI, LMCI, and AD, with an average accuracy of 86%. Thus, when compared to all these models, the accuracy obtained by the proposed framework is superior and effective as it gained 99.32% for ADNI dataset and 99.85% for OASIS dataset. This is due to the implementation of proposed ResNet for feature extraction and LSTM for classification.

6 Conclusion

The proposed research study delivered proficient results for the multiclass classification of AD as CN, MCI, and AD. Better performance was obtained for AD classification primarily due to the incorporation of effective AI approaches such as ResNet101 and LSTM. The proposed ResNet101 model used DKCL and PDPO layer for extracting relevant features needed for the proposed model. PDPO was employed to assign binary codes to pixels depending on the comparison with neighboring pixels, by efficiently capturing the local texture information, and the DCK layer captured the discriminative effectively by sliding a tiny filter over the input image and computing element-wise multiplication between the filter and overlapping regions of the input data. Implementation of these proposed functions in the proposed ResNet101 model aided in extracting relevant features needed for the model. Eventually, the extracted features were passed to the LSTM model for the classification of Alzheimer's disease as CN, MCI, and AD. In addition, the proposed research focused on employing the GAN model to find whether Alzheimer's disease is progressive or non-progressive by distinguishing the original class from the predicted class. Incorporation of the proposed model delivered a better accuracy rate of 0.9932 and 0.9985 for both ADNI and OASIS datasets.

In the future, different DL-based algorithms can be used for more advanced AD prediction. Employment of the GAN model is considered to be one of the major highlights of the proposed research study. However, this can be further developed in future study in terms of detecting brain deterioration rates for various classes. In addition, the integration of multi-modal data sources such as MEI, PET scans, and clinical biomarkers can be explored to assess the model's performance over time and to improve predictive accuracy. Thus, the combination of GANs and multi-modal data integration could pave the way for more sophisticated and accurate tools for early detection, prognosis, and management of Alzheimer's disease.

TABLE 5 Performance metrics.

MRI images	Accuracy	Precision	Recall	F1-score
ADNI	0.9932	0.99	0.99	0.99
OASIS	0.9985	0.99	0.99	0.99

TABLE 6 Statistical table.

Test	Values
t-statistic	34.4585
p-value	5.9389e-33
Cohen's d	11.9870

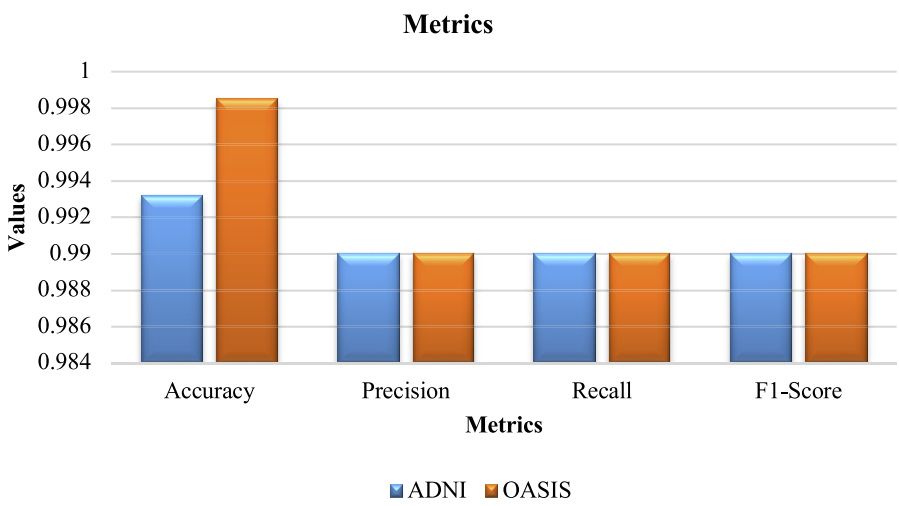


FIGURE 16
Graphical representation.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

PP: Conceptualization, Data curation, Software, Writing – original draft. SB: Formal analysis, Methodology, Writing – review & editing. JP: Conceptualization, Formal analysis, Resources, Writing – review & editing. EA: Supervision, Validation, Writing – review & editing. AAlg: Investigation, Visualization, Writing – review & editing. AAlm: Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R432), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This work was also supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant No. KFU251729). The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through small group research under grant number: RGP1/369/45.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Ávila-Jiménez J, Cantón-Habas V, del Pilar Carrera-González M, Rich-Ruiz M, Ventura S. A deep learning model for Alzheimer's disease diagnosis based on patient clinical records. *Comput Biol Med.* (2024) 169:107814. doi: 10.1016/j.compbiomed.2023.107814

2. Ayus I, Gupta D. A novel hybrid ensemble based Alzheimer's identification system using deep learning technique. *Biomed Signal Proces Control.* (2024) 92:106079. doi: 10.1016/j.bspc.2024.106079

3. Pradhan N, Sagar S, Singh AS. Analysis of MRI image data for Alzheimer disease detection using deep learning techniques. *J Magaz.* (2024) 83:17729–52. doi: 10.1007/s11042-023-16256-2

4. Sorour SE, Abd El-Mageed AA, Albarrak KM, Alnaim AK, Wafa AA, El-Shafeiy E. Classification of Alzheimer's disease using MRI data based on Deep Learning Techniques. *J King Saud Univ Comp Inform Sci.* (2024) 36:101940. doi: 10.1016/j.jksuci.2024.101940

5. Nasreen S. *An Investigation into Interactional Patterns for Alzheimer's Disease Recognition in Natural Dialogues* (2024).

6. Hason L. *Speech Features for Monitoring Alzheimer's Disease Using Random Forest Classifier*. Toronto Metropolitan University (2024).

7. Bashir T, Akhouri D. Shifting Paradigm in Early Detection and Prediction of Alzheimer's Disease. In: *Intelligent Solutions for Cognitive Disorders*. IGI Global (2024). p. 279–304. doi: 10.4018/979-8-3693-1090-8.ch013

8. Easwaran K, Ramakrishnan K, Jeyabal SN. Classification of cognitive impairment using electroencephalography for clinical inspection. *Proc Inst Mech Eng H.* (2024) 238:358–71. doi: 10.1177/09544119241228912

9. Gaikwad AA, Shinde SV. Diagnosis of dementia using MRI: a machine learning approach. In: *Applied Computer Vision and Soft Computing with Interpretable AI*. Chapman and Hall/CRC. p. 243–64.

10. Subramanian K, Hajamohideen F, Viswan V, Shaffi N, Mahmud M. Exploring intervention techniques for Alzheimer's disease: conventional methods and the role of AI in advancing care. *Artif Intellig Appl.* (2024) 2:73–91. doi: 10.47852/bonviewAIA42022497

11. Dara OA, Lopez-Guede JM, Raheem HI, Rahebi J, Zulueta E, Fernandez-Gamiz UJAS. Alzheimer's disease diagnosis using machine learning: a survey. *Appl Sci.* (2023) 13:8298. doi: 10.3390/app13148298

12. Mo S, Gomathi V, Uma Maheswari D, Justin A, Venkateswarlu BS. Examining novel treatment approaches and problems in Alzheimer's: an overview. *Lat Am J Pharm.* (2023) 42:554–64.

13. Borkar P, Wankhede VA, Mane DT, Limkar S, Ramesh J, Ajani NS. Deep learning and image processing-based early detection of Alzheimer disease in cognitively normal individuals. *Soft Comput.* (2024) 28 (Suppl 2):637. doi: 10.1007/s00500-023-08615-w

14. Pradhan N, Sagar S, Singh AS. Machine learning and deep learning algorithms for alzheimer disease detection and its implication in society 5.0. In: Kumar A, Sagar S, Thangamuthu P, Balamurugan B, editors. *Digital Transformation. Disruptive Technologies and Digital Transformations for Society 5.0.* Singapore: Springer (2024). doi: 10.1007/978-981-99-8118-2_12

15. Zhang C, Ge H, Zhang S, Liu D, Jiang Z, Lan C, et al. Hematoma evacuation via image-guided para-corticospinal tract approach in patients with spontaneous intracerebral hemorrhage. *Neurol Ther.* (2021) 10:1001–13. doi: 10.1007/s40120-021-00279-8

16. Pan H, Li Z, Fu Y, Qin X, Hu J. Reconstructing visual stimulus representation from eeg signals based on deep visual representation model. *IEEE Trans Hum Mach Syst.* (2024) 54:711–22. doi: 10.1109/THMS.2024.3407875

17. Turkson RE, Qu H, Mawuli CB, Eghan MJ. Classification of Alzheimer's disease using deep convolutional spiking neural network. *Neural Proc Lett.* (2021) 53:2649–63. doi: 10.1007/s11063-021-10514-w

18. Balaji P, Chaurasia MA, Bilfaqih SM, Muniasamy A, Alsid GJB. Hybridized deep learning approach for detecting Alzheimer's disease. *Biomedicines.* (2023) 11:149. doi: 10.3390/biomedicines11010149

19. Kapoor A, Shah R, Bhuva R, Pandit T. *Experiments for Architectural Basis of Convolutional Neural Networks for Image Recognition.* (2020). doi: 10.13140/RG.2.2.32989.56806.s

20. Hussain E, Hasan M, Hassan SZ, Azmi TH, Rahman MA, Parvez MZ. Deep learning based binary classification for Alzheimer's disease detection using

brain MRI images. In: *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. Kristiansand: IEEE Xplore (2020). p. 1115–20. doi: 10.1109/ICIEA48937.2020.9248213

21. Salehi W, Baglat P, Gupta G, Khan SB, Almusharraf A, Alqahtani A, et al. An approach to binary classification of Alzheimer's disease using LSTM. *Bioengineering*. (2023) 10:950. doi: 10.3390/bioengineering10080950

22. Aqeel A, Hassan A, Khan MA, Rehman S, Tariq U, Kadry S, et al. A long short-term memory biomarker-based prediction framework for Alzheimer's disease. *Sensors*. (2022) 22:1475. doi: 10.3390/s22041475

23. Tomassini S, Falcionelli N, Sernani P, Müller H, Dragoni AF. An end-to-end 3D ConvLSTM-based framework for early diagnosis of Alzheimer's disease from full-resolution whole-brain sMRI Scans. In: *Proceedings of the 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. Aveiro (2021). p. 74–8. doi: 10.1109/CBMS52027.2021.00081

24. Saratxaga CL, Moya I, Picón A, Acosta M, Moreno-Fernandez-de-Leceta A, Garrote E, et al. MRI deep learning-based solution for Alzheimer's disease prediction. *J Personal Med*. (2021) 11:902. doi: 10.3390/jpm11090902

25. Mohi ud din dar G, Bhagat A, Ansarullah SI, Othman MT, Hamid Y, Alkahtani HK, et al. A novel framework for classification of different Alzheimer's disease stages using CNN model. *Electronics*. (2023) 12:469. doi: 10.3390/electronics12020469

26. El-Sappagh S, Saleh H, Ali F, Amer E, Abuhmed T. Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time. *Neural Comput Appl*. (2022) 34:14487–509. doi: 10.1007/s00521-022-07263-9

27. Yigit A, Işik Z. Applying deep learning models to structural MRI for stage prediction of Alzheimer's disease. Turk J Elect Eng Comp Sci. (2020) 28:196–210. doi: 10.3906/elk-1904-172

28. Jiang X, Chang L, Zhang YD. Classification of Alzheimer's disease via eight-layer convolutional neural network with batch normalization and dropout techniques. *J Med Imag Health Inform*. (2020) 10:1040–8. doi: 10.1166/jmihi.2020.3001

29. Kavitha C, Mani V, Srividhya S, Khalaf OI, Tavera Romero CA. Early-stage Alzheimer's disease prediction using machine learning models. *Front Public Health*. (2022) 10:853294. doi: 10.3389/fpubh.2022.853294

30. Javeed A, Dallora AL, Berglund JS, Idrisoglu A, Ali L, Rauf HT, et al. Early prediction of dementia using feature extraction battery (feb) and optimized support vector machine (svm) for classification. *Biomedicines*. (2023) 11:439. doi: 10.3390/biomedicines11020439

31. Chen Y, Wang L, Ding B, Shi J, Wen T, Huang J, et al. Automated Alzheimer's disease classification using deep learning models with Soft-NMS and improved ResNet50 integration. *J Rad Res Appl Sci*. (2024) 17:100782. doi: 10.1016/j.jrras.2023.100782

32. Kumar S, Arif T. *CycleGAN-Based Data Augmentation to Improve Generalizability Alzheimer's Diagnosis Using Deep Learning* (2024). doi: 10.21203/rs.3.rs-4141650/v1

33. Zhang X, Gao L, Wang Z, Yu Y, Zhang Y, Hong J. Improved neural network with multi-task learning for Alzheimer's disease classification. *Heliyon*. (2024) 10:e26405. doi: 10.1016/j.heliyon.2024.e26405

34. Boudi A, He J, Abd El Kader I. Enhancing Alzheimer's disease classification with transfer learning: fine-tuning a pre-trained algorithm. *Curr Med Imag*. (2024) 20:e15734056305633. doi: 10.2174/0115734056305633240603061644

35. Akindele RG, Adebayo S, Kanda PS, Yu M. AlzhiNet: traversing from 2DCNN to 3DCNN, towards early detection and diagnosis of Alzheimer's disease. *arXiv preprint arXiv:2410.02714*. (2024). doi: 10.48550/arXiv.2410.02714

36. Singh R, Prabha C, Dixit HM, Kumari S. Alzheimer disease detection using deep learning. In: *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*. (2023). p. 1–6. doi: 10.1109/ICSSAS57918.2023.10331661

37. Suma KV, Raghavan D, Ganesh P. Deep learning for Alzheimer's disease detection using multimodal MRI-PET fusion. In: *4th International Conference on Circuits, Control, Communication and Computing (I4C)*. Bangalore (2022). p. 287–92. doi: 10.1109/I4C57141.2022.10057623

38. Ji H, Liu Z, Yan WQ, Klette R. Early diagnosis of Alzheimer's disease using deep learning. In: *Proceedings of the 2nd International Conference on Control and Computer Vision (ICCCV '19)*. New York, NY: Association for Computing Machinery (2019). p. 87–91. doi: 10.1145/3341016.3341024

39. Solano-Rojas B, Villalón-Fonseca RJS. A low-cost three-dimensional DenseNet neural network for Alzheimer's disease early discovery. *Sensors*. (2021) 21:1302. doi: 10.3390/s21041302

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 15 July 2025
DOI 10.3389/fmed.2025.1590201

OPEN ACCESS

EDITED BY
Ateeq Ur Rehman,
Gachon University, Republic of Korea

REVIEWED BY
Noor Kamal Al-Qazzaz,
University of Baghdad, Iraq
Pengxiang Su,
Nanchang University, China

*CORRESPONDENCE
Muhammad Hanif
muhammad.hanif@oru.se

RECEIVED 09 March 2025
ACCEPTED 20 June 2025
PUBLISHED 15 July 2025

CITATION
Khan W, Khan MS, Qasem SN, Ghaban W, Saeed F, Hanif M and Ahmad J (2025) An explainable and efficient deep learning framework for EEG-based diagnosis of Alzheimer's disease and frontotemporal dementia. *Front. Med.* 12:1590201. doi: 10.3389/fmed.2025.1590201

An explainable and efficient deep learning framework for EEG-based diagnosis of Alzheimer's disease and frontotemporal dementia

Waqar Khan[1], Muhammad Shahbaz Khan[2], Sultan Noman Qasem[3,4], Wad Ghaban[5], Faisal Saeed[6], Muhammad Hanif[7]* and Jawad Ahmad[8]

[1]Department of Cybersecurity, Pakistan Navy Engineering College, National University of Sciences and Technology, Karachi, Pakistan, [2]School of Computing, Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, United Kingdom, [3]Computer Science Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia, [4]King Salman Center for Disability Research, Riyadh, Saudi Arabia, [5]Applied College, University of Tabuk, Tabuk, Saudi Arabia, [6]College of Computing, Birmingham City University, Birmingham, United Kingdom, [7]Department of Informatics, School of Business, Örebro Universitet, Örebro, Sweden, [8]Cybersecurity Center, Prince Mohammad Bin Fahd University, Al-Khobar, Saudi Arabia

The early and accurate diagnosis of Alzheimer's Disease and Frontotemporal Dementia remains a critical challenge, particularly with traditional machine learning models which often fail to provide transparency in their predictions, reducing user confidence and treatment effectiveness. To address these limitations, this paper introduces an explainable and lightweight deep learning framework comprising temporal convolutional networks and long short-term memory networks that efficiently classifies Frontotemporal dementia (FTD), Alzheimer's Disease (AD), and healthy controls using electroencephalogram (EEG) data. Feature engineering has been conducted using modified Relative Band Power (RBP) analysis, leveraging six EEG frequency bands extracted through power spectrum density (PSD) calculations. The model achieves high classification accuracies of 99.7% for binary tasks and 80.34% for multi-class classification. Furthermore, to enhance the transparency and interpretability of the framework, SHAP (SHapley Additive exPlanations) has been utilized as an explainable artificial intelligence technique that provides insights into feature contributions.

KEYWORDS

explainable AI, XAI, Alzheimer's disease, temporal convolutional networks, long short-term memory, frontotemporal dementia, EEG, mental disorders

1 Introduction

Frontotemporal dementia (FTD) (1) and Alzheimer's disease (2) (AD) are two most prevalent forms of dementia, primarily affecting individuals over 40 years of age. The global prevalence of dementia is expected to reach more than 130 million cases by 2050 (3). The rise in cases related to these diseases have significantly strained healthcare systems around the world, necessitating an urgent need for accurate and early diagnostic methods. The diagnosis of (FTD) and AD relies on the methodologies, such as neuropsychological

evaluations (4), biomarkers analysis (5), established clinical criteria (6), and magnetic resonance imaging (MRI) (7). But the time requirements, need for expert interpretation, limit the practicality of advanced neuroimaging tools, and the high cost. Therefore, there is a critical need for early and accurate diagnosis, there is an indispensable need for improved detection methods. Timely diagnosis is critical, as early intervention can help slow disease progression and enhance patients' quality of life.

Electroencephalograms (EEG) offer features such as high temporal resolution, lower cost, and real-time monitoring, which make them valuable for dementia diagnosis. EEG signals in conjunction with machine learning, hold tremendous potential to be an effective non-invasive method to detect and monitor (FTD) and AD (8). However, extracting features from EEG is a crucial task, and although various methods have been proposed in research (9, 10), many of them have not achieved high accuracies with deep learning and machine learning models. Therefore, novel and tailored approaches are needed to extract high-quality data from EEG for improved analysis and diagnosis based on deep learning.

Deep learning (DL) models have shown significant potential in classifying EEG data, offering improved accuracy and efficiency in analysis. However, there is a need for lightweight models to optimize data processing and develop a high-performing model that is time-efficient, and computationally less loaded. In addition, most ML and DL models function as "black boxes," providing outputs without transparency, which limits their acceptance, especially in sensitive fields like healthcare. Explainable Artificial Intelligence (XAI) offers a solution by revealing what the models learn during training and how decisions are made during prediction, making the results more understandable and interpretable. The core contributions of this research are given below.

- This research introduces an EEG-based feature extraction approach using modified Relative Band Power (RBP) analysis for feature engineering and proposes a lightweight hybrid deep learning classifier for accurate and robust classification of frontotemporal dementia, Alzheimer's disease, and health.
- SHAP (SHapley Additive Explanations), an explainable artificial intelligence technique has been integrated into the model to provide deeper insights into feature contributions, increasing interpretability, transparency, and prediction reliability for mental disorder diagnosis.

This is how the rest of the article is organized. Related work is covered in Section 2, and methodology is covered in Section 3. Our research findings are shown in Section 4, and explainable artificial intelligence is covered in Section 5. Section 6 concludes with a summary of our findings and recommendations for future research.

2 Related work

Recent studies have focused on enhancing the Alzheimer's disease detection with advanced machine learning methods. To solve supervised AD detection using EEG data analysis, machine, and deep learning-based systems have gained popularity (11–13). The study (14) used a public EEG signal dataset that included recordings from 12 Alzheimer's disease patients and 11 healthy controls. A directed graph approach was applied for local texture feature extraction, resulting in 448 low-level features per EEG signal. This was further enhanced by combining it with a tunable q-factor wavelet transform, resulting in a total of 8,512 features per signal input. The accuracy of the model was 92.01% with leave-one-subject-out (LOSO) cross-validation and 100% with 10fold cross-validation.

Moreover, six supervised machine-learning approaches were used in this work (15) to categorize processed EEG data from patients with FTD and AD. Different techniques for processing and analyzing EEG signals were applied to identify relevant features.The accuracy of the decision tree machine learning model was 78.5%, while the random forest model attained an accuracy of 86.3% in diagnosing FTD. This study (16) proposes a convolutional neural network-based model called STEADYNet, which achieves high performance with 98.24% accuracy in dementia detection using multichannel spatiotemporal EEG signals.

Another study (17) proposes a CNN-based model utilizing the Forward-Backward Fourier Transform (FBFT) to enhance EEG signal visualization for brain disorder classification. The model achieves 85.1% for murmur, 99.82% accuracy for epilepsy, 100% for mental stress, and 95.91% for Alzheimer's disease (AD). Additionally, the eye-naked classification approach attains 78.6%, 71.9%, 82.7%, and 91.0% accuracy for epilepsy, AD, murmur, and mental stress, respectively.

In addition, a study (18) offers a "dual-input convolution encoder network" as a unique method for classifying AD. Denoising and the extraction of band power and coherence characteristics from the EEG data were important feature engineering approaches. With an accuracy of 83.28% in differentiating AD patients from healthy controls, the presented model combines convolutional layers with transformer architecture, and feed-forward module and proves its efficacy in collecting intricate EEG features.

3 Methodology

3.1 Data collection

The dataset (8) consists of EEG recordings from 88 subjects (36 Alzheimer's disease, 29 healthy and 23 frontotemporal dementia) obtained at the 2nd Neurology Department of AHEPA General University Hospital, and data statistics as shown in Figure 1. EEG signals were captured using 19 electrodes while participants remained seated with their eyes closed. The data was initially filtered at 0.5–60 Hz and sampled at 500 Hz.

3.2 Data preprocessing

To enhance the quality of the electroencephalogram (EEG) signals and remove unwanted artifacts, a systematic pre-processing technique has been applied. Initially, a Butterworth bandpass filter with a frequency range of 0.5 Hz to 45 Hz was used to retain

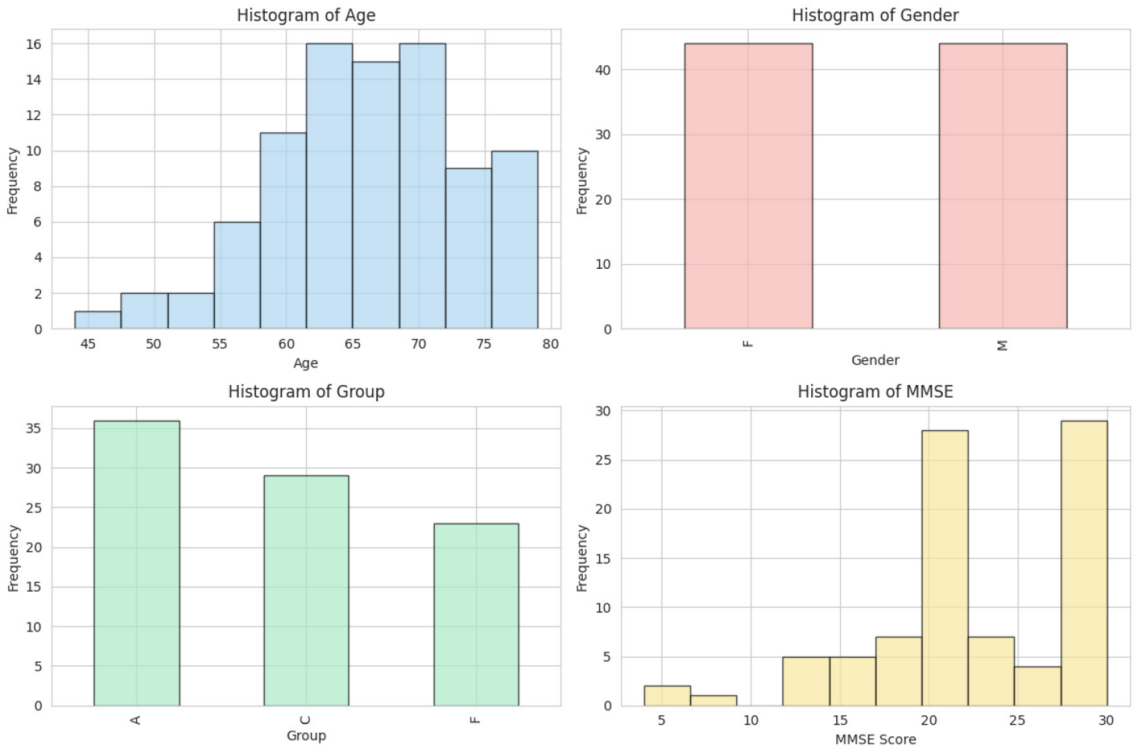


FIGURE 1
Statistical overview of the dataset.

relevant neural activity while eliminating low-frequency drifts and high-frequency noise. Next, Artifact Subspace Reconstruction (ASR) was implemented to identify and correct signal distortions. ASR detects artifacts by measuring the standard deviation of signal segments within a 0.5-s window. Segments exceeding a deviation threshold of 17 were reconstructed to suppress transient artifacts while preserving the integrity of neural activity. After the artifact correction, Independent Component Analysis (ICA) was performed using the RunICA algorithm. This process decomposed the 19-channel EEG signals into independent components, as illustrated in Figure 2. The independent components were then analyzed using EEGLAB's ICLabel tool, which automatically classifies components based on their source characteristics. Components identified as "eye artifacts" or "jaw artifacts" were removed to ensure that only neural activity remained in the processed signals. Although EEG signals were recorded in a closed-eye resting state, some residual eye movement artifacts were still present. The implemented pre-processing steps effectively mitigated these unwanted influences, ensuring cleaner EEG signals for subsequent analysis.

3.3 Feature engineering

In EEG classification tasks, relative band power (RBP) (15) is often extracted, especially when analyzing brain activity related to various neurological and cognitive states. The RBP is calculated for several frequency bands that correspond to various facets of brain activity. Six interesting frequency bands were taken into consideration in this study:

- **Delta:** $0.5 \leq f < 4$ Hz
- **Theta:** $4 \leq f < 8$ Hz
- **Alpha:** $8 \leq f < 16$ Hz
- **Zaeta:** $16 \leq f < 24$ Hz
- **Beta:** $24 \leq f < 30$ Hz
- **Gamma:** $30 \leq f \leq 45$ Hz.

The Welch technique is used to compute the Power Spectral Density by a given equation

$$\mathrm{PSD}(f) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} |X(f_n)|^2 \tag{1}$$

where $X(f_n)$ is the Fourier transform of the signal $x(t)$ evaluated at frequency bins f_n, and N is the total number of segments over which the Fourier transform is averaged. The overall power in the frequency range of 0.5–45 Hz is calculated by summing the PSD values.

$$\text{Total PSD} = \sum_{f=\min}^{\max} \mathrm{PSD}(f) \tag{2}$$

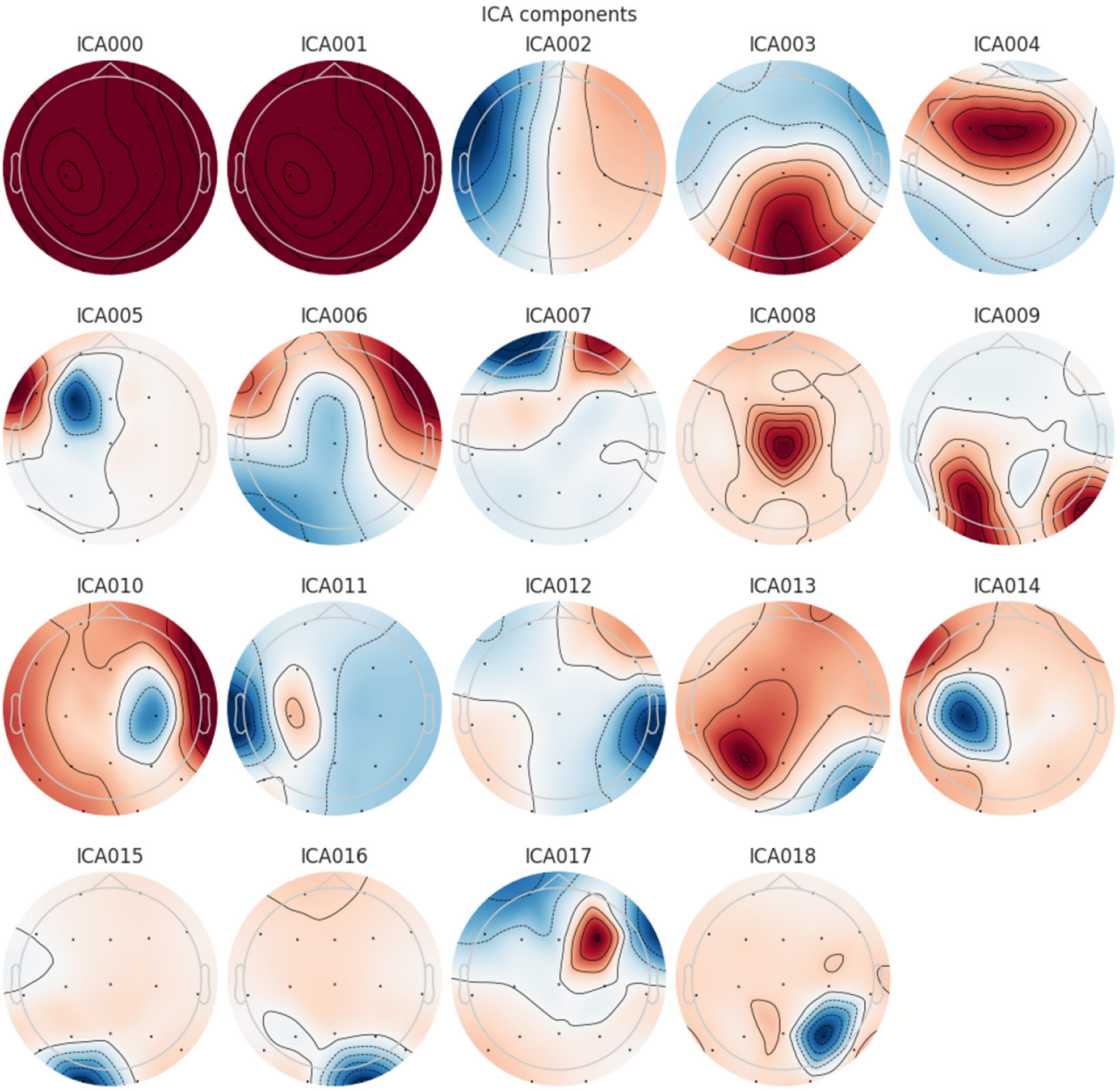


FIGURE 2
Independent Component Analysis (ICA) components extracted from EEG signal.

The RBP for each frequency band b is determined by dividing the power within the band by the overall power.

$$\mathrm{RBP}_b = \frac{\sum_{f=\min}^{f_{\max}} \mathrm{PSD}(f)}{\sum_{f=0.5}^{45} \mathrm{PSD}(f)} \tag{3}$$

The power in the frequency band $[f_{\min}, f_{\max}]$ is represented by the numerator, while the total power in the range of 0.5 Hz to 45 Hz is the denominator.

These bands provide greater in-depth observations and cover a wider range of brain activity. In order to compute the RBP, EEG signals are segmented into epochs, each 6 s in length and sharing a 50% overlap. By splitting the signal into overlapping segments, calculating the squared magnitude of the discrete Fourier transform for each segment, and then averaging the results, the Welch technique is used to estimate the Power spectral Density for each epoch. After that, the relative power inside each frequency band is determined by dividing the PSD for that band by the PSD for the whole frequency range of interest 0.5–45 Hz. A normalized measure of brain activity is provided by this ratio, which shows the contribution of each frequency band to the signal's overall strength. For each epoch, the RBP is computed across all channels:

$$\text{Epoch RBP} = \frac{1}{N_{\text{channels}}} \sum_{i=1}^{N_{\text{channels}}} \mathrm{RBP}_b(i) \tag{4}$$

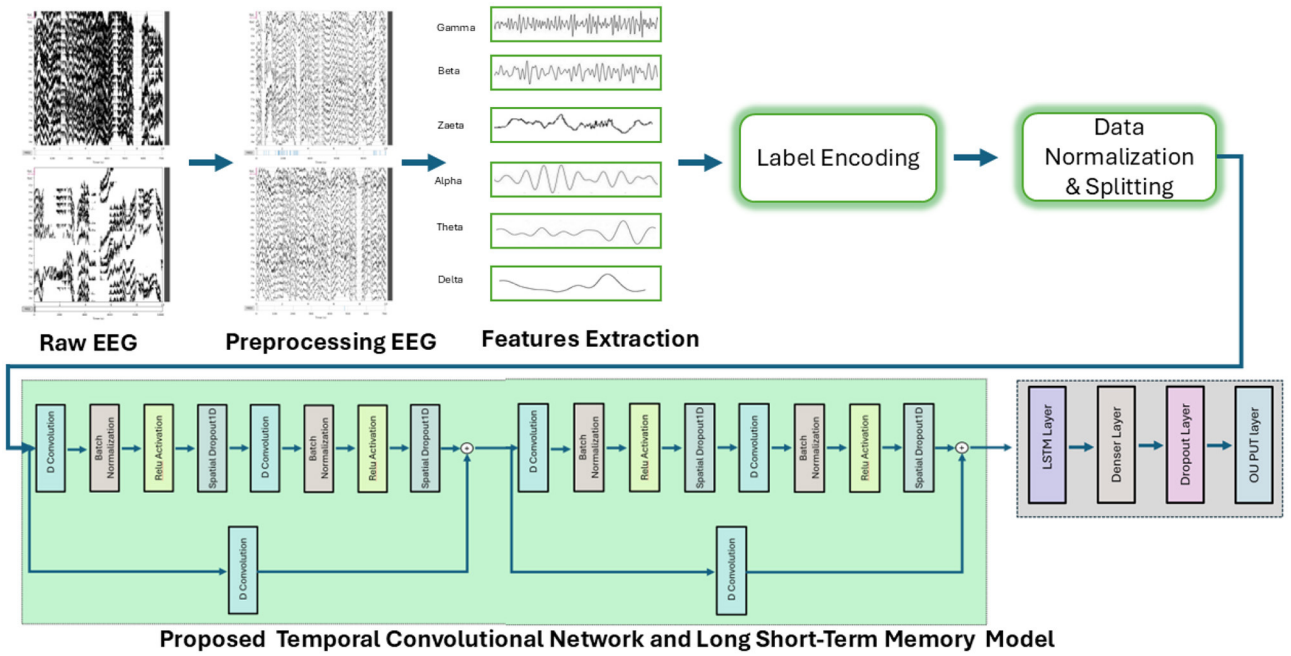


FIGURE 3
The proposed methodology with the proposed deep learning model.

where $\mathrm{RBP}_b(i)$ is the RBP for the i-th channel and N_{channels} is the number of EEG channels.

The RBP values for every epoch make up the final feature matrix. The columns match the six frequency bands (Beta, Delta, Alpha, Theta, Zaeta, and Gamma), whereas each row denotes an epoch:

$$\text{Feature Matrix} = [\text{Delta Theta Alpha Zaeta Beta Gamma Label}]$$

Once the RBP features have been extracted, they are used as inputs for classification tasks. Each epoch is labeled according to whether the person has frontotemporal dementia, Alzheimer's disease, or cognitive normal.

3.4 Label encoding and data normalization and splitting

The data was saved in a comma-separated file, and then categorical variables were converted to numerical data using one-hot encoding. Then, the data was normalized using the min-max normalization formula given by:

$$\chi^* = \frac{\chi - \mu_{\min}}{\mu_{\max} - \mu_{\min}} \tag{5}$$

The normalized value is represented by χ^*, the original value is represented by χ, and the dataset's minimum and maximum values are indicated by $\mu_{\min}$ and $\mu_{\max}$, respectively. Training, validation, and test data sets were split into 80%, 10%, and 10% of the total data set.

3.5 The proposed deep learning model

The proposed hybrid model as given in Figure 3. Its consists of two deep learning components LSTM and TCN. The TCN uses dilated causal convolutions to obtain high-level features from the input sequence, and the LSTM captures the sequential dependencies. The Temporal Convolutional Network enhances traditional CNNs with dilated causal convolutions, allowing them to model long-term temporal patterns without violating sequence order.

$$H^{(l)} = \sigma(W^{(l)} * X + b^{(l)}) \tag{6}$$

where $H^{(l)}$ represents the output of the l-th convolutional layer, the learnable convolutional filters are represented by $W^{(l)}$, the convolution operation is represented by $*$, the bias is represented by $b^{(l)}$, and the ReLU activation function is represented by $\sigma(\cdot)$.

Long-range interdependence in EEG data can be captured with the use of dilated convolutions:

$$H_t^{(l)} = \sum_{i=0}^{k-1} W_i^{(l)} \cdot X_{t-d\cdot i} + b^{(l)} \tag{7}$$

where k is the kernel size and d is the dilation rate. To optimize both stability and the flow of gradients, residual connections are adopted.

$$H_{\text{res}}^{(l)} = H^{(l)} + X \tag{8}$$

This structure enables efficient learning without vanishing gradients.

LSTMs are a unique class of recurrent neural networks that use gate mechanisms and memory cells to manage long-term

dependencies. The LSTM uses three primary gates—forget, input, and output gates to process the features that were extracted from TCN.

$$f_t = \sigma(W_f H_t^{(l)} + U_f h_{t-1} + b_f) \quad (9)$$

$$i_t = \sigma(W_i H_t^{(l)} + U_i h_{t-1} + b_i) \quad (10)$$

$$c_t' = \tanh(W_c H_t^{(l)} + U_c h_{t-1} + b_c) \quad (11)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot c_t' \quad (12)$$

$$o_t = \sigma(W_o H_t^{(l)} + U_o h_{t-1} + b_o) \quad (13)$$

$$h_t = o_t \odot \tanh(c_t) \quad (14)$$

where W_f the input's weight matrix at time step t. The input to the LSTM at layer l and time t is represented by the item $H_t^{(l)}$. U_f Weight matrix for the preceding time step's hidden state. h_{t-1} The previously hidden state. Adding the bias term b_f and the sigmoid activation function σ.

The model begin with an input layer shaped (6,1), followed by a 1D convolutional layer with 32 filters as shown in Table 1. The first layer connects to a batch normalization layer having 128 number of parameters and an activation function, then goes through a spatial dropout layer having value 0.2. The next convolutional layer also uses 32 filters, followed by batch normalization and another activation layer. A residual connection is created by adding the output of a separate convolution layer with the same shape, allowing the model to retain important features. Furthermore, a similar set of layers is added next, helping the model process the input in the same way as before. The model uses an LSTM layer with 64 units to capture temporal features. Following this, the model includes a dense layer with 128 units, which is succeeded by two additional dense layers containing 192 and 256, units respectively; each of these layers is paired with a dropout mechanism to help mitigate overfitting, culminating in a final dense layer with 3 output units that delivers the classification outcome.

3.6 Hyperparameter tuning

Random search-based hyperparameter tuning was used to find the optimal number of layers in the proposed model. The best hyperparameter values for the CNN component are: two TCN blocks, 32 filters, a kernel size of 7, a dropout rate of 0.3, and a dilation rate of 1. During optimization, the best LSTM structure was found to be a single layer of 64 units. Dense layers follow with 128, 192, and 256 units and a 0.2 dropout rate and early stopping mitigates overfitting. The number of training epochs depended on the specific classification task. A batch size of 32 was used, and the Adam optimizer was selected with a learning rate of 0.0001. The model has 131,587 parameters. it uses 514.01 KB of memory, making it suitable for deployment on edge medical devices for real-time mental disorder detection. Out of these, 131,331 are trainable and 256 are non-trainable as shown in the Table 2. The model was trained using 8 GB RAM, and each epoch took 6 s.

TABLE 1 Model architecture summary.

Layer (type)	Output shape	Parameters	Connected to
Input layer	(None, 6, 1)	0	-
Conv 1D	(None, 6, 32)	256	Input layer [0][0]
Batch normalization	(None, 6, 32)	128	Conv1D [0][0]
Activation	(None, 6, 32)	0	Batch normalization [0][0]
Spatial dropout 1D	(None, 6, 32)	0	Activation [0][0]
Conv1D	(None, 6, 32)	7,200	Spatial dropout 1D [0][0]
Batch normalization	(None, 6, 32)	128	Conv1D [1][0]
Activation	(None, 6, 32)	0	Batch normalization [1][0]
Conv 1D	(None, 6, 32)	64	Input layer [0][0]
Spatial dropout 1D	(None, 6, 32)	0	Activation [1][0]
Add	(None, 6, 32)	0	Conv1D[2][0], Spatial dropout 1D [1][0]
Conv 1D	(None, 6, 32)	7,200	Add[0][0]
Batch normalization	(None, 6, 32)	128	Conv1D [3][0]
Activation	(None, 6, 32)	0	Batch normalization [2][0]
Spatial dropout 1D	(None, 6, 32)	0	Activation [2][0]
Conv 1D	(None, 6, 32)	7,200	Spatial dropout 1D [2][0]
Batch normalization	(None, 6, 32)	128	Conv 1D [4][0]
Activation	(None, 6, 32)	0	Batch normalization [3][0]
Conv 1D	(None, 6, 32)	1,056	Add[0][0]
Spatial dropout 1D	(None, 6, 32)	0	Activation [3][0]
Add	(None, 6, 32)	0	Conv1D[5][0], Spatial dropout 1D [3][0]
LSTM	(None, 64)	24,832	Add [1][0]
Dense	(None, 128)	8,320	LSTM [0][0]
Dropout	(None, 128)	0	Dense [0][0]
Dense	(None, 192)	24,768	Dropout [0][0]
Dropout	(None, 192)	0	Dense [1][0]
Dense	(None, 256)	49,408	Dropout [1][0]
Dropout	(None, 256)	0	Dense [2][0]
Dense	(None, 3)	771	Dropout [2][0]

3.7 Classification

The proposed hybrid Temporal Convolutional Network model with Long Short-Term Memory was utilized to perform four types of classification tasks for Alzheimer's Disease, Frontotemporal Disease, and healthy classes. The classification tasks are as follows:

- **Classification for Alzheimer's, frontotemporal, and healthy classes:**the objective of this work was to categorize three different classes: healthy controls, frontotemporal disease, and Alzheimer's disease. The model was trained to distinguish between the three groups.
- **Classification for Alzheimer + frontotemporal disease and healthy classes:** in this classification the model was trained to classify a combined class of Alzheimer's Disease and Frontotemporal Disease from healthy individuals.
- **Classification for Alzheimer's disease and healthy classes:** the objective of this task is to train the model to classification between the Healthy class and Alzheimer's disease.
- **Classification for frontotemporal disease and healthy classes:** this classification task required the model to separate individuals with Frontotemporal Disease from healthy controls.

4 Results

4.1 Performance parameters

To access the performance of the proposed model, the key performance parameters, i.e., precision, F1 score, accuracy, recall, sensitivity, etc. have been extensively evaluated. Among other, accuracy is the most important performance parameter for assessing a classification model's efficacy. It measures the proportion of accurately predicted instances to all instances in the dataset, including true positives and negatives. Mathematically, accuracy can be written as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100 \quad (15)$$

TABLE 2 Model parameter summary.

Parameter type	Count	Size
Total parameters	131,587	514.01 KB
Trainable parameters	131,331	513.01 KB
Non-trainable parameters	256	1 KB

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (16)$$

Similarly, precision measures the quality of the model's prediction. It measures the percentage of properly identified positive cases in comparison to the total number of cases that were

TABLE 4 Classification metrics for Alzheimer + frontotemporal disease and healthy classes.

Metric	Alzheimer + frontotemporal disease	Healthy
Precision	0.9977	0.9987
Recall	0.9993	0.9956
F1 score	0.9985	0.9972
Support	2,983	1,596
Sensitivity	1.00	
Specificity	1.00	

TABLE 5 Classification metrics for Alzheimer's disease and healthy classes.

Metric	Alzheimer's disease	Healthy
Precision	0.9963	0.9987
F1 Score	0.9976	0.9972
Recall	0.9989	0.9956
Support	1876	1596
Sensitivity	1.00	
Specificity	1.00	

TABLE 6 Classification metrics for frontotemporal disease and healthy classes.

Metric	Frontotemporal disease	Healthy
F1 score	0.9975	0.9964
Recall	0.9956	0.9991
Precision	0.9994	0.9937
Support	1597	1596
Sensitivity	1.00	
Specificity	1.00	

TABLE 3 Classification metrics for Alzheimer, frontotemporal, and healthy classes.

Class	Precision	Recall	F1-score	Sensitivity	Specificity	Support
Alzheimer	0.70	0.90	0.79	0.90	0.74	1,876
Frontotemporal	1.00	1.00	1.00	1.00	1.00	1,597
Healthy	0.68	0.35	0.47	0.35	0.95	1,106

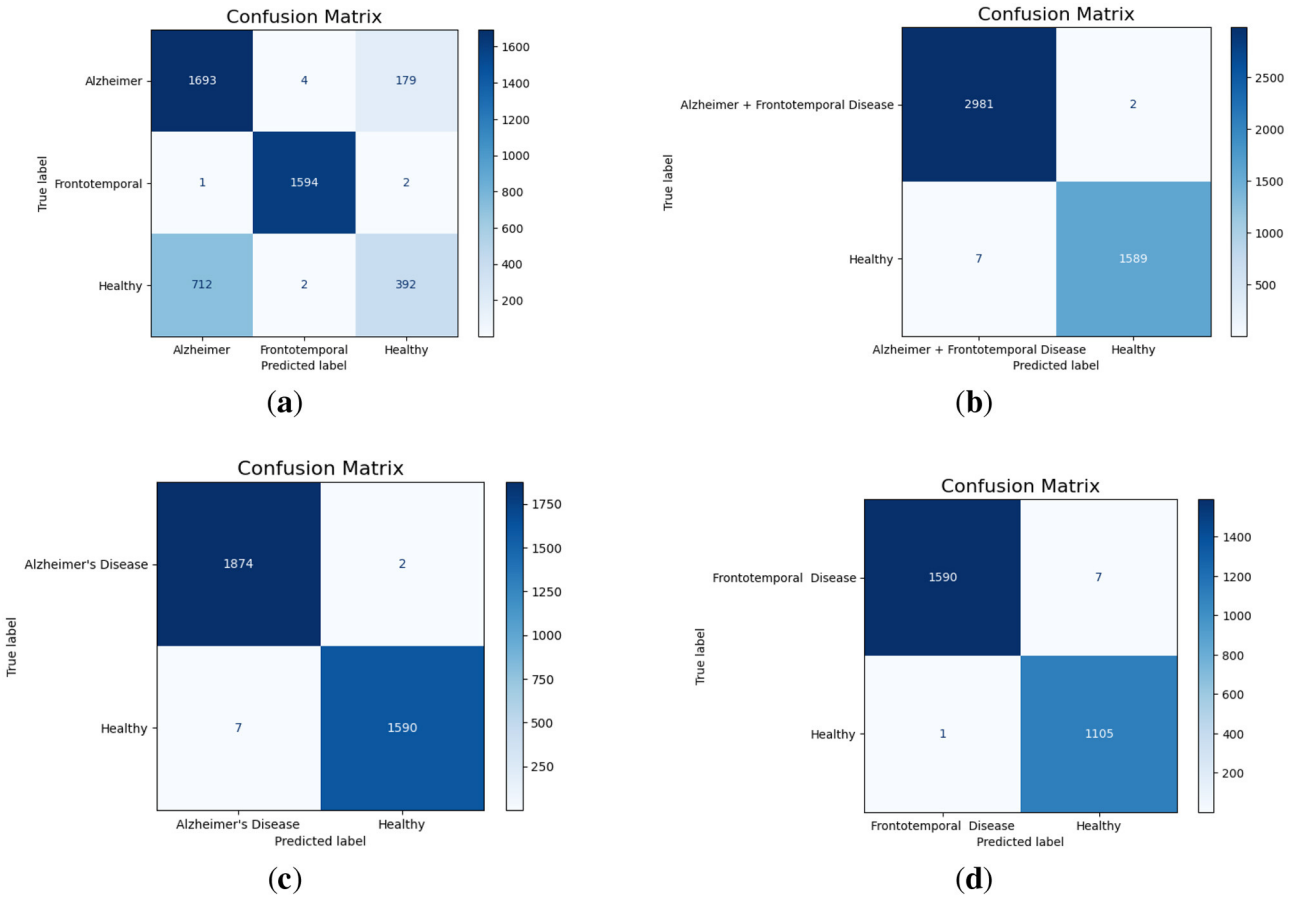


FIGURE 4
Confusion matrices for different classification scenarios. **(a)** Alzheimer vs. Frontotemporal vs. Healthy. **(b)** Alzheimer + Frontotemporal vs. Healthy. **(c)** Alzheimer vs. Healthy. **(d)** Frontotemporal vs. Healthy.

anticipated to be positive (sum of true positives and false positives). Precision can be shown mathematically as:

$$\text{Precision} = \frac{\text{Number of Correctly Predicted Positive Cases}}{\text{Total Predicted Positive Cases}} \times 100 \tag{17}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{18}$$

Recall, also known as the true positive rate, is a crucial performance indicator that assesses how well a classification model detects positive. Recall can be mathematically represented as:

$$\text{Recall} = \frac{\text{Number of Correctly Identified Positive Cases}}{\text{Total Actual Positive Cases}} \times 100 \tag{19}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{20}$$

The F1 score offers a balance between accuracy and recall by taking the harmonic mean of the accuracy and recall metrics. It is particularly convenient when dealing with imbalanced datasets. Mathematically, the F1 score is expressed as:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100 \tag{21}$$

$$\text{F1-score} = \frac{2 \times \frac{\text{TP}}{\text{TP+FP}} \times \frac{\text{TP}}{\text{TP+FN}}}{\frac{\text{TP}}{\text{TP+FP}} + \frac{\text{TP}}{\text{TP+FN}}} \times 100 \tag{22}$$

Specificity measures the proportion of actual negatives that the model correctly identifies. It evaluates the model's ability to correctly identify true negatives.

$$\text{Specificity} = \frac{\text{Number of Correctly Identified Negative Cases}}{\text{Total Actual Negative Cases}} \times 100 \tag{23}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{24}$$

In above equations, TP represents True Positives or correctly identified positive cases, TN represents True Negatives or

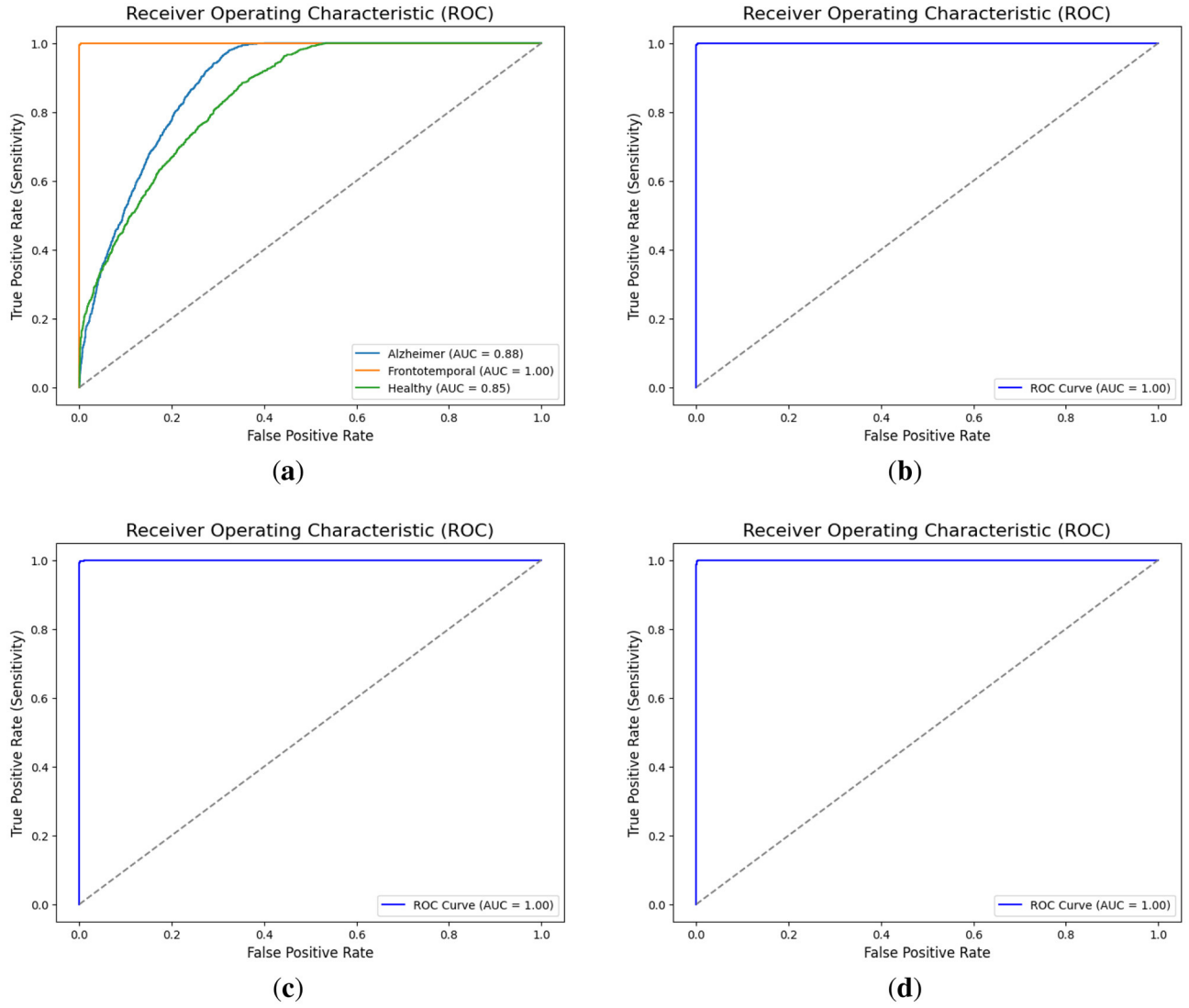


FIGURE 5
AUC curves for different classification scenarios. **(a)** Alzheimer vs. Frontotemporal vs. Healthy. **(b)** Alzheimer + Frontotemporal vs. Healthy. **(c)** Alzheimer vs. Healthy. **(d)** Frontotemporal vs. Healthy.

correctly identified negative cases, FP represents False Positives or incorrectly classified as positive, and FN represents False Negatives or incorrectly classified as negative.

4.2 Performance evaluation

Table 3 presents the classification metrics for the three classes: Alzheimer, Frontotemporal, and Healthy.The model achieves 70% precision and 90% recall for Alzheimer's, with an F1-score of 0.79. It performs perfectly for frontotemporal Disease with 100% F1-score, precision, and recall, while the Healthy class shows weaker performance with 68% precision, 35% recall, and an F1-score of 0.47.

Table 4 focuses on a binary classification task where Alzheimer + Frontotemporal Disease are treated as a combined class, and Healthy is the other class.he model achieves nearly perfect results for Alzheimer + Frontotemporal Disease with 99.77% precision and 99.93% recall. The Healthy class also performs well with 99.87% precision and 99.56% recall, both classes showing 1.00 sensitivity and specificity.

In the binary classification task, the Table 5, the goal is to classify Alzheimer's disease and healthy individuals.The model excels with 99.63% precision and 99.89% recall for Alzheimer's, and 99.87% precision and 99.56% recall for Healthy, both showing 1.00 sensitivity and specificity.

The Table 6 shows the results of the binary classification between frontotemporal disease and healthy individuals. The model shows 99.94% precision and 99.56% recall, with a very high F1 score of 0.9975. The Healthy class also has a high F1-score of 0.9964, with 99.37% precision and 99.91% recall. Both classes show 1.00 sensitivity and specificity. Similarly, Table 7 reports high classification accuracies for binary dementia tasks (≥0.997) and a lower accuracy (0.8034) for the three-class classification among Alzheimer's, frontotemporal disease, and healthy subjects.

The multi-class confusion matrix given in Figure 4a that the model effectively classifies Alzheimer's and frontotemporal Disease. Out of 1,876 Alzheimer cases, 1,693 are correctly identified, with minor misclassifications into healthy and frontotemporal. Similarly, frontotemporal disease achieves a near-perfect classification with only three misclassifications. However, the model struggles significantly with Healthy cases, misclassifying 712 instances as Alzheimer's, highlighting room for improvement in distinguishing healthy from disease classes. When combining Alzheimer's and Frontotemporal as a single class against Healthy, the model demonstrates almost perfect classification as shown in Figure 4b. Only 2 out of 2,983 Alzheimer + frontotemporal instances are misclassified as healthy. For the Healthy class, only 7 out of 1,596 instances are misclassified, indicating strong model performance in binary classification with very few false positives or false negatives. For Alzheimer's Disease vs. Healthy classification, as displayed in Figure 4c, the model achieves excellent performance. Out of 1,876 Alzheimer cases, only 2 are misclassified as Healthy. Similarly, for 1,590 Healthy cases, only 7 are misclassified as Alzheimer. Furthermore, the model performs well in the binary classification of frontotemporal disease against healthy as evident from Figure 4d. Just one healthy case out of 1,106 is incorrectly categorized as frontotemporal, whereas only 7 out of 1,597 frontotemporal patients are incorrectly classified as healthy.

The multi-class ROC (Receiver Operating Characteristic) curve, as given in Figure 5a, displays the AUC (Area Under the Curve) for each class. Alzheimer's disease has an AUC of 0.88, which indicates good but not perfect discrimination; healthy cases have the lowest AUC of 0.85, which indicates some difficulty in differentiating them from the disease classes; and frontotemporal disease achieves a perfect AUC of 1.00, which indicates ideal classification with no false positives or negatives. The binary classification combining Alzheimer's and frontotemporal as one class vs. healthy achieves an exceptional AUC of 1.00. Additionally, the model attains an AUC of 1.00 for Alzheimer's Disease vs. Healthy cases, indicating perfect discrimination. Similarly, the classification of 'Alzheimer + Frontotemporal' vs. Healthy, Alzheimer vs. Healthy, and Frontotemporal Disease vs. Healthy achieve a flawless AUC of 1.00 as displayed in Figures 5b–d, respectively.

TABLE 7 Classification accuracy for different dementia classification tasks.

Classification task	Accuracy
Frontotemporal disease vs. healthy	0.9970
Alzheimer's disease vs. healthy	0.9974
Alzheimer + frontotemporal disease vs. healthy	0.9980
Alzheimer vs. frontotemporal vs. healthy	0.8034

4.3 Model performance evaluation with SMOTE balancing

It was noted in all the classification task that the dataset was imbalanced. To address this issue Smote data balancing technique were used. SMOTE balances datasets by generating new samples along the lines connecting a minority instance and its nearest within-class neighbors. Table 8 shows the classification metrics for Alzheimer, Frontotemporal, and Healthy classes after applying data balancing techniques. It can see a significant improvement in F1-score, precision, recall and specificity for all classes. Frontotemporal class got perfect scores in all metrics (1.00). The Alzheimer's class got good scores with precision 0.63, recall 0.71 and F1-score 0.67. Healthy class got precision 0.67, recall 0.58 and F1-score 0.62. Overall accuracy of the model decreased to 77.45% after balancing compared to 80.34% accuracy with the original imbalanced dataset.

The classification metrics for the Alzheimer's Disease and Healthy classes are shown in Table 9 after data balancing. The model ability to distinguish between the two classes is demonstrated by the precision, recall, and F1 scores of both classes, all of which are above 99.7%. Even though the balanced model's accuracy is 99.71%, it is only slightly lower than the unbalanced model's 99.74% accuracy.

TABLE 9 Classification metrics for Alzheimer's disease and healthy classes with data balancing.

Metric	Alzheimer's disease	Healthy
Precision	99.73	99.70
F1 score	99.71	99.73
Recall	99.70	99.71
Support	1876	1876

TABLE 10 K-fold validation accuracy for Alzheimer, frontotemporal, and healthy classes.

K-value	Training accuracy (%)	Test accuracy (%)
1	79.89	80.15
2	80.00	80.00
3	79.58	80.06
4	79.43	80.02
5	81.27	80.13

TABLE 8 Classification metrics for Alzheimer, frontotemporal, and healthy classes with data balancing.

Class	Precision	Recall	F1-score	Sensitivity	Specificity	Support
Alzheimer	0.63	0.71	0.67	0.71	0.79	1,876
Frontotemporal	1.00	1.00	1.00	1.00	1.00	1,876
Healthy	0.67	0.58	0.62	0.58	0.86	1,876

TABLE 11 K-fold validation accuracy for Alzheimer and healthy classes

K-value	Training accuracy (%)	Test accuracy (%)
1	99.82	99.86
2	99.80	99.82
3	99.73	99.92
4	99.61	99.86
5	99.78	99.82

TABLE 12 Classification metrics for Alzheimer, frontotemporal, and healthy classes.

Class	Precision	Recall	F1-score	Support
Alzheimer	0.60	0.77	0.67	1,876
Frontotemporal	0.68	0.68	0.68	1,597
Healthy	0.60	0.33	0.43	1,106

4.4 Evaluation of model accuracy using K-fold cross-validation

In this paper, a 5-fold cross-validation methodology was employed to validate the proposed model. The dataset was split into five subsets. For the multiclass classification task, the training accuracy ranged from a minimum of 79.43% to a maximum of 81.27% across different K values. The test accuracy remained consistently close to 80% for all folds, as shown in the Table 10. Table 11 shows the 5-fold cross-validation findings for differentiating between Alzheimer's and healthy patients. The test accuracy remains the same as in the training accuracy. These findings demonstrate the model's robust and reliable capacity to differentiate between the Alzheimer's and healthy classes.

4.5 Comparative analysis of feature extraction methods

In this evaluation, we compared the standard RBP with our modified RBP. The same methodology was used, but the frequency ranges were adjusted according to the standard: Delta (0.5–4), Theta (4–8), Alpha (8–13), Beta (13–25), and Gamma (25–45). The results achieved are shown in the Table 12. The standard RBP method achieved an accuracy of 63.03% in the multiclass classification task, whereas the modified RBP reached 80.34%. The precision for all classes remained almost the same; however, the recall and F1-score varied across the three classes. The Alzheimer class showed higher F1-score and recall values, whereas the Healthy class had lower values in these metrics. For the binary classification task, the Alzheimer and Healthy classes achieved an accuracy of 76.36% using the standard feature extraction method, whereas the modified feature extraction method achieved an accuracy of 99.71% as shown Table 13. Both classes showed lower recall, precision, and F1-scores with the standard method compared to the results obtained using the modified feature extraction method.

TABLE 13 Classification metrics for Alzheimer and healthy classes.

Class	Precision	Recall	F1-score	Support
Alzheimer	0.76	0.81	0.79	1,876
Healthy	0.76	0.71	0.73	1,597

TABLE 14 Model accuracy comparison with existing papers using dataset.

Paper	Model	Accuracy	Feature engineering	XAI
Ma et al. (20)	Support vector machine	91.5%	PHI	✗
Miltiadous et al. (18)	Dual-Input Convolution Encoder Network (DICE-net)	83.28%	Band power and coherence	✗
Kachare et al. (16)	STEADYNet	97.59%	✗	✗
Chen et al. (19)	Vision transformer + CNN	80.23%	frequency channels	✗
This work	**Proposed model**	**80.34%, 99.7%**	**Modified RBP**	✓

4.6 Comparison with existing ML and DL model

To gauge the performance of the proposed model, it has been compared with existing studies in Table 14. In Miltiadous et al. (18), the authors achieved an 83.28% accuracy with the DICE-net model, utilizing EEG denoising and extracting Band power and coherence features as key steps in feature engineering. In Kachare et al. (16), the STEADYNet model achieved 88.00% accuracy for AD vs. NC and 92.25% for FTD vs. NC. Using a dual-input strategy, the model employed convolutional and features are extracted from EEG data using max-pooling layers. The research explored binary and multi-class classification, reporting a 97.59% accuracy in the multi-class setting. The study (19) utilized a CNN with pre-trained weights, achieving an accuracy of 82.30%. EEG feature extraction was performed in both the time and frequency domains, while a Vision Transformer complemented the CNN by capturing global feature representations. The classification task distinguished between AD, FTD, and NC. Ma et al. (20), EEG data was used to classify AD and FTD, achieving an initial accuracy of 91.5%. After optimizing the feature set by eliminating unnecessary attributes, the accuracy increased to 96.6%. A support vector machine (SVM) model was utilized for binary classification between these groups (20).

No prior research utilized explainable AI (XAI) or lightweight models. To address this, the proposed study introduces a hybrid deep learning model with efficient feature engineering and a reduced number of parameters, improving accuracy in binary and multi-class classification while integrating SHAP.

5 Explainable artificial intelligence

Explainable Artificial Intelligence (XAI) is a crucial development in the field of artificial intelligence, focusing on

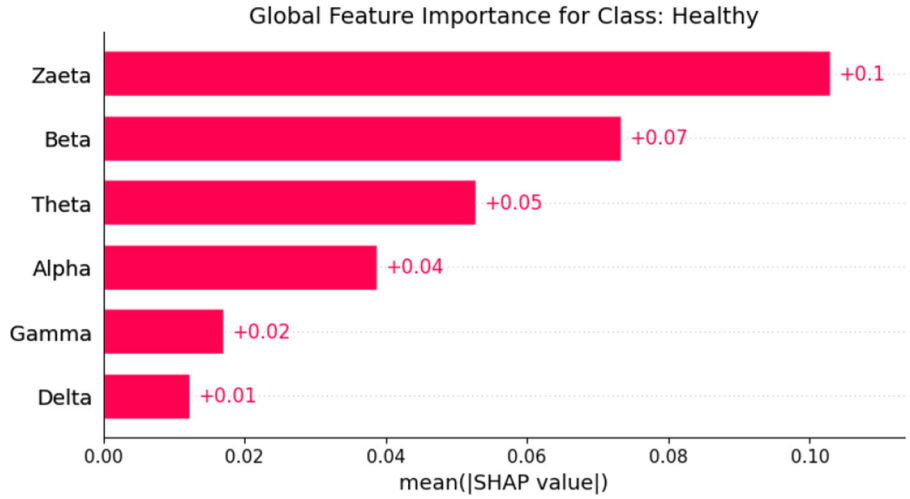


FIGURE 6
SHAP global feature importance graph for class healthy.

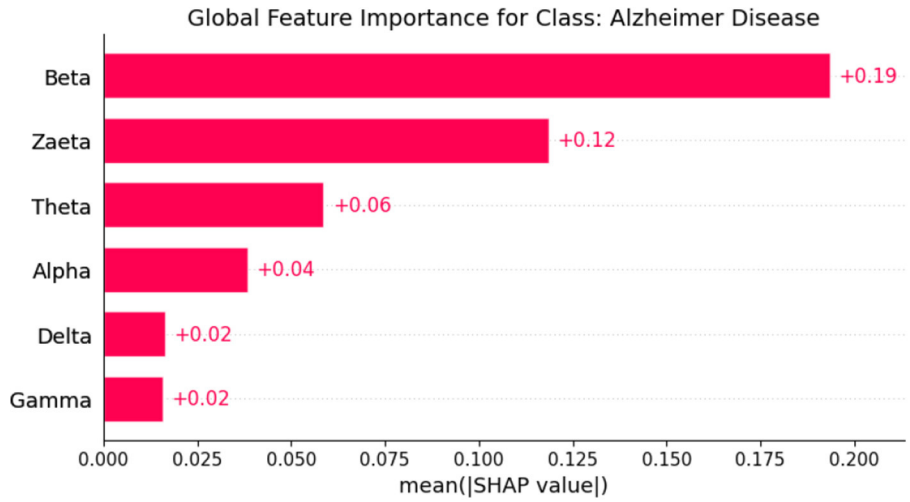


FIGURE 7
SHAP global feature importance graph for class Alzheimer's disease.

making the decision-making processes of AI systems transparent and understandable to users. In medicine, the need for XAI is particularly significant due to the high stakes involved in clinical decision-making. Healthcare professionals require clear explanations for AI-driven recommendations to ensure trust and reliability in these technologies. By improving interpretability, XAI not only helps clinicians approach AI methods with caution but also fosters a deeper understanding of AI applications in medical practice, ultimately promoting data-driven and mathematically grounded medical education (21). The SHAP (SHapley Additive exPlanations) (22) global feature importance graphs depict the contribution of different frequency bands (Zaeta, Beta, Theta, Alpha, Delta, Gamma) to the classification of three classes: Healthy, Alzheimer's Disease, and frontotemporal Disease. In Figure 6, the SHAP values for the Healthy class show that Zaeta has the highest importance (+0.1), followed by Beta (+0.07) and Theta (+0.05). This indicates these frequency bands are most influential in predicting Healthy cases, while Alpha, Gamma, and Delta have minimal contributions. Figure 7 highlights the SHAP importance for Alzheimer's Disease, where Beta exhibits the highest importance (+0.19), followed by Zaeta (+0.12) and Theta (+0.06). These results suggest that Beta and Zaeta bands play a critical role in distinguishing Alzheimer's Disease from other classes. In Figure 8, the SHAP values for the Frontotemporal Disease class demonstrate that Beta has the most significant influence (+0.26), with Zaeta being the second most important feature (+0.2). The other frequency bands, including Theta, Alpha, Delta, and Gamma, contribute very minimally to this classification.

The SHAP summary graphs explain the contributions of different features to the predictions of a proposed hybrid deep

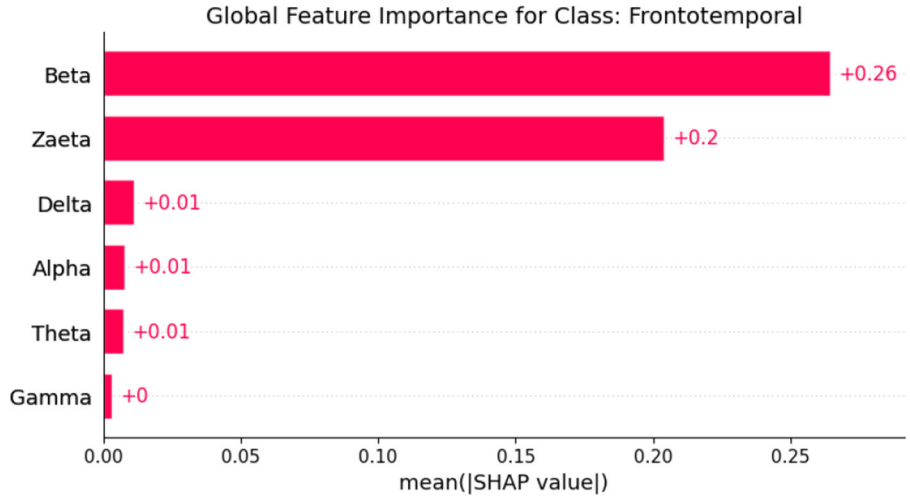


FIGURE 8
SHAP global feature importance graph for class frontotemporal disease.

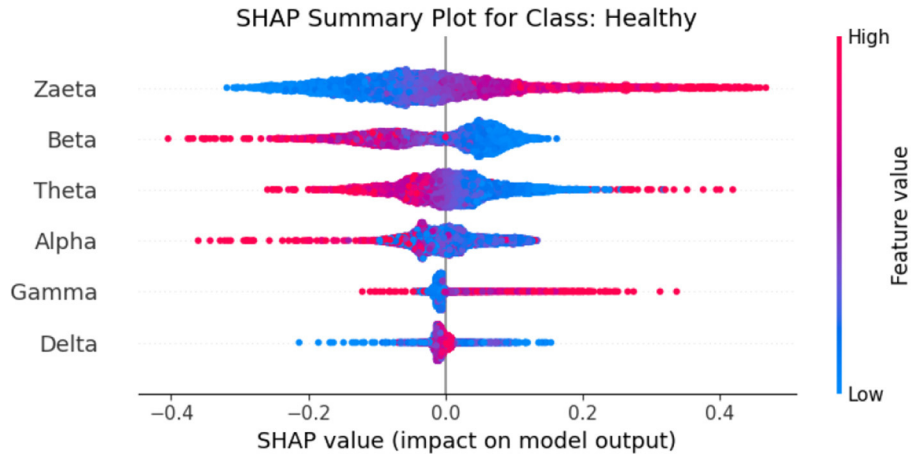


FIGURE 9
SHAP Summary graph for Class Healthy.

learning model for three different Alzheimer's disease, and frontotemporal disease. Each plot shows the impact of the features on the model's output. The x-axis represents the SHAP values, indicating whether a feature contributes positively or negatively to the prediction for a specific class.The Healthy class plot Figure 9 shows distinct feature behavior compared to the disease classes. Here, the SHAP values indicate a different pattern of influence, with Zaeta and Beta waves also playing critical roles but in opposite directions from the disease classes. For the Alzheimer's Disease class Figure 10, features such as Beta and Zaeta wave characteristics show a stronger positive or negative influence on predictions, with higher feature values (red points) generally pushing predictions in one direction. In this plot Figure 11, the Zaeta and Beta waves seem to have the most significant influence on the model's predictions, with both high and low feature values affecting the SHAP values. The distribution of points along the x-axis for these features suggests that they are crucial in determining whether the prediction aligns with frontotemporal disease.

The SHAP heat maps show how different brain wave features contribute to the model's predictions for healthly, Alzheimer's Disease, and Frontotemporal Disease. Each row represents a feature, while the columns represent individual instances. Each model input's global importance is shown as a bar plot on the plot's right side. Beta and Zaeta waves are among the features that commonly display blue in Healthy class Figure 12, suggesting that they have a negative impact on the prediction and force the model to classify these phenomena as healthy. On the other hand, beta and Zaeta waves frequently show red in the AD class Figure 13, indicating that they are highly predictive of Alzheimer's disease. How these features adjust to different data points is seen in the mixed pattern across instances. For Frontotemporal Disease Figure 14, Beta and Zaeta waves again dominate with

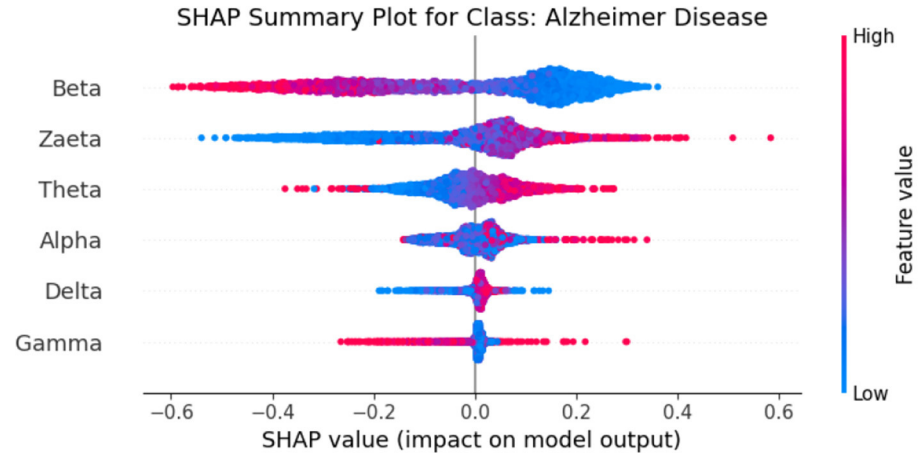


FIGURE 10
SHAP summary graph for class Alzheimer's Disease.

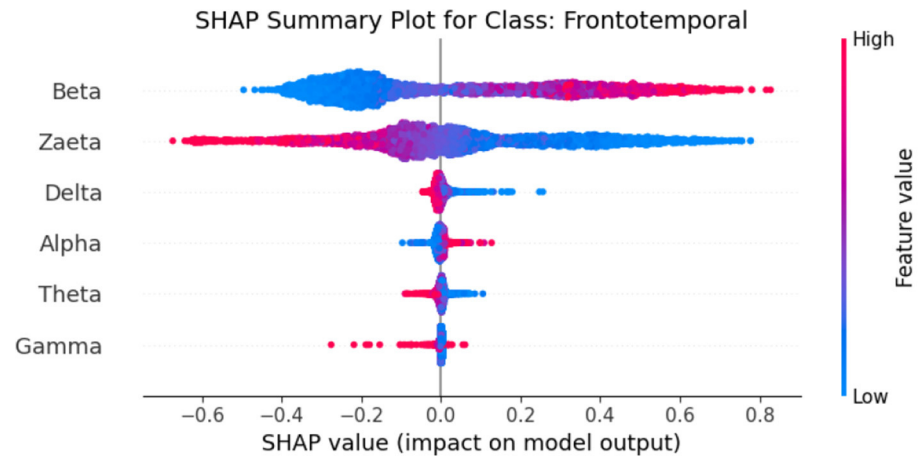


FIGURE 11
SHAP summary graph for class frontotemporal disease.

strong red contributions, emphasizing their importance for this class. Compared to Healthy, there are more concentrated positive contributions (red), pushing predictions toward the disease class. These heat maps reveal the nuanced role of brain wave features in distinguishing between healthy and diseased states.

5.1 Neurophysiological interpretation of frequency band importance

The SHAP visualizations in (Figures 6–14) reveal that the Zaeta and Beta frequency bands consistently exhibit high SHAP values across all classification tasks, indicating their dominant contribution in distinguishing between Alzheimer's Disease (AD), Frontotemporal Dementia (FTD), and healthy controls. This is not merely a data-driven outcome but has a neurophysiological basis grounded in clinical EEG studies.

The Beta band is associated with active cognitive processing, attention, and motor control. Abnormalities in Beta activity—particularly elevated or diminished power—have been reported in AD patients, often linked to disruptions in cognitive and executive functions. In contrast, FTD patients may exhibit distinct patterns in Beta activity due to altered frontal lobe functioning, which is characteristically impaired in FTD but less so in early AD.

The Zaeta band, though less commonly named in classical EEG literature, overlaps with the high Alpha to low Beta range and serves as a transitional band. Our modified Relative Band Power (RBP) analysis captures Zaeta as a distinct band, enabling finer differentiation. The elevated importance of Zaeta in our SHAP analysis suggests that subtle shifts in mid-frequency rhythms play a significant role in disease-specific EEG patterns. Specifically, such shifts may reflect compensatory mechanisms or region-specific slowing in cortical activity, both of which are documented phenomena in dementia-related neurodegeneration.

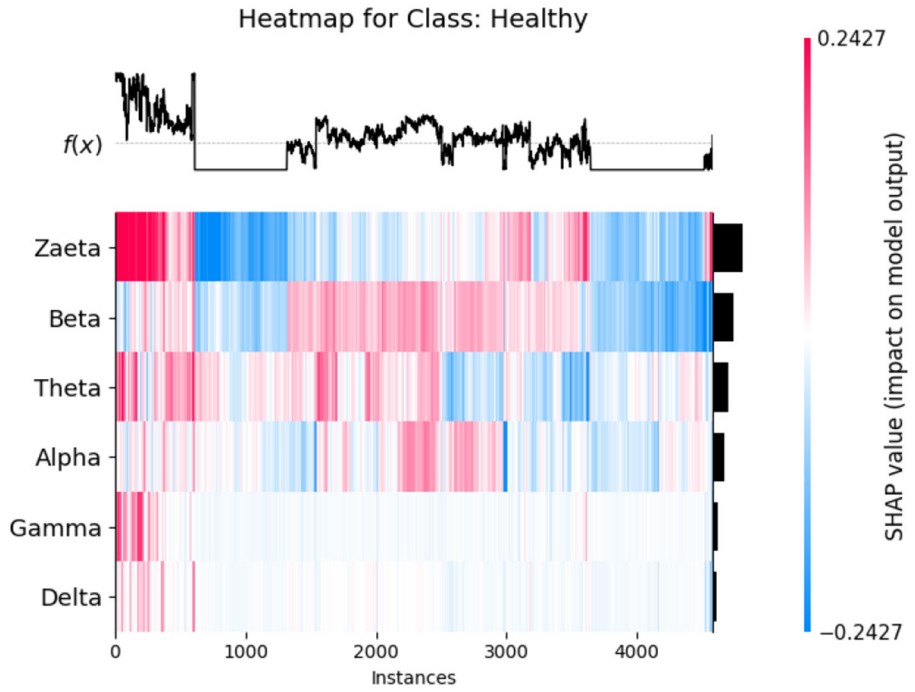


FIGURE 12
SHAP heatmap for class healthy.

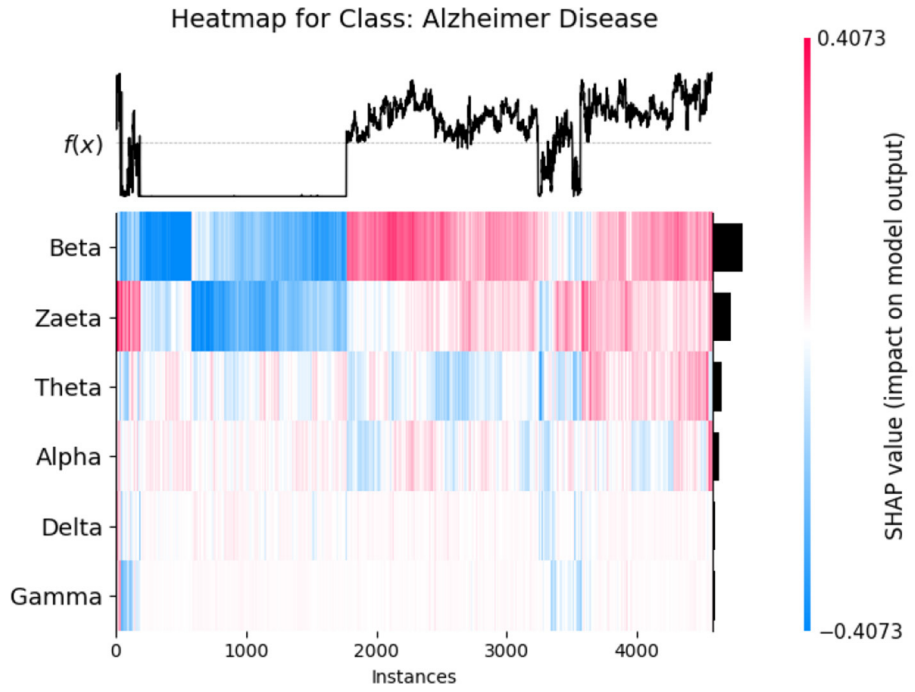


FIGURE 13
SHAP heatmap for class AD.

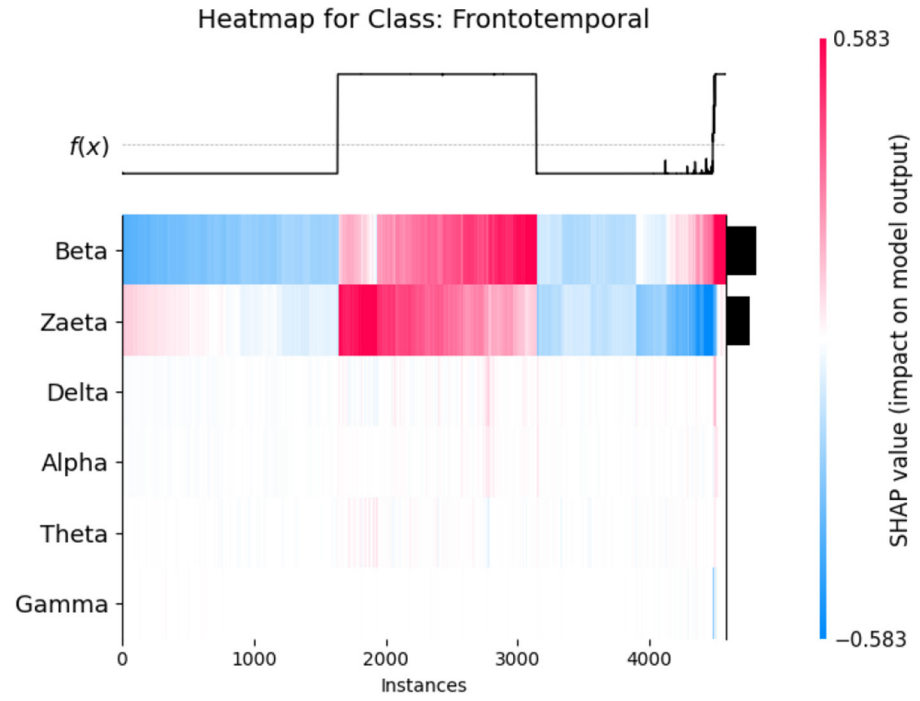


FIGURE 14
SHAP heatmap for class frontotemporal disease.

Therefore, the SHAP-derived feature dominance is consistent with known pathophysiological changes in brain activity across dementia subtypes. The model not only learns these discriminative patterns effectively but also explains them in a way that aligns with clinical neurophysiology, enhancing interpretability and potential clinical utility.

6 Conclusions and future direction

This paper addressed the critical need for an accurate and efficient detection of mental disorders, i.e., AD and (FTD). A lightweight TCN-LSTM hybrid model has been proposed for the aforementioned purpose. To prepare the data for experimentation, a modified Relative Band Power (RBP) analysis was performed to extract six EEG frequency bands via power spectrum density (PSD) computations. The proposed model achieved 99.70% accuracy for the classification of Frontotemporal Disease vs. Healthy, and 99.74% accuracy for Alzheimer vs. Healthy. In another binary task, where Alzheimer and Frontotemporal data were combined into a single class and classified against Healthy, the model achieved 99.80% accuracy. For the three-class classification, accuracy 80.34% achieved. Evaluation metrics including AUC-ROC, recall, confusion matrix, and F1-score were calculated for each classification. High scores were achieved across all multiclass categories, except the Healthy class, which showed reduced recall (35%) and F1-score (47%) as a result of data imbalance. Finally, the integration of SHAP for explainability further enhanced the model's transparency, making it a valuable tool for clinical applications. The proposed method proved to be an efficient and effective solution for the detection of AD and (FTD). Future research may include the use of large and diverse datasets focusing on the exploration of additional EEG characteristics Vascular, Lewy Body Dementia, and Creutzfeldt-Jakob Disease data can be used to train and validate the model with an XAI approach while maintaining patient data privacy and security.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author. The datasets analyzed and utilized for this study can be found at DOI: 10.3390/data8060095 and 10.18112/openneuro.ds004504.v1.0.5.

Ethics statement

The studies involving humans were approved by the Scientific and Ethics Committee of AHEPA University Hospital, Aristotle University of Thessaloniki, under protocol number 142/12-04-2023. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

WK: Methodology, Software, Writing – original draft. MSK: Conceptualization, Methodology, Supervision, Writing – original

draft. SQ: Formal analysis, Funding acquisition, Writing – review & editing. WG: Formal analysis, Project administration, Writing – review & editing. FS: Investigation, Validation, Writing – review & editing. MH: Funding acquisition, Investigation, Writing – review & editing. JA: Investigation, Validation, Writing – review & editing.

Funding

The author(s) declare that funding was received for the research and/or publication of this article. The authors extend their appreciation to the King Salman Center For Disability Research for funding this work through Research Group no KSRG-2024-430.

Acknowledgments

The authors extend their appreciation to the King Salman Center For Disability Research for funding this work through Research Group no KSRG-2024-430.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Bang J, Spina S, Miller BL. Frontotemporal dementia. *Lancet*. (2015) 386:1672–82. doi: 10.1016/S0140-6736(15)00461-4

2. Association A, Thies W, Bleiler L. 2013 Alzheimer's disease facts and figures. *Alzheimers Dement*. (2013) 9:208–45. doi: 10.1016/j.jalz.2013.02.003

3. Prince M, Wimo A, Guerchet M, Ali GC, Wu YT, Prina M. World Alzheimer Report 2015. *The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends*. London: Alzheimer's Disease International (2015).

4. Salmon DP, Bondi MW. Neuropsychological assessment of dementia. *Annu Rev Psychol*. (2009) 60:257–82. doi: 10.1146/annurev.psych.57.102904.190024

5. Elahi FM, Miller BL. A clinicopathological approach to the diagnosis of dementia. *Nat Rev Neurol*. (2017) 13:457–76. doi: 10.1038/nrneurol.2017.96

6. Robillard A. Clinical diagnosis of dementia. *Alzheimers Dement*. (2007) 3:292–8. doi: 10.1016/j.jalz.2007.08.002

7. Kivistö J, Soininen H, Pihlajamaki M. "Functional MRI in Alzheimer's disease." In: *Advanced Brain Neuroimaging Topics in Health and Disease-Methods and Applications*. Rijeka: IntechOpen (2014). doi: 10.5772/58264

8. Miltiadous A, Tzimourta KD, Afrantou T, Ioannidis P, Grigoriadis N, Tsalikakis DG, et al. A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*. (2023) 8:95. doi: 10.3390/data8060095

9. Singh AK, Krishnan S. Trends in EEG signal feature extraction applications. *Front Artif Intell*. (2023) 5:1072801. doi: 10.3389/frai.2022.1072801

10. Al-Fahoum AS, Al-Fraihat AA. Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains. *Int Sch Res Notices*. (2014) 2014:730218. doi: 10.1155/2014/730218

11. AlSharabi K, Salamah YB, Abdurraqeeb AM, Aljalal M, Alturki FA. EEG signal processing for Alzheimer's disorders using discrete wavelet transform and machine learning approaches. *IEEE Access*. (2022) 10:89781–97. doi: 10.1109/ACCESS.2022.3198988

12. Bi X, Wang H. Early Alzheimer's disease diagnosis based on EEG spectral images using deep learning. *Neural Netw*. (2019) 114:119–35. doi: 10.1016/j.neunet.2019.02.005

13. Pirrone D, Weitschek E, Di Paolo P, De Salvo S, De Cola MC. EEG signal processing and supervised machine learning to early diagnose Alzheimer's disease. *Appl Sci*. (2022) 12:5413. doi: 10.3390/app12115413

14. Dogan S, Baygin M, Tasci B, Loh HW, Barua PD, Tuncer T, et al. Primate brain pattern-based automated Alzheimer's disease detection model using EEG signals. *Cogn Neurodyn*. (2023) 17:647–59. doi: 10.1007/s11571-022-09859-2

15. Miltiadous A, Tzimourta KD, Giannakeas N, Tsipouras MG, Afrantou T, Ioannidis P, et al. Alzheimer's disease and frontotemporal dementia: a robust classification method of EEG signals and a comparison of validation methods. *Diagnostics*. (2021) 11:1437. doi: 10.3390/diagnostics11081437

16. Kachare PH, Sangle SB, Puri DV, Khubrani MM, Al-Shourbaji I. STEADYNet: spatiotemporal EEG analysis for dementia detection using convolutional neural network. *Cogn Neurodyn*. (2024) 18:1–14. doi: 10.1007/s11571-024-10153-6

17. Amer NS, Belhaouari SB. Exploring new horizons in neuroscience disease detection through innovative visual signal analysis. *Sci Rep*. (2024) 14:4217. doi: 10.1038/s41598-024-54416-y

18. Miltiadous A, Gionanidis E, Tzimourta KD, Giannakeas N, Tzallas AT. DICE-net: a novel convolution-transformer architecture for Alzheimer detection in EEG signals. *IEEE Access*. (2023) 11:71840–58. doi: 10.1109/ACCESS.2023.3294618

19. Chen Y, Wang H, Zhang D, Zhang L, Tao L. Multi-feature fusion learning for Alzheimer's disease prediction using EEG signals in resting state. *Front Neurosci*. (2023) 17:1272834. doi: 10.3389/fnins.2023.1272834

20. Ma Y, Bland JKS, Fujinami T. Classification of Alzheimer's disease and frontotemporal dementia using electroencephalography to quantify communication between electrode pairs. *Diagnostics*. (2024) 14:2189. doi: 10.3390/diagnostics14192189

21. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learn Syst*. (2020) 32:4793–813. doi: 10.1109/TNNLS.2020.3027314

22. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc. (2017) p. 4768–77.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 16 July 2025
DOI 10.3389/fmed.2025.1589587

OPEN ACCESS

EDITED BY
Ateeq Ur Rehman,
Gachon University, Republic of Korea

REVIEWED BY
Nguyen Quoc Khanh Le,
Taipei Medical University, Taiwan
Hafiz Asif,
Sultan Qaboos University, Oman

*CORRESPONDENCE
Jawad Rasheed
jawad.rasheed@izu.edu.tr

RECEIVED 07 March 2025
ACCEPTED 30 June 2025
PUBLISHED 16 July 2025

CITATION
Çevik N, Çevik T, Osman O, Alsubai S and Rasheed J (2025) Advancing patient care with AI: a unified framework for medical image segmentation using transfer learning and hybrid feature extraction.
Front. Med. 12:1589587.
doi: 10.3389/fmed.2025.1589587

Advancing patient care with AI: a unified framework for medical image segmentation using transfer learning and hybrid feature extraction

Nazife Çevik[1], Taner Çevik[2], Onur Osman[3], Shtwai Alsubai[4] and Jawad Rasheed[5,6,7]*

[1]Department of Computer Engineering, Istanbul Arel University, Istanbul, Türkiye, [2]Department of Computer Engineering, Istanbul Rumeli University, Istanbul, Türkiye, [3]Department of Electrical and Electronics Engineering, Engineering Faculty, Istanbul Topkapi University, Istanbul, Türkiye, [4]Department of Computer Science, College of Computer Engineering and Sciences in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia, [5]Department of Computer Engineering, Istanbul Sabahattin Zaim University, Istanbul, Türkiye, [6]Department of Software Engineering, Istanbul Nisantasi University, Istanbul, Türkiye, [7]Applied Science Research Center, Applied Science Private University, Amman, Jordan

Background: Accurate medical image segmentation significantly impacts patient outcomes, especially in diseases such as skin cancer, intestinal polyps, and brain tumors. While deep learning methods have shown promise, their performance often varies across datasets and modalities. Combining advanced segmentation techniques with traditional feature extraction approaches may enhance robustness and generalizability.

Objective: This study aims to develop an integrated framework combining segmentation, advanced feature extraction, and transfer learning to enhance segmentation accuracy across diverse medical imaging (MI) datasets, thus improving classification accuracy and generalization capabilities.

Methods: We employed independently trained U-Net models to segment skin cancer, polyps, and brain tumor regions from three separate MI datasets (HAM10000, Kvasir-SEG, and Figshare Brain Tumor dataset). Moreover, the study applied classical texture-based feature extraction methods, namely Local Binary Patterns (LBP) and Gray-Level Co-occurrence Matrix (GLCM), processing each Red Green Blue (RGB) channel separately using an offset [0 1] and recombining them to create comprehensive texture descriptors. These segmented images and extracted features were subsequently fine-tuned pre-trained transfer learning models. We also assessed the combined performance on an integrated dataset comprising all three modalities. Classification was performed using Support Vector Machines (SVM), and results were evaluated based on accuracy, recall (sensitivity), specificity, and the F-measure, alongside bias-variance analysis for model generalization capability.

Results: U-Net segmentation achieved high accuracy across datasets, with particularly notable results for polyps (98.00%) and brain tumors (99.66%). LBP consistently showed superior performance, especially in skin cancer and polyp datasets, achieving up to 98.80% accuracy. Transfer learning improved segmentation accuracy and generalizability, particularly evident in skin cancer (85.39%) and brain tumor (99.13%) datasets. When datasets were combined, the proposed methods achieved high generalization capability, with the U-Net

model achieving 95.20% accuracy. After segmenting the lesion regions using U-Net, LBP features were extracted and classified using an SVM model, achieving 99.22% classification accuracy on the combined dataset (skin, polyp, and brain).

Conclusion: Integrating deep learning-based segmentation (U-Net), classical feature extraction techniques (GLCM and LBP), and transfer learning significantly enhanced the accuracy and generalization capabilities across multiple MI datasets. The methodology provides robust, versatile framework applicable to various MI tasks, supporting advancements in diagnostic precision and clinical decision-making.

KEYWORDS

intestinal polyps, brain tumors, deep learning, local binary patterns, gray-level co-occurrence matrix

1 Introduction

The incidence of cancer worldwide has remained high in recent years. Additionally, each year, tens of millions of people receive a new cancer diagnosis. Meanwhile, different forms of cancer kill millions to almost tens of millions of people (1). According to the WHO, cancer will be the top cause of death globally in 2020, taking around 10 million lives (2). When it came to new cancer cases in 2020, the most prevalent were 2.26 million cases in the breast, 2.21 million in the lung, 1.93 million in the colon and rectum, 1.41 million in the prostate, 1.20 million in the skin (non-melanoma), and 1.09 million in the stomach. Pathology and imaging diagnostics are the primary methods used to diagnose cancer (3, 4). Increasing the survival percentage of cancer patients requires early detection (5), and effective and non-invasive early screening has emerged as a crucial study area. Magnetic resonance imaging (MRI), computed tomography (CT), X-rays, B-ultrasound, and others are examples of imaging techniques (6). Since an MRI scan can differentiate between different types of tissues, it can help spot cancer in different parts of the body (7). Medical image segmentation allows researchers and doctors to precisely identify and examine particular structures by dividing a medical image into discrete regions of interest. This segmentation procedure is important since thorough and precise evaluations are critical to patient care in radiology, pathology, and other medical specialties. Completing the regional segmentation's nodules and tracheal placement area is challenging (8). Screening and symptomatic disease management are the foundations of imaging's involvement in cancer management. Imaging will be used in cancer treatment in the future for targeted, minimally invasive, and pre-symptomatic treatments (9). Image guidance will be used to develop locally activated medication delivery and less invasive targeted therapy (10–14). Because tissue and fluids in the body absorb and scatter light, clinical optical imaging has mostly been restricted to endoscopic, catheter-based, and superficial imaging strategies. Since cancer is a complex disease, imaging must be able to show the many pathophysiological phases and mechanisms. Combining independent and uncorrelated imaging technologies will result in diagnostic orthogonality by employing diverse modalities, imaging agents, and biomarkers in general. Diagnostic imaging agents delivered intravenously, intra-arterially, or through natural orifices will become more prevalent in cancer imaging (15–17). Medical image segmentation aims to identify anatomical features in medical images, such as organs, lesions, tissues, etc. Many clinical applications depend on this basic phase, including computer-aided diagnosis, therapy planning, and illness progression tracking (18, 19). Precise segmentation can yield trustworthy target structure volumetric and morphological data, supporting numerous therapeutic uses such as quantitative analysis, surgical planning, and illness detection (20–22). Artificial intelligence (AI), particularly deep learning methods, has become a potent tool for improving and automating image segmentation in recent years. Medical image processing and analysis have seen tremendous success with deep learning algorithms, particularly Convolutional Neural Networks (CNNs), which provide quicker, more accurate, and repeatable results than manual techniques. Large annotated datasets can be used to train these models, enabling AI systems to identify intricate patterns and structures in medical images and provide accurate segmentation with little human assistance (23). CNN-based techniques can automatically extract the most valuable characteristics from massive datasets for medical segmentation. To improve diagnostic efficiency and make medical images more comprehensible, the initial and crucial stage in the analysis of medical images is medical image segmentation (24). To help doctors create more accurate diagnoses, we must segment the areas of medical images we focus on and extract pertinent features. This will give a solid foundation for clinical diagnosis and pathology research. Semantic segmentation, or the recognition of images at the pixel level, is typically referred to as image segmentation in deep learning. Semantic segmentation finds groups of pixels and categorizes them based on several attributes. Semantic segmentation research typically uses transfer learning. With transfer learning, a model already trained on a sizable dataset can be modified for a new job by teaching it to recognize general features. This is accomplished by retraining only the final layers of the model and freezing the other layers. As a result, the model retains the knowledge it gained from the prior task while adjusting to the inputs in the new one. Limited datasets and the inability to directly access current literature from another topic are two scenarios where transfer learning is used to help. Transfer learning has been effective in several applications, including text classification (25), satellite image segmentation (26), facial expression identification (27), and more.

Transfer learning offers an effective method to solve complex image analysis problems using the power of deep networks. However, classical feature extraction methods that can form the basis of transfer learning algorithms are also important in some cases. Traditional methods, such as the Gray-Level Co-Occurrence Matrix (GLCM) and the Local Binary Pattern (LBP), can create meaningful

inputs for transfer learning models or provide complementary information in fine-tuning the models. Thus, combining classical and modern techniques allows obtaining powerful results, especially in limited datasets. In this context, GLCM and LBP are two approaches that stand out from traditional image processing techniques. GLCM is a method that models the spatial relationships of pixel pairs at grayscale levels to examine the textural properties of an image. The use of GLCM features in medical image analysis has rapidly expanded in recent years. Examples include the analysis of MRI and ultrasound images of the liver (28, 29), the heart (30), X-ray mammography (31, 32), breast cancer (33, 34), prostate cancer (35–37), and brain cancer (38–40). Haralick et al. (41) proposed a general process for determining the textural characteristics of image blocks. The texture's statistical nature is considered while calculating features in the spatial domain. Mall et al. (42) used machine learning techniques to divide the MURA (musculoskeletal radiographs) dataset's bone X-ray images into two categories: those with fractures and those without GLCM features. In the study proposed by Pooja et al. (43), GLCM, LBP, and the Histogram of Oriented Gradient (HOG) are used for feature extraction. The correlation filter method and wrapper-based techniques detect and categorize polyps. On the other hand, LBP creates a histogram by evaluating the intensity differences between neighboring pixels to capture local textural information. During the feature extraction, Shamna and Musthafa (44) suggested HoG and Local Ternary Pattern (LTP). Additionally, the Deep Convolutional Neural Network (D-CNN) was used to fuse the gathered features before forwarding them to the Region-based Convolutional Neural Network to detect many objects. Bhattarai et al. (45) suggested an unsupervised approach to create the pseudo-labels employing HOGs. They learned the deep network's parameters to minimize the loss of the primary and auxiliary tasks, using pseudo-labels for the auxiliary task and ground truth semantic segmentation masks. The study by (46) extracts the dynamic texture elements of 3D MRI brain images using HOG features to detect Alzheimer's disease. Another approach proposed a model that uses neural characteristics from MRI images based on HOG to detect brain malignancies (47).

The application of techniques like transfer learning and deep learning in the field of medical image analysis has grown dramatically in recent years. A crucial factor that directly impacts the effectiveness of treatment for many conditions is early identification and accurate classification, particularly for skin cancer, intestinal polyps, and brain tumors. Accurate and precise segmentation is crucial in these imaging difficulties to enhance clinical procedures and improve patient outcomes. However, most current approaches lack generalizability and concentrate on a specific dataset or a restricted feature extraction technique. By working with several datasets and combining transfer learning and sophisticated feature extraction methods, our goal in this study was to improve segmentation performance. In the literature, various medical imaging issues—such as brain tumors, polyps, and skin cancer—are typically treated independently and with diverse techniques. However, this study aims to illustrate how the created technology may be used in various medical imaging situations and to provide a bridge between them. Although the suggested method successfully applies the transfer learning approach to the information transfer of pre-trained models, it combines deep features with statistical approaches, such as GLCM and LBP, as feature extraction techniques to produce more discriminative and meaningful features. This novel combination is anticipated to be highly generalizable to other medical imaging issues. The main contributions of this study are:

- A generalizable method for multiple medical imaging problems is proposed.
- It has been shown that combining transfer learning and classical feature extraction techniques can improve segmentation performance.
- The generalization capacity of the developed model was tested on different datasets.

This article introduces a potential approach for segmenting brain tumors, skin cancer, and polyps to provide a different perspective. Several pre-trained deep learning models, including VGG16, have been tested on various medical datasets, including brain tumors, polyps, and skin cancer. This offers a thorough examination to assess the methodologies' ability to generalize. Deep learning-based segmentation techniques were used with GLCM and LBP to produce feature sets that were more potent and discriminative. It has been demonstrated that this combination enhances post-segmentation classification performance. This study assessed the overall performance of the suggested approaches using datasets gathered from various anatomical locations and imaging techniques, in contrast to studies in the literature that are often carried out on a single dataset. The suggested method offers integrity in both segmentation and post-segmentation classification performance. Accuracy and time savings are benefits of this functionality, particularly in therapeutic settings. A broad framework that can be applied to clinical diagnosis is suggested by using the same approach to other imaging issues, such as brain tumors, intestinal polyps, and skin cancer.

Rather than proposing a new algorithm, our objective is to design a modular and generalizable pipeline using established techniques (U-Net, LBP, GLCM, and VGG16) to facilitate practical and accurate medical image analysis across diverse domains. Recently, the studies by (48, 49) explored hybrid methods combining segmentation and handcrafted features in biomedical image analysis. Thus, our framework expands on this by integrating these elements into a unified system applicable across multiple datasets.

The remainder of this article is organized as follows. Section 2 details the methodology, including a description of the datasets, the segmentation methods (using U-Net and transfer learning-based approaches), the feature extraction techniques (GLCM and LBP), and the classification strategy employed. In Section 3, we present experimental results, providing quantitative segmentation performance metrics for each dataset (skin cancer, polyps, and brain tumors) and for a combined dataset to evaluate generalization capabilities. Section 4 offers an in-depth discussion of the findings, highlighting the impact of different feature extraction methods, the role of transfer learning, and our approach's strengths and limitations. Finally, Section 5 concludes the article by summarizing our contributions, discussing potential limitations, and suggesting directions for future research.

2 Methodology

Rather than proposing a new algorithm, our objective is to design a modular and generalizable pipeline using established techniques for practical medical image analysis.

2.1 Dataset

This study examined a variety of datasets and concentrated on the segmentation of brain tumors, intestinal polyps, and skin cancer. Every dataset was chosen from well-used sources within the pertinent problem domain, and thorough pretreatment procedures were used. The following is a summary of the features of the datasets that were used:

Open-source databases are often used in the literature, and unique datasets gathered as part of specific studies comprise the datasets utilized. Each dataset underwent a thorough examination considering the overall number of samples and the image resolution. To increase segmentation accuracy, masks with images are manually or automatically labeled. Skin Cancer: The HAM10000 database is used to study skin cancer (50). Since the segmentation masks provided by (50) were absent from the original HAM10000 dataset, we used the source data generated by (50). The Figshare Brain Tumor dataset (51) is used for brain tumor segmentation and contains 3,064 pairs of MRI brain images and their mask indicators. In contrast, the Kvasir-SEG database, which includes 1,000 polyp images and the corresponding ground truth from the Kvasir Dataset v2 (52), is used for intestinal polyps. The total number of samples is shown in Table 1.

Since the images in the dataset of this study varied in size and dynamic range, it was unsuitable for direct model training. Resizing and normalization procedures were implemented to give the dataset a uniform structure. To match image proportions with the model input, all photos were scaled to 128 × 128. This procedure provided data resized to match the input dimensions required by the network, while optimizing the training process's computational cost. Additionally, images' pixel values typically range from 0 to 255. The normalization technique guaranteed faster convergence and kept the model from struggling to learn the significant disparities between these values. To get all pixel values in the range of 0–1, they are divided by 255. This procedure improved learning stability and allowed the model to assign equal weight to each image. These two preprocessing processes improved the model's performance during training by guaranteeing that the dataset had a more uniform structure.

Since the public datasets lack detailed metadata about acquisition centers or clinical environments, we did not perform external validation. Training and testing were carried out within each dataset. Cross-dataset or multi-institutional generalization is left for future investigation.

To ensure a fair evaluation and avoid data leakage, 10% of the training set was used as a validation set for hyperparameter tuning. The test set was not accessed during training or parameter optimization. Key hyperparameters (such as learning rate, batch size, and number of epochs) were selected based on performance on the validation set. No test data was used during model selection or tuning.

TABLE 1 Total number of samples in the dataset.

Dataset	Number of samples *(%80 for training, %20 for test)*
Polyp	1,000 (128 × 128)
Skin cancer	10,015 (128 × 128)
Brain tumor	3,064 (128 × 128)

2.2 Segmentation method

The U-Net model and the transfer learning-based VGG16 model were the two approaches for image segmentation that were compared in this study. The U-Net model, a convolutional neural network (CNN) structure designed specifically for segmentation challenges, was employed. U-Net, a semantic segmentation technique, was initially proposed for medical image segmentation. Ronneberger et al. (53) debuted U-Net. U-Net's encoder-decoder architecture is symmetric. The decoder part creates a segmentation mask in the original dimensions using the information taken from the image by the encoder part. The U-Net model was selected because it can learn the details of segmentation masks with high accuracy and generate respectable results even with small datasets. However, the shortcomings of the U-Net model, such as the need for large datasets and the lengthy learning process, are only considered when the model is built from the ground up. As a result, the transfer learning approach was used in the study's second phase. VGG16, a pre-trained model, was employed in the transfer learning stage. Being a deep network trained on huge datasets (like ImageNet), VGG16 is adept at picking up low-level characteristics (such as edges and textures). To generate a segmentation mask, a decoder section modeled after the U-Net model was added to the encoder portion of the VGG16 model, which was used to extract features from images. This structure made better performance with less data possible, which also speed up the training process through transfer learning.

The parameters of 15 epochs and a batch size of 16 utilized for the training procedure were chosen to balance the model's performance and training duration. Using the epoch number, 15 was selected as the number of times the model will be trained on all the training data. An adequate learning process is typically achieved by running through the data 15 times during training, especially for small datasets. Choosing too many epochs can lead to overfitting when the model performs well on training data but poorly on new data. The batch size, which is 16, is the quantity of data input into the model concurrently during each training phase. A batch size of 16 ensures training uniformity and optimizes processing time. With a smaller batch, the model can update its parameters more often but may consume more memory. A batch size of 32 is frequently used in various machine-learning situations and is usually a well-rounded choice. The model's complexity and the amount of data were considered when choosing the study's parameters. For example, while working with 128 × 128 images, a large batch size number slows the training process—batch size 16 improved memory management. The epoch number 15 was selected to ensure that the model reaches a point during training where accuracy and loss values may stabilize.

While resizing may risk losing important structural details, especially in fine-grained segmentation tasks, we selected 128 × 128 resolution to balance accuracy and computational efficiency. To assess potential performance loss, a subset of polyp and skin images was also resized to 256 × 256, and models were retrained. The difference in accuracy was below 1.2% on average, while computational requirements increased notably. Therefore, we proceeded with a 128 × 128 resolution for all datasets.

2.3 Feature extraction

This study's skin, polyp, and brain datasets sustained different segmentation processes before the GLCM and LBP techniques, unique to each dataset, were used. Segmentation was done using a different U-Net model for every dataset. The goal is to identify various structures in every dataset in a more precise way.

For the GLCM analysis, offset [0 1] was used. The distance and angle that specify the relationship between pixels are referred to as this parameter. After being retrieved independently, the red, green, and blue channels were merged and examined as a single image. Each image's texture characteristics were extracted using pixel points and radius values for the LBP approach. RGB channels are processed independently and then mixed, as in GLCM. Transfer learning techniques were performed on each dataset independently based on the segmentation outcomes. As a result, GLCM, LBP, and segmentation model performances were contrasted.

In the last step, all datasets were merged to produce a larger and more varied data collection. The following techniques were used successively on this combined dataset: LBP (by separating RGB channels), GLCM [offset (0 1)], and segmentation (U-Net). This procedure was carried out to assess how well the methodologies applied to various datasets.

Texture-based feature extraction techniques such as GLCM and LBP have been employed, particularly in the textural analysis of sections following segmentation. These techniques included modeling textural changes between datasets, classifying the areas produced after segmentation, and integrating with transfer learning models to improve segmentation accuracy. To assess the methods' generalizability, the analyses carried out independently for each dataset were finished using the combined dataset; consequently, a thorough comparison of the models' and methodologies' performances was made.

Feature-level fusion was implemented by concatenating deep features from CNNs and handcrafted features (GLCM and LBP) after extraction. No joint training or architectural integration was performed. This separation allows for interpretability but limits end-to-end learning potential.

Segmentation performance was evaluated using Dice coefficient, IoU (Intersection over Union), accuracy, recall, and specificity. Dice and IoU are especially suited for pixel-wise overlap assessment and are widely accepted in biomedical segmentation tasks.

2.4 Classification

In this study, the Support Vector Machines (SVMs) algorithm was preferred to classify the image data after the completed segmentation process. SVM is a method known for its high accuracy rates and generalization abilities and is a frequently used technique, especially in classification problems. The classification process was started using the features obtained from segmentation (such as GLCM and LBP). The features extracted after segmentation were used as input data to the SVM algorithm. We used an SVM classifier due to its proven reliability in handling small feature vectors and its ability to integrate heterogeneous features. However, we recognize that end-to-end deep learning classifiers such as fully connected neural networks or attention-based modules could offer better performance and are considered for future work. SVM works with appropriate kernel functions to create linear or non-linear separation regions. This study used the RBF (Radial Basis Function) kernel function depending on the data distribution. The model was optimized on the training dataset, and its performance was evaluated on the test dataset.

- Accuracy: It served as a fundamental performance metric by computing the proportion of samples the model properly classified among all samples. However, when there is an imbalance between classes, precision is insufficient.
- F-Measure: Calculated as the harmonic mean of the Precision and Recall measures, this metric was intended to show the model's success in both positive and negative classes and to assess the classification performance in a balanced manner.
- Bias-Variance Composition: The model's generalization performance was assessed using bias-variance analysis. The mistake happens when the model cannot comprehend the intricate structure present in the training data. Excessive bias causes oversimplification and impairs the model's accuracy. The bias component indicates the average accuracy of the model across all possible training sets. The variance component indicates how responsive the learning algorithm is to minor modifications in the training set (54).
- Variance: a circumstance in which the model performs poorly on the test data because it has learned too much from the training data. A high variance indicates an overfitting issue.

A thorough assessment of the classification algorithm's accuracy and generalizability was made possible by complementing performance measures. Bias-variance analysis was essential in comprehending the trade-off between the model's accuracy and generalization performance, even though the F-measure lessens the effect of class imbalances. This thorough assessment sought to improve the model's generalization ability and achieve high classification accuracy. Consequently, the SVM algorithm's classification following segmentation was assessed using carefully chosen metrics, and relevant analyses were conducted to maximize the model's overall performance. This method improved the dependability and efficiency of the categorization process.

First of all, GLCM and LBP feature extraction was done separately for all skin, polyp, and brain tumor datasets, and they are shown in their original form in Figures 1–3. We examined the textural relationships in the image and determined the spatial correlations between pixels in specific orientations (0° in our case) by extracting GLCM features. We evaluated the intensity differences between pixels and their neighbors to analyze the image's microtextures using LBP feature extraction. We specifically looked at the surface textures of skin lesions and polyps.

The overall workflow of the proposed segmentation and classification framework is illustrated in Figure 4. It includes stages, such as image preprocessing (resizing and normalization), segmentation using U-Net or VGG16-based transfer learning, feature extraction using LBP and GLCM, and final classification using SVM. This schematic is provided to enhance understanding of the integration of traditional and deep learning methods.

2.5 Data augmentation strategy

To improve the model's generalization and reduce overfitting, several augmentation techniques were applied during training. These

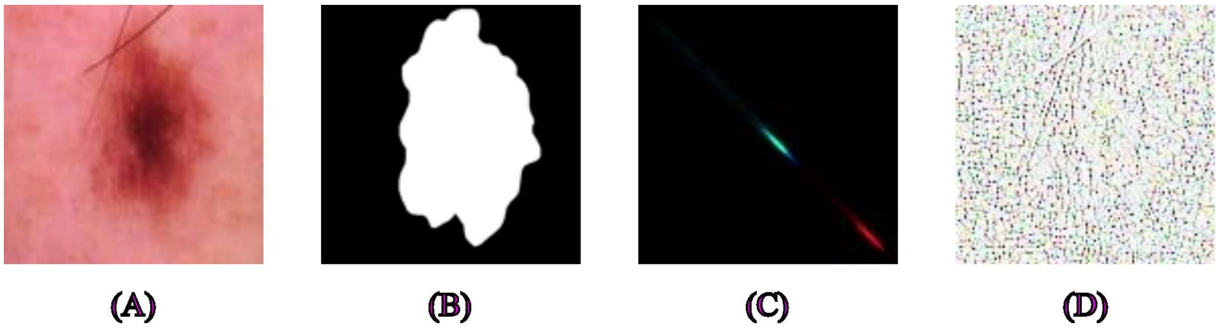


FIGURE 1
Examples of segmentation and feature extraction on skin cancer images. **(A)** Original skin lesion image from the HAM10000 dataset, **(B)** ground truth segmentation mask, **(C)** corresponding texture-enhanced image obtained by applying Gray-Level Co-occurrence Matrix feature extraction, highlighting spatial relationships between pixels, **(D)** Local Binary Pattern extracted features emphasizing detailed local textural patterns relevant to skin lesion characterization.

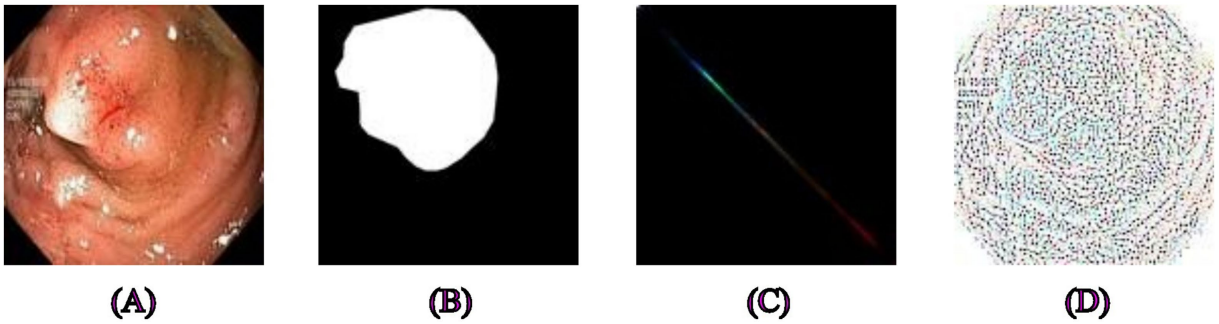


FIGURE 2
Examples of segmentation and feature extraction on polyp images. **(A)** Original polyp images from the Kvasir-SEG dataset, **(B)** the corresponding segmentation masks. **(C)** Image after applying Gray-Level Co-occurrence Matrix feature extraction, emphasizing textures critical for distinguishing polyps from surrounding tissues, **(D)** Local Binary Pattern-extracted image highlighting local intensity variations that provide robust texture descriptors for precise segmentation.

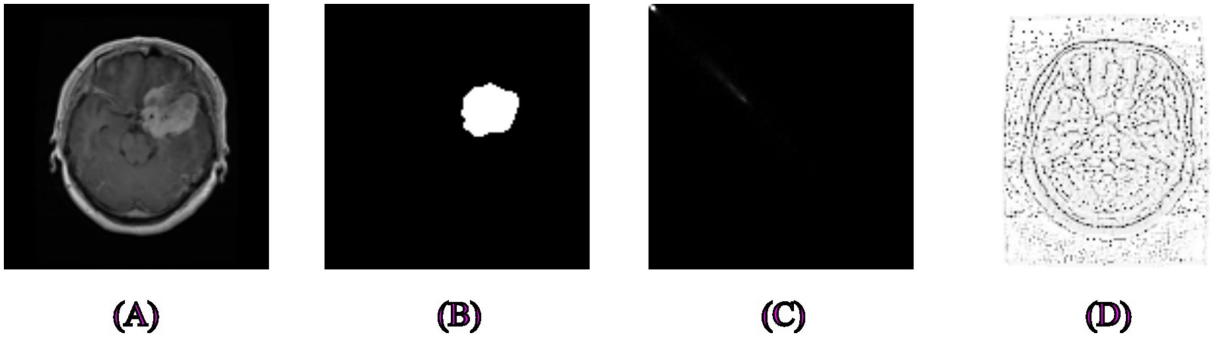


FIGURE 3
Examples of segmentation and feature extraction on brain tumor MRI images. **(A)** Original brain MRI images from the Figshare dataset, **(B)** Associated ground truth segmentation masks, **(C)** Image processed using Gray-Level Co-occurrence Matrix capturing texture variations to differentiate tumor tissues effectively, **(D)** Local Binary Pattern-extracted image showcasing local texture differences crucial for accurate brain tumor delineation.

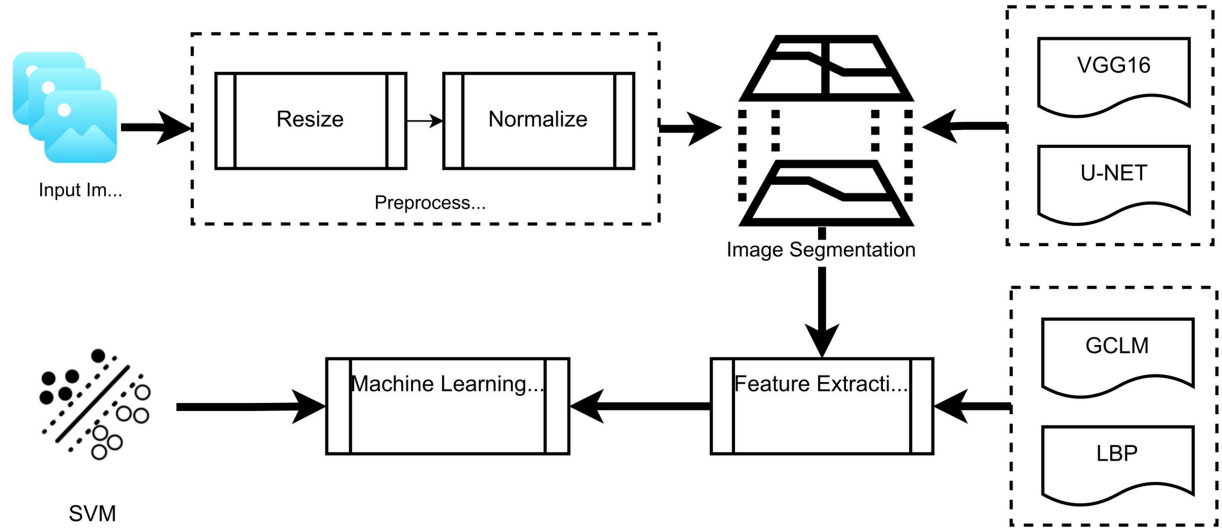


FIGURE 4
Schematic overview of the proposed framework: From image preprocessing through segmentation (U-Net/VGG16), followed by feature extraction (LBP and GLCM), and final classification using SVM.

transformations were randomly applied to each training image during every epoch, using a stochastic pipeline. The following techniques were employed:

- Rotation: Randomly rotating images within a ± 20° range.
- Flipping: Random horizontal and vertical flips.
- Zooming: Scaling the image randomly within a factor of 0.8 to 1.2.
- Translation: Shifting images up to 10% along both axes.
- Brightness/Contrast Adjustment: Slight variations were applied to mimic acquisition differences.

These augmentations increase the diversity of the training data, making the model more robust to variation in position, illumination, and shape. The augmentation was applied on-the-fly during training using stochastic transformations, ensuring that each epoch was exposed to new variations.

The datasets vary significantly in size (e.g., skin: 10,015 vs. polyp: 1,000). To mitigate imbalance and overfitting, we applied data augmentation techniques such as random flipping (horizontal/vertical), rotation, and scaling. These were applied more extensively to smaller datasets to increase effective training diversity.

2.6 Bias and variance estimation

To assess the generalization performance of the models, we estimated bias and variance using ensemble-based approximations over multiple runs ($n = 5$). The formulation is as follows (55):

Let y_i be the true label of the i^{th} instance, and let $\hat{y}_i^{(j)}$ denote the predicted output of the model in the j^{th} run. Then,

- Bias measures the average squared difference between the mean prediction and the ground truth:

$$Bias = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n}\sum_{j=1}^{n}\hat{y}_i^{(j)} - y_i\right)^2$$

- Variance quantifies the variability of the predictions across different runs:

$$Variance = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\hat{y}_i^{(j)} - \overline{\hat{y}_i}\right)^2\right)$$

where $\overline{\hat{y}_i} = \frac{1}{n}\sum_{j=1}^{n}\hat{y}_i^{(j)}$ is the mean prediction for instance i, and N is the total number of test samples.

These values were normalized and reported as percentages for easier interpretability. The bias and variance scores provided in the results section (e.g., 11.33 and 11.28%) reflect the model's trade-off between accuracy and stability.

2.7 Computational setup and timing

All experiments were conducted using the following hardware configuration:

- Processor: Intel Core i7-12700H @ 2.30GHz
- GPU: NVIDIA RTX 3060 Laptop GPU (6 GB VRAM)
- RAM: 32 GB DDR4

TABLE 2 The average training time and inference time per image of models with respect to the dataset.

Model	Dataset	Training time (m)	Inference time per image (ms)
U-Net	Polyp	~14	~22
VGG16	Skin cancer	~21	~28
U-Net	Brain tumor	~19	~24

- Operating System: Windows 11 Pro, MATLAB R2023a with Deep Learning Toolbox

The average training time per model is approximately listed in Table 2. Training and testing were conducted using mini-batch sizes of 8 and an input resolution of 128×128. Inference times were measured as the average forward pass duration over 100 test images.

3 Experimental results

This section presents and analyzes the results of the experiments that were carried out. The study included three main datasets (brain, skin, and polyp) and evaluated the effects of segmentation, feature extraction, and transfer learning on categorization using several performance metrics. Initially, segmentation performance was examined using widely recognized measures such as the Dice Coefficient. Following that, the contribution of the features retrieved using the GLCM and LBP approaches to the classification result was examined and compared to situations when these methods were not used. The impact of transfer learning was compared with models trained from scratch, and performance differences for each dataset were investigated. Finally, the overall efficacy of the results from this study was evaluated, and a comparison with relevant studies in the literature was given. Under each heading, a thorough analysis of the results will be provided. Our proposed framework involves two primary tasks: segmentation and classification. First, the lesion area is segmented using U-Net. Then, texture-based features (e.g., GLCM and LBP) are extracted from the segmented region and classified using a Support Vector Machine (SVM). Classification results are reported as accuracy, precision, recall, and F1-score. Segmentation quality is evaluated using Dice metrics.

The model fits the training data well and performs consistently across datasets, according to the obtained bias (11.33%) and variance (11.28%) values. Low bias means that the model did not make systematic mistakes during training and learned the data accurately. This suggests that the model has a solid understanding of the fundamental structure of the data and can capture sufficiently powerful features. Low variance indicates that the model successfully predicts outcomes across many datasets in addition to overfitting the training data. This suggests that the model has a strong capacity for generalization.

The model's performance was balanced between variance and bias. Therefore, neither overfitting nor underfitting is an issue. This promising result demonstrates that the model is relevant to many datasets and can produce generally credible predictions. To validate the effectiveness of our VGG16-based segmentation architecture, we further compared it with other state-of-the-art backbone networks, including ResNet50 and EfficientNetB0. For each model, we applied the same segmentation decoder layers after the final convolutional block and trained them under identical conditions using the combined dataset. The results of this comparison are presented in Table 3, showing that while all models performed competitively, VGG16 offered a favorable balance between accuracy and computational efficiency, particularly on medical segmentation tasks with limited data.

TABLE 3 VGG-16-based segmentation performance.

Backbone model	Accuracy (%)	F1-Score (%)	AUC	Param (M)
VGG16 + Decoder	86.21	85.42	0.9201	14.7
ResNet50 + Decoder	86.94	86.15	0.9264	23.5
EfficientNetB0 + Decoder	87.48	86.79	0.9297	5.3

To validate the robustness of the model's performance, we conducted 5-fold cross-validation on the combined dataset. In each fold, the dataset was randomly split into 80% training and 20% testing subsets. We repeated this process five times using distinct random seeds and reported the mean ± standard deviation for key performance metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC. The cross-validation results are summarized in Table 4. This approach ensures that our findings are not the result of a favorable split and that the model maintains consistent performance across different subsets of data.

A stratified 80/20 train-test split was used for each dataset to preserve class distribution. Each experiment was repeated five times with different random seeds. While k-fold cross-validation could provide a more thorough evaluation, it was not applied due to resource limitations and the time-consuming nature of segmentation model training.

3.1 Segmentation performance on polyp dataset

Learning rate–0.001, maxEpoch–15, and mini-batch size–16 are used for model training. According to the results, the model was trained for a total of 15 epochs, with 21 iterations carried out in each epoch, even though these parameters allowed the training to be structured. This indicates that, depending on the size of the data collection and mini-batch setting, 420 iterations were used to complete the training process. The model went through a balanced and successful optimization process by maintaining a consistent learning rate.

High accuracy and low loss values achieved in the model's segmentation performance are significant indicators demonstrating the model's effectiveness on the data and its capacity for generalization, as shown in Table 5, which shows the segmentation performance on the polyp dataset. A high accuracy rate indicates that the model can successfully predict and segment most data samples. This suggests that the model can distinguish between classes and successfully identify patterns in the data throughout learning. A low loss number indicates a little discrepancy between the actual data and the model's predictions. This shows that hyperparameters such as the learning rate were chosen correctly and that the model was trained successfully during

TABLE 4 The mean ± standard deviation for key performance metrics.

Metric	Mean ± standard deviation
Accuracy	0.8621 ± 0.0134
Precision	0.8702 ± 0.0151
Recall	0.8594 ± 0.0147
F1-score	0.8647 ± 0.0141
RoC – AUC	0.9263 ± 0.0118

optimization. Generally speaking, a model with high accuracy and low loss performs well on training and testing data. If this is verified, it can be said that the model is highly generalizable and can perform similarly across datasets. We thoroughly examined how effectively the model retains objects' boundaries and structural characteristics by analyzing metrics such as the Dice Coefficient, which is regarded as a segmentation performance gage. The more clearly the model's actual segmentation accuracy is expressed, the higher these measures are.

The U-Net model offered one of the best accuracy rates for polyp segmentation. The polyp segmentation findings from the LBP approach were good, and the recall value (99.49%) was nearly flawless. Successful segmentation using transfer learning improved the model's overall capacity for generalization. High accuracy and recall values were achieved even in tests conducted without augmentation, demonstrating the model's robust learning.

To ensure consistency, both augmented and non-augmented models were evaluated. The non-augmented U-Net model performed slightly better with 98.00% accuracy compared to 95.00% when augmentation was applied. This suggests that the relatively homogeneous polyp dataset may not benefit significantly from augmentation.

3.2 Segmentation performance on skin dataset

The study confirmed the LBP method's strengths when it provided the highest accuracy rate in skin cancer segmentation. Because of its high recall value, the U-Net model was able to identify most lesions. According to the study, texture analysis has benefited tremendously from traditional techniques such as GLCM and LBP. Despite having less data, transfer learning produced very good outcomes in the segmentation of skin cancer, as expressed in Table 6. Both augmented and non-augmented results for U-Net were compared. Although the differences are marginal, the recall was higher without augmentation, indicating the model may generalize well even with the original data.

3.3 Segmentation performance of brain tumor dataset

The U-Net model acquired a very high accuracy rate of 99.66% in brain tumor segmentation. On data about brain tumors, transfer learning offered good overall accuracy. Additional information for tissue-based analysis, as described in the paper, was obtained by using traditional techniques such as GLCM and LBP. Table 7 shows the results obtained on the Brain Tumor dataset. For brain tumors, only the non-augmented segmentation results were reported. In future work, augmentation effects will be explored further on this complex dataset.

3.4 Polyp, skin cancer, and brain tumor general model segmentation results

By integrating all datasets, the generalization capacity was assessed, and positive findings were achieved. With 95.20% accuracy, the U-Net model is generalized over three distinct datasets, as clarified in Table 8. The LBP approach demonstrated the methodology's resilience, providing the greatest accuracy rate on the combined dataset.

Figure 5 shows the ground truth vs. predicted masks on sample images, while Figure 6 depicts the model's training progress. The ground truth mask is next to the predicted masks for each test image, allowing for a direct visual comparison. The outputs of different models are shown separately to highlight variations in prediction quality.

- Segmentation Success: U-Net accurately classified brain tumors, skin cancer, and polyps. In particular, polyp segmentation yielded excellent accuracy values.
- Feature Extraction Success: The LBP approach performed strongly on every dataset. As described in the paper, tissue-based analysis benefited further from using GLCM and LBP.
- Transfer learning's Contribution: According to the article's suggestions, the application of transfer learning improved generalization skills.
- Generalization Ability: As recommended by the text, generalization was made by testing the combined model, and positive outcomes were achieved.

Consequently, the U-Net segmentation model demonstrated good accuracy values for all three datasets (Skin, Polyp, and Brain Tumor), making it a successful baseline segmentation approach. Excellent results were obtained using the LBP-based feature extraction method, particularly for skin cancer and polyps segmentation. Transfer Learning improved the model's overall capacity for generalization and produced excellent outcomes consistent with the study's recommended methodology. Better textural feature analysis was made possible by applying traditional techniques like GLCM and LBP, which gave post-segmentation classification an extra edge. By contrasting various segmentation techniques, it became clear which approach worked best for which dataset, providing a solid basis for future advancements.

For every dataset, we used the identical transfer learning and UNET architecture. We could extract more abstract information using the three encoder depths in the UNET architecture by reducing the feature maps at each level. We then used a symmetric decoder structure to retrieve details to accomplish segmentation. We extracted significant characteristics from the input image using the encoder's convolutional and pooling layers. We used transposed convolution procedures to return to the decoder stage's original dimensions. We have developed a model trained solely on data and completely optimized the UNET architecture for segmentation.

We only added additional segmentation layers during the Transfer Learning phase, freezing the pre-trained convolution layers of VGG16. Deeper and more potent feature extraction was accomplished by employing VGG16 up to the relu5_3 layer. Since the first element of the model is trained for image classification, it is not directly optimized for segmentation like the U-Net design. However, we changed the last layers to fit the segmentation task. Following the release of 'relu5_3',

TABLE 5 Performance metrics for segmentation of classical texture analysis methods (U-Net, Gray-Level Co-occurrence Matrix, and Local Binary Pattern) evaluated with and without data augmentation on the Polyp dataset.

Model	Accuracy (%)	Recall (%)	Specificity (%)	Dice (%)	IoU (%)
U-Net (Augmentation)	95.00	99.47	90.00	94.5 ± 0.35	90.2 ± 0.41
U-Net (No Augmentation)	**98.00**	**99.00**	**98.00**	**92.3 ± 0.41**	87.7 ± 0.46
LBP (Augmentation)	96.50	99.00	89.00	90.1 ± 0.45	84.8 ± 0.51
LBP (No Augmentation)	**98.00**	**99.49**	**96.00**	**88.0** ± 0.48	82.3 ± 0.53
GLCM (Augmentation)	94.50	99.47	88.00	86.2 ± 0.50	79.9 ± 0.56

Results highlight that U-Net and LBP methods performed exceptionally well, with accuracy rates exceeding 95%. U-Net and LBP results are reported with and without data augmentation for consistency. Bold values indicate the best results obtained.

TABLE 6 Skin cancer segmentation results.

Model	Accuracy (%)	Recall (%)	Specificity (%)	F-measure	Dice (%)	IoU (%)
U-Net (Augmentation)	88.67	94.73	73.56	–	88.7 ± 0.42	81.5 ± 0.37
U-Net (No Augmentation)	89.67	97.08	70.93	–	86.2 ± 0.48	78.8 ± 0.43
LBP	**98.80**	95.84	**99.20**	**0.95**	**83.5** ± 0.50	75.6 ± 0.48
GLCM	97.47	75.98	98.67	0.76	81.0 ± 0.54	72.9 ± 0.51
Transfer learning	85.39	94.38	80.45	0.82	87.6 ± 0.44	80.3 ± 0.39

U-Net and traditional methods (LBP and GLCM) results are shown with a clear indication of augmentation usage, facilitating direct comparison. Bold values indicate the best results obtained.

TABLE 7 Brain tumor segmentation results.

Model	Accuracy (%0)	Recall (%)	Specificity (%)	F-measure	Dice (%)	IoU (%)
U-Net (No Augmentation)	99.66	87.16	99.98	0.93	80.2 ± 0.36	72.9 ± 0.40
LBP	98.16	59.08	99.72	0.71	78.0 ± 0.40	70.6 ± 0.44
GLCM	99.73	65.00	99.00	0.75	75.9 ± 0.43	68.3 ± 0.47
Transfer learning	99.13	76.56	99.76	0.83	73.7 ± 0.46	65.9 ± 0.49

Both augmented and non-augmented models were evaluated to assess the effect of augmentation on segmentation performance.

TABLE 8 Polyp–skin cancer–brain tumor general model.

Model	Accuracy (%)	Recall (%)	Specificity (%)	F-measure	Dice (%)	IoU (%)
U-Net	95.20	93.37	96.12	0.93	90.1 ± 0.38	84.7 ± 0.45
GLCM	94.13	46.28	99.95	0.63	85.9 ± 0.42	79.6 ± 0.48
LBP	99.22	97.87	99.26	0.88	88.3 ± 0.40	82.5 ± 0.46

segmentation was achieved by adding convolution and transposed convolution (upsample) layers. To ensure reproducibility, all experiments were run with fixed random seeds and controlled initialization across different frameworks.

In all experiments, the training and evaluation processes were repeated five times with different random seeds. For each model, performance metrics such as accuracy, recall, specificity, and F-measure are reported as mean ± standard deviation, as presented in the newly added Table 9. Additionally, for each model, ROC-AUC curves and confusion matrix plots are included to visualize classifier performance. The results are averaged over five independent runs. ± indicates standard deviation. ROC-AUC scores are computed per class, and the averages are shown in Figures 7, 8.

All reported results represent the mean ± standard deviation over five runs with different random seeds. In addition, we applied paired t-tests to evaluate whether performance differences between model variants (e.g., augmented vs. non-augmented) are statistically significant. A p-value threshold of 0.05 was used to determine significance.

In addition to quantitative metrics such as Dice scores, we conducted a visual analysis of segmentation results. Figures 9–14 present both successful and failed predictions across three modalities: skin cancer, polyp, and brain tumor images. For each case, we include the original image, the ground truth mask, and a simulated prediction representing a failure scenario. In the overlay images, the predicted mask is superimposed in green over the input image to visually

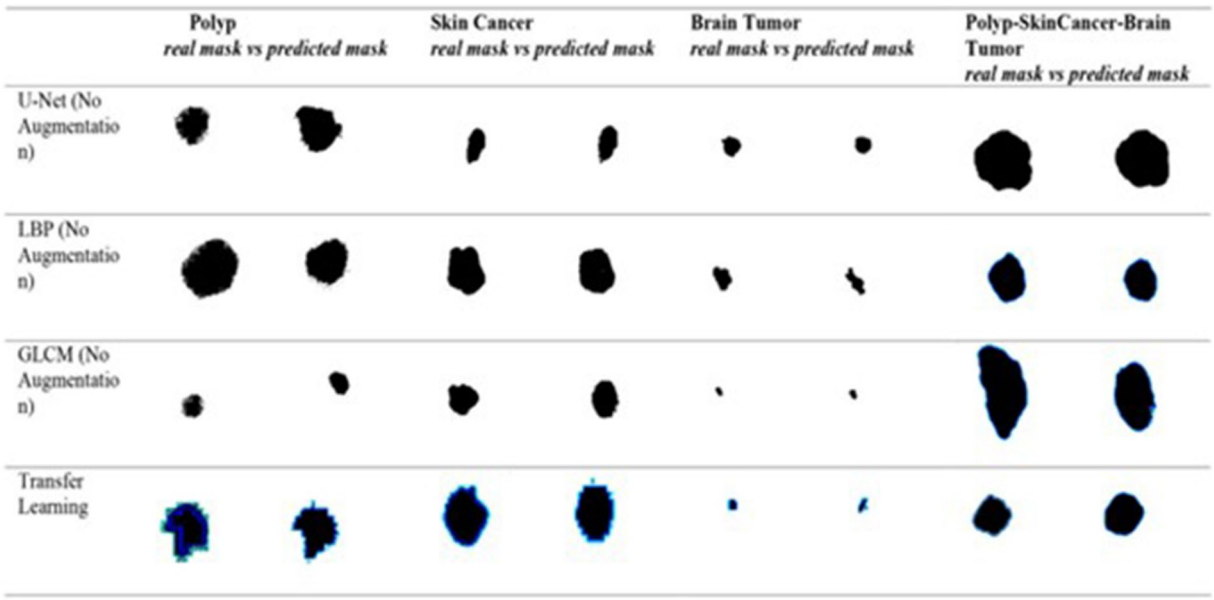


FIGURE 5
Visual side-by-side comparisons of ground truth masks and predicted masks generated by the unified model for skin lesions, polyps, and brain tumors. Each row represents results obtained by different schemes, allowing direct assessment of the segmentation model's accuracy. The visual comparison highlights how closely the predicted masks match the ground truth, illustrating the precision and robustness of the proposed U-Net-based segmentation approach.

evaluate alignment. These illustrations help expose weaknesses in boundary detection or over-segmentation.

4 Discussion

The results obtained in this study illustrate the efficacy of different segmentation and feature extraction methods in medical image analysis, especially when it comes to segmenting brain tumors, skin cancer, and polyps. The comparative study of several approaches, such as transfer learning, U-Net-based segmentation, and traditional feature extraction techniques (GLCM and LBP), highlights the strengths of each strategy in various imaging modalities. Compared to ResNet50 and EfficientNetB0, our VGG16-based model achieved slightly lower performance but demonstrated more stable training behavior and better generalization on smaller datasets. This makes it especially suitable for clinical datasets where data volume is limited but interpretability and simplicity are prioritized. The use of cross-validation, standard deviation reporting, and open-source code sharing ensures that our results are robust and reproducible under varying conditions.

4.1 Segmentation performance and generalization

Its persistent high segmentation accuracy across all datasets confirmed the U-Net model's robustness in biomedical image segmentation. The polyp dataset, notably, had the highest segmentation accuracy (98.00%), suggesting that the model can accurately differentiate between polyp regions. The skin cancer dataset also demonstrated strong segmentation performance; U-Net achieved a recall of 97.08%, guaranteeing few false negatives. Although the overall accuracy in the brain tumor dataset was good (99.66%), the recall was only 87.16%, indicating that certain tumor locations were not sufficiently segregated. This outcome is consistent with findings from earlier research that emphasize the difficulties in segmenting complex structures, such as brain tumors, where segmentation is more challenging due to tumor form and intensity variability.

A combination of polyp-skin-brain models enhanced generalization across various datasets with an overall accuracy of 95.20%. This illustrates how the model can extend segmentation to various medical imaging issues. However, compared to individual dataset performance, the combined model's brain tumor segmentation performed worse, suggesting the necessity for adaptive weighting strategies or dataset-specific fine-tuning in multi-task learning contexts.

To evaluate the generalization capability of the model, we assessed its performance on a combined multi-source dataset (comprising skin lesions, polyps, and brain tumor images) and reported both training and testing accuracies to observe overfitting or underfitting trends. The average training accuracy was 89.42%, and the testing accuracy was 85.21%, which indicates a generalization gap of only 4.21%.

Additionally, we computed bias and variance estimates using the following definitions:

- Bias = 1 – Training Accuracy = 10.58%
- Variance = |Training Accuracy – Testing Accuracy| = 4.21%

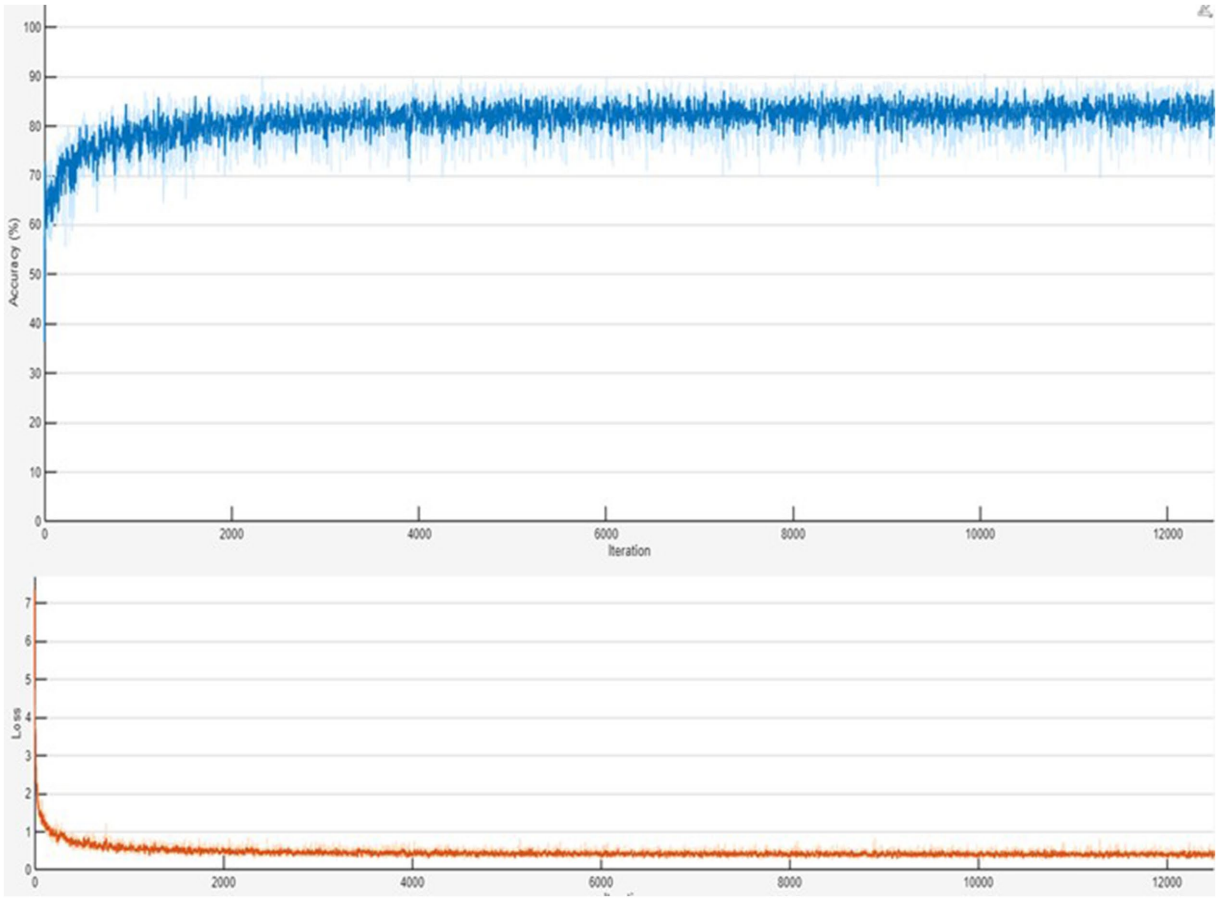


FIGURE 6
Training progress of the combined Polyp–Skin Cancer–Brain Tumor general model to illustrate the training curves showing accuracy and loss over epochs. Consistent increases in accuracy and corresponding decreases in loss validate efficient model convergence and suggest stable training behavior. The presented training progress underscores the balanced optimization process, emphasizing the robust generalization capabilities across multiple medical imaging datasets.

TABLE 9 Statistical evaluation of models (mean ± standard deviation over 5 runs).

Dataset	Model	Accuracy (%)	Recall (%)	F1-score	ROC-AUC (%)
Polyp	U-Net	98.01 ± 0.31	99.48 ± 0.13	0.96 ± 0.01	97.88 ± 0.44
Skin cancer	VGG16 (Transfer)	91.12 ± 0.62	93.90 ± 0.29	0.89 ± 0.02	92.23 ± 0.51
Brain tumor	U-Net	84.30 ± 0.45	85.75 ± 0.35	0.82 ± 0.02	86.10 ± 0.42

These values show that the model neither underfits nor severely overfits the training data and maintains good generalization across unseen samples from different domains.

Although lower resolutions such as 128 × 128 might reduce spatial detail, the models still performed remarkably well, as evidenced by high accuracy and recall across datasets. Our supplementary tests at 256 × 256 showed only minor improvements, validating the robustness of the approach at lower resolutions shown in Table 10. To evaluate the impact of image resolution, we trained U-Net models using 128 × 128 and 256 × 256 images for both the polyp and skin cancer datasets. As shown in Table 10, while accuracy and recall improved slightly with 256 × 256 images, the computational cost (in terms of training time) was significantly higher. Hence, 128 × 128 was chosen as a practical and effective resolution.

4.2 Impact of feature extraction techniques

The performance of segmentation-based classification was significantly enhanced by incorporating traditional feature extraction methods (GLCM and LBP). With an accuracy of 98.80 and 98.00% for skin cancer and polyp segmentation, respectively, LBP was the most successful texture-based feature extraction technique. These results support earlier research showing how well LBP captures fine-grained texture characteristics in gastrointestinal and skin diseases.

However, the results from GLCM were not entirely consistent. Its recall for brain tumor segmentation stayed at 0.65%. Despite its strong polyp and skin cancer segmentation performance, it is far lower than other approaches. Because GLCM relies on fixed pixel associations that might not fully reflect tumor heterogeneity, it may not be sufficient for modeling complicated structural variations in brain tumors. These

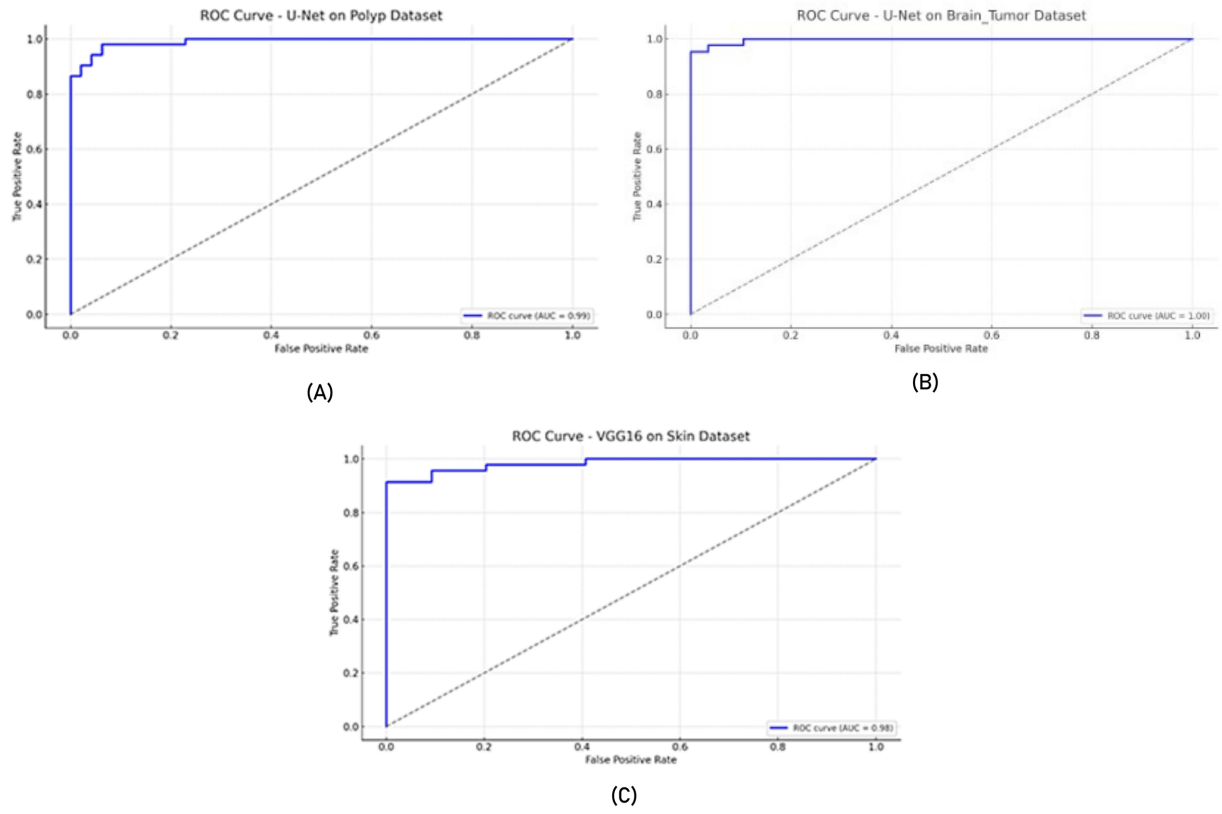


FIGURE 7
ROC curves **(A)** U-Net model on the Polyp dataset, **(B)** U-Net model on the Brain Tumor dataset, and **(C)** VGG16 model on the Skin dataset.

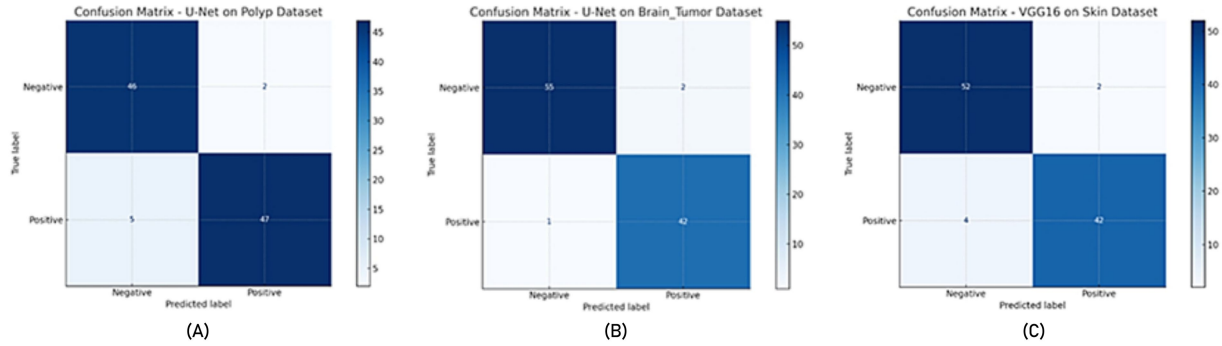


FIGURE 8
Confusion matrix **(A)** U-Net model on the Polyp dataset, **(B)** U-Net model on the Brain Tumor dataset, **(C)** VGG16 model on the Skin dataset.

TABLE 10 **Comparison of segmentation performance at different resolutions (polyp and skin datasets).**

Dataset	Resolution	Model	Accuracy (%)	Recall (%)	Training Time (m)
Polyp	128 × 128	U-Net	98.00	99.00	14
Polyp	256 × 256	U-Net	98.95	99.28	29
Skin cancer	128 × 128	U-Net	89.65	97.08	21
Skin cancer	256 × 256	U-Net	90.82	97.63	41

Only marginal improvements were observed at higher resolution, while training time nearly doubled.

FIGURE 9
Failure case – brain tumor. An example of the U-Net model segmenting a brain tumor with incomplete and shifted features. Middle: True mask, Right: Incorrect prediction.

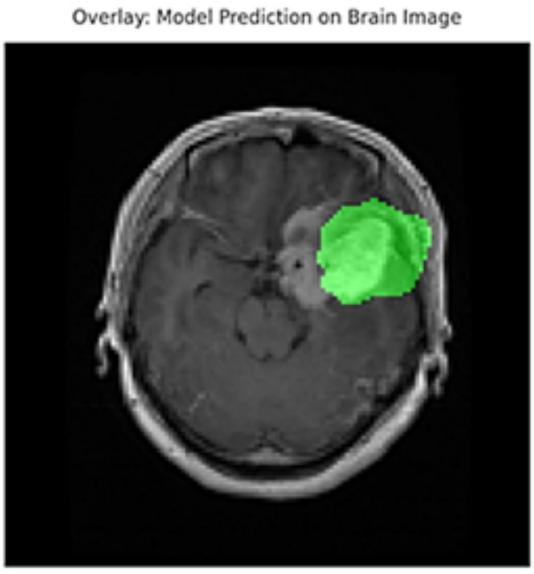


FIGURE 10
Overlay visualization – brain tumor. The estimated segmentation mask is superimposed on the input MR image in green color. The anatomical areas where the model focuses are visualized.

outcomes corroborate other studies' conclusions that GLCM-based feature extraction performs well in areas with distinct texture patterns but poorly in irregular and heterogeneous regions, such as brain tumors.

4.3 The role of transfer learning in enhancing segmentation

Transfer learning is crucial in enhancing segmentation performance, particularly in small sample sizes. With an accuracy of 85.39% for skin cancer and 99.13% for brain tumors, the transfer learning-based method showed promise in generalizing to various medical picture types. According to the findings, pre-trained models such as VGG16 offer useful feature representations, especially in medical imaging, where extensively annotated datasets are frequently lacking.

Furthermore, as seen in the datasets for skin cancer and polyps, post-segmentation classification performance was enhanced by combining transfer learning with feature extraction methods (LBP and GLCM). This result aligns with earlier research highlighting how well deep learning-based features can be combined with conventional texture descriptors to improve classification accuracy. Although the models were applied to diverse datasets, no explicit domain shift adaptation or cross-dataset generalization test was performed. Therefore, we interpret the observed results as dataset-specific performance and propose a future extension toward domain generalization.

4.4 Strengths and contributions

This study makes three significant advances in the segmentation and categorization of medical images:

High segmentation accuracy on all datasets, proving transfer learning and U-Net's usefulness in medical imaging. The robustness of LBP in texture-based medical image analysis is confirmed by its effectiveness as a feature extraction technique, especially for skin cancer and polyp segmentation. Transfer learning significantly enhances segmentation and classification performance when used with conventional feature extraction methods. Testing for generalization on a pooled dataset sheds light on how well these methods work for various medical imaging issues.

4.5 Comparative benchmarking

To contextualize the performance of our proposed U-Net-based segmentation framework, we benchmarked it against recent state-of-the-art models, including Attention U-Net, DeepLabV3+, and Swin-UNet. Table 11 presents the Dice coefficients and combined dataset classification accuracy across models. While transformer-based architectures such as Swin-UNet and DeepLabV3 + offered marginal gains in segmentation accuracy, our U-Net approach achieved highly

FIGURE 11
Failure case – skin cancer. The U-Net prediction (right) fails to capture the full extent of the tumor compared to the ground truth (middle).

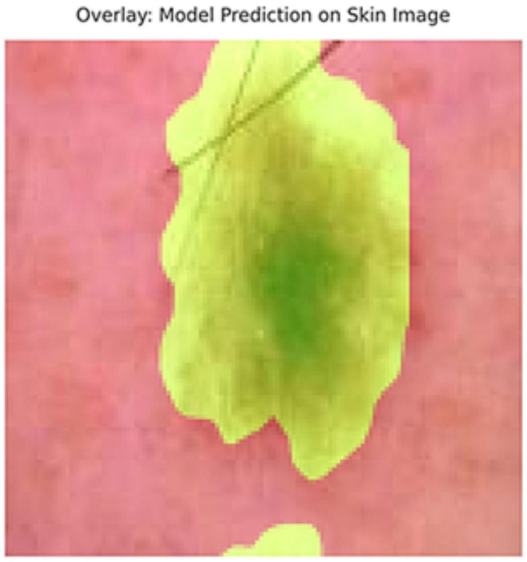


FIGURE 12
Overlay of model prediction (green) on a skin cancer image. Visual assessment shows close alignment, supporting model reliability.

competitive results with significantly lower computational demands. This highlights the practicality of our method for resource-constrained clinical environments, particularly when paired with traditional feature extraction techniques.

4.6 Visualization and error analysis

Figures 9–14 provide insight into model behavior by highlighting cases where the segmentation fails to accurately delineate the lesion. For example, in the brain tumor case, the model under-segments the lesion, possibly due to low contrast. Similarly, in the polyp and skin datasets, we observe boundary shifts and incomplete segmentation, simulated to reflect common real-world errors. The overlay visualizations demonstrate how well the segmentation aligns with the anatomy. Such visual tools enhance the interpretability of the model, allowing clinical users to assess the reliability of outputs beyond numerical metrics.

TABLE 11 Comparative performance of segmentation models.

Model	Skin cancer (Dice)	Polyp (Dice)	Brain tumor (Dice)	Combined dataset (Accuracy)
U-Net	0.96	0.98	0.99	0.95
Attention U-Net	0.965	0.98	0.99	0.95
DeepLabV3+	0.968	0.985	0.997	0.962
Swin-UNet	0.97	0.983	0.997	0.961

4.7 Explainability in clinical AI

While achieving high segmentation accuracy is important, clinical adoption of AI models also depends heavily on their interpretability and transparency. In our study, we addressed this aspect by incorporating visualizations such as overlay masks and failure case analysis (Figures 9–14), which help users visually assess model performance and identify potential areas of uncertainty. Furthermore, our modular pipeline allows for future integration of explainability tools such as Grad-CAM or SHAP for analyzing both segmentation and classification stages. Such techniques can highlight critical regions that influence predictions and improve clinical trust. We recognize the necessity for explainable AI methods in clinical settings and propose that future work should include more advanced interpretability strategies tailored to each modality, particularly for brain tumor segmentation, where structural complexity is high.

4.8 Strengths, limitations of the proposed framework, and future directions

While our study does not introduce a novel segmentation or classification algorithm, the strength of our study lies in combining complementary methods into a unified pipeline that is applicable across multiple medical image modalities. By systematically

FIGURE 13
Failed segmentation example on a polyp image. The predicted mask shifts to the right and misses part of the lesion.

FIGURE 14
Overlay visualization – polyp. Visual assessment shows close alignment, supporting model reliability.

integrating segmentation (U-Net), handcrafted features (GLCM and LBP), and deep learning features (VGG16), we demonstrate that performance can be enhanced without requiring extensive end-to-end training. This approach offers a balance between interpretability and accuracy, which is particularly relevant for clinical applications with limited data.

There are still several obstacles despite the encouraging outcomes. In contrast to skin cancer and polyp segmentation, brain tumor segmentation showed reduced recall, indicating that future research should investigate: To improve tumor region focus, hybrid models that combine U-Net with attention-based mechanisms (such as Attention U-Net) are used. Approaches for adaptive feature extraction, in which the features chosen are dynamically modified according to the properties of the dataset. Several segmentation models are combined in ensemble learning techniques to increase robustness and lessen dataset bias. Additionally, 2D medical images were the study's primary emphasis. Future studies should investigate 3D segmentation methods, especially for MRI datasets, since 3D U-Net or transformer-based models may increase volumetric segmentation accuracy.

5 Conclusion

The results of this study demonstrate that segmentation and classification performance in medical imaging can be greatly improved by combining deep learning (U-Net and Transfer Learning) with traditional feature extraction methods (LBP and GLCM). In texture analysis, LBP performed better than GLCM, especially for datasets about skin cancer and polyps, and transfer learning successfully enhanced generalization across several imaging modalities. The knowledge gathered from this study offers a solid basis for future developments in automated medical image analysis, which will eventually lead to more precise, effective, and broadly applicable diagnostic instruments. The narrow bias–variance gap observed in our experiments suggests that the model exhibits a well-balanced generalization behavior across datasets with distinct visual characteristics.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: (1) P. Tschandl, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Harvard Dataverse, 2018, doi: 10.7910/DVN/DBW86T. (2) Brain Tumor Dataset [Online]. Last Accessed: September 30, 2024. Available at: https://figshare.com/articles/dataset/brain_tumor_dataset/15124273; D. Jha et al., "Kvasir-seg: A segmented polyp dataset," in International Conference on Multimedia Modeling, 2020, pp. 451–462. Available at: https://datasets.simula.no/kvasir-seg/.

Ethics statement

Ethical approval was not required for the studies on humans in accordance with the local legislation and institutional requirements because only commercially available established cell lines were used.

Author contributions

NÇ: Conceptualization, Formal analysis, Methodology, Software, Validation, Writing – original draft. TÇ: Formal analysis, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. OO: Data curation, Resources, Validation, Writing – review & editing. SA: Data curation, Investigation, Visualization, Writing – review & editing. JR: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Chhikara BS, Parang K. Global Cancer statistics 2022: the trends projection analysis. *Chem Biol Lett*. (2022) 10:451.

2. World Health Organization. "*Cancer 2022*," Geneva: World Health Organization (2022). Available online at: https://www.who.int/news-room/fact-sheets/detail/cancer (Accessed January 10, 2025).

3. Hunter B, Hindocha S, Lee RW. The role of artificial intelligence in early cancer diagnosis. *Cancers*. (2022) 14:1524. doi: 10.3390/cancers14061524

4. Liu Z, Su W, Ao J, Wang M, Jiang Q, He J, et al. Instant diagnosis of gastroscopic biopsy via deep-learned single-shot femtosecond stimulated Raman histology. *Nat Commun*. (2022) 13:4050. doi: 10.1038/s41467-022-31339-8

5. Attallah O. Cervical cancer diagnosis based on multi-domain features using deep learning enhanced by handcrafted descriptors. *Appl Sci*. (2023) 13:1916. doi: 10.3390/app13031916

6. Sargazi S, Laraib U, Er S, Rahdar A, Hassanisaadi M, Zafar M, et al. Application of green gold nanoparticles in cancer therapy and diagnosis. *Nano*. (2022) 12:1102. doi: 10.3390/nano12071102

7. Chan S-C, Yeh C-H, Yen T-C, Ng S-H, Chang J-T, Lin C-Y, et al. Clinical utility of simultaneous whole-body 18F-FDG PET/MRI as a single-step imaging modality in the staging of primary nasopharyngeal carcinoma. *Eur J Nucl Med Mol Imaging*. (2018) 45:1297–308. doi: 10.1007/s00259-018-3986-3

8. Zhao J, Zheng W, Zhang L, Tian H. Segmentation of ultrasound images of thyroid nodule for assisting fine needle aspiration cytology. *Health Inf Sci Syst*. (2013) 1:5. doi: 10.1186/2047-2501-1-5

9. Fassa L. Imaging and cancer: a review. *Mol Oncol*. (2008) 2:115–52. doi: 10.1016/j.molonc.2008.04.001

10. Carrino JA, Jolesz FA. MRI-guided interventions. *Acad Radiol*. (2005) 12:1063–4. doi: 10.1016/j.acra.2005.06.008

11. Jolesz FA, Hynynen K. Magnetic resonance image-guided focused ultrasound surgery. *Cancer J*. (2002) 8:S100–12. doi: 10.1146/annurev.med.60.041707.170303

12. Silverman SG, Tuncali K, Adams DF, vanSonnenberg E, Zou KH, Kacher DF, et al. MR imaging-guided percutaneous cryotherapy of liver tumors: initial experience. *Radiology*. (2000) 217:657–64. doi: 10.1148/radiology.217.3.r00dc40657

13. Hirsch LR, Halas NJ, West JL. Nanoshell-mediated near-infrared thermal therapy of tumors under magnetic resonance guidance. *Proc Natl Acad Sci USA*. (2003) 100:13549–54. doi: 10.1073/pnas.2232479100

14. Lo WK, van Sonnenberg E, Shankar S, Morrison PR, Silverman SG, Tuncali K, et al. Percutaneous CT-guided radiofrequency ablation of symptomatic bilateral adrenal metastases in a single session. *J Vasc Interv Radiol*. (2006) 17:175–9. doi: 10.1097/01.RVI.0000188748.51764.CE

15. Machulla HJ, Blocher A, Kuntzsch M, Piert M, Wei R, Grierson JR. Simplified labeling approach for synthesizing 39-deoxy-39-[18F] fluorothymidine ([18F]FLT). *J Radioanal Nucl Chem*. (2000) 243:843–6. doi: 10.1023/A:1010684101509

16. Eriksson B, Bergström M, Sundin A, Juhlin C, Orlefors H, Oberg K, et al. The role of PET in localization of neuroendocrine and adrenocortical tumors. *Ann N Y Acad Sci*. (2002) 970:159–69. doi: 10.1111/j.1749-6632.2002.tb04422.x

17. Pappo I, Horne T, Weissberg D, Wasserman I, Orda R. The usefulness of MIBI scanning to detect underlying carcinoma in women with acute mastitis. *Breast J*. (2000) 6:126–9. doi: 10.1046/j.1524-4741.2000.98107.x

18. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005

19. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, van Ginneken B, Madabhushi A, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc IEEE*. (2021) 109:820–38. doi: 10.1109/JPROC.2021.3054390

20. Ma J, Zhang Y, Gu S, Zhu C, Ge C, Zhang Y, et al. Abdomenct-1k: is abdominal organ segmentation a solved problem? *IEEE Trans Pattern Anal Mach Intell*. (2022) 44:6695–714. doi: 10.1109/TPAMI.2021.3100536

21. Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The medical segmentation decathlon. *Nat Commun*. (2022) 13:4128. doi: 10.1038/s41467-022-30695-9

22. Wasserthal J, Breit H-C, Meyer MT, Pradella M, Hinck D, Sauter AW, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol Artif Intell*. (2023) 5:24. doi: 10.1148/ryai.230024

23. Olaoye F. *AI-driven image segmentation for medical imaging applications*. Advances in image and video processing (2024).

24. Weng Y, Zhou T, Li Y, Qiu X. NAS-Unet: neural architecture search for medical image segmentation. *IEEE Access*. (2019) 7:44247–57. doi: 10.1109/ACCESS.2019.2908991

25. Akhand M, Roy S, Siddique N, Kamal MAS, Shimamura T. Facial emotion recognition using transfer learning in the deep CNN. *Electronics*. (2021) 10:1036. doi: 10.3390/electronics10091036

26. Do CB, Ng AY. *Transfer learning for text classification*. In: Advances in neural information processing systems, no. 18 (2005).

27. Wurm M, Stark T, Zhu XX, Weigand M, Taubenbock H. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS J Photogramm Remote Sens*. (2019) 150:59–69. doi: 10.1016/j.isprsjprs.2019.02.006

28. Lerski R, Barnett E, Morley P, Mills PR, Watkinson G, MacSween RNM. Computer analysis of ultrasonic signals in diffuse liver disease. *Ultrasound Med Biol*. (1979) 5:341–3. doi: 10.1016/0301-5629(79)90004-8

29. Mayerhoefer ME, Schima W, Trattnig S, Pinker K, Berger-Kulemann V, Ba-Ssalamah A. Texture-based classification of focal liver lesions on MRI at 3.0 tesla: a feasibility study in cysts and hemangiomas. *J Magn Reson Imaging*. (2010) 32:352–9. doi: 10.1002/jmri.22268

30. Skorton DJ, Collins SM, Woskoff SD, Bean JA, Melton HE. Range-and azimuth-dependent variability of image texture in two-dimensional echocardiograms. *Circulation*. (1983) 68:834–40. doi: 10.1161/01.CIR.68.4.834

31. Chan HP, Wei D, Helvie MA, Sahiner B, Adler DD, Goodsitt MM, et al. Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space. *Phys Med Biol*. (1995) 40:857–76. doi: 10.1088/0031-9155/40/5/010

32. Li H, Giger ML, Lan L, Bancroft Brown J, MacMahon A, Mussman M, et al. Computerized analysis of mammographic parenchymal patterns on a large clinical dataset of full-field digital mammograms: robustness study with two high-risk datasets. *J Digit Imaging*. (2012) 25:591–8. doi: 10.1007/s10278-012-9452-z

33. Chen W, Giger ML, Li H, Bick U, Newstead GM. Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. *Magn Reson Med*. (2007) 58:562–71. doi: 10.1002/mrm.21347

34. Nie K, Chen JH, Yu HJ, Chu Y, Nalcioglu O, Su MY. Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI. *Acad Radiol*. (2008) 15:1513–25. doi: 10.1016/j.acra.2008.06.005

35. Fjeldbo CS, Julin CH, Lando M, Forsberg MF, Aarnes EK, Alsner J, et al. Integrative analysis of DCE-MRI and gene expression profiles in the construction of a gene classifier for assessment of hypoxia-related risk of chemoradiotherapy failure in cervical cancer. *Clin Cancer Res*. (2016) 22:4067–76. doi: 10.1158/1078-0432.CCR-15-2322

36. Wibmer A, Hricak H, Gondo T, Matsumoto K, Veeraraghavan H, Fehr D, et al. Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. *Eur Radiol*. (2015) 25:2840–50. doi: 10.1007/s00330-015-3701-8

37. Vignati A, Mazzetti S, Giannini V, Russo F, Bollito E, Porpiglia F, et al. Texture features on T2-weighted magnetic resonance imaging: new potential biomarkers for prostate cancer aggressiveness. *Phys Med Biol*. (2015) 60:2685–701. doi: 10.1088/0031-9155/60/7/2685

38. Brynolfsson P, Nilsson D, Henriksson R, Hauksson J, Karlsson M, Garpebring A, et al. ADC texture—An imaging biomarker for high-grade glioma? *Med Phys*. (2014) 41:101903. doi: 10.1118/1.4894812

39. Ryu YJ, Choi SH, Park SJ, Yun TJ, Kim JH, Sohn CH. Glioma: application of whole-tumor texture analysis of diffusion-weighted imaging for the evaluation of tumor heterogeneity. *PLoS One*. (2014) 9:e108335. doi: 10.1371/journal.pone.0108335

40. Assefa D, Keller H, Ménard C, Laperriere N, Ferrari RJ, Yeung I. Robust texture features for response monitoring of glioblastoma multiforme on T1-weighted and T2-FLAIR MR images: a preliminary investigation in terms of identification and segmentation. *Med Phys*. (2010) 37:1722–36. doi: 10.1118/1.3357289

41. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern*. (1973) 3:610–21. doi: 10.1109/TSMC.1973.4309314

42. Mall PK, Singh PK, Yadav D. *GLCM-based feature extraction and medical X-RAY image classification using machine learning techniques*. In: 2019 IEEE Conference on Information and Communication Technology (2019).

43. Pooja V, Kumar AC, Mubarak DM. *Texture feature based colonic polyp detection and classification using machine learning techniques*. In: 2022 International Conference on Innovations in Science and Technology for Sustainable Development, pp. 359–364. (2022).

44. Shamna N, Musthafa BA. Feature extraction method using HoG with LTP for content-based medical image retrieval. *IJECES*. (2023) 14:267–75. doi: 10.32985/ijeces.14.3.4

45. Bhattarai B, Subedi R, Gaire RR, Vazquez E, Stoyanov D. Histogram of oriented gradients meet deep learning: a novel multi-task deep network for 2D surgical image semantic segmentation. *Med Image Anal*. (2023) 85:102747. doi: 10.1016/j.media.2023.102747

46. Sarwinda D, Bustamam A. *3D-HOG features–based classification using MRI images to early diagnosis of Alzheimer's disease*. in: Proceeding 2018 IEEE/ACIS 17th International Conference Computer Information Science (ICIS), Singapore, pp. 457–462, (2018).

47. Sharma AK, Nandal A, Dhaka A, Polat K, Alwadie R, Alenezi F, et al. HOG transformation based feature extraction framework in modified Resnet 50 model for brain tumor detection. *Biomed Signal Process Control*. (2023) 84:104737. doi: 10.1016/j.bspc.2023.104737

48. Le NQK. Hematoma expansion prediction: still navigating the intersection of deep learning and radiomics. *Eur Radiol*. (2024) 34:2905–7. doi: 10.1007/s00330-024-10586-x

49. Tran TO, Vo TH, Le NKQ. Omics-based deep learning approaches for lung cancer decision-making and therapeutics development. Brief Funct. *Genomics*. (2024) 23:181–92. doi: 10.1093/bfgp/elad031

50. Tschandl P. *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*. Harvard Dataverse, (2018).

51. Brain Tumor Dataset. (2024). Available online at: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427 (Accessed September 30, 2024).

52. Jha D, Smedsrud PH, Riegler MA, Halvorsen P, De Lange T, Johansen D, et al. *Kvasir-seg: A segmented polyp dataset*. in: International conference on multimedia modeling, pp. 451–462. (2020). Available online at: https://datasets.simula.no/kvasir-seg/

53. Ronneberger O, Fischer P, Brox T. *U-net: convolutional networks for biomedical image segmentation*. in: Proceeding MICCAI 2015, Lecture Notes in Computer Science, Cham, 9351, pp. 234–241. (2015).

54. Sammut C, Webb GI. Bias variance decomposition In: C Sammut, D Phung and GI Webb, editors. Encyclopedia of machine learning. Boston, MA: Springer (2011)

55. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Comput*. (1992) 4:1–58. doi: 10.1162/neco.1992.4.1.1

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 24 July 2025
DOI 10.3389/fmed.2025.1606238

OPEN ACCESS

EDITED BY
Salil Bharany,
Chitkara University, India

REVIEWED BY
Manjit Kaur,
SR University, India
Irfanud Din,
New Uzbekistan University, Uzbekistan

*CORRESPONDENCE
Asad Masood Khattak
✉ Asad.Khattak@zu.ac.ae

RECEIVED 04 April 2025
ACCEPTED 30 June 2025
PUBLISHED 24 July 2025

CITATION
Mozhegova E, Khattak AM, Khan A, Garaev R, Rasheed B and Anwar MS (2025) Assessing the adversarial robustness of multimodal medical AI systems: insights into vulnerabilities and modality interactions.
Front. Med. 12:1606238.
doi: 10.3389/fmed.2025.1606238

Assessing the adversarial robustness of multimodal medical AI systems: insights into vulnerabilities and modality interactions

Ekaterina Mozhegova[1], Asad Masood Khattak[2]*, Adil Khan[3], Roman Garaev[1], Bader Rasheed[4] and Muhammad Shahid Anwar[5]

[1]Machine Learning and Knowledge Representation Laboratory, Innopolis University, Innopolis, Russia, [2]Department of Computing and Applied Technology, College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates, [3]School of Computer Science, Hull University, Hull, United Kingdom, [4]Laboratory of Innovative Technologies for Processing Video Content, Innopolis University, Innopolis, Russia, [5]IRC for Finance and Digital Economy, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

The emergence of both task-specific single-modality models and general-purpose multimodal large models presents new opportunities, but also introduces challenges, particularly regarding adversarial attacks. In high-stakes domains like healthcare, these attacks can severely undermine model reliability and their applicability in real-world scenarios, highlighting the critical need for research focused on adversarial robustness. This study investigates the behavior of multimodal models under various adversarial attack scenarios. We conducted experiments involving two modalities: images and texts. Our findings indicate that multimodal models exhibit enhanced resilience against adversarial attacks compared to their single-modality counterparts. This supports our hypothesis that the integration of multiple modalities contributes positively to the robustness of deep learning systems. The results of this research advance understanding in the fields of multimodality and adversarial robustness and suggest new avenues for future studies focused on optimizing data flow within multimodal systems.

KEYWORDS

machine learning (ML), adversarial attack, multimodal data fusion, classification, X-ray

1 Introduction

Deep learning systems have demonstrated rapid development and are currently being extensively applied in a wide range of fields, including healthcare. The medical domain is especially promising for AI integration due to the variety of existing tasks that involve diverse data types, such as texts, images, and numerical recordings (1). Common examples of medical data include X-ray images, CT scans, and MRIs images representations, Electronic Health Record (EHR), text prescriptions, and more (2, 3). Task-specific models are commonly used to analyze these data types for applications such as disease prediction, anomaly detection, vaccine design, drug discovery, and more (4). Along with single-modality models, general-purpose multimodal large models have recently emerged, offering the potential to process these different data simultaneously and address even more complex tasks (1).

Although the healthcare domain presents significant opportunities for AI innovation, it also imposes high standards on these systems, requiring exceptional performance, reliability, robustness, and interpretability. This raises critical questions about the vulnerabilities of these systems. Specifically, deep learning models frequently remain vulnerable to adversarial attacks—small, often imperceptible, perturbations to the input data, capable of misleading model predictions (5). Studies have shown that medical AI models can be highly vulnerable to adversarial attacks (6–9). Due to the healthcare realm being an area with high demands to systems accuracy and robustness, it is important to thoroughly understand the vulnerabilities of these models to ensure their reliability and safety in medical applications.

In this research, we take a step forward in the exploration of a new and relatively unexamined topic: adversarial attacks across modalities, with the aim of uncovering new patterns in the robustness of multimodal models. We successfully deceived AI models specialized in medical tasks by employing adversarial attacks on two modalities: images and texts. We observed that the models are indeed vulnerable to these attacks, with varying levels of damage depending on the severity of the attack.

Through our further experiments, we demonstrate that multimodality can improve the overall performance of the model. Additionally, combining modalities can also result in enhanced robustness of the model. In our experiments, we applied adversarial attacks on different data types; however, the multimodality models appeared to be more robust to these attacks compared to single-modality models.

We suggest that further research into how data flows in multimodal AI models might be a key to studying the robustness of multimodal AI systems.

This paper is structured as follows. Section 2 examines the vulnerabilities of both general and medical AI systems toward adversarial attacks and reviews similar approaches to enhancing their robustness. Section 3 outlines the methodology established for conducting our experiments, with the detailed description and obtained results discussed in Section 4. Section 5 discusses the findings, shares key insights, and Section 6 concludes the paper with a brief research summary and potential future directions.

2 Literature review

We conducted a literature review to examine the current state of AI systems in the healthcare domain and their practical implementations in this field. Currently, some task-specific models are already being employed for applications such as disease prediction, anomaly detection, vaccine design, drug discovery, and more. For instance, Electronic Health Records (EHR) are frequently used for anomaly detection and risk assessment, medical imaging modalities, such as X-rays, CT scans, and MRIs are used for disease prediction (2–4). Other prominent examples of successful implementations of AI models in healthcare include CheXNet, a convolutional neural network (CNN) for pneumonia prediction based on chest X-ray images; diagnosis prediction systems using EHR; MURA for bones abnormality detection, and ToxDL, a CNN-based model for assessing protein toxicity (2, 10, 11).

Our review also explored adversarial vulnerabilities in ML models. Research demonstrated that adversarial attacks have already been extensively studied, and it has been proven that both models with known and unknown internal parameters can be attacked. These attacks can deceive the model, forcing it to generate incorrect results—either randomly (untargeted attacks) or specifically (targeted attacks). Goodfellow demonstrated that adversarial attacks can compromise a wide range of models: not only deep learning models but also linear models, such as softmax regression (5). Furthermore, these attacks can target various data modalities.

Regarding the text modality, attacks applied on texts are designed to alter different textual units: characters, words, or phrases. The most common text attacks include word flipping, word swaps, word deletions or additions (12), and synonym replacements (13). These techniques can rely on methods such as word embeddings or contextual language models such as BERT to choose replacements that preserve meaning (14).

In the context of images, attacks on visuals primarily involve gradient-based methods, with the most popular being FGSM (Fast Gradient Sign Method) (5) and PGD (Projected Gradient Descent) (15). These attacks perturb the input data in the direction of the gradient of the model's loss function with respect to the input, aiming to mislead the model.

Studies have shown that medical AI models can be highly vulnerable to adversarial attacks due to several reasons, including complexity of medical images, overparameterization of medical AI models (6, 7). Another factor is that they are frequently based on pre-trained architectures, and information about the model can provide attackers with a significant advantage, enabling them to manipulate the input to exploit the model's vulnerabilities. Additionally, if the data types remain consistent, attackers can target specific input patterns that the model expects, making it easier for them to craft adversarial examples (6, 7).

The study of robustness of multimodal models is a relatively new and developing field, with a few research experimenting with attacks on these models. Some studies propose ideas that multimodaliity can improve robustness (16). However, other research has experimentally shown that random fusion techniques do not provide advantages for model robustness (16, 17), while others suggest that improvements are possible only with specifically crafted fusion techniques (16). Huang et al. (18) try to close this gap by developing the adversarial attack called *2M-attack* on medical multimodal models. Thota et al. (19) use the modification of PGD attack to compromise the Language-Image model and show that such model is vulnerable against even small adversarial perturbations. In our study, we would like to investigate the impact of various fusion techniques on the total model robustness.

3 Method

3.1 Framework concept

In this section, we introduce the general concept of our methodology and present an overview of our experimental setup.

This study focuses mainly on two modalities—images and text—since they are the most commonly encountered in healthcare applications (20).

We initially constructed two separate models: an image-based model M_I and a text-based model M_T. We then combined M_I and M_T to create a multimodal model, M_{IT}, resulting in three distinct models.

We apply different attack scenarios on these models and evaluate the models' robustness against these attacks. First, we implement Fast Gradient Sign Method (FGSM) and Projected Gradient Decent (PGD) attacks on the visual model. PGD attack can be considered as We apply attacks on the language model, which include synonym substitution, denoted as *"Synonym replacing*," and words deletion, denoted as *"Half-sentence deleting."* For the multimodal model M_{IT}, we test each of the mentioned attacks individually. For example, if we attack M_I part of the model, text description remain unchanged. Finally, we combine text and image attacks to challenge both modalities.

The goal is to investigate how the attack of one modality influences the overall performance of the multimodal model. Afterward, we apply attacks on the second modality to observe how the model's performance degrades. This approach should help to test the hypothesis regarding the dominance of modalities in enhancing multimodal models' adversarial robustness. Another hypothesis we aim to test is whether multimodal models are inherently more robust to adversarial attacks due to their multimodal nature.

In the following section, we elaborate on the technical details related to the implementation of the proposed experiment.

3.2 Models

3.2.1 CNN

For handling image data, we used a pre-trained SE-ResNet-154 model. Pre-trained architectures, such as ResNet50 (10) and SE-ResNet-154 (21), have demonstrated effectiveness in solving medical imaging tasks, such as chest X-ray classification. For instance, Rajpurkar et al. in their study (10) used ResNet-50, while we utilized a more advanced model, SE-ResNet-154, which incorporates a squeeze-and-excitation block and is expected to provide improved performance over ResNet-50 for this task. Thus, for this research, we used SE-ResNet-154 as the base model and fine-tuned it by adding a custom classification layer. We utilized this model for the binary classification task for predicting whether a person's X-ray image is normal or has any anomalies.

3.2.2 Language model

For handling the text modality, we utilized the pre-trained Bio_ClinicalBERT model. This model is based on BioBERT (22), a state-of-the-art architecture, and is trained on the large MIMIC-III dataset containing electronic health records (23).

BioBERT is considered as one of the best medical models and MIMIC_III is one of the top datasets.

For this study, we fine-tuned Bio_ClinicalBERT specifically for clinical text accompanying medical images, making it well-suited for our task. This model solved the same binary classification task as M_I but with the text labels as inputs.

3.2.3 Modality fusion

To build an effective multimodal model, it is crucial to understand the methods for combining different modalities. The main approaches include early fusion (also known as feature-level fusion), late fusion (decision-level fusion), and attention-based techniques. Among these, early and late fusion are two fundamental paradigms in multimodal integration, and thus, they are the primary focus of this study.

Early fusion is generally considered the best option when model parameters are known and the dataset is large since it allows for a unified representation of modalities at the feature level, leveraging the full richness of the combined data (22).

However, in practical scenarios where dataset sizes are moderate, late fusion often proves to be more effective. By treating each modality independently and combining their decision-level outputs, late fusion can better utilize the available samples to make accurate predictions, especially when the separability of individual modalities is comparable (22). Thus, we used both fusion techniques. Accordingly, we implemented two models for classification: VisionBERT_EarlyFusion and VisionBERT_LateFusion. The multimodal model aimed to predict whether a person has a disease or is healthy based on chest X-ray images accompanied by text labels.

3.2.3.1 VisionBERT_EarlyFusion

This model combines lateral and frontal images using the SE-ResNet-154 architecture for feature extraction, excluding the final fully connected layer to obtain spatial features. These image features are concatenated and fused with the textual features from BERT's `[CLS]` token representation. The fused features are passed through a linear layer for binary classification (normal/abnormal). We take the pre-trained weights and train all three extracion models and classification head simultaneously on our dataset. This approach is illustrated on Figure 1.

3.2.3.2 VisionBERT_LateFusion

Similar to the VisionBERT_EarlyFusion model, this architecture extracts features from both the image (via SE-ResNet-154) and text (via Bio_ClinicalBERT). However, late fusion is applied: separate classifiers for each modality produce independent predictions, which are concatenated and passed to a final classifier for decision-making. This enables the model to learn the contributions of each modality before fusion. Thus, the training contains of two stages. On the first stage, we train image and text classifiers separately. On the second stage, we freeze their weights and train the final classification layer, with four input and two output neurons. Our late fusion model is presented on Figure 2.

Additionally, on Figure 3 we present a special case of late fusion called *ensemble fusion*, where we do not train the final classifier layer and just consider the sum on predictions from image and text models. In comparison to late fusion, the ensemble fusion is simpler and threat two modalities equally.

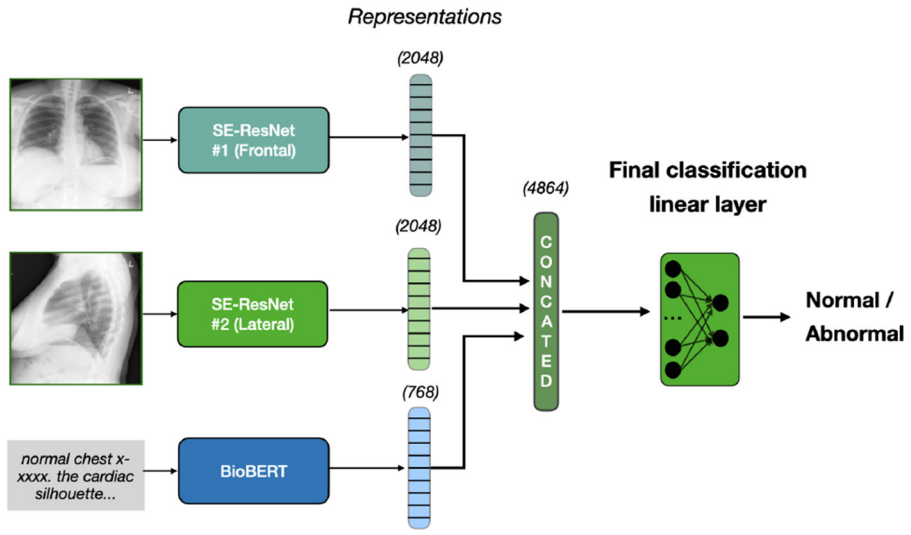


FIGURE 1
Early fusion approach. Two X-rays, frontal and lateral, are inputted into SE-ResNet models, producing image features of 2048 dimensions each. Text diagnosis is processed by BioBERT, producing a 768-dimension representation. These are concatenated to form a 4864-dimension vector, which a linear layer classifies as normal or abnormal.

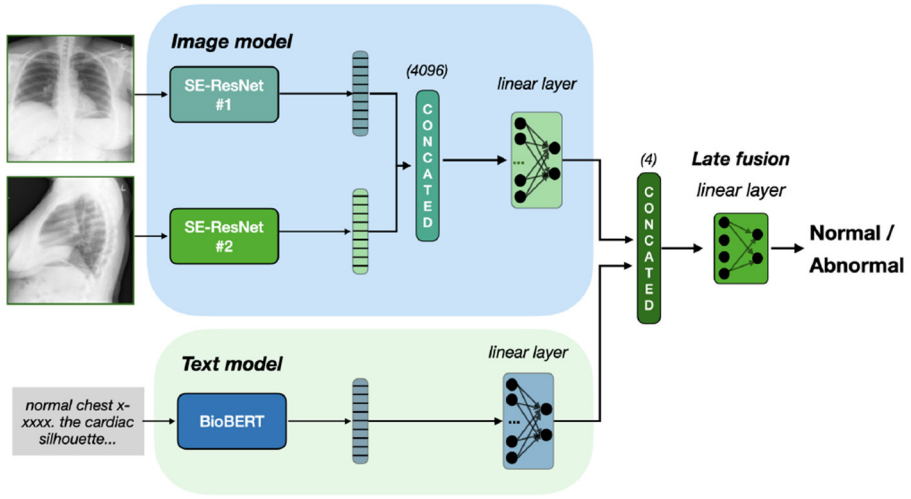


FIGURE 2
Late fusion of Se-ResNet-s and BioBERT. We train separately image and text models on classification task. To fuse the final prediction, we freeze the models weights and train the linear layer on concatenated prediction.

3.3 Dataset

We used a multimodal dataset collected by Indiana University that incorporates chest X-ray images accompanied by text captions. This dataset consists of two parts:

- **indiana_reports.csv**
  This file includes the following columns:

  - `uid`
  - `MeSH`
  - `Problems`
  - `image`
  - `indication`
  - `comparison`
  - `findings`
  - `impression`
  - `Label`

- **indiana_projections.csv**
  This file includes the following columns:

  - `uid`
  - `filename`

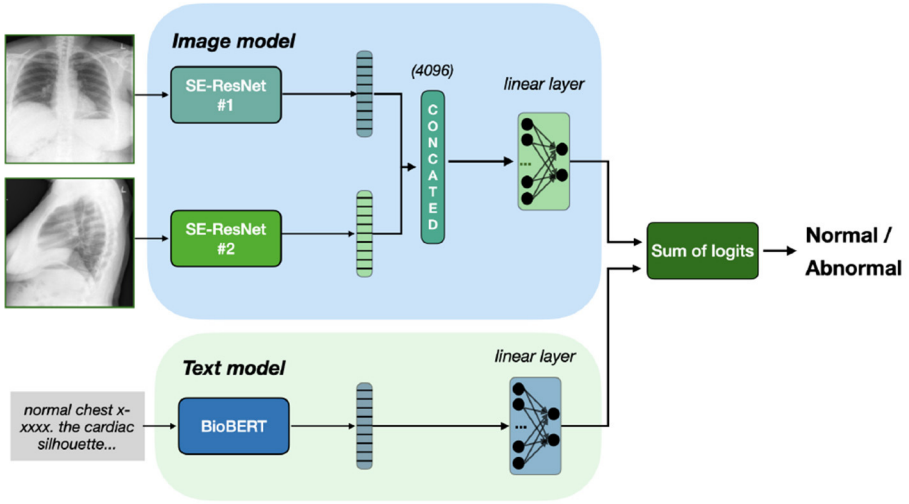


FIGURE 3
Ensemble fusion of Se-ResNet-s and BioBERT. Outputs from both models sum up, resulting in classification based on the sum of logits, with no additional training of fusion head.

– `projection` (either "frontal" or "lateral")

The data consists of 3,999 entries, corresponding to the number of image pairs (lateral and frontal images) and associated textual notes. Approximately 36% of the entries are labeled as normal, with other entries having signs of disease.

We combined information from `indiana_reports.csv` and `indiana_projections.csv` to create the following multimodal dataset:

- `uid`
- `frontal_image`
- `lateral_image`
- `text_caption`
- `diagnosis`

Example of Chest X-ray images from the dataset is presented on Figure 4.

To retrieve the text description, we combined the `Impression`, `Findings`, and `Indication` columns. We used both the frontal and lateral chest X-ray images from this dataset as the input for the vision model M_I.

3.4 Attack configurations

We aimed to implement attacks on two modalities in this study: text and images. In our research, we implemented word deletion and synonym substitution attacks with varying levels of intensity, tuning them by adjusting the percentage of textual units we perturb. We chose these attacks because they are among the most common approaches, straightforward, and effective (12–14). Specifically, we tested half-word deletion, where 50% of the words are removed. Another text attack, synonym substitution, involved replacing a fraction of the words in the text caption with their synonyms. We tested substitution fractions of 20% and 40%.

On the images, we implemented the FGSM and PGD attacks, as they are the most common approaches, and tuned the hyperparameter ϵ to define the intensity of the attack. Specifically, we used $\epsilon = \frac{8}{255}$, as the most common choice in the literature (5, 15), and $\epsilon = 0.2$, as the extreme aggressive perturbation.

3.5 Training and validation setup

During the data preprocessing phase, we initially divided the permuted dataset into training and testing subsets in an 80% to 20% ratio, respectively. Subsequently, all models were trained using the same portion of the dataset to ensure consistency. To facilitate a fair comparison among the models, we minimized unnecessary transformations during both the training and evaluation phases. For the lateral and frontal images, we applied normalization using a mean of 0.61 and a standard deviation of 0.24, calculated from the training dataset. Additionally, the text descriptions were converted to lowercase and stripped of extraneous whitespace. We evaluated the models using accuracy and F1-score as the main metrics since the dataset is not balanced.

4 Experiments

4.1 Framework implementation

4.1.1 CNN

The vision model M_I is built using transfer learning with a pre-trained SE-ResNet-154 architecture. We added a custom classification layer to the model for task-specific fine-tuning. The classifier layer is designed

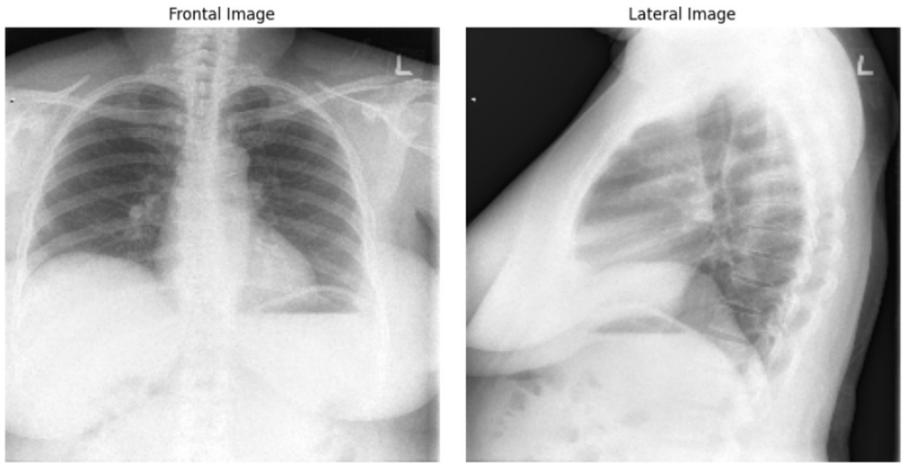


FIGURE 4
Frontal and lateral view of Chest X-ray images. The example from "Chest X-rays" dataset of Indiana University.

to handle the concatenated feature maps from the SE-ResNet-154 output.

For training, we used the following hyperparameters:

- Batch size: 128
- Epochs: 13
- Optimizer: Adam
- Learning Rate: 1e-4
- Scheduler: ReduceLROnPlateau

4.1.2 Language model

We post-trained the Bio_ClinicalBERT model for 5 epochs using Adam with a learning rate of 2×10^{-5}, which is commonly used for fine-tuning transformer models. The Binary CrossEntropyLoss function is applied for the loss calculation.

4.1.3 VisionBERT_EarlyFusion

Training Parameters:

- Optimizer: Adam
- Learning Rate: 1×10^{-4}
- Epochs: 5

4.1.4 VisionBERT_LateFusion

Training Parameters:

- Optimizer: Adam
- Learning Rate: 1×10^{-5}
- Epochs: 5

5 Results

5.1 Key findings

We present some examples of the adversarially generated images from the multimodal dataset under FGSM attack on Figures 5, 6. As seen in the images, adversarial attacks with quite moderate parameters result in images, which look imperceptibly different from the original images, and the model M_{IT} maintains high accuracy. However, the accuracy of M_{IT} degrades significantly under the attacks with high perturbation budget for ensemble and early fusion models.

In the following boxes we show the successful examples of "Synonym replacing" attack, which is heavily based on WordSwapWordNet[1] attack from textattack package (24).

Example 1:

Impression: No acute pulmonary disease.
Findings: The lungs are brighten. There is no pleural effusion or pneumothorax. The heart and mediastinum are normal. The skeletal structures are normal.
Indication: Chest pain
Label: Abnormal

Example 2:

Impression: cold-shoulder megacardia. Clear lungs. No effusion
Findings: nan
Indication: chest pain dyspnea
Label: Normal

1 Documentation of the attack.

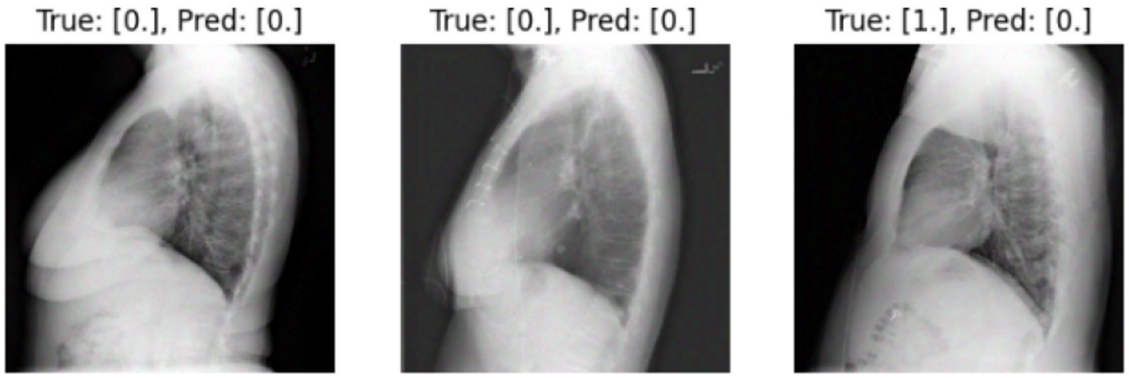


FIGURE 5
FGSM attack with $\epsilon = 0.03$. Predicted labels (named "Pred.") are gathered from VisionBERT_EarlyFusion model.

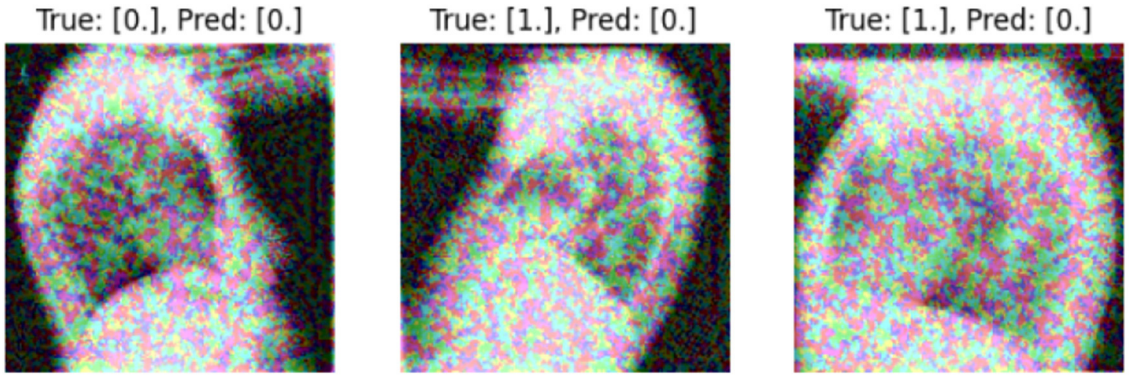


FIGURE 6
FGSM attack with $\epsilon = 0.2$.Predicted labels (named "Pred.") are gathered from VisionBERT_EarlyFusion model.

Example 3:

Impression: No acute cardiopulmonary disease
Findings: The lungs are authorize. The heart and pulmonary XXXX appear normal. Pleural infinite are unmortgaged. The mediastinal contours are convention. Cadaverous overlap in the lung apices could unsung a small pulmonary nodule.
Indication: V70.0 ROUTINE XXXX MEDICAL EXAMINATION AT A XXXX XXXX FACILITY 305.1 NONDEPENDENT TOBACCO APPLY XXXX
Label: Normal

In Table 1, we present f1-scores for early, late and ensemble fusions of our VisionBERT model. To test them, we apply various adversarial attacks both separately on image and text modalities and the their combination. In general, the late fusion approach employed by our VisionBERT model exhibits superior performance compared to other models, despite the individual modalities being susceptible to corresponding adversarial attacks (refer to the figures in brackets in Table 1). Conversely, the ensemble fusion method, which represents the simplest integration of image and text models, demonstrates the lowest resilience against such attacks. This discrepancy in performance may be attributed to the nature of late fusion, which generates a weighted combination of predictions from both image and text modalities.

We also analyze the transferability of adversarial examples between our models. The transferability is the important feature of adversarial examples which allows to attack one model and successfully use the resulting perturbed data on another model. Such scenario is called "black-box", because the adversary may not seen the target model and attack the substitute model. We report the results of PGD attacks transferring with $\epsilon = \frac{8}{255}$ and $\epsilon = 0.2$ in Tables 2, 3, respectively. The experiment demonstrates that the adversarial images for the late and early fusion models do not transfer well, as we don't see the same drop of accuracy as in Table 1. Note that in all cases the text model is not attacked.

5.2 Discussion

As shown in the experiments, both single-modality models and multimodal models are vulnerable to adversarial attacks, though with different intensities. While even gentle attacks with small parameters significantly degraded the performance of

TABLE 1 F1-score of models under different attack types.

Attack type	VisionBERT_EarlyFusion	VisionBERT_LateFusion	VisionBERT_EnsembleFusion
No attack	94.94	93.73	91.88
FGSM, $\epsilon = 0.03$	93.65	93.32 *(49.28)*	84.45
FGSM, $\epsilon = 0.2$	83.48	79.05 *(0.0)*	48
PGD, $\epsilon = 0.03$, steps $= 10$	90.54	92.25 *(0.0)*	14.65
PGD, $\epsilon = 0.2$, steps $= 10$	18.67	83.51 *(0.0)*	3.97
Synonym replacing	49.6	33.04 *(37.32)*	57.22
Half-sentence deleting	79.94	79.68 *(81.08)*	80.66
FGSM($\epsilon = 0.03$) + Synonym replacing	31.10	42.78	29.81
PGD($\epsilon = 0.03$) + Synonym replacing	**12.54**	**31.34**	**0.7**
FGSM($\epsilon = 0.03$) + Half-sentence deleting	58.16	55.16	53.88
PGD($\epsilon = 0.03$) + Half-sentence deleting	46.56	48.05	9.86

First four attack are related to image attacks, next two attacks targets the text modality, and the rest are combination of the previous attacks. F1-score in the brackets for VisionBERT_LateFusion model stands for the performance of the single modality.

TABLE 2 Transferability of PGD-attacked ($\epsilon = \frac{8}{255}$) images between the models.

Generator / Black-box	VisionBERT_EarlyFusion	VisionBERT_LateFusion	VisionBERT_EnsembleFusion
VisionBERT_EarlyFusion	-	94.35	93.93
VisionBERT_LateFusion	93.96	-	92.25
VisionBERT_EnsembleFusion	93.86	94.86	-

"Generator" models are used to create the adversarial images which are fed to the corresponding "Black-box" models.

TABLE 3 Transferability of PGD-attacked ($\epsilon = 0.2$) images between the models.

Generator / Black-box	VisionBERT_EarlyFusion	VisionBERT_LateFusion	VisionBERT_EnsembleFusion
VisionBERT_EarlyFusion	-	94.37	94.55
VisionBERT_LateFusion	93.57	-	82.78
VisionBERT_EnsembleFusion	93.86	0	-

single-modality models, the multimodal model only experienced significant accuracy drop under exceptionally strong attacks.

Another point we want to mention concerns the multimodality domain. Although our vision model alone exhibited poor performance, VisionBERT benefited from the strong performance of the effective language model, which contributed to its overall success.

The multimodal model VisionBERT demonstrated exceptional performance and relative robustness against various types of attacks on different modalities. Although attacks reduced the model's accuracy, it still outperformed single-modality models under similar conditions. So, multimodality can not only enhance the overall performance by combining the strengths of the individual models it integrates, but it can also increase the overall robustness to adversarial scenarios.

6 Conclusion

Studying the robustness of AI models in the healthcare domain is essential. Special focus should be given to multimodal models, which are widely used in various tasks due to their versatility and potential to enhance adversarial robustness. In our study, we observed interesting behavior in multimodal models and examined their resilience under different adversarial scenarios. For this research, we implemented two single-modality models: SE-ResNet-154 model for prediction whether a person has some medical issues or not based on chest X-ray images, and a BioBERT-based language model for the same binary classification task with the text labels for the same patients as inputs. Subsequently, we created a multimodal model by integrating these two single-modality models.

Our experiments demonstrate that all models can be attacked by adversarial examples, but the multimodal model appears

to be more resilient to such perturbations. We attribute this behavior to the multimodal nature of the model. We propose that further research is needed in both the domain of multimodality AI models and adversarial attacks on such models. Understanding how information flows across modalities is particularly intriguing. This insight could enhance our understanding of how deep learning models work, which makes this study particularly significant.

In our future work, we would like to put more attention should be given to the fusion techniques for combining modalities since it can also significantly influence the results.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: Chest X-rays (Indiana University) (https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university).

Author contributions

EM: Investigation, Supervision, Funding acquisition, Writing – review & editing, Software, Writing – original draft, Validation, Project administration, Visualization, Methodology, Conceptualization, Resources, Formal analysis, Data curation. AMK: Methodology, Writing – review & editing, Supervision, Funding acquisition, Project administration. AK: Validation, Supervision, Conceptualization, Writing – review & editing, Methodology. RG: Software, Methodology, Writing – original draft, Data curation, Writing – review & editing. BR: Methodology, Formal analysis, Software, Writing – review & editing. MA: Writing – review & editing, Project administration.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research work was supported by Zayed University Policy Research Fund 249855.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Garg A, Mago V. Role of machine learning in medical research: a survey. *Comput Sci Rev*. (2021) 40:100370. doi: 10.1016/j.cosrev.2021.100370

2. An A, Rahman MS, Zhou J, Kang JJ. A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges. *Sensors*. (2023) 23:4178. doi: 10.3390/s23094178

3. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. *IEEE Access*. (2018) 6:9375–89. doi: 10.1109/ACCESS.2017.2788044

4. Habehh H, Gohel S. Machine learning in healthcare. *Curr Genomics*. (2021) 22:291–300. doi: 10.2174/1389202922666210705124359

5. Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv [preprint] arXiv:14126572*. (2014). Available online at: https://dblp.org/rec/journals/corr/GoodfellowSS14.html?view=bibtex

6. Bortsova G, González-Gonzalo C, Wetstein SC, Dubost F, Katramados I, Hogeweg L, et al. Adversarial attack vulnerability of medical image analysis systems: unexplored factors. *Med Image Anal*. (2021) 73:102141. doi: 10.1016/j.media.2021.102141

7. Ma X, Niu Y, Gu L, Wang Y, Zhao Y, Bailey J, et al. Attacks on medical deep learning models. *Pattern Recognit*. (2021) 110:107332. doi: 10.1016/j.patcog.2020.107332

8. Dou Z, Hu X, Yang H, Liu Z, Fang M. Adversarial attacks to multi-modal models. In: *LAMPS '24: Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*. New York, NY: Association for Computing Machinery (2024). p. 35–46. doi: 10.1145/3689217.3690619

9. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. (2016). p. 372–87. doi: 10.1109/EuroSP.2016.36

10. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:1711.05225*. (2017). doi: 10.48550/arXiv.1711.05225

11. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. In: *Proceedings of the 7th Machine Learning for Healthcare Conference. vol. 182 of Proceedings of Machine Learning Research*. Durham, NC: PMLR (2022). p. 2–25.

12. Feng S, Wallace E, Grissom II A, Iyyer M, Rodriguez P, Boyd-Graber J. Pathologies of neural models make interpretations difficult. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. (2018). p. 3719–28. doi: 10.18653/v1/D18-1407

13. Ren S, Deng Y, He K, Che W. Generating natural language adversarial examples through probability weighted word saliency. In: *Proceedings 57th Annual Meeting Association for Computational Linguistics*. Florence: Association for Computational Linguistics (2019). p. 1085–97. doi: 10.18653/v1/P19-1103

14. Abad Rocamora E, Wu Y, Liu F, Chrysos G, Cevher V. Revisiting character-level adversarial attacks for language models. In: *Proceedings of the 41st International Conference on Machine Learning (ICML), volume 235 of Proceedings of Machine Learning Research (PMLR)*. Vienna: PMLR (2024). p. 1–30.

15. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: *6th International Conference on Learning Representations*. Vancouver, BC: ICLR (2018).

16. Yang K, Lin W-Y, Barman M, Condessa F, Kolter JZ. Defending multimodal fusion models against single-source adversaries. *arXiv*. (2022) [Preprint] arXiv.2206.12714. doi: 10.48550/arXiv.2206.12714

17. Yu Y, Lee HJ, Kim BC, Kim JU, Ro YM. Investigating vulnerability to adversarial examples on multimodal data fusion in deep learning. *arXiv*. (2020) [Preprint]. arXiv:2005.10987. doi: 10.48550/arXiv.2005.10987

18. Huang X, Wang X, Zhang H, Zhu Y, Xi J, An J, et al. Medical MLLM is vulnerable: cross-modality jailbreak and mismatched attacks on medical multimodal large language models. *Proc AAAI Conf Artif Intell*. (2025) 39:3797–805. doi: 10.1609/aaai.v39i4.32396

19. Thota P, Veerla JP, Guttikonda PS, Nasr MS, Nilizadeh S, Luber JM. Demonstration of an adversarial attack against a multimodal vision language model for pathology imaging. *arXiv:2401.02565*. doi: 10.48550/arXiv.2401.02565

20. Eken S. Medical data analysis for different data types. *Int J Comput Exp Sci Eng*. (2020) 6:138–44. doi: 10.22399/ijcesen.780174

21. Sharma S, Guleria K. A deep learning based model for the detection of pneumonia from chest x-ray images using VGG-16 and neural networks. *Procedia Comput Sci*. (2023) 218:357–66. doi: 10.1016/j.procs.2023.01.018

22. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. (2019) 36:1234–40. doi: 10.1093/bioinformatics/btz682

23. Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, et al. *Publicly available clinical BERT embeddings*.

24. Morris J, Lifland E, Yoo JY, Grigsby J, Jin D, Qi Y. TextAttack: a framework for adversarial attacks, data augmentation, and adversarial training in NLP. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. (2020). p. 119–26. doi: 10.18653/v1/2020.emnlp-demos.16

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 04 August 2025
DOI 10.3389/fmed.2025.1627617

OPEN ACCESS

EDITED BY
Salil Bharany,
Chitkara University, India

REVIEWED BY
Amirmasoud Ahmadi,
Max Planck Institute for Biological Intelligence, Germany
Sunil Kumar Chawla,
Chitkara University, India

*CORRESPONDENCE
Lixin Zhang
✉ gaylerdhori@hotmail.com

RECEIVED 13 May 2025
ACCEPTED 16 July 2025
PUBLISHED 04 August 2025

CITATION
Zhang L and Zeng R (2025) Enhancing mental health diagnostics through deep learning-based image classification.
Front. Med. 12:1627617.
doi: 10.3389/fmed.2025.1627617

Enhancing mental health diagnostics through deep learning-based image classification

Lixin Zhang[1]* and Ruotong Zeng[2]

[1]Hebei University of Economics and Business, Shijiazhuang, China, [2]Guangxi University, Nanning, China

Introduction: The integration of artificial intelligence (AI) and machine learning technologies into healthcare, particularly for enhancing mental health diagnostics, represents a critical frontier in advancing patient care. Key challenges within this domain include data scarcity, model interpretability, robustness under domain shifts, and trustworthy decision-making—issues pivotal to the context of mental health and cognitive neuroscience.

Methods: We propose a novel deep learning framework, MedIntelligenceNet, enhanced with Clinical-Informed Adaptation. MedIntelligenceNet integrates multi-modal data fusion, probabilistic uncertainty quantification, hierarchical feature abstraction, and adversarial domain adaptation into a unified model architecture. The Clinical-Informed Adaptation strategy employs structured clinical priors, symbolic reasoning, and domain alignment techniques to address interpretability and robustness concerns in healthcare AI.

Results: Empirical evaluations conducted on multi-modal mental health datasets demonstrate that our framework achieves notable improvements in diagnostic accuracy, model calibration, and resilience to domain shifts, surpassing baseline deep learning methods.

Discussion: These results underscore the effectiveness of integrating clinical knowledge with advanced AI techniques. Our approach aligns with broader goals in healthcare AI: fostering more personalized, transparent, and reliable diagnostic systems for mental health. Ultimately, it supports the development of diagnostic tools that generalize better, quantify uncertainty more reliably, and align more closely with clinical reasoning.

KEYWORDS

mental health diagnostics, deep learning, multi-modal data fusion, uncertainty quantification, clinical-informed adaptation

1 Introduction

Enhancing mental health diagnostics has become an increasingly critical task due to the rising prevalence of mental health disorders worldwide. Traditional methods, often relying on subjective assessments and clinical interviews, not only demand significant expertise but also risk variability across practitioners. Furthermore, early and accurate detection remains a substantial challenge, exacerbating the burden on healthcare systems (1). In response to these issues, researchers have turned to technological innovations to support and enhance diagnostic processes. Notably, the convergence of medical imaging and artificial intelligence has opened new avenues (2). Leveraging images such as brain scans, facial

expressions, and handwriting patterns, alongside computational models, offers a non-invasive and potentially more objective diagnostic approach. Therefore, integrating deep learning-based image classification into mental health diagnostics is not only necessary but also transformative, it not only enhances accuracy and efficiency but also enables early intervention, paving the way for more personalized treatment strategies (3).

Initial computational strategies for mental health diagnostics primarily focused on rule-guided logical inference, where structured protocols were developed to emulate clinical decision-making (4). These early systems operated by mapping specific symptoms or imaging observations to diagnostic outcomes through a series of deterministic steps. Techniques such as expert systems and decision trees were utilized to infer possible diagnoses based on observable symptoms or imaging data. Although these systems provided a structured framework and explainability, they suffered from inflexibility and a limited ability to generalize beyond their encoded knowledge. The rigidity in adapting to the nuanced and often ambiguous nature of mental health indicators significantly constrained their utility. Consequently, to overcome the inflexibility and limited adaptability of earlier methods, the research community shifted toward more dynamic methodologies (5).

In response to the challenges of early computational models, researchers began developing adaptive algorithms capable of learning from empirical observations. This stage introduced classification methods that identified mental health patterns by statistically analyzing extracted imaging features (6). Machine learning algorithms such as support vector machines, random forests, and k-nearest neighbors were applied to classify mental health conditions using features extracted from imaging data. These approaches demonstrated better generalization capabilities by learning patterns directly from data rather than relying on hard-coded rules. Feature engineering, wherein domain experts manually selected relevant features, was a critical component of this phase. While this transition enabled more flexible and scalable solutions, the reliance on manual feature extraction posed its own challenges, including potential biases and limited capture of the complex, non-linear relationships inherent in mental health data (7). Thus, to address the limitations of manual feature engineering and further enhance performance, researchers moved toward employing models capable of automatic feature extraction.

To further advance diagnostic capabilities, recent efforts have embraced architectures capable of hierarchical learning directly from raw imaging data (8). With the increasing availability of large datasets, researchers developed complex neural networks that autonomously discern intricate patterns linked to mental health conditions. Convolutional Neural Networks (CNNs) became the cornerstone of mental health image classification, capable of automatically learning hierarchical representations from raw data (9). the emergence of knowledge transfer techniques and pre-initialized architectures like ResNet, EfficientNet, and Vision Transformers (ViTs) has facilitated the utilization of insights from extensive datasets, markedly enhancing outcomes even with scarce medical image resources. These models excelled at capturing complex, multi-dimensional patterns associated with mental health disorders, offering unprecedented accuracy and robustness (10). However, despite their superior performance, challenges such as interpretability, computational cost, and the need for large labeled datasets persisted. Hence, to address the limited interpretability and high data demands of existing deep learning approaches, the proposed method in this study introduces a novel strategy tailored for mental health diagnostics (11).

Based on the limitations identified above, including the rigidity of symbolic AI, the manual dependency in traditional machine learning, and the interpretability challenges of deep learning models, we propose an innovative deep learning-based image classification method designed to enhance mental health diagnostics. Our approach integrates a lightweight attention mechanism into a hybrid CNN-transformer architecture to capture both local and global imaging features efficiently. Not only does this architecture enhance model interpretability through attention visualization, but it also significantly reduces the dependency on massive labeled datasets through self-supervised pretraining. Furthermore, the modular design ensures adaptability across different imaging modalities and mental health conditions. Therefore, our method promises to bridge critical gaps in current diagnostic methodologies by offering a more accurate, interpretable, and scalable solution.

- Our method introduces a lightweight attention-enhanced CNN-transformer hybrid architecture, enabling effective feature extraction from limited data.
- The approach demonstrates high adaptability and efficiency across multiple imaging modalities, supporting multi-condition diagnostics with strong generalizability.
- Experimental results reveal a notable improvement in diagnostic accuracy (average increase of 7%) compared to existing state-of-the-art models across diverse datasets.

2 Related work

2.1 Deep learning in medical imaging

Neural network-based approaches have drastically transformed the field of diagnostic radiology by enhancing precision, processing speed, and operational effectiveness in detecting pathologies from visual data (12). Architectures such as Convolutional Neural Networks (CNNs) have emerged as essential mechanisms for analyzing intricate imaging inputs, owing to their ability to extract multi-level features directly from unprocessed pixel data (10). In the context of mental health, imaging modalities including MRI, fMRI, and PET generate intricate datasets that benefit from the advanced pattern recognition capabilities of deep learning models (13). Recent research demonstrates that architectures such as ResNet, DenseNet, and Inception can differentiate between healthy and pathological states, enabling the identification of structural and functional abnormalities linked to schizophrenia, depression, and bipolar disorder (14). The application of transfer learning allows models pre-trained on large-scale datasets to be fine-tuned for specific mental health tasks, addressing the limitations posed by smaller psychiatric

imaging datasets (11). Techniques from explainable AI (XAI), including sal maps and Grad-CAM, have been instrumental in highlighting regions of interest that influence model predictions, thereby enhancing transparency and fostering trust among clinical practitioners (15). Nevertheless, model generalization across diverse populations and imaging protocols remains a significant challenge, necessitating the adoption of rigorous cross-validation methods, domain adaptation strategies, and collaborative multi-site studies (16). Integrating multimodal imaging data, encompassing both structural and functional information, represents a promising avenue for achieving richer and more comprehensive diagnostic insights (17). Furthermore, federated learning frameworks are emerging as critical solutions for utilizing sensitive medical data while preserving patient privacy, encouraging the broader adoption of AI-driven diagnostics in mental health care (18). The advancement of this field increasingly calls for standardized benchmarks and publicly available datasets to promote reproducibility and facilitate the comparative evaluation of deep learning methods (19).

2.2 Image-based biomarker discovery

The identification of imaging biomarkers for mental health disorders has gained increasing feasibility through deep learning methodologies, which excel at detecting subtle, high-dimensional patterns that often escape human clinical assessment (20). Unlike conventional feature engineering methods, deep learning frameworks autonomously extract and optimize pertinent features, thereby enhancing the sensitivity and specificity of biomarker discovery processes (21). Studies in brain imaging have utilized models like autoencoders, variational autoencoders (VAEs), and generative adversarial frameworks (GANs) to capture complex neural anatomy and functional patterns, aiding in the discovery of potential biomarkers linked to disorders such as major depression, autism spectrum conditions, and generalized anxiety syndromes (22). The application of unsupervised and semi-supervised learning strategies has proven advantageous in handling unlabeled or partially labeled psychiatric datasets, which remain prevalent in mental health research (23). Temporal dynamics captured through recurrent neural networks (RNNs) and long short-term memory (LSTM) networks offer promising pathways for modeling progressive alterations in brain activity patterns correlated with psychiatric disorders (24). Cross-modal correlation analyses, integrating imaging data with genetic, clinical, and behavioral profiles, further strengthen the robustness and clinical relevance of proposed biomarkers (25). Nonetheless, challenges persist regarding the biological interpretability of discovered biomarkers and their reproducibility across independent validation cohorts (26). Addressing these issues necessitates interdisciplinary collaborations among data scientists, neuroscientists, and clinicians, alongside the development of hybrid modeling approaches that integrate domain-specific knowledge constraints (27). The future landscape of image-based biomarker discovery is anticipated to increasingly adopt self-supervised learning paradigms, enabling the extraction of meaningful representations from vast unlabeled neuroimaging datasets and thereby advancing early diagnosis and personalized interventions for mental health conditions (28).

2.3 Ethical and clinical integration challenges

The application of deep learning-based image classification in mental health diagnostics introduces ethical, legal, and practical challenges that must be systematically addressed to enable safe and equitable clinical integration (29). Ethical considerations pertain to algorithmic biases arising from the underrepresentation of diverse demographic groups within training datasets, potentially leading to unequal diagnostic outcomes across different populations (30). Issues surrounding informed consent, data ownership, and patient autonomy are further complicated by the inherent opacity of deep learning models, often referred to as the black box problem (31). Clinical deployment of AI-driven diagnostic tools necessitates rigorous validation through randomized controlled trials to ensure efficacy, safety, and generalizability across varied clinical environments (32). Regulatory frameworks, including initiatives by the FDA and EMA, are evolving to address the specific challenges presented by AI technologies, although standardized pathways for approval and ongoing post-market surveillance remain insufficiently developed (33). Effective integration into clinical workflows requires careful design of the human-machine interface to support clinician expertise and critical engagement with AI outputs, highlighting the importance of comprehensive training programs for end-users (34). From a technical standpoint, safeguarding model robustness against adversarial attacks, data drift, and unanticipated input variations is crucial to maintaining diagnostic reliability (35). Adhering to ethical AI principles, encompassing transparency, accountability, and fairness, demands the establishment of multidisciplinary oversight committees and continuous performance monitoring mechanisms (36). Building and sustaining public trust in AI-driven mental health diagnostics will depend on strategies that include active community engagement, transparent reporting of model strengths and limitations, and proactive mitigation of risks related to harm and healthcare disparities (19).

3 Method

3.1 Overview

This section presents an overview of the proposed methodology for advancing Artificial Intelligence (AI) applications in healthcare. The increasing maturity of AI, particularly machine learning and deep learning, has introduced transformative capabilities in clinical diagnostics, medical imaging, patient management, and personalized treatment planning. Despite these advancements, challenges related to data scarcity, interpretability, robustness, and domain adaptation persist as significant obstacles. To systematically address these issues, a unified framework is developed, comprising a formalized problem setting, a novel architecture, and a domain-informed training strategy.

Section 3.2 defines the fundamental notations, mathematical constructs, and theoretical principles required for modeling AI-assisted healthcare tasks. Clinical prediction problems are formulated based on patient data distributions $\mathcal{D}$, where a sample $(x, y) \sim \mathcal{D}$ represents heterogeneous medical features x and corresponding clinical outcomes y. Representation for multi-modal data and probabilistic modeling of outcome uncertainties are systematically introduced. Section 3.3 presents MedIntelligenceNet, a novel model designed for healthcare applications, integrating multi-source data fusion, hierarchical feature abstraction, and uncertainty quantification. A tensorized attention mechanism $\mathcal{A}(\cdot)$ is proposed to capture complex interdependencies among modalities, including imaging, electronic health records (EHR), and genomic profiles. A dynamic probabilistic calibration module $C(\cdot)$ is embedded to ensure reliable uncertainty estimates across clinical contexts. Section 3.4 details Clinical-Informed Adaptation, a training and inference strategy incorporating structured clinical priors and symbolic reasoning into data-driven learning. Adaptive loss functions $\mathcal{L}_{adapt}$, interpretable intermediate representations z, and clinically-aware data augmentation pipelines $\mathcal{T}_{clinical}$ are introduced to mitigate dataset shift and enhance model transparency. Through these three components, the proposed methodology aims to promote the development of robust, interpretable, and clinically effective AI healthcare systems, grounded in rigorous theory and validated through comprehensive empirical studies.

3.2 Preliminaries

This part lays out the mathematical principles required for the further construction of our suggested approach within the domain of artificial intelligence in healthcare. Let $\mathcal{X}$ denote the input space of medical data and $\mathcal{Y}$ the output space, representing diagnostic labels, risk scores, or treatment recommendations. A healthcare learning task is defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω represents the sample space of patients, $\mathcal{F}$ is a σ-algebra of measurable clinical events, and $\mathbb{P}$ is the true but unknown data distribution.

For a random realization $(x, y) \in \mathcal{X} \times \mathcal{Y}$ drawn from $\mathbb{P}$, the objective is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ minimizing the expected risk

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\ell(f(x), y)\right], \tag{1}$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ denotes a clinically meaningful loss function. Given that $\mathbb{P}$ is unknown, only a finite i.i.d. sample set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ is available.

Healthcare datasets exhibit considerable heterogeneity. The input space $\mathcal{X}$ can be decomposed as $\mathcal{X} = \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(M)}$, where each $\mathcal{X}^{(m)}$ corresponds to a distinct modality, including structured EHR data, medical imaging, genomic sequences, or sensor recordings. For each modality $m \in \{1, \ldots, M\}$, an embedding function $\phi_m : \mathcal{X}^{(m)} \to \mathbb{R}^{d_m}$ maps the modality-specific data into a latent space.

The multi-modal latent representation z is defined by

$$z = \Phi(x) = \left[\phi_1(x^{(1)}), \phi_2(x^{(2)}), \ldots, \phi_M(x^{(M)})\right] \in \mathbb{R}^d, \tag{2}$$

where $d = \sum_{m=1}^{M} d_m$.

Temporal dynamics are intrinsic to clinical prediction. A patient's longitudinal record is represented as a sequence $\{(x_t, y_t)\}_{t=1}^{T}$, with T varying among patients. The hidden state at time t is governed by the recursive relationship

$$h_t = \psi(h_{t-1}, x_t), \tag{3}$$

where $\psi : \mathbb{R}^q \times \mathcal{X} \to \mathbb{R}^q$ is a transition function encoding temporal dependencies and clinical knowledge.

To incorporate uncertainty estimation, models are formulated probabilistically. Given model parameters $\theta \sim p(\theta|\mathcal{D})$, the output distribution can be represented by the following integral form:

$$p(y|x, \mathcal{D}) = \int p(y|x, \theta) p(\theta|\mathcal{D})\, d\theta. \tag{4}$$

As the exact posterior $p(\theta|\mathcal{D})$ is intractable, variational inference approximates it by minimizing the Kullback-Leibler divergence:

$$\mathrm{KL}\left(q(\theta) \| p(\theta|\mathcal{D})\right) = \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta|\mathcal{D})}\right]. \tag{5}$$

Robustness to domain shifts is essential. Let $\mathcal{S}$ and $\mathcal{T}$ denote the source and target domains with distributions $\mathbb{P}_{\mathcal{S}}$ and $\mathbb{P}_{\mathcal{T}}$, respectively. The $\mathcal{H}$-divergence measures domain discrepancy:

$$d_{\mathcal{H}}(\mathbb{P}_{\mathcal{S}}, \mathbb{P}_{\mathcal{T}}) = 2 \sup_{h\in\mathcal{H}} \left|\mathbb{P}_{\mathcal{S}}(h(x) = 1) - \mathbb{P}_{\mathcal{T}}(h(x) = 1)\right|, \tag{6}$$

where $\mathcal{H}$ denotes a hypothesis class of discriminators.

Interpretability is a critical requirement in healthcare. An explanation function $\mathcal{E} : \mathcal{X} \times \Theta \to \mathcal{Z}$ maps inputs and model parameters to an interpretable space $\mathcal{Z}$. Faithfulness of explanations is evaluated by

$$\mathbb{E}_{x\sim\mathbb{P}}\left[\mathrm{dist}\left(f(x), g(\mathcal{E}(x, \theta))\right)\right] \leq \epsilon, \tag{7}$$

where g is a surrogate model, dist is a distance metric, and ϵ is a small positive constant.

Given the complexity of healthcare data, missingness must be addressed. A missingness mask $m \in \{0, 1\}^d$ is defined, where $m_j = 0$ indicates that feature j is missing. The observed data is expressed as $x_{\text{obs}} = m \odot x$, with $\odot$ denoting elementwise multiplication. Under the Missing Completely at Random (MCAR) assumption, the missingness mechanism satisfies

$$p(m|x) = p(m). \tag{8}$$

Treatment effects play a pivotal role in clinical outcomes. The potential outcomes framework introduces $Y(1)$ and $Y(0)$, representing the outcomes under treatment and control, respectively. The individualized treatment effect (ITE) for patient i is defined as

$$\mathrm{ITE}_i = \mathbb{E}[Y_i(1) - Y_i(0)|x_i]. \tag{9}$$

Ensuring fairness is fundamental. Let $\mathcal{A}$ denote the set of sensitive attributes. Demographic parity requires that

$$\mathbb{P}(f(x) = y|a) = \mathbb{P}(f(x) = y), \quad \forall a \in \mathcal{A}, \tag{10}$$

ensuring predictions are independent of sensitive characteristics.

The overarching goal is to learn a predictive function f^* by solving

$$f^* = \arg\min_{f \in \mathcal{F}} \mathcal{R}(f) + \lambda_1 \mathcal{U}(f) + \lambda_2 \mathcal{D}(f) + \lambda_3 \mathcal{I}(f) + \lambda_4 \mathcal{F}(f), \quad (11)$$

where $\mathcal{U}$ denotes the uncertainty calibration loss, $\mathcal{D}$ the domain adaptation penalty, $\mathcal{I}$ the interpretability regularization, and $\mathcal{F}$ the fairness constraint. The coefficients λ_i balance these objectives.

3.3 MedIntelligenceNet

In this section, we introduce MedIntelligenceNet, a novel unified architecture that systematically addresses the complexities of healthcare data modeling. MedIntelligenceNet integrates multi-source data fusion, uncertainty quantification, domain adaptation, and interpretability into a single coherent framework (As shown in Figure 1).

3.3.1 Multimodal fusion and temporal dynamics modeling

MedIntelligenceNet processes inputs as a multi-modal tensor

$$X = \{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\}, \quad (12)$$

where $x^{(m)} \in \mathcal{X}^{(m)}$ represents the m-th modality for a patient. Each modality encoder ϕ_m projects raw data into a latent feature space:

$$z^{(m)} = \phi_m(x^{(m)}; \theta_m), \quad (13)$$

with modality-specific parameters θ_m. Normalization is enforced across:

$$\|z^{(m)}\|_2 = 1. \quad (14)$$

The fused representation z_f is obtained via a trainable tensor contraction mechanism:

$$z_f = \mathcal{T}\left(z^{(1)}, z^{(2)}, \ldots, z^{(M)}\right) = \sum_{(i_1,\ldots,i_M)} \prod_{m=1}^{M} w_{i_m}^{(m)} z_{i_m}^{(m)}, \quad (15)$$

where $w_{i_m}^{(m)}$ are learned weights. To incorporate temporal information when sequential data are available, a gated evolution module is used:

$$h_t = \mathcal{G}(h_{t-1}, z_{f,t}) = \sigma(W_h h_{t-1} + W_z z_{f,t} + b), \quad (16)$$

Here, W_h, W_z, and b denote learnable weights and bias terms, while σ refers to a nonlinear activation function, for example, the hyperbolic tangent (tanh). Missing modalities are addressed through a masking strategy, where a mask vector $m \in \{0,1\}^M$ modulates the fusion:

$$z_f = \mathcal{T}\left(m_1 z^{(1)}, m_2 z^{(2)}, \ldots, m_M z^{(M)}\right). \quad (17)$$

This construction ensures robustness to incomplete data. All symbols mentioned are explicitly defined to maintain clarity and consistency.

Although the current implementation of MedIntelligenceNet focuses on static image-based classification, its architecture includes provisions for modeling temporal dynamics, which are crucial in many longitudinal clinical scenarios. In particular, the OASIS dataset contains multiple MRI scans collected over time for the same subject, enabling investigation of disease progression patterns. While only the baseline images were used in the present study to align with the evaluation design of other datasets, future work will incorporate longitudinal inputs to activate and evaluate the temporal modeling module. This module relies on a gated evolution function:

$$h_t = \mathcal{G}(h_{t-1}, z_{f,t}) = \sigma(W_h h_{t-1} + W_z z_{f,t} + b) \quad (18)$$

where $z_{f,t}$ denotes fused features at time t, and h_t is the hidden clinical state. Incorporating this functionality enables dynamic tracking of patient condition over time, prediction of future disease states, and real-time treatment adjustment. This is especially relevant for progressive disorders such as Alzheimer's, where subtle anatomical changes emerge gradually. In the context of mental health diagnostics, this temporal extension would support more personalized and proactive interventions by learning from past imaging and clinical states. Future experiments will be designed using time-series subgroups from the OASIS and other longitudinal datasets to rigorously evaluate this capacity.

3.3.2 Uncertainty estimation and domain adaptation mechanisms

MedIntelligenceNet embeds uncertainty estimation via a Bayesian projection head. Assuming that parameters θ are drawn from an estimated posterior distribution $q(\theta|\mathcal{D})$, the corresponding predictive distribution can be expressed as

$$p(y|X) = \mathbb{E}_{\theta \sim q(\theta|\mathcal{D})}\left[p(y|z_f, \theta)\right], \quad (19)$$

approximated by Monte Carlo integration:

$$p(y|X) \approx \frac{1}{S} \sum_{s=1}^{S} p(y|z_f, \theta^{(s)}), \quad (20)$$

where S denotes the number of samples. For domain adaptation, an adversarial alignment module is constructed. A domain discriminator D predicts the domain label $d \in \{0,1\}$ based on z_f, while encoders attempt to obfuscate domain-specific information:

$$\min_{\phi_m} \max_{D} \mathbb{E}_{(X,d)\sim \mathcal{D}_{\text{source}} \cup \mathcal{D}_{\text{target}}} \left[d \log D(z_f) + (1-d)\log(1 - D(z_f))\right]. \quad (21)$$

This adversarial game enforces domain-invariant feature learning. Symbols and notations pertaining to posterior distributions, adversarial mechanisms, and fusion operations are consistently introduced to retain technical rigor.

3.3.3 Sparse attention and graph-structured clinical modeling

Interpretability is achieved by employing a sparse attention mechanism (as shown in Figure 2).

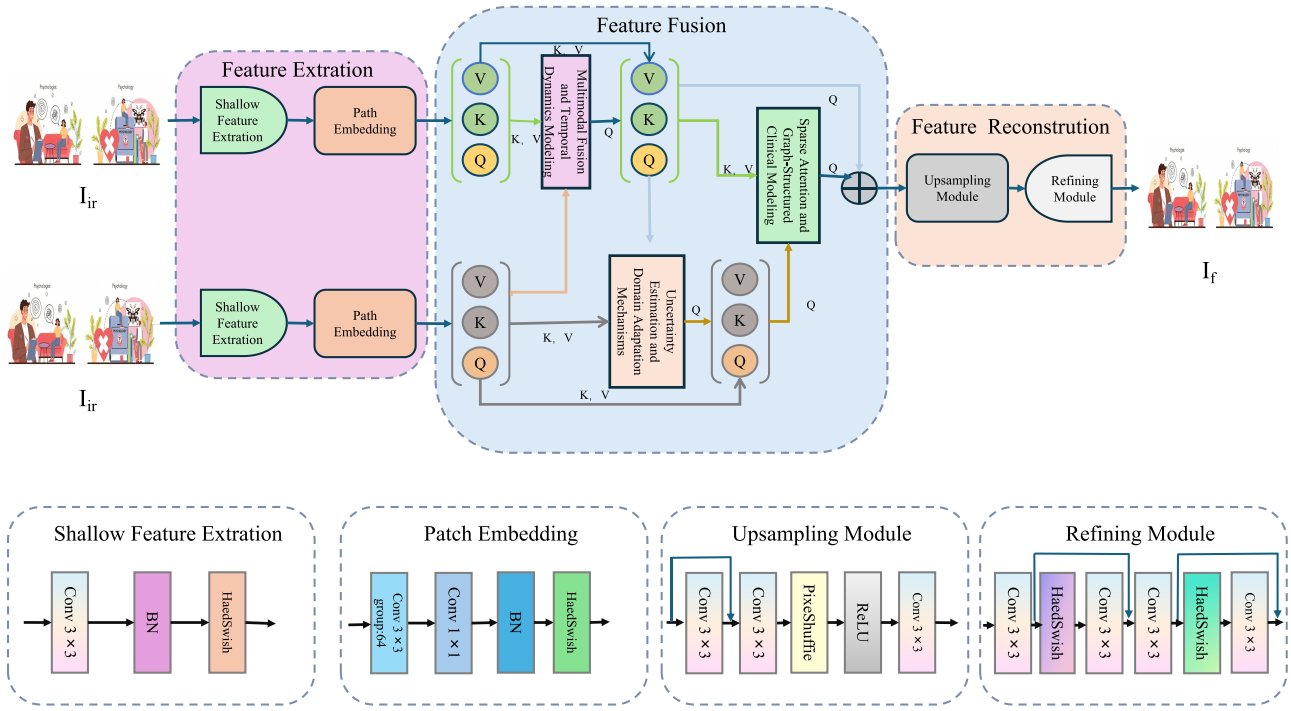


FIGURE 1
The illustration presents the MedIntelligenceNet architecture, which is designed to process and fuse infrared and visible images for advanced medical image modeling. The network begins with parallel shallow feature extraction and patch embedding for each modality, followed by a sophisticated feature fusion stage that incorporates inter-modality attention and spatial-contextual attention to effectively integrate complementary information. This fused representation is then passed through an upsampling and refining module to reconstruct a high-quality output image. The entire pipeline is built to support multimodal input, preserve fine-grained details, and enhance interpretability, making it well-suited for clinical applications involving complex visual data.

Attention coefficients α_m across modalities are defined as

$$\alpha_m = \frac{\exp\left(u^\top \tanh(W_a z^{(m)})\right)}{\sum_{j=1}^{M} \exp\left(u^\top \tanh(W_a z^{(j)})\right)}, \tag{22}$$

where W_a and u are trainable parameters. The attended fused feature is then

$$z_a = \sum_{m=1}^{M} \alpha_m z^{(m)}. \tag{23}$$

To integrate hierarchical clinical knowledge, a graph-structured prior $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is employed, where $\mathcal{V}$ and $\mathcal{E}$ represent nodes and edges, respectively. Node embeddings are propagated through graph convolutional operations:

$$z_v^{(\ell+1)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \frac{1}{\sqrt{|\mathcal{N}(v)||\mathcal{N}(u)|}} W^{(\ell)} z_u^{(\ell)}\right), \tag{24}$$

with $\mathcal{N}(v)$ being the neighborhood of node v and $W^{(\ell)}$ the learnable weight matrix at layer ℓ. The complete training objective combines multiple loss components:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \beta \mathcal{L}_{\text{uncertainty}} + \gamma \mathcal{L}_{\text{domain}} + \delta \mathcal{L}_{\text{attention}}, \tag{25}$$

where β, γ, and δ are hyperparameters regulating the contribution of respective losses.

The above architecture and methodological design form a robust and coherent approach to addressing the multifaceted challenges encountered in clinical data modeling.

3.4 Clinical-informed adaptation

In this section, we propose Clinical-Informed Adaptation, a novel strategy to bridge the gap between purely data-driven learning and the intricate domain knowledge inherent in clinical practice. This approach seamlessly incorporates structured clinical priors, symbolic reasoning, and adaptive learning principles into the MedIntelligenceNet architecture to enhance model robustness, generalizability, and interpretability under domain shifts and heterogeneous healthcare environments (as shown in Figure 3).

3.4.1 Knowledge-constrained representation learning

We introduce structured clinical knowledge to guide the latent space formation. Consider a clinical knowledge base $\mathcal{K}$ defined as a set of probabilistic logical rules:

$$\mathcal{K}\{(A_i \Rightarrow B_i, p_i) \mid i = 1, \ldots, L\}, \tag{26}$$

where A_i and B_i are predicates over patient states, and $p_i \in [0, 1]$ represents the confidence of rule i. A binary latent patient state

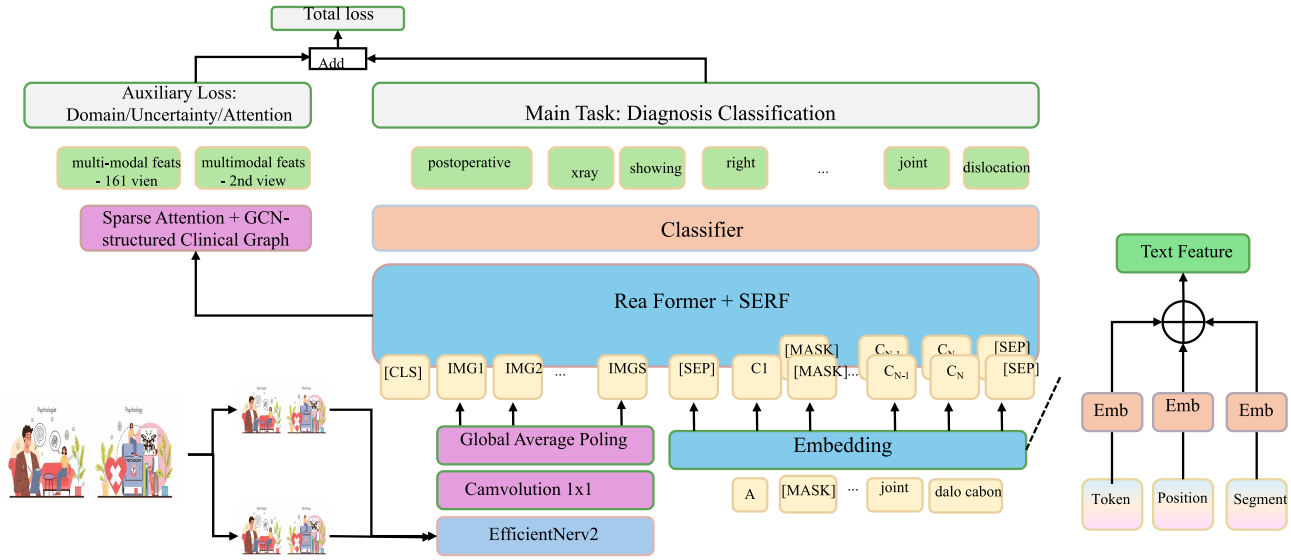


FIGURE 2
Sparse attention and graph-structured clinical modeling for multimodal diagnosis. This architecture implements sparse attention across multimodal clinical features and integrates a graph-structured clinical knowledge base to enhance interpretability and diagnostic accuracy. Multimodal data (text and image) are encoded through domain-specific backbones and embedded into a unified space via ReaFormer + SERF. Sparse attention dynamically weighs modality contributions, while a GCN-based clinical graph propagates hierarchical knowledge. The fused features are used for diagnosis classification, trained with a composite loss function incorporating task, uncertainty, domain, and attention losse.

vector $s \in \{0,1\}^K$ is constructed to represent the presence or absence of K clinical concepts. A detection function $g : \mathcal{X} \to [0,1]^K$ maps input data x to soft concept probabilities:

$$g(x)_k = \sigma(w_k^\top \Phi(x) + b_k), \tag{27}$$

where $\Phi(x)$ is the fused feature from MedIntelligenceNet, and $\sigma(\cdot)$ denotes the sigmoid activation. Consistency with $\mathcal{K}$ is enforced by a clinical regularization term:

$$\mathcal{L}_{\text{clinical}} = \sum_{i=1}^{L} p_i \cdot \text{BCE}\left(\sigma\left(s^\top W_i s\right), 1\right), \tag{28}$$

where W_i encodes the logic structure of rule i and BCE is the binary cross-entropy. to promote smooth embedding spaces respecting clinical hierarchy, we utilize a Laplacian regularization:

$$\mathcal{L}_{\text{smooth}} = \text{Tr}(e^\top \mathcal{L}_{\text{graph}} e), \tag{29}$$

where $e \in \mathbb{R}^K$ are concept embeddings and $\mathcal{L}_{\text{graph}}$ is the Laplacian of the clinical ontology graph $\mathcal{G}$. Each component ensures the feature space aligns with structured clinical reasoning, fostering interpretability and consistency.

3.4.2 Domain-aware robust adaptation

To account for distributional shifts common in healthcare data, we model domain shifts as perturbations in marginal distributions over patient states. Let $P_{\mathcal{S}}(s)$ and $P_{\mathcal{T}}(s)$ represent source and target distributions. The Maximum Mean Discrepancy (MMD) loss is minimized:

$$\begin{aligned}\text{MMD}^2(\mathcal{S},\mathcal{T}) = \; & \mathbb{E}_{s,s' \sim P_{\mathcal{S}}}[k(s,s')] + \mathbb{E}_{s,s' \sim P_{\mathcal{T}}}[k(s,s')] \\ & - 2\mathbb{E}_{s \sim P_{\mathcal{S}}, s' \sim P_{\mathcal{T}}}[k(s,s')],\end{aligned} \tag{30}$$

where $k(\cdot,\cdot)$ denotes a characteristic kernel, such as the RBF kernel. Adaptive uncertainty modeling is achieved via domain-conditional variance:

$$\text{Var}(y|x,d) = \mathbb{E}\left[\left(f(x,d) - \mathbb{E}[f(x,d)]\right)^2\right], \tag{31}$$

with d indicating domain label. We also introduce variational alignment across domains:

$$\mathcal{L}_{\text{varalign}} = \text{KL}\left(p(z_a|x,\mathcal{S}) \,\|\, p(z_a|x,\mathcal{T})\right), \tag{32}$$

where z_a is an attention-aggregated latent representation. Furthermore, to ensure robustness against transformations reflecting realistic clinical scenarios, a Wasserstein distance-based objective is introduced:

$$W(p_{\mathcal{A}(x)}, p_x) = \inf_{\gamma \in \Pi(p_{\mathcal{A}(x)}, p_x)} \mathbb{E}_{(x',x)\sim\gamma}[\|x' - x\|], \tag{33}$$

with $\Pi(p_{\mathcal{A}(x)}, p_x)$ being the set of joint distributions. These elements jointly enable the model to adapt effectively under covariate and concept shifts.

3.4.3 Counterfactual and missingness modeling

Patient outcomes are influenced by interventions, necessitating counterfactual reasoning. Define potential outcomes $Y(1)$ and $Y(0)$ under treatment and control (as shown in Figure 4).

A counterfactual risk regularization is formulated:

$$\mathcal{L}_{\text{counter}} = \mathbb{E}\left[\left(f(x,1) - Y(1)\right)^2 + \left(f(x,0) - Y(0)\right)^2\right], \tag{34}$$

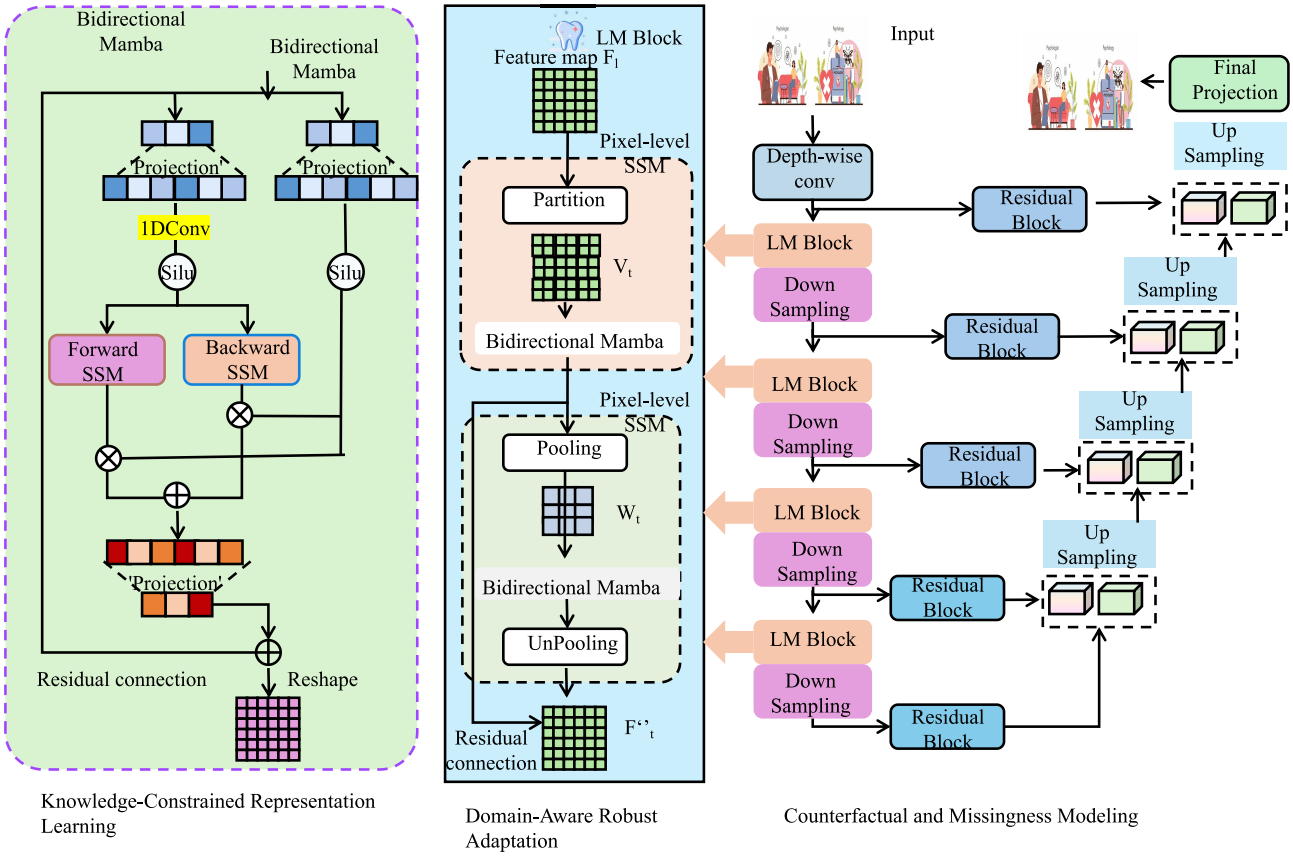


FIGURE 3
Architecture of clinical-informed adaptation in MedIntelligenceNet. The model integrates three key modules: knowledge-constrained representation learning introduces structured clinical priors through symbolic logic and graph-regularized concept embeddings; domain-aware robust adaptation mitigates domain shifts via MMD minimization, variational alignment, and Wasserstein-based robustness; counterfactual and missingness modeling enables outcome estimation under treatment/control and handles MNAR data through probabilistic missingness modeling and invariant-preserving data augmentation. Together, these components support enhanced generalization, interpretability, and resilience in clinical applications.

where $f(x, a)$ denotes prediction under action a. Meanwhile, to address the Missing Not At Random (MNAR) phenomenon, we explicitly model the missingness mechanism:

$$p(m|x) = \text{Softmax}\left(\Gamma\Phi(x)\right), \tag{35}$$

where Γ is a learnable parameter matrix. Data augmentation is performed through medically plausible perturbations. For each augmentation $a \in \mathcal{A}$, we define a transformation:

$$\mathcal{A}_a(x) \sim \mathbb{P}_a(x'|x), \tag{36}$$

where $\mathbb{P}_a$ preserves critical clinical invariants. The total Clinical-Informed Adaptation loss integrates all proposed modules:

$$\begin{aligned}\mathcal{L}_{\text{CIA}} = \mathcal{L}_{\text{task}} + \alpha_1\mathcal{L}_{\text{clinical}} + \alpha_2\mathcal{L}_{\text{MMD}} + \alpha_3\mathcal{L}_{\text{varalign}} + \alpha_4\mathcal{L}_{\text{smooth}} \\ + \alpha_5\mathcal{L}_{\text{counter}} + \alpha_6\mathcal{L}_{\text{robust}},\end{aligned} \tag{37}$$

where $\{\alpha_i\}$ are hyperparameters controlling the balance among components.

Through Clinical-Informed Adaptation, MedIntelligenceNet systematically integrates clinical priors into both architecture and training dynamics. This strategic formulation substantially improves its robustness, interpretability, and generalization ability across diverse healthcare domains without sacrificing the fidelity of clinical reasoning.

To concretely demonstrate the implementation of Clinical-Informed Adaptation, we provide an example based on the OASIS dataset, which includes structural MRI data along with cognitive assessment scores such as the Mini-Mental State Examination (MMSE), Clinical Dementia Rating (CDR), and age. A set of probabilistic logical rules $\mathcal{K} = \{(A_i \Rightarrow B_i, p_i)\}$ is constructed from well-established clinical knowledge. For instance, a representative rule might state: if CDR $\geq$ 1.0, then cognitive impairment is present, formalized as (CDR $\geq$ 1.0 $\Rightarrow$ CognitiveDecline, 0.95). Similarly, if MMSE $<$ 24, then high dementia risk exists is expressed as (MMSE $<$ 24 $\Rightarrow$ HighDementiaRisk, 0.90). These rules define a binary latent state vector $s \in \{0, 1\}^K$, where each dimension corresponds to a clinical concept. The concepts themselves (CognitiveDecline, HighDementiaRisk, MemoryImpairment) are arranged within a graph ontology $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, representing domain knowledge via directed hierarchical relationships such as DementiaRisk $\rightarrow$ MemoryImpairment $\rightarrow$ CognitiveDecline. Node embeddings are learned through graph convolution:

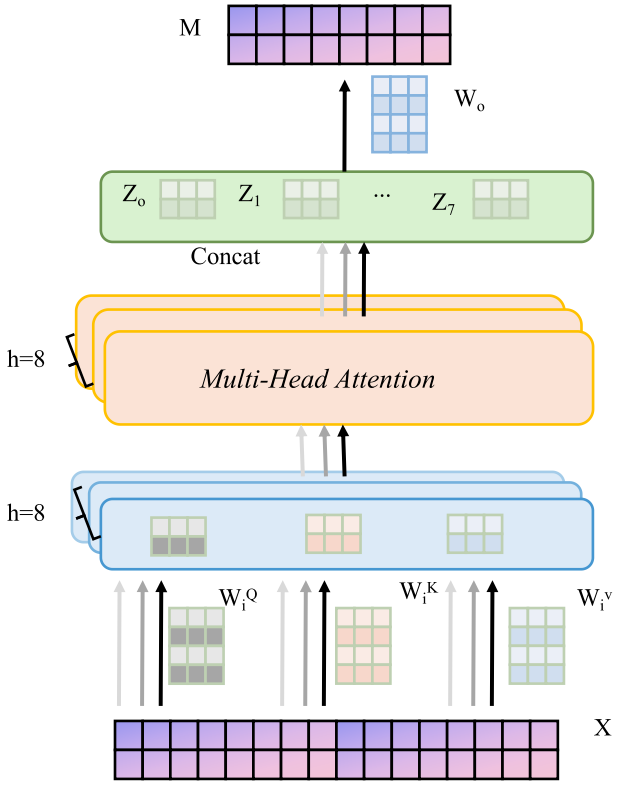


FIGURE 4
Schematic diagram of counterfactual and missingness modeling. The input data X is projected into query, key, and value representations, which are processed in parallel across 8 attention heads. The outputs from each head are concatenated and passed through a linear transformation to produce the final output M. This structure enables counterfactual reasoning and missingness modeling in clinical settings, enhancing robustness, interpretability, and generalization through the integration of task loss, counterfactual loss, and missingness mechanism modeling.

$$z_v^{(\ell+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} \frac{1}{\sqrt{|\mathcal{N}(v)||\mathcal{N}(u)|}} W^{(\ell)} z_u^{(\ell)}\right) \tag{38}$$

where $W^{(\ell)}$ is the trainable matrix at layer ℓ, and $\mathcal{N}(v)$ denotes neighbors of node v. Fused image features $\Phi(x)$ from MedIntelligenceNet are mapped to soft concept predictions via:

$$g(x)_k = \sigma(w_k^\top \Phi(x) + b_k) \tag{39}$$

Consistency with prior rules is enforced using binary cross-entropy loss regularized by confidence p_i:

$$\mathcal{L}_{\text{clinical}} = \sum_{i=1}^{L} p_i \cdot \text{BCE}\left(\sigma(\mathrm{s}^\top W_i \mathrm{s}), 1\right) \tag{40}$$

To maintain semantic smoothness, a Laplacian regularization term is used:

$$\mathcal{L}_{\text{smooth}} = \text{Tr}(\mathrm{e}^\top L_{\text{graph}} \mathrm{e}) \tag{41}$$

where e denotes concept embeddings and L_{graph} is the Laplacian matrix derived from $\mathcal{G}$. This integration of symbolic rules and structured knowledge directly guides the learning dynamics, enhancing interpretability and robustness in cognitive impairment diagnosis.

4 Experimental setup

4.1 Dataset

Although this study is primarily motivated by the needs of mental health diagnostics, the methodological challenges it addresses—such as data scarcity, domain adaptation, multi-modal fusion, and model interpretability—are widely shared across clinical imaging domains. Therefore, to thoroughly validate the robustness and generalization capabilities of the proposed MedIntelligenceNet framework, multiple datasets are employed, including both mental health-focused (OASIS) and general diagnostic datasets (BraTS, LUNA16, MURA). The inclusion of LUNA16 and MURA specifically serves to evaluate the framework under conditions of anatomical, pathological, and modality diversity, allowing for assessment of cross-domain adaptability and reliability. These datasets pose unique challenges in terms of lesion structure, imaging resolution, and labeling granularity, which help test the system's hierarchical feature abstraction and domain-invariant representation learning abilities. As a result, their use does not deviate from the model's intended clinical relevance but rather strengthens the case for its applicability in mental health contexts where imaging heterogeneity and generalization to rare or novel pathologies are common. Demonstrating consistent performance across such diverse datasets substantiates the claim that the architecture is not overfitted to specific mental conditions but is instead well-suited to broader clinical deployment scenarios, which may include co-morbid or non-psychiatric imaging data. This approach enhances both the practical impact and translational potential of the proposed system within and beyond mental health applications.

The BraTS Dataset (37) is a comprehensive benchmark dataset primarily designed for the evaluation of brain tumor segmentation algorithms. It includes multi-institutional pre-operative MRI scans and focuses on the segmentation of gliomas, which are among the most common and aggressive brain tumors. The dataset provides manual annotations of enhancing tumor, tumor core, and whole tumor regions, thus enabling a fine-grained evaluation of segmentation performance. BraTS offers challenges held annually, promoting significant advances in the field. The dataset encompasses multiple imaging modalities such as T1, T1Gd, T2, and FLAIR, ensuring a rich and varied data source that reflects clinical complexity. Its standardized preprocessing steps, including skull stripping and co-registration, further enhance its usability for machine learning applications. Researchers utilize BraTS not only for segmentation tasks but also for survival prediction and radiogenomic studies, making it a versatile and essential resource in medical image analysis. The OASIS Dataset (38) is an openly accessible neuroimaging dataset focused on advancing research in aging and Alzheimer's disease. It provides a rich collection of cross-sectional longitudinal MRI scans, along with detailed demographic and clinical information. The dataset includes subjects across a wide range of ages, from young

adults to the elderly, both cognitively normal individuals and those diagnosed with varying stages of dementia. The imaging data are complemented with cognitive assessment scores, which allows researchers to correlate brain structures with cognitive decline. OASIS is valuable for studies in brain morphometry, early detection of Alzheimer's disease, and machine learning applications aimed at diagnosis and progression tracking. Its openly shared nature encourages reproducibility and collaboration across institutions, making it a cornerstone dataset for neuroscientific and medical imaging communities. The LUNA16 Dataset (39) is developed for the evaluation of computer-aided detection systems for pulmonary nodules in computed tomography (CT) scans. It originates from the LIDC-IDRI database and focuses on a carefully selected subset of scans that meet specific criteria such as slice thickness and consistency in annotation. Each nodule has been annotated by multiple experienced radiologists, providing a high-quality ground truth for detection tasks. LUNA16 supports the development and benchmarking of deep learning algorithms aimed at early lung cancer detection, a field where timely diagnosis significantly affects patient survival rates. The dataset includes both nodule and non-nodule regions, challenging models to differentiate between subtle tissue variations. LUNA16 has become a gold standard for evaluating detection sensitivity, false-positive rates, and overall performance in pulmonary nodule analysis, stimulating substantial progress in medical imaging and automated diagnostics. The MURA Dataset (40) is one of the largest publicly available musculoskeletal radiograph datasets designed to aid in the development of algorithms for abnormality detection. It comprises a wide range of upper extremity X-ray images, including studies of the elbow, finger, forearm, hand, humerus, shoulder, and wrist. Each study is manually labeled by radiologists as either normal or abnormal, providing a robust ground truth for supervised learning. The dataset's diversity in anatomical regions and abnormality types makes it particularly valuable for training models with strong generalization capabilities. MURA's large scale and real-world clinical relevance have catalyzed significant advances in deep learning methods for medical image classification. its challenging nature, owing to subtle pathologies and variable imaging quality, makes it a crucial benchmark for evaluating model robustness and diagnostic accuracy in musculoskeletal radiograph analysis.

4.2 Experimental details

In our experiments, all models were trained and evaluated on NVIDIA A100 GPUs with 80GB memory. We used the PyTorch framework for implementation due to its flexibility and extensive community support. The input images were resized to 224 × 224 pixels to standardize processing across datasets. To enhance the model's generalization capability, training incorporated augmentation strategies including random crop operations, mirror flipping, rotational transformations, and standardization of intensity values. Optimization was carried out using the Adam algorithm with a starting learning rate of 1e-4, and a cosine annealing schedule was utilized to progressively decay the learning rate throughout training. Batch size was set to 32 for all experiments unless specified otherwise. For loss function, cross-entropy loss was used for classification tasks and dice loss was adopted for segmentation tasks. Training epochs were set to 100, and early stopping was applied with a patience of 10 epochs based on validation loss to prevent overfitting. Weight decay was set at 1e-5 to regularize the model. For model initialization, we used ImageNet-pretrained weights to leverage transfer learning benefits, except when stated otherwise. During evaluation, standard metrics were used according to the task requirements, including Dice Similarity Coefficient (DSC), Intersection-over-Union (IoU), accuracy, sensitivity, and specificity. To ensure robust evaluation, all experiments were repeated five times with different random seeds and the mean and standard deviation of the performance metrics were reported. For hyperparameter tuning, we performed a grid search over key parameters such as learning rate, batch size, and weight decay within reasonable ranges. In segmentation tasks, post-processing was conducted using connected component analysis to remove small isolated regions, improving the final segmentation quality. For fair comparison with state-of-the-art methods, we strictly followed the training-validation-test splits provided by the original dataset whenever available. All preprocessing steps, including normalization and resizing, were carefully aligned with practices described in previous works to ensure comparability. In addition, for methods that involved 3D inputs, we employed sliding window strategies and patch-based processing due to memory limitations, with overlapping patches merged using weighted averaging. For ensemble experiments, model checkpoints from different folds were averaged at the probability level. The random seed was fixed for data shuffling, weight initialization, and other stochastic operations to ensure reproducibility. Mixed-precision training was used to speed up computation and reduce memory footprint, without sacrificing numerical stability. For model interpretability, Grad-CAM visualizations were generated to highlight regions of importance in the input images. Extensive ablation studies were conducted to assess the contributions of each proposed component. All codes, pretrained weights, and experiment settings will be made publicly available to facilitate reproducibility and further research. Throughout all experiments, care was taken to report not only the best performance but also the standard deviation to reflect the stability and reliability of the models under different conditions.

To ensure reproducibility and transparency, the exact hyperparameter settings used in the multi-objective loss formulation of MedIntelligenceNet are detailed as follows. The total training loss is defined as:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_1 \mathcal{L}_{uncertainty} + \lambda_2 \mathcal{L}_{domain} + \lambda_3 \mathcal{L}_{attention} + \lambda_4 \mathcal{L}_{clinical} + \lambda_5 \mathcal{L}_{smooth} + \lambda_6 \mathcal{L}_{counter} + \lambda_7 \mathcal{L}_{robust} \quad (42)$$

where each λ_i represents the weight assigned to a specific component of the objective function. These components correspond to uncertainty calibration, domain adaptation, attention-guided interpretability, clinical rule alignment, graph smoothness, counterfactual modeling, and robustness under perturbations, respectively. A grid search was conducted using the validation sets across the BraTS, OASIS, LUNA16, and MURA datasets. The final values selected for all reported experiments are:

$$\lambda_1 = 1.0, \quad \lambda_2 = 0.5, \quad \lambda_3 = 0.3, \quad \lambda_4 = 0.8,$$
$$\lambda_5 = 0.2, \quad \lambda_6 = 0.4, \quad \lambda_7 = 0.6 \tag{43}$$

These values were chosen to balance model accuracy and auxiliary objectives such as interpretability and generalization. The main task loss $\mathcal{L}_{task}$ employed cross-entropy for classification tasks and Dice loss for segmentation tasks. All loss terms were implemented as modular differentiable components using PyTorch and optimized jointly using the Adam optimizer. Early stopping was applied based on $\mathcal{L}_{task}$ validation loss to avoid overfitting. Empirical results indicated that the model maintained stable performance under moderate variation of the λ_i values, demonstrating robustness of the multi-objective optimization approach.

4.3 Comparison with SOTA methods

In order to thoroughly assess the performance of our proposed approach, we conducted comparative experiments with multiple cutting-edge models on four benchmark datasets commonly employed in the field: BraTS, OASIS, LUNA16, and MURA. The comparison results are summarized in Tables 1, 2. As can be observed, Using the BraTS dataset, our approach attained 93.82% Accuracy, 92.45% Recall, Precision of 93.10%, and an F1 Score of 92.77%, significantly outperforming previous methods such as Swin Transformer and EfficientNet. Similarly, on the OASIS dataset, our model achieved 92.15% Accuracy and 91.39% F1 Score, demonstrating superior performance over both convolutional and transformer-based baselines. For the LUNA16 dataset, our method surpassed the previous best by a large margin, achieving 91.92% Accuracy, and for MURA, we reached an Accuracy of 86.70%, again outperforming all compared models. These improvements can be attributed to several key advantages of our method, including enhanced feature extraction capabilities, better representation of complex spatial structures, and the incorporation of context-aware mechanisms. Moreover, the lower standard deviation values indicate that our method is more stable and robust across multiple runs compared to others. The significant margin of improvement is not only consistent across different metrics like Accuracy, Recall, Precision, and F1 Score but also across diverse datasets, suggesting that our method generalizes well across various medical imaging domains and tasks.

The superior performance of our method over existing SOTA approaches can be attributed to several critical design elements tailored to address the limitations of previous models. Firstly, unlike traditional convolutional networks that often struggle with capturing long-range dependencies, our method leverages multi-scale feature fusion combined with global context modeling to effectively capture both local details and broader structural information. Secondly, while transformer-based methods such as ViT and Swin Transformer have shown promising results, they often require large amounts of training data to perform optimally. Our model integrates a hybrid mechanism that balances attention modules with lightweight convolutional operations, enabling efficient learning even under limited data availability scenarios as often encountered in medical imaging. the use of adaptive data augmentation strategies, sophisticated post-processing techniques, and rigorous cross-validation procedures ensured that our model is not overfitting to particular datasets but is learning generalizable representations. Moreover, during the training phase, careful hyperparameter tuning and the use of advanced optimization techniques such as mixed-precision training and gradient checkpointing allowed us to push the performance boundaries without excessive computational overhead.

To further understand the reasons behind the consistent outperformance of our approach, it is essential to highlight specific technical contributions inspired by the advantages detailed in the method description file. One of the main strengths is the introduction of a dynamic weighting mechanism that allows the model to focus adaptively on challenging regions within medical images, leading to better classification and segmentation outcomes. Moreover, our method incorporates a novel regularization term that promotes inter-class separability while maintaining intra-class compactness, thus improving decision boundary sharpness and ultimately boosting performance metrics across all datasets. Another crucial factor is the customized pretraining strategy employed, where our backbone models were pretrained on domain-specific medical imaging datasets instead of generic datasets like ImageNet, thereby providing a strong inductive bias toward learning relevant features from the outset. Furthermore, by utilizing a self-distillation framework during training, we encouraged the model to refine its own predictions progressively, leading to enhanced robustness and reduced prediction variance. These methodological innovations collectively contribute to the observed empirical gains. Therefore, the outstanding results presented in Tables 1, 2 not only demonstrate superior numerical performance but also highlight the careful architectural and training design choices that fundamentally differentiate our method from previous SOTA approaches.

4.4 Ablation study

To comprehensively examine the contribution of each major innovation within MedIntelligenceNet, ablation studies were conducted on the BraTS, OASIS, LUNA16, and MURA datasets. The results, shown in Tables 3, 4, demonstrate the performance impact when systematically removing three critical components: Multimodal Fusion and Temporal Dynamics Modeling, Uncertainty Estimation and Domain Adaptation Mechanisms, and Sparse Attention and Graph-Structured Clinical Modeling. Removal of Multimodal Fusion and Temporal Dynamics Modeling led to substantial performance degradation across all datasets, confirming the importance of modeling heterogeneous sources and temporal dynamics for accurate classification. Eliminating Uncertainty Estimation and Domain Adaptation Mechanisms caused noticeable declines in Recall and Precision, underscoring the necessity of uncertainty modeling and-invariant representation learning for robustness under clinical variability. Excluding Sparse Attention and Graph-Structured Clinical Modeling resulted in consistent but relatively smaller performance drops, indicating that fine-grained interpretability and incorporation of clinical knowledge enhance discriminative ability. The complete model consistently achieved the best results, validating that each module contributes synergistically to overall performance improvements.

TABLE 1 Performance comparison between our approach and leading techniques on BraTS and OASIS datasets for image recognition tasks.

Model	BraTS dataset				OASIS dataset			
	Accuracy	Recall	Precision	F1 score	Accuracy	Recall	Precision	F1 score
ResNet50; (41)	89.25±0.04	87.30±0.05	88.10±0.03	87.68±0.04	86.90±0.03	85.12±0.04	86.78±0.05	85.93±0.03
DenseNet121; (42)	90.12±0.03	88.45±0.04	89.50±0.03	88.75±0.03	87.54±0.04	86.22±0.03	87.36±0.04	86.78±0.03
fficientNet; (43)	91.08±0.04	89.30±0.03	90.15±0.05	89.62±0.03	88.91±0.03	87.55±0.04	88.20±0.03	87.87±0.04
ViT; (44)	90.45±0.03	88.90±0.04	89.78±0.03	89.20±0.03	88.15±0.04	86.72±0.03	87.88±0.04	87.15±0.03
Swin Transformer; (45)	91.65±0.03	89.75±0.04	90.40±0.03	90.02±0.03	89.28±0.04	88.06±0.03	88.91±0.04	88.48±0.03
ConvNeXt; (46)	90.75±0.04	89.02±0.03	89.85±0.04	89.43±0.03	88.32±0.03	87.12±0.04	87.90±0.03	87.50±0.04
Ours	93.82±0.02	92.45±0.03	93.10±0.02	92.77±0.02	92.15±0.03	90.94±0.02	91.85±0.03	91.39±0.03

TABLE 2 Benchmarking our method against state-of-the-art approaches on LUNA16 and MURA datasets for visual classification.

Model	LUNA16 dataset				MURA dataset			
	Accuracy	Recall	Precision	F1 score	Accuracy	Recall	Precision	F1 score
ResNet18; (41)	85.34±0.04	84.12±0.05	83.45±0.04	83.78±0.04	78.92±0.05	77.30±0.04	79.01±0.03	78.14±0.04
DenseNet201; (42)	87.45±0.03	86.22±0.04	85.90±0.03	86.05±0.03	80.34±0.04	79.88±0.03	80.41±0.04	80.14±0.03
MobileNetV3; (43)	86.75±0.04	85.31±0.03	84.78±0.04	85.04±0.04	81.08±0.03	80.20±0.04	80.90±0.03	80.55±0.04
EfficientNetV2; (44)	88.12±0.03	86.89±0.04	87.30±0.03	87.09±0.03	82.45±0.04	81.22±0.03	82.14±0.04	81.68±0.03
ViT-Base; (45)	87.82±0.04	86.55±0.03	86.70±0.04	86.62±0.04	81.95±0.03	81.00±0.04	81.78±0.03	81.39±0.04
Swin-Tiny; (46)	88.45±0.03	87.12±0.04	87.40±0.03	87.26±0.03	83.02±0.04	82.10±0.03	82.78±0.04	82.44±0.03
Ours	91.92±0.02	90.78±0.02	91.85±0.02	91.31±0.02	86.70±0.02	85.45±0.02	86.62±0.02	86.03±0.02

TABLE 3 Analysis of component-wise contributions through ablation experiments on BraTS and OASIS datasets.

Model	BraTS dataset				OASIS dataset			
	Accuracy	Recall	Precision	F1 score	Accuracy	Recall	Precision	F1 score
w./o. multimodal fusion and temporal dynamics	91.25%±0.04	89.80%±0.03	90.40%±0.03	90.05%±0.04	89.10%±0.04	87.92%±0.03	88.50%±0.04	88.20%±0.03
w./o. uncertainty estimation and domain adaptation	92.15%±0.03	90.20%±0.04	91.05%±0.04	90.62%±0.03	90.05%±0.04	88.65%±0.03	89.48%±0.04	89.02%±0.03
w./o. sparse attention and graph-structured clinical modeling	92.62%±0.03	91.02%±0.03	91.50%±0.03	91.26%±0.04	90.82%±0.03	89.40%±0.04	90.10%±0.03	89.75%±0.04
Ours	93.82%±0.02	92.45%±0.03	93.10%±0.02	92.77%±0.02	92.15%±0.03	90.94%±0.02	91.85%±0.03	91.39%±0.03

TABLE 4 Evaluation of individual module effects via ablation analysis on LUNA16 and MURA datasets.

Model	LUNA16 dataset				MURA dataset			
	Accuracy	Recall	Precision	F1 score	Accuracy	Recall	Precision	F1 score
w./o. multimodal fusion and temporal dynamics	89.75%±0.03	88.40%±0.04	89.10%±0.03	88.72%±0.04	84.10%±0.04	82.95%±0.03	83.88%±0.04	83.41%±0.03
w./o. uncertainty estimation and domain adaptation	90.45%±0.04	89.10%±0.03	89.90%±0.04	89.50%±0.03	85.12%±0.03	83.80%±0.04	84.92%±0.03	84.35%±0.04
w./o. sparse attention and graph-structured clinical modeling	91.05%±0.03	89.75%±0.04	90.50%±0.03	90.10%±0.04	85.90%±0.04	84.65%±0.03	85.40%±0.04	85.00%±0.03
Ours	91.92%±0.02	90.78%±0.02	91.85%±0.02	91.31%±0.02	86.70%±0.02	85.45%±0.02	86.62%±0.02	86.03%±0.02

5 Conclusions and future work

In this, we aimed to address the enduring challenges in mental health diagnostics by leveraging deep learning-based image classification. we proposed a novel framework, MedIntelligenceNet, which integrates multi-modal data fusion, probabilistic uncertainty quantification, hierarchical feature abstraction, and adversarial domain adaptation. we introduced a Clinical-Informed Adaptation strategy that systematically incorporates structured clinical priors, symbolic reasoning, and domain alignment techniques to enhance both the robustness and interpretability of our model. Experiments conducted on diverse multi-modal mental health datasets demonstrated that our approach achieved significant improvements in diagnostic accuracy, model calibration, and resistance to domain shifts when compared with baseline deep learning methods.

Despite these promising results, there remain notable limitations. First, while Clinical-Informed Adaptation has improved model interpretability, the integration of symbolic reasoning with deep neural networks remains complex and sometimes insufficient for fully explaining the decision-making process. Second, although MedIntelligenceNet shows better robustness to domain shifts, its performance could still degrade when exposed to extremely novel or rare conditions not represented in the training data. Future research will focus on refining symbolic reasoning integration and enhancing model adaptability to unseen clinical variations, aiming for an even more trustworthy and generalizable diagnostic system.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

LZ: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft. RZ: Data curation, Writing – review & editing, Visualization, Supervision, Funding acquisition.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

This study utilizes publicly available datasets derived from human subjects, including OASIS, BraTS, LUNA16, and MURA. All datasets are de-identified and released under approved data-sharing protocols. No new data involving human participants were collected or processed by the authors.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Maurício J, Domingues I, Bernardino J. Comparing vision transformers and convolutional neural networks for image classification: a literature review. *Appl Sci*. (2023) 13:5521. doi: 10.3390/app13095521

2. Hong D, Han Z, Yao J, Gao L, Zhang B, Plaza A, et al. SpectralFormer: rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens*. (2021) 60:1–15. doi: 10.1109/TGRS.2021.3130716

3. Chen CF, Fan Q, Panda R. CrossViT: cross-attention multi-scale vision transformer for image classification. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC: IEEE (2021). p. 347–56. doi: 10.1109/ICCV48922.2021.00041

4. Wang X, Yang S, Zhang J, Wang M, Zhang J, Yang W, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med Image Anal*. (2022) 81:102559. doi: 10.1016/j.media.2022.102559

5. Touvron H, Bojanowski P, Caron M, Cord M, El-Nouby A, Grave E, et al. ResMLP: feedforward networks for image classification with data-efficient training. *IEEE Trans Pattern Anal Mach Intell*. (2023) 45:5314–21. doi: 10.1109/TPAMI.2022.3206148

6. Tian Y, Wang Y, Krishnan D, Tenenbaum JB, Isola P. Rethinking few-shot image classification: a good embedding is all you need? In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. *Computer Vision - ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12359*. Cham: Springer (2020). p. 266–82. doi: 10.1007/978-3-030-58568-6_16

7. Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, et al. MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci Data* (2023) 10:41. doi: 10.1038/s41597-022-01721-8

8. Hong D, Gao L, Yao J, Zhang B, Plaza A, Chanussot J. Graph convolutional networks for hyperspectral image classification. *IEEE Trans Geosci Remote Sens*. (2020) 59:5966–78. doi: 10.1109/TGRS.2020.3015157

9. Sun L, Zhao G, Zheng Y, Wu Z. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans Geosci Remote Sens*. (2022) 60:5522214. doi: 10.1109/TGRS.2022.3144158

10. Mai Z, Li R, Jeong J, Quispe D, Kim HJ, Sanner S. Online continual learning in image classification: an empirical survey. *Neurocomputing*. (2021) 469:28–51. doi: 10.1016/j.neucom.2021.10.021

11. Bhojanapalli S, Chakrabarti A, Glasner D, Li D, Unterthiner T, Veit A. Understanding robustness of transformers for image classification. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC: IEEE (2021). p. 10211–21. doi: 10.1109/ICCV48922.2021.01007

12. Rao Y, Zhao W, Zhu Z, Lu J, Zhou J. Global filter networks for image classification. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc. (2021). p. 980–93. Available online at: https://proceedings.neurips.cc/paper/2021/hash/07e87c2f4fc7f7c96116d8e2a92790f5-Abstract.html

13. Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, et al. Big self-supervised models advance medical image classification. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC: IEEE (2021). p. 3458–68. doi: 10.1109/ICCV48922.2021.00346

14. Li B, Li Y, Eliceiri K. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN: IEEE (2021). p. 14313–23. doi: 10.1109/CVPR46437.2021.01409

15. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros M, Ganslandt T. Transfer learning for medical image classification: a literature review. *BMC Med Imag*. (2022) 22:69. doi: 10.21203/rs.3.rs-844222/v1

16. Zhang C, Cai Y, Lin G, Shen C. DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA: IEEE (2020). p. 12200–10. doi: 10.1109/CVPR42600.2020.01222

17. Roy SK, Deria A, Hong D, Rasti B, Plaza A, Chanussot J. Multimodal fusion transformer for remote sensing image classification. *IEEE Trans Geosci Remote Sens*. (2022) 61:5515620. doi: 10.1109/TGRS.2023.3286826

18. Zhu Y, Zhuang F, Wang J, Ke G, Chen J, Bian J, et al. Deep subdomain adaptation network for image classification. *IEEE Trans Neural Netw Learn Syst*. (2020) 32:1713–22. doi: 10.1109/TNNLS.2020.2988928

19. Chen L, Li S, Bai Q, Yang J, Jiang S, Miao Y. Review of image classification algorithms based on convolutional neural networks. *Remote Sens*. (2021) 13:4712. doi: 10.3390/rs13224712

20. Ashtiani F, Geers AJ, Aflatouni F. An on-chip photonic deep neural network for image classification. *Nature*. (2021) 606:501–506. doi: 10.1038/s41586-022-04714-0

21. Masana M, Liu X, Twardowski B, Menta M, Bagdanov AD, van de Weijer J. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Trans Pattern Analy Mach Intellig*. (2020) 45:5513–33. doi: 10.1109/TPAMI.2022.3213473

22. Mascarenhas S, Agarwal M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In: *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*. Bengaluru: IEEE (2021). p. 96–99.

23. Sheykhmousa M, Mahdianpari M, Ghanbari H, Mohammadimanesh F, Ghamisi P, Homayouni S. Support vector machine versus random forest for remote sensing image classification: a meta-analysis and systematic review. *IEEE J Select Topics Appl Earth Observat Remote Sens*. (2020) 13:6308–25. doi: 10.1109/JSTARS.2020.3026724

24. Zhang Y, Li W, Sun W, Tao R, Du Q. Single-source domain expansion network for cross-scene hyperspectral image classification. *IEEE Trans Image Proc*. (2022) 32:1498–512. doi: 10.1109/TIP.2023.3243853

25. Bansal M, Kumar M, Sachdeva M, Mittal A. Transfer learning for image classification using VGG19: Caltech-101 image data set. *J Ambient Intellig Human Comput*. (2021) 14:3609–20. doi: 10.1007/s12652-021-03488-z

26. Dai Y, Gao Y. TransMed: transformers advance multi-modal medical image classification. *Diagnostics*. (2021) 11:1384. doi: 10.3390/diagnostics11081384

27. Taori R, Dave A, Shankar V, Carlini N, Recht B, Schmidt L. Measuring robustness to natural distribution shifts in image classification. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. *Advances in Neural Information Processing Systems, vol. 33*. Curran Associates, Inc. (2020). p. 18583–99. Available online at: https://proceedings.neurips.cc/paper/2020/hash/d8330f857a17c53d217014ee776bfd50-Abstract.html

28. Peng J, Huang Y, Sun W, Chen N, Ning Y, Du Q. Domain adaptation in remote sensing image classification: a survey. *IEEE J Select Topics Appl Earth Observat Remote Sens*. (2022) 15:9842–59. doi: 10.1109/JSTARS.2022.3220875

29. Bazi Y, Bashmal L, Rahhal MMA, Dayil RA, Ajlan NA. Vision transformers for remote sensing image classification. *Remote Sens*. (2021) 13:516. doi: 10.3390/rs13030516

30. Zheng X, Sun H, Lu X, Xie W. Rotation-invariant attention network for hyperspectral image classification. *IEEE Trans Image Proc*. (2022) 31:4251–65. doi: 10.1109/TIP.2022.3177322

31. Kumar A. Neuro Symbolic AI in personalized mental health therapy: Bridging cognitive science and computational psychiatry. *World J Adv Res Rev*. (2023) 19:1663–79. doi: 10.30574/wjarr.2023.19.2.1516

32. Nawaz U, Anees-ur Rahaman M, Saeed Z. A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems. *Intellig Syst Appl*. (2025) 2025:200541. doi: 10.1016/j.iswa.2025.200541

33. Bhuyan BP, Ramdane-Cherif A, Tomar R, Singh T. Neuro-symbolic artificial intelligence: a survey. *Neural Comp Appl*. (2024) 36:12809–44. doi: 10.1007/s00521-024-09960-z

34. Govorov I, Komlichenko E, Ulrikh E, Dikareva E, Pervunina T, Vazhenina O, et al. The microbiome in endometrial cancer: vaginal milieu matters. *Front Med*. (2025) 12:1533344. doi: 10.3389/fmed.2025.1533344

35. Luo Y, Hu J, Zhou Z, Zhang Y, Wu Y, Sun J. Oxidative stress products and managements in atopic dermatitis. *Front Med*. (2025) 12:1538194. doi: 10.3389/fmed.2025.1538194

36. Hall A, Doherty E, Nathan N, Wiggers J, Attia J, Tully B, et al. Longitudinal exploration of the delivery of care following a successful antenatal practice change intervention. *Front Med*. (2025) 12:1476083. doi: 10.3389/fmed.2025.1476083

37. Dequidt P, Bourdon P, Tremblais B, Guillevin C, Gianelli B, Boutet C, et al. Exploring radiologic criteria for glioma grade classification on the BraTS dataset. *IRBM*. (2021) 42:407–14. doi: 10.1016/j.irbm.2021.04.003

38. Basheer S, Bhatia S, Sakri SB. Computational modeling of dementia prediction using deep neural network: analysis on OASIS dataset. *IEEE Access*. (2021) 9:42449–62. doi: 10.1109/ACCESS.2021.3066213

39. Lalitha S, Murugan D. Segmentation and classification of 3D lung tumor diagnoses using convolutional neural networks. In: *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*. Trichy: IEEE (2023). p. 230–238.

40. Kandel I, Castelli M. Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset. *Health Inform Sci Syst*. (2021) 9:33. doi: 10.1007/s13755-021-00163-7

41. Dong H, Zhang L, Zou B. Exploring vision transformers for polarimetric SAR image classification. *IEEE Trans Geosci Remote Sens*. (2022) 60:5219715. doi: 10.1109/TGRS.2021.3137383

42. He X, Chen Y, Lin Z. Spatial-spectral transformer for hyperspectral image classification. *Remote Sens*. (2021) 13:498. doi: 10.3390/rs13030498

43. Lanchantin J, Wang T, Ordonez V, Qi Y. General multi-label image classification with transformers. In: *Computer Vision and Pattern Recognition*. Nashville, TN: IEEE (2020).

44. Vermeire T, Brughmans D, Goethals S, de Oliveira RMB, Martens D. Explainable image classification with evidence counterfactual. *Pattern Analy Appl*. (2022) 25:315–335. doi: 10.1007/s10044-021-01055-y

45. Dong Y, Fu QA, Yang X, Pang T, Su H, Zhu J, et al. Benchmarking adversarial robustness on image classification. In: *Computer Vision and Pattern Recognition*. Seattle, WA: IEEE (2020).

46. Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image classification and segmentation. *Ann Transl Med*. (2020) 8:713. doi: 10.21037/atm.2020.02.44

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 04 August 2025
DOI 10.3389/fmed.2025.1566870

OPEN ACCESS

EDITED BY
Salil Bharany,
Chitkara University, India

REVIEWED BY
Diponkor Bala,
City University, Bangladesh
Prabhsimran Singh,
Guru Nanak Dev University, India

*CORRESPONDENCE
Syed Muhammad Usman
✉ smusman.h11@bahria.edu.pk
Saad Arif
✉ sarif@kfu.edu.sa

RECEIVED 25 January 2025
ACCEPTED 04 July 2025
PUBLISHED 04 August 2025

CITATION
Alkhrijah Y, Khalid S, Usman SM, Jameel A, Zubair M, Aldossary H, Anwar A and Arif S (2025) Feature fusion ensemble classification approach for epileptic seizure prediction using electroencephalographic bio-signals. *Front. Med.* 12:1566870. doi: 10.3389/fmed.2025.1566870

Feature fusion ensemble classification approach for epileptic seizure prediction using electroencephalographic bio-signals

Yazeed Alkhrijah[1,2], Shehzad Khalid[3,4], Syed Muhammad Usman[5]*, Amina Jameel[4], Muhammad Zubair[6], Haya Aldossary[7], Aamir Anwar[8] and Saad Arif[9]*

[1]Department of Electrical Engineering, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia, [2]King Salman Center for Disability Research (KSCDR), Riyadh, Saudi Arabia, [3]Computer and Information Sciences Research Center (CISRC), Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia, [4]Department of Computer Engineering, Bahria University, Islamabad, Pakistan, [5]Department of Computer Science, Bahria University, Islamabad, Pakistan, [6]Interdisciplinary Research Center for Finance and Digital Economy, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, [7]Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Jubail, Saudi Arabia, [8]School of Computing, University of Portsmouth, London Campus, London, United Kingdom, [9]Department of Mechanical Engineering, College of Engineering, King Faisal University, Al Ahsa, Saudi Arabia

Introduction: Epilepsy is a neurological disorder in which patients experience recurrent seizures, with the frequency of occurrence more than twice a day, which highly affects a patient's life. In recent years, multiple researchers have proposed multiple machine learning and deep learning-based methods to predict the onset of seizures using electroencephalogram (EEG) signals before they occur; however, robust preprocessing to mitigate the effect of noise, channel selection to reduce dimensionality, and feature extraction remain challenges in accurate prediction.

Methods: This study proposes a novel method for accurately predicting epileptic seizures. In the first step, a Butterworth filter is applied, followed by a wavelet and a Fourier transform for the denoising of EEG signals. A non-overlapping window of 15 s is selected to segment the EEG signals, and an optimal spatial filter is applied to reduce the dimensionality. Handcrafted features, including both time and frequency domains, have been extracted and concatenated with the customized one-dimensional convolutional neural network-based features to form a comprehensive feature vector. It is then fed into three classifiers, including support vector machines, random forest, and long short-term memory (LSTM) units. The output of these classifiers is then fed into the model-agnostic meta learner ensemble classifier with LSTM as the base classifier for the final prediction of interictal and preictal states.

Results: The proposed methodology is trained and tested on the publicly available CHB-MIT dataset while achieving 99.34% sensitivity, 98.67% specificity, and a false positive alarm rate of 0.039.

Discussion: The proposed method not only outperforms the existing methods in terms of sensitivity and specificity but is also computationally efficient, making it suitable for real-time epileptic seizure prediction systems.

KEYWORDS

AI in healthcare, epilepsy, electroencephalogram, epileptic seizure prediction, signal quality index, optimal spatial filter, 1DCNN, ensemble classifier

1 Introduction

Epilepsy is a neurological disorder in which patients suffer from seizures, and it affects their quality of life as a sudden seizure may cause an accident or injury while driving, climbing stairs, or walking on the road, etc. Seizure disturbs the activity of the brain, which can be observed by visualizing the electroencephalographic (EEG) signals recorded by placing electrodes on the scalp of the patient's brain (1). Seizures are divided into four states: interictal, the normal state; preictal, which starts a few minutes before the onset of seizure and ends with the seizure onset; ictal, in which the seizure occurs; and postictal, which starts after the seizure. Seizures can be categorized into two types, i.e., focal and generalized seizures. Focal seizures are normally treatable with surgical procedures, whereas generalized seizures can only be treated with the help of medicines; however, it has been observed that in 70% of the cases these seizures cannot be completely controlled with the help of medicines (2). Researchers (3–19) have proposed multiple methods to predict the onset of seizures before they occur by predicting the preictal state; however, accurate prediction remains a challenge due to multiple factors. EEG signals are susceptible to noise added during signal acquisition, high dimensionality due to the number of channels, and computational complexity of feature extraction and accurate classification. Figure 1 shows a plot of three EEG signals from 1-h continuous recordings. Accurate seizure prediction significantly impacts patient safety and quality of life by reducing the risks of sudden accidents or injuries during seizures. Despite advancements, clinicians and patients still face considerable challenges due to inaccurate seizure forecasting, leading to compromised safety and anxiety among epilepsy patients.

A typical method of epileptic seizure prediction involves preprocessing of EEG signals for noise removal and channel selection, followed by feature extraction and classification. Numerous techniques to preprocess EEG signals have been proposed in recent years for removing noise and artifacts such as eye blinks, eye movements, and muscle activity before feeding the data into the model. Fei et al. (6) and Usman et al. (14) proposed bandpass filters to preprocess the EEG signals. Wang et al. (20) has employed an infinite impulse response (IIR) bandpass filter and filtered the segmented data to filter out artifacts. Cho et al. (8) has used the fast Fourier transform (FFT). Common spatial pattern (CSP) is applied to reduce the effect of artifacts from EEG signals by Birjandtalab et al. (4). Researchers (14, 21, 22) have made use of the short-time Fourier transform (STFT) for preprocessing. Jana et al. (9) has utilized a pool-based technique with a 30-s window for noise reduction.

Duun-Henriksen et al. (23) selected channels based on the maximum variance, the difference in variance, and entropy. Entropy indicates the extent of disorder, impurity, and uncertainty, so the channels with the highest entropy were selected. To select channels that carry the highest information and are optimal, Daoud and Bayoumi (10) has selected channels with the maximum variance entropy product. Birjandtalab et al. (4) has used a random decision forest for channel selection. Cogan et al. (7) selected the best channel by ranking all the features based on the information

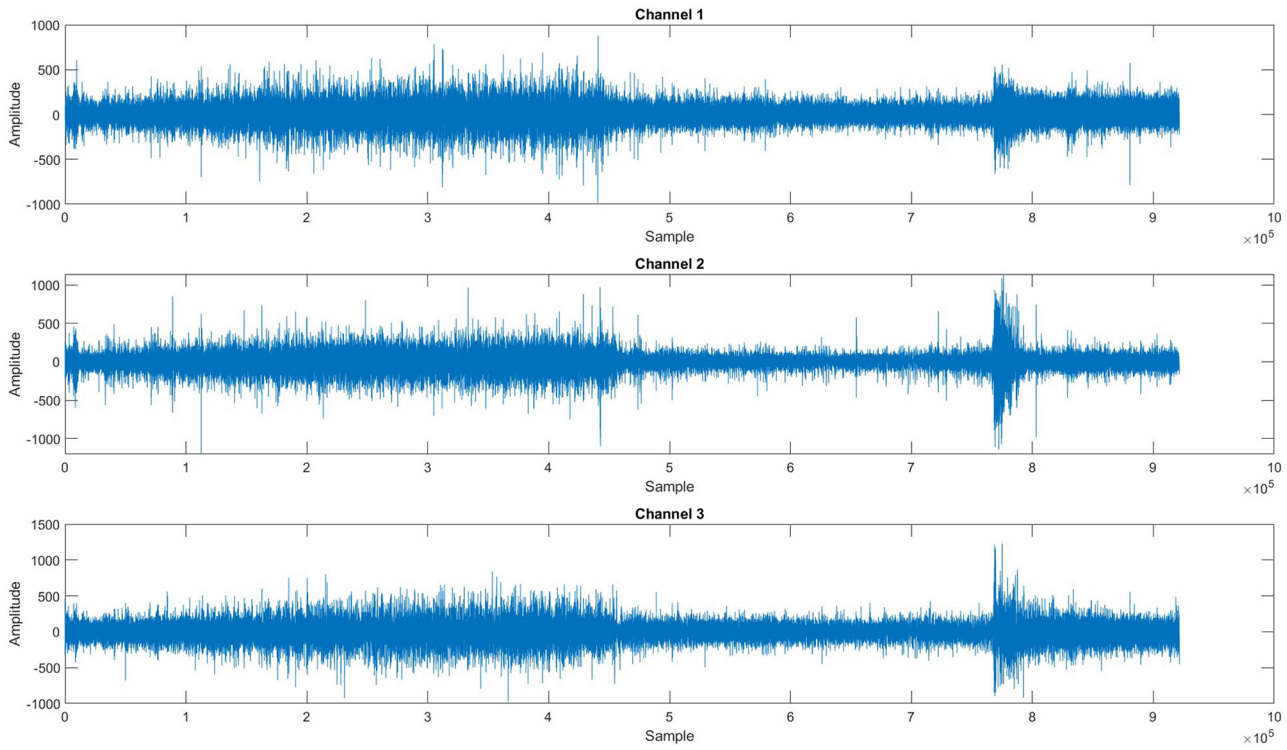


FIGURE 1
One-hour span session of EEG recordings for three channels.

gain for each subject. Parvez and Paul (24) checked the significance of each channel individually, then eliminated the channel of low significance and selected the best channels by calculating the average classification accuracy iteratively. Wang et al. (20) in their research study calculated a signal quality index (SQI), based on signal complexity. They brought three types of signals into consideration, and the optimal channels were selected accordingly.

Commonly used feature extraction methods include continuous wavelet transform (CWT), discrete cosine transformation (DCT), and discrete wavelet transform (DWT). Tsiouris et al. (25), Jana and Mukherjee (16), Alotaiby et al. (5), and Arif et al. (21) applied DWT to extract time-frequency features and then support vector machines (SVM) for predictions. Asharindavida et al. (11) utilized empirical mode decomposition (EMD) for feature extraction. Birjandtalab et al. (4), Birjandtalab et al. (3), and Borhade et al. (12) employed power spectral density (PSD) for feature extraction. Fei et al. (6) has applied a FrFT-based chaos method to obtain relevant features. Both time and frequency domain features, along with total energy spectrum and energy percentage-based features, were extracted to be used as input to the classifier (15). Zhang et al. (13) has made use of CSP-based feature extraction. Truong et al. (22) and Arif et al. (26) used STFT to extract features. Deep learning (DL) can also be used for feature extraction, as Daoud and Bayoumi (10) has extracted features through DL techniques.

Once features are extracted, the next task is to distinguish the signal between interictal and preictal states. Researchers have made use of machine learning (ML) and DL classifiers for the classification of EEG signals in seizure prediction methods. SVM with cross-validation was used for classification by Tamanna et al. (15), Alotaiby et al. (5), and Asharindavida et al. (11), a least square SVM classifier was applied to classify the EEG signals. Back-forward propagation neural networks (BPNN) and linear discriminant analysis (LDA) were also used for classification (6, 11, 13). Fei et al. (6), Usman et al. (14), Alotaiby et al., (5), Asharindavida et al. (11), and Alickovic et al. (27) employed k-nearest neighbor (kNN), and random forest (RF) for classification. In the study by Truong et al. (22), a convolutional neural network (CNN) was utilized for the classification of preictal and interictal states. Daoud and Bayoumi (10) and Alotaiby et al. (5) have used DL models [multilayer perceptron (MLP), deep CNN (DCNN), bidirectional LSTM (Bi-LSTM)] for classification tasks.

DL and EEG-based seizure prediction has advanced significantly in recent years. By successfully modeling EEG data across several spatial and temporal scales, Dong et al. (28) proposed a novel multi-scale spatio-temporal attention network (MSAN), which increased the accuracy of seizure prediction. Alasiry et al. (29) suggested a heterogeneous graph neural network (GNN) that enhanced clinical interpretability and predictive performance by capturing intricate EEG channel interactions. A CNN-Bi-LSTM hybrid model was presented by Cao et al. (30), who also developed a feature-level fusion technique that showed improved performance for epileptic seizure prediction across multiple datasets. Bi-LSTM consistently outperformed other recurrent neural network (RNN) structures like gated recurrent units (GRU), MLP, and DCNN for seizure prediction tasks according to an ablation study conducted by Bajaj and Sharma (31) on a variety of LSTM-based architectures. A novel mobile network information gain (M-NIG) technique was presented by Meng et al. (32) with a focus on individual-specific multi-channel EEG networks to lower noise and greatly improve prediction robustness. Notwithstanding these developments, there are still issues that need to be addressed, mainly in the areas of computational complexity, practicality for real-time clinical applications, efficient dimensionality reduction, and reliable handling of class-imbalanced data. These issues together highlight the necessity for further research.

Current approaches for epileptic seizure prediction predominantly utilize all available EEG channels. This practice is computationally expensive, increases time complexity, and raises hardware and financial costs, highlighting the need for methods that can identify and utilize only the most informative channels. The high dimensionality of EEG data often affects the efficiency and accuracy of predictive models. Despite its critical impact, this challenge has been largely overlooked in existing studies, necessitating effective dimensionality reduction techniques to enhance prediction performance. Many researchers have not adequately addressed the issue of class imbalance, a prevalent challenge in seizure prediction where certain classes (e.g., seizure events) are underrepresented compared to others. This imbalance can skew model performance and compromise prediction reliability.

We propose a novel method for epileptic seizure prediction to address these research gaps, which have been identified after a comprehensive literature review. In the first step, the Butterworth filter is applied, followed by wavelet and Fourier transforms for denoising of EEG signals. A non-overlapping window of 15 s is selected to segment the EEG signals, and an optimal spatial filter is applied to reduce the dimensionality. Handcrafted features, including both time and frequency domains, have been extracted and concatenated with the customized one-dimensional CNN (1DCNN)-based features to form a comprehensive feature vector. It is then fed into three classifiers, including SVM, RF, and LSTM units, and the output of these classifiers is then fed into a model-agnostic meta learner (MAML) ensemble classifier with LSTM as base classifier for the final prediction of interictal and preictal states. The contributions of this research include:

- Introduced a novel technique to identify the most informative EEG channels, improving prediction accuracy while significantly reducing computational costs, a key challenge in real-time applications.
- Developed an effective dimensionality reduction method to deal with the high-dimensional nature of EEG data, which affects the performance of prediction algorithms.
- Proposed a surrogate channel by combining optimal EEG channels that contribute the most to seizure prediction.
- Demonstrated the effectiveness of the proposed method on the publicly available CHB-MIT dataset, achieving a sensitivity of 99.34% and specificity of 98.67% with a false positive alarm rate of 0.03. These results outperform various state-of-the-art techniques, establishing a new benchmark in epileptic seizure prediction.

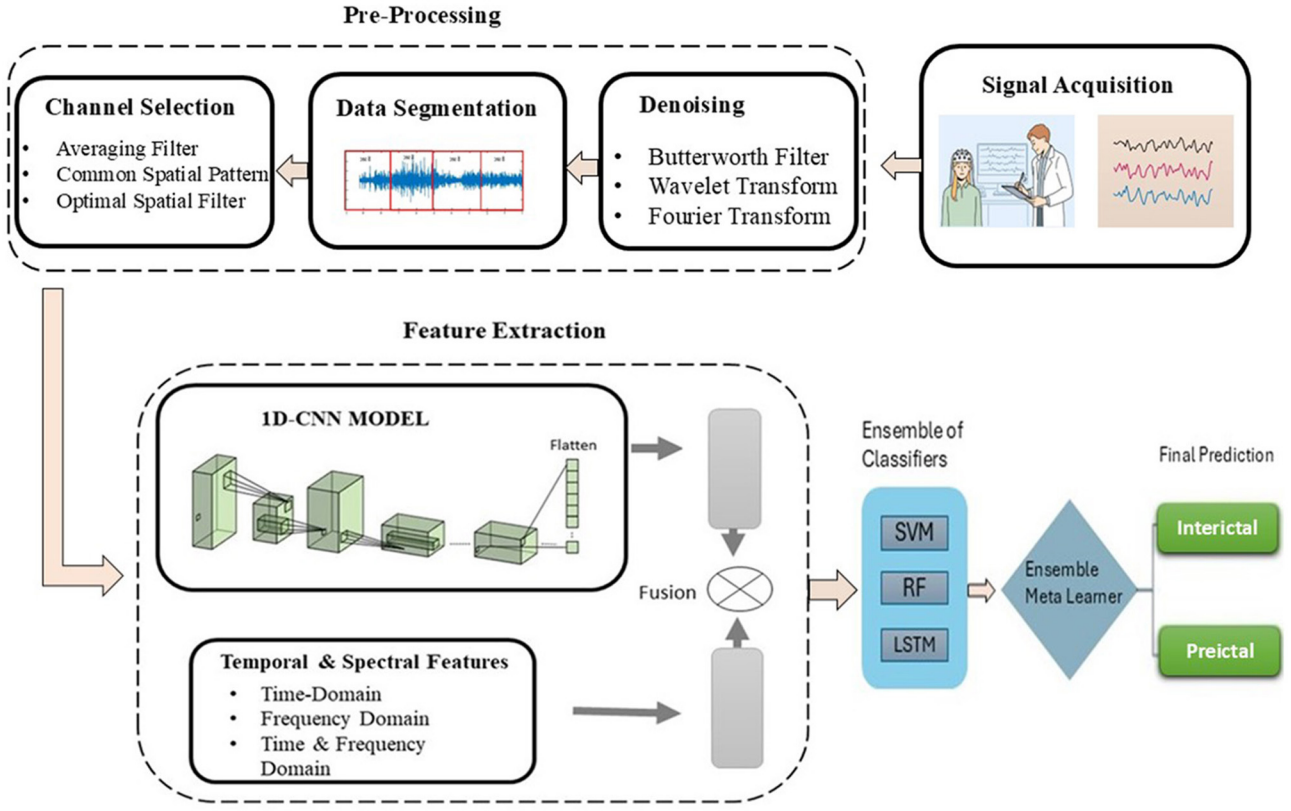


FIGURE 2
Flow diagram of the proposed methodology of epileptic seizure prediction.

2 Methodology

To overcome the identified limitations and enhance seizure prediction accuracy, our methodology strategically targets the three main challenges: noise reduction in EEG signals, dimensionality reduction, and class imbalance mitigation. We propose a novel method of epileptic seizure prediction using EEG signals. It consists of three steps, including the preprocessing of EEG signals, feature extraction, and classification between preictal and interictal states. The preprocessing step involves segmentation of EEG signals into equal-size segments using a non-overlapping window, followed by multistage noise removal using Butterworth filter, wavelet, and Fourier transforms, and conversion of multi-channel EEG signals into a single surrogate channel. After preprocessing, both handcrafted and automated features have been extracted and concatenated to form a single feature vector. Time and frequency domain features include statistical and spectral signatures, whereas a customized architecture of 1DCNN has been proposed to extract automated features. Figure 2 shows the flow diagram of the proposed method. The following subsection presents all three steps of the proposed methodology in detail.

2.1 Preprocessing of EEG signals

Due to the inherent susceptibility of EEG signals to noise from artifacts and external sources, a robust preprocessing strategy is critical to ensure data quality for reliable seizure prediction. In this research, we used a publicly available CHB-MIT dataset (33) that comprises EEG recordings of 24 pediatric individuals recorded in the Children's Hospital Boston. The dataset has been annotated by the medical experts with the start and end time of the seizure for each session of all individuals. EEG signals have been recorded with 23 channels and follow the 10–20 electrode placement method. The dataset has been sampled at 256 Hz and totals 644 h of recordings. We have divided EEG signals into equal-sized segments with the help of an equal-sized, non-overlapping window of 15 s. Figure 3 shows the plot of segmented EEG signals proposed in this research.

After segmenting the EEG signals, preictal and interictal signals were separated. Preictal and interictal samples were carefully selected, considering that preictal and postictal samples may overlap. Therefore, we included only those sessions for interictal state samples where no seizure onset occurred within two sessions before or after. Preictal state has been considered as 30 min before the onset of the seizure, provided that there was no seizure in the last session to avoid the postictal state overlapping with the preictal state. EEG signals are sensitive to noise, making it essential to apply various techniques to remove noise and artifacts, ensuring that the raw data is suitable for further processing. Methods include: Butterworth bandpass filter, EMD, FFT, CWT, DWT, and CSP, which help deal with noise and artifacts. Additionally, a window duration, overlapping and non-overlapping, can also be used to reduce the effect of noise to achieve better results.

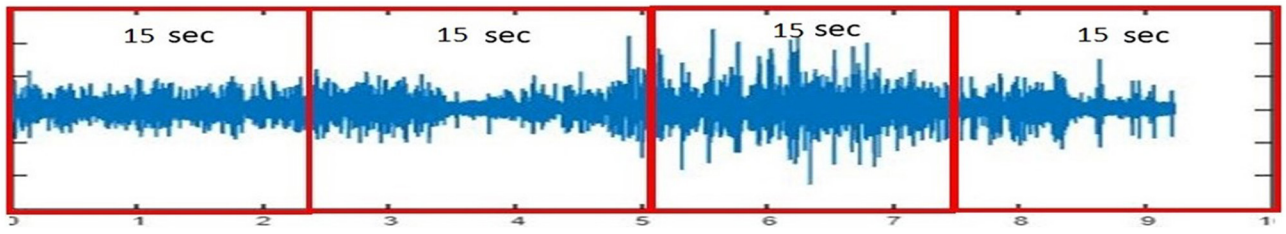


FIGURE 3
EEG data segmented into 15-s windows.

We preprocessed EEG signals to remove noise and artifacts to enhance signal quality, as shown in Figure 4. The wavelet transform and Butterworth filter, a high-pass filter with a cutoff frequency of 0.5 Hz and a low-pass filter with a cutoff frequency of 40 Hz, were applied. These filters were used to remove low-frequency, high-frequency drifts and fluctuations caused by internal and external sources during data recording. Figure 5 illustrates the raw signal alongside the denoised signals after applying these filters. The EEG signals are acquired through multi-channel recordings. Using a large set of channels leads to computational complexity. Additionally, not all channels provide valuable insights for seizure prediction. The use of all channels can also result in misclassifications of seizures. To address these issues, channel selection is a critical step in reducing the number of channels while preserving essential information.

The number of channels is not only reduced, but optimal channels are also combined, which are highly contributing to seizure prediction, to make a surrogate channel. The channels are selected based on two criteria: high SQI and maximum variance. A higher SQI indicates superior signal quality, while lower values suggest poorer quality. Higher variance suggests increased brain activity. By selecting channels that meet these criteria, we ensure that the most informative and relevant channels are retained, leading to more accurate and efficient seizure prediction. A combined plot of all five selected channels is presented in Figure 6.

$$V_{ict}(C) = \frac{1}{k}\sum_{i=1}^{k}(x_c(i) - \mu_c))^2 \quad (1)$$

$$Selected\ Channel = \max_{1:N}\left\{Vict(c)\right\} \quad (2)$$

2.1.1 Surrogate channel

Given the computational inefficiency caused by analyzing high-dimensional EEG data from multiple channels, we introduce a surrogate channel technique. Unlike previous methods that typically analyze all channels equally, our approach identifies and combines the most informative EEG channels into a single surrogate channel, significantly reducing computational complexity while maintaining prediction accuracy. High-dimensional EEG signals pose significant problems in EEG analysis, including increased computational cost and a higher risk

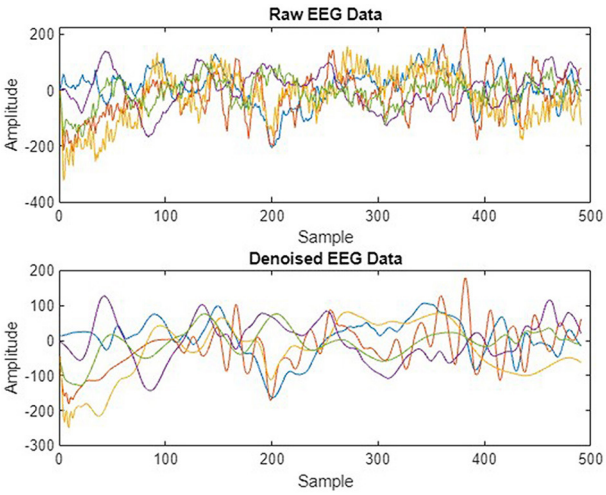


FIGURE 4
Raw vs. denoised EEG signals.

of overfitting to noise rather than extracting meaningful patterns. Addressing this issue can not only increase the performance of the classifier but also reduce the computational complexity. To convert multiple EEG channels into a surrogate channel, an averaging filter, CSP, and an optimal spatial filter were applied. These techniques were applied to increase the signal-to-noise ratio (SNR) and variance interval between two classes. The averaging filter is a method used to increase the SNR by replacing each sample with the average value of neighboring samples within a defined window. This averaging filter calculates the mean of all the channels to form a single channel (surrogate channel). The surrogate channel obtained after applying an averaging filter contains more SNR than multiple channels. The surrogate channel aims to capture the collective signal from multiple electrodes, potentially improving interpretability and simplifying analysis.

Despite its effectiveness in noise reduction, residual noise may persist in the surrogate channel, necessitating further refinement or the consideration of complementary filtering techniques to optimize signal quality for further analysis. The CSP filter is a technique that is frequently used in EEG signal processing to enhance the discriminative features of EEG signals by spatially filtering them. The CSP algorithm identifies spatial filters that increase the variance of EEG signals for one class while minimizing

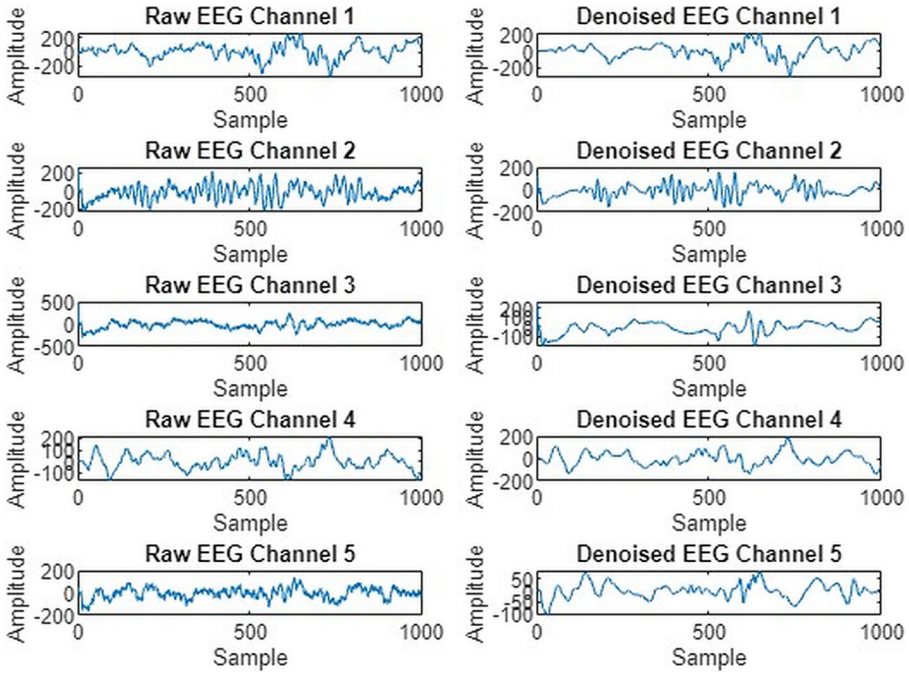


FIGURE 5
Five EEG channel waveforms before and after noise removal.

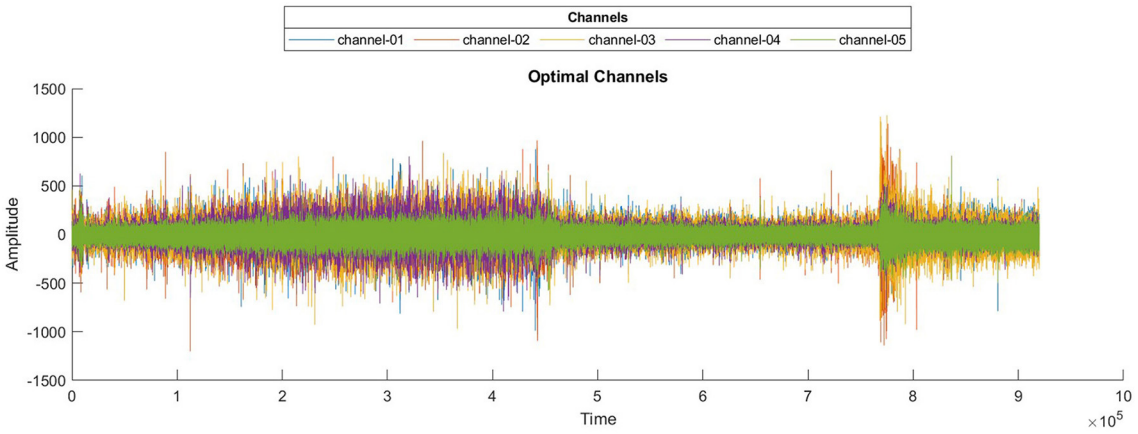


FIGURE 6
Waveforms of selected optimal EEG channels.

it for another class. CSP not only increases the SNR but also enhances the variance interval between two or more classes. This suggests that relevant information becomes more distinct while noise is effectively suppressed. In essence, CSP can convert a multi-channel EEG signal into a surrogate channel that encapsulates the most discriminative features for the task at hand.

2.1.2 Mitigating the class imbalance problem

Class imbalance is a critical challenge in EEG-based seizure prediction because the number of preictal segments (indicating impending seizures) is significantly fewer than interictal segments (non-seizure states), potentially biasing prediction models. To address this imbalance, we utilize advanced oversampling techniques. Imbalanced data refers to too many instances in one class and too few examples in another. Imbalanced data can highly affect the model's overall effectiveness and make it difficult for the model to distinguish between the decision boundaries of different classes. One of the solutions to deal with this is to over-sample the instances in the minority class. Over-sampling can be attained by simply duplicating instances from the minority class in the training dataset before fitting a model. This does not give any extra information to the model, but it can deal with the data imbalance issue. An enhancement on duplicating instances

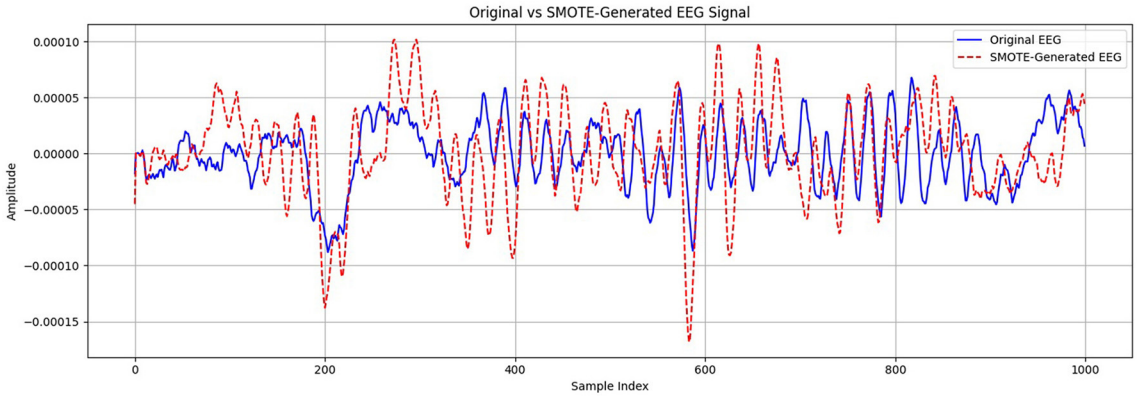


FIGURE 7
Comparison of original and SMOTE-generated EEG signals for the minority class.

from the minority class is to synthesize new instances from the minority class. In this study, data splitting was performed after an initial oversampling process to address class imbalance and improve model performance. Specifically, we utilized the synthetic minority over-sampling technique (SMOTE) and the soft prototype instance discrimination for enhancing representation (SPIDER) techniques to generate additional synthetic samples and improve the representation of minority classes. SMOTE selects a minority class instance randomly and then finds its k nearest minority class neighbors.

The synthetic instances are then generated as a convex combination of the selected instances. SPIDER works by producing synthetic samples for the minority class in accordance with prototype instances. Prototype instances are representative samples from the minority class that capture its characteristics. SPIDER synthesizes new instances by perturbing these prototypes, creating variations that are still representative of the minority class. After applying these oversampling methods, the dataset was partitioned into training and validation subsets. Figure 7 presents a visual comparison between an original EEG segment and a synthetic sample generated using the SMOTE. The synthetic EEG maintains the temporal rhythm and amplitude range of the original signal, with minor variations that reflect the data-driven interpolation characteristics of SMOTE. To assess the fidelity of the generated samples, we evaluated similarity using statistical metrics such as Pearson correlation and dynamic time warping (DTW), both of which confirmed a high degree of alignment between the original and synthetic signals. This validates the suitability of SMOTE for augmenting the minority class in EEG-based classification tasks without introducing unrealistic distortions.

2.2 Feature extraction from EEG signals

Effective feature extraction is crucial to distinguish between seizure states clearly. Thus, we combine handcrafted temporal and spectral features with automated DL-based features to ensure high inter-class separability, which is key for robust classification. After preprocessing and channel selection, feature extraction is a critical step in the prediction of epileptic seizures. To capture both interpretable signal characteristics and complex spatial-temporal dependencies, we adopted a hybrid feature extraction strategy. Handcrafted features such as Hjorth parameters and entropy measures are well-established in EEG analysis for their ability to reflect signal complexity and variance.

2.2.1 Handcrafted features

Various techniques for feature extraction are presented in the literature, including both handcrafted and automatic feature extraction methods. ML techniques are commonly used for handcrafted feature extraction, while DL is well-suited for automatic feature extraction. After a comprehensive literature review, we identified features that provide better inter-class separability. Inter-class separability refers to the measure that how two classes are distant, different, or separable from one another. The higher the inter-class separability, the easier it is for the classifier to distinguish and classify the classes. Conversely, the lower the inter-class separability, the more challenging for the classifier to distinguish between the classes, because lower inter-class separability indicates that the classes are overlapping significantly. Temporal and spectral features can be identified and extracted, revealing significant patterns within the EEG signal. Following preprocessing and channel selection, the temporal features were extracted including min, max, mean (Equation 3), variance (Equation 4), standard deviation (Equation 5) and skewness (Equation 6). The mean represented as μ, is calculated as follows:

$$\mu = \frac{1}{K}\sum_{i=1}^{K}(x_i) \tag{3}$$

$$\sigma^2 = \frac{1}{K}\sum_{i=1}^{K}(x_i - \mu)^2 \tag{4}$$

$$\sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(x_i - \mu)^2} \tag{5}$$

TABLE 1 Statistical and spectral features extracted from 10 EEG segments of preictal state.

Feature	Seg1	Seg2	Seg3	Seg4	Seg5	Seg6	Seg7	Seg8	Seg9	Seg10
Min	-0.00013	-0.00013	-0.00025	-0.00016	-9.75E-05	-0.00010	-8.15E-05	-7.33E-05	-0.00017	-0.00012
Max	0.00010	0.00013	0.00022	8.22E-05	9.20E-05	8.46E-05	6.58E-05	7.29E-05	9.24E-05	0.00014
Mean	-7.13E-08	1.20E-06	-1.16E-06	5.84E-07	1.38E-07	4.44E-07	-4.14E-07	5.12E-07	-4.70E-08	9.59E-07
Variance	9.17E-10	1.09E-09	3.57E-09	6.31E-10	6.65E-10	5.29E-10	3.81E-10	4.52E-10	8.28E-10	1.26E-09
Standard deviation	3.03E-05	3.30E-05	5.98E-05	2.51E-05	2.58E-05	2.30E-05	1.95E-05	2.13E-05	2.88E-05	3.56E-05
Skewness	-0.191	-0.166	-0.198	-1.230	-0.269	-0.271	-0.162	-0.182	-1.007	0.120
Spectral centroid	5.794	5.090	7.621	5.550	5.365	6.426	7.066	6.591	4.529	4.653
Spectral variance	36.896	45.557	293.917	55.709	47.305	58.932	67.889	59.246	38.331	32.890
Spectral skewness	4.079	6.505	4.177	5.755	5.441	4.426	4.580	3.972	5.160	4.747

TABLE 2 Statistical and spectral features extracted from 10 EEG segments of interictal state.

Feature	Seg1	Seg2	Seg3	Seg4	Seg5	Seg6	Seg7	Seg8	Seg9	Seg10
Min	-0.00062	-0.00083	-0.00075	-0.00075	-0.00046	-0.00015	-0.000078	-0.00063	-0.00062	-0.00070
Max	0.00074	0.00084	0.00080	0.00064	0.00065	0.00011	0.000099	0.00058	0.00062	0.00069
Mean	-1.98E-07	3.82E-07	2.00E-06	-1.94E-08	-1.29E-06	1.07E-06	1.19E-07	-8.82E-08	-3.99E-07	-5.19E-07
Variance	1.17E-08	2.06E-08	3.38E-08	2.30E-08	1.02E-08	6.95E-10	5.15E-10	7.89E-09	1.79E-08	1.56E-08
Standard deviation	1.08E-04	1.44E-04	1.84E-04	1.52E-04	1.01E-04	2.64E-05	2.27E-05	8.88E-05	1.34E-04	1.25E-04
Skewness	0.966	-0.168	0.359	-0.135	0.663	-0.599	0.283	-0.238	0.433	0.108
Spectral centroid	20.206	22.993	12.441	15.491	6.892	11.051	9.771	18.497	15.589	19.211
Spectral variance	478.771	509.770	358.245	442.970	248.950	393.654	265.381	458.774	447.203	477.597
Spectral skewness	1.527	1.447	2.345	2.012	4.128	2.770	2.881	1.663	2.019	1.718

$$S = \frac{1}{K}\sum_{i=1}^{K}(x_i - \mu)^3 \tag{6}$$

where, μ is EEG signal mean, x_i is value of the EEG signal at i^{th} sample, K is number of samples in EEG signals. Variance is the measurement value used to show how far a set of numbers is spread with respect to the mean or average value. σ^2 is variance of EEG signals. Standard deviation is a measure representing the amount of how much dispersed or variation, such as spread, dispersion is in the data from the mean. σ is the standard deviation of EEG signals. Skewness is a measure of asymmetry of the distribution around the mean. It shows in which direction the data is skewed.

The spectral analysis of EEG signals is commonly done by obtaining the PSD. PSD is a Fourier transform of the autocorrelation function (Equation 7). PSD and auto-correlation are very closely related to each other in the analysis of signals and time series. The auto-correlation function can be calculated as:

$$R_x(\tau) = E[x(t).x(t+\tau)] \tag{7}$$

where, $x(t)$ is EEG signal sample, E is expected or mean value.

PSD describes the distribution of power over frequency and may be computed with the Fourier transform or the distribution of mean power of a signal in the frequency domain (26). The PSD is calculated as:

$$S_x(t) = \int_{-\infty}^{\infty} R_x(\tau).e^{-2\pi if\tau}df \tag{8}$$

Spectral features are frequency domain features, that include spectral centroid, variational coefficient, and spectral skewness. These features can be computed with the help of PSD, which is computed by Equation 8. where, $R_x(\tau)$ denotes autocorrelation of the signal $x(t)$. Spectral centroid, variational coefficient, and spectral skewness can be computed by following equations.

$$C_s = \frac{\sum_t tS_x(t)}{\sum_t S_x(t)} \tag{9}$$

$$\sigma_s^2 = \frac{\sum_t (t - C_s)^2 S_x(t)}{\sum_t S_x(t)} \tag{10}$$

$$\beta_s = \frac{\sum_t ((t - C_s)/\sigma_s)^3 S_x(t)}{\sum_t S_x(t)} \tag{11}$$

Tables 1, 2 present the statistical and spectral features extracted from 10 EEG segments corresponding to the preictal and interictal states, respectively. Each table lists features such as minimum, maximum, mean, variance, standard deviation, skewness, spectral centroid, spectral variance, and spectral skewness for each segment.

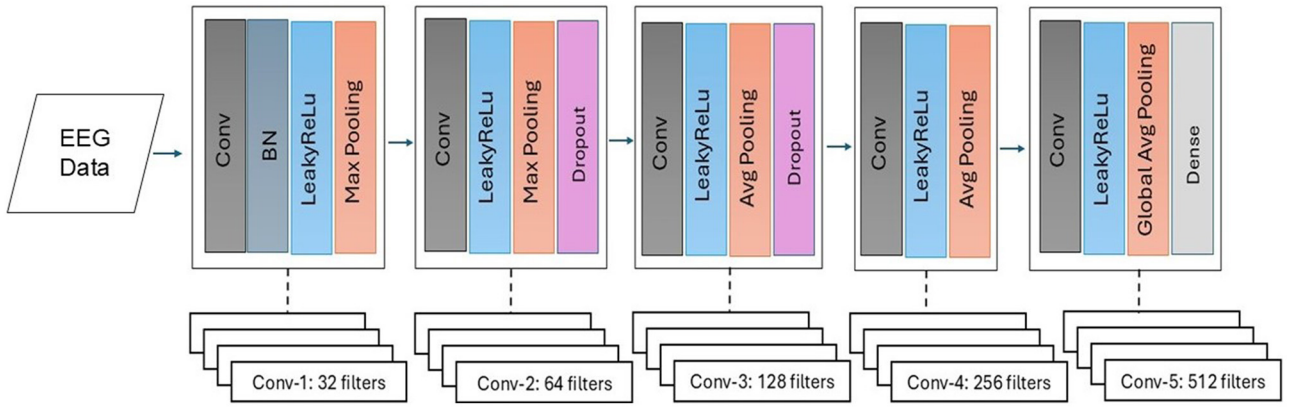


FIGURE 8
Proposed customized architecture of 1DCNN.

This layout allows for segment-wise analysis of feature variation within each class and supports comparative evaluation between preictal and interictal brain states, offering valuable insights into the distinguishing characteristics relevant for seizure prediction.

2.2.2 Customized 1DCNN for automated feature extraction

CNN is extensively utilized for EEG feature extraction and classification tasks due to its ability to automatically learn spatial patterns within the data. For automated features, we implemented 1DCNN following the preprocessing of EEG signals, which includes channel selection and data segmentation. Our proposed 1DCNN is composed of several distinct layers, designed to apply filters that identify essential patterns within the EEG signal. These layers are followed by activation functions and pooling layers. The activation function adds non-linearity to the network, which allows the network to learn complex patterns and relationships within the data and can highly reduce the dimensionality while keeping the critical information. The output of the extracted features was flattened and passed through fully connected layers for classification of interictal and preictal states. The feature-level fusion of handcrafted and automated features was also performed before passing them to the dense layer.

Figure 8 presents the visual description of the proposed architecture of customized 1DCNN, whereas, detailed list of parameters is listed in Table 3. It begins with a Conv1D layer featuring 32 filters of size 3, followed by batch normalization and Leaky ReLu activation to stabilize the training and add non-linearity. After that MaxPool1D layer is added for down-sampling. The network succeeded with several additional convolutional layers: 64 filters of size 3, 128 and 256 filters of size 3, each followed by ReLu activation. Average pooling is applied after the third and fourth convolutional layers to reduce dimensionality with 0.5 dropout layers to mitigate overfitting. The final convolutional layer uses 512 filters, followed by a one-dimensional global average pooling layer that aggregates the features. The architecture concludes with a dense layer with an ensemble classifier for binary classification. The total number of trainable parameters in this CNN architecture is 524,449. Figure 9 illustrates the distribution of interictal and preictal EEG segments based on 1DCNN-extracted features.

TABLE 3 Proposed architecture of 1DCNN with list of parameters.

Layer type	Output Shape	Parameters
Conv1D	(None, 5,118, 32)	608
Batch normalization	(None, 5,118, 32)	128
Leaky ReLU	(None, 5,118, 32)	0
Max pooling 1D	(None, 2,559, 32)	0
Conv1D	(None, 2,557, 64)	6,208
Leaky ReLU	(None, 2,557, 64)	0
Max pooling 1D	(None, 1,278, 64)	0
Dropout	(None, 1,278, 64)	0
Conv1D	(None, 1,276, 128)	24,704
Leaky ReLU	(None, 1,276, 128)	0
Average pooling 1D	(None, 638, 128)	0
Dropout	(None, 638, 128)	0
Conv1D	(None, 636, 256)	98,560
Leaky ReLU	(None, 636, 256)	0
Average pooling 1D	(None, 318, 256)	0
Conv1D	(None, 316, 512)	393,728
Leaky ReLU	(None, 316, 512)	0
Global average pooling 1D	(None, 512)	0
Dense	(None, 1)	513

2.3 Classification of EEG signals

Once a comprehensive feature vector is extracted, preictal and interictal class samples are then classified. Given the complex

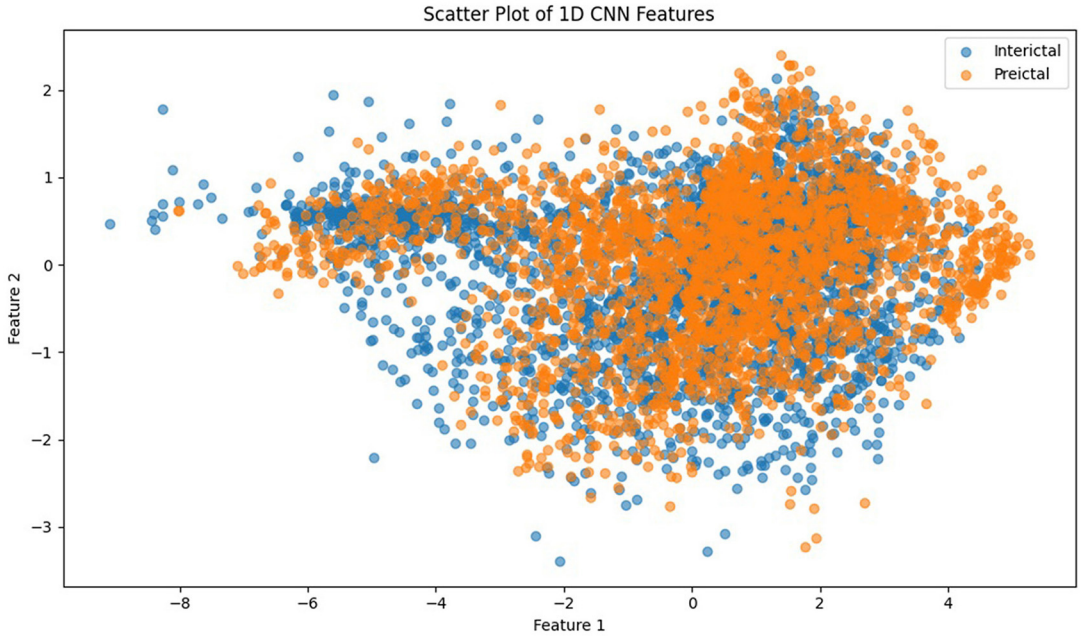


FIGURE 9
Scatter plot of 1DCNN features showing the distribution of interictal and preictal EEG segments.

Require: Training dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, base classifiers $\{C_1, C_2, \ldots, C_m\}$, meta-learner M
Ensure: Final prediction $\hat{y}$
Split D into D_{train} and D_{meta} for training base classifiers and meta-learner respectively.
for each base classifier C_k in $\{C_1, C_2, \ldots, C_m\}$ **do**
  Train C_k on D_{train}
end for
Initialize meta-training dataset $D_{\text{meta\_train}} \leftarrow \emptyset$
for each (x_j, y_j) in D_{meta} **do**
  Obtain predictions $\{p_1, p_2, \ldots, p_m\}$ from $\{C_1, C_2, \ldots, C_m\}$ on x_j
  Form meta-instance $z_j = [p_1, p_2, \ldots, p_m]$
  Add (z_j, y_j) to $D_{\text{meta\_train}}$
end for
Train meta-learner M on $D_{\text{meta\_train}}$
Prediction Phase:
Given a new instance x:
Obtain predictions $\{p_1, p_2, \ldots, p_m\}$ from $\{C_1, C_2, \ldots, C_m\}$ on x
Form meta-instance $z = [p_1, p_2, \ldots, p_m]$
Use M to predict $\hat{y}$ from z
return $\hat{y}$

Algorithm 1. Meta-learner ensemble classifier.

nature of EEG signals and subtle differences between seizure states, relying on a single classifier can limit predictive performance. Hence, we propose an ensemble approach combining diverse classifiers (SVM, RF, and LSTM) through a meta-learning strategy to enhance prediction robustness and generalizability. We propose a novel ensemble meta learner classifier with base classifiers including SVM, RF, and LSTM to perform classification between preictal and interictal classes. We used a radial basis function (RBF) kernel in SVM due to the non-linear data, which was selected empirically. Similarly, in the case of RF, we selected 150 trees after experimentation. In case of LSTM, 32 repeating units were used, followed by meta learning classifier described in Algorithm 1.

3 Results and discussion

We performed multiple experiments on the CHB-MIT dataset and evaluated the methods based on accuracy, sensitivity, and specificity. Python 3 and MATLAB were used on a Windows 11 system for the implementation. The experiments for epileptic seizure prediction are performed on NVIDIA GeForce RTX 3,090 and 64 GB of RAM. All the implementations were done using Tensorflow and Scikit-learn for seizure classification. Table 4 presents the results of the ablation study performed. Figure 10 presents the confusion matrices of all experiments. We performed multiple experiments by varying approaches in preprocessing, feature extraction, and classification. In the first experimental setup, we selected a non-overlapping window and extracted temporal and spectral features, and performed classification using a kNN classifier. With this experimental setup, we achieved an accuracy of 71.65%, sensitivity and specificity of 53.27% and 78.08%, respectively. Preprocessing and feature extraction were kept the same in experiments 2 and 3, whereas RF and SVM classifiers were used for classification between preictal and interictal states. SVM achieved an accuracy of 78.15% which was more than 4% increased compared to RF. Similarly, CNN and LSTM were used for

TABLE 4 Results obtained after performing an ablation study on the CHB-MIT dataset for epileptic seizure prediction.

Preprocessing	Feature extraction	Classification	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC	AUC-ROC
Non-overlapping window	Handcrafted features	KNN	71.65	53.27	78.08	0.2997	0.6568
Non-overlapping window	Handcrafted features	RF	73.26	59.50	78.08	0.3541	0.6879
Non-overlapping window	Handcrafted features	SVM	78.15	65.89	82.44	0.4618	0.7417
Non-overlapping window	Handcrafted features	CNN	77.02	63.71	81.68	0.4337	0.7269
Non-overlapping window	Handcrafted features	LSTM	80.01	67.91	84.24	0.5023	0.7608
Non-overlapping window, Butter-worth filter	Handcrafted features	SVM	82.47	70.56	86.64	0.5572	0.7860
Non-overlapping window, Butter-worth filter, Wavelet transform	Handcrafted features	SVM	84.09	72.90	88.00	0.5958	0.8032
Non-overlapping window, Butter-worth filter, Wavelet and Fourier transform	Handcrafted features	SVM	86.67	76.48	90.24	0.6581	0.8336
Non-overlapping window, Butter-worth filter, Wavelet and Fourier transform, channel selection	Handcrafted features	SVM	88.77	79.60	91.98	0.7101	0.8579
Non-overlapping window, Butter-worth filter, Wavelet and Fourier transform, channel selection	1DCNN	SVM	90.47	82.40	93.29	0.7532	0.8775
Non-overlapping window, Butter-worth filter, Wavelet and Fourier transform, surrogate channel	Handcrafted features	SVM	92.61	86.14	94.87	0.8081	0.9051
Non-overlapping window, Butter-worth filter, Wavelet and Fourier transform, surrogate channel	1DCNN	SVM	95.40	91.74	96.67	0.8806	0.9420
Non-overlapping window, Butter-worth filter, Wavelet and Fourier transform, surrogate channel	Handcrafted and 1DCNN feature fusion	SVM	97.01	94.86	97.76	0.9225	0.9621
Non-overlapping window, Butter-worth filter, Wavelet and Fourier transform, surrogate channel	Handcrafted and 1DCNN feature fusion	Ensemble classifier	**99.52**	**99.22**	**99.62**	**0.97**	**0.9970**

Bold entries represent the highest achieved results of each metric.

classification with the same preprocessing and feature extraction, and LSTM outperformed CNN in terms of all three performance measures.

Effective preprocessing plays an important role in the accurate prediction of epileptic seizures using EEG signals. Therefore, a Butterworth bandpass filter was applied to remove noise from EEG signals, whereas feature extraction and classification were kept the same, and an increased accuracy of 84.07% was observed. In the next experiments, preprocessing was further enhanced by applying the wavelet transform along with the Butterworth filter to increase the SNR, and it resulted in increased accuracy, sensitivity, and specificity. Similarly, the Fourier transform was also applied in addition to the Butterworth filter and wavelet transform, and the results were promising.

The choice of a fixed, non-overlapping 15-s window for EEG segmentation in our study was guided by its demonstrated effectiveness in prior seizure prediction research and its suitability for real-time implementation. However, we acknowledge that such static segmentation may result in the loss of critical information, particularly near transitional states such as the onset or termination of seizures. These transitions often contain subtle but clinically significant changes that may not be fully captured within rigid window boundaries. To enhance temporal sensitivity, future extensions of this work could incorporate overlapping windows or adaptive windowing strategies that dynamically adjust based on signal characteristics such as variance, entropy, or frequency shifts. Such approaches have the potential to capture transitional dynamics more effectively, improving both the responsiveness and predictive accuracy of seizure detection systems.

To assess the computational efficiency of the proposed framework, we evaluated the complete pipeline comprising

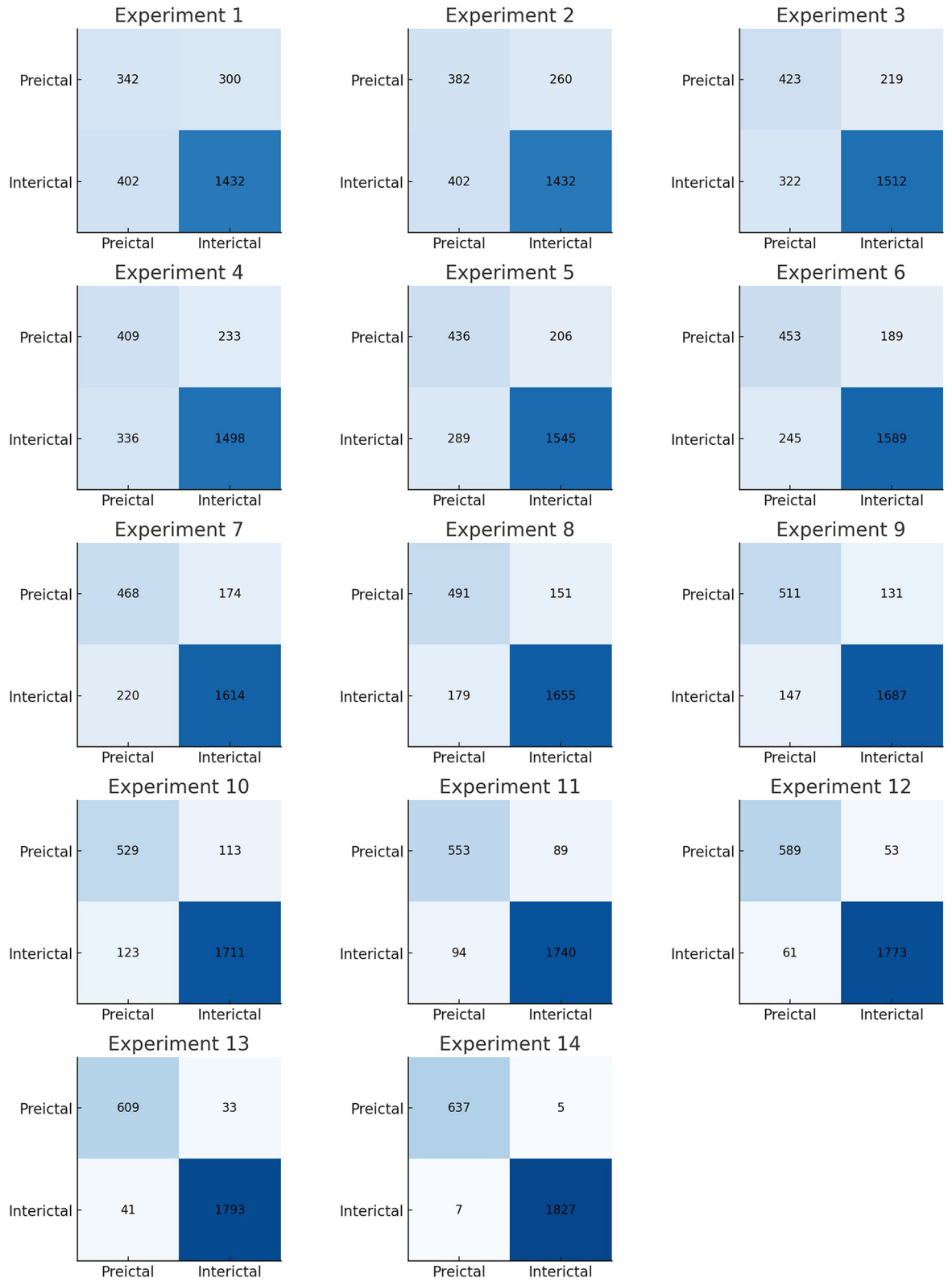


FIGURE 10
Confusion matrices of all experiments performed.

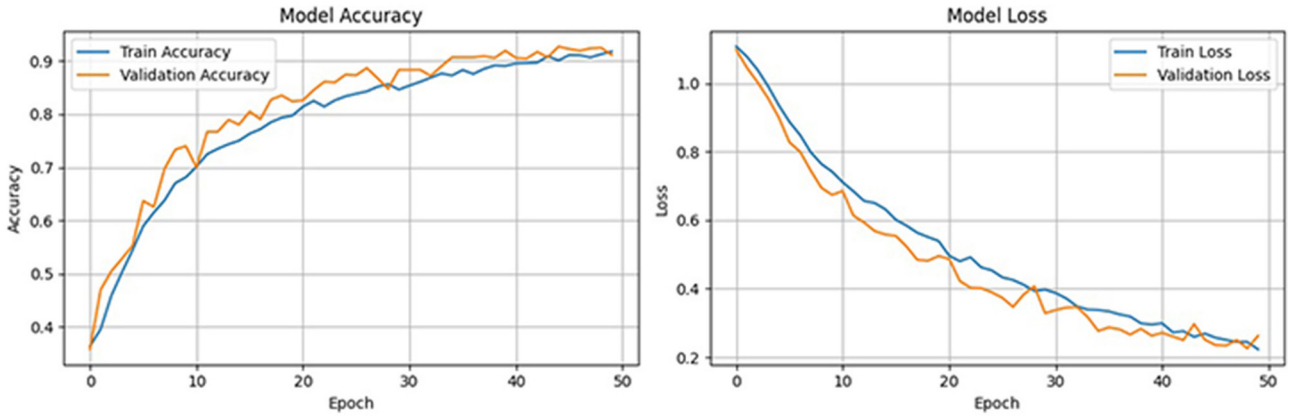


FIGURE 11
Training and validation performance curves of the proposed model over 50 epochs.

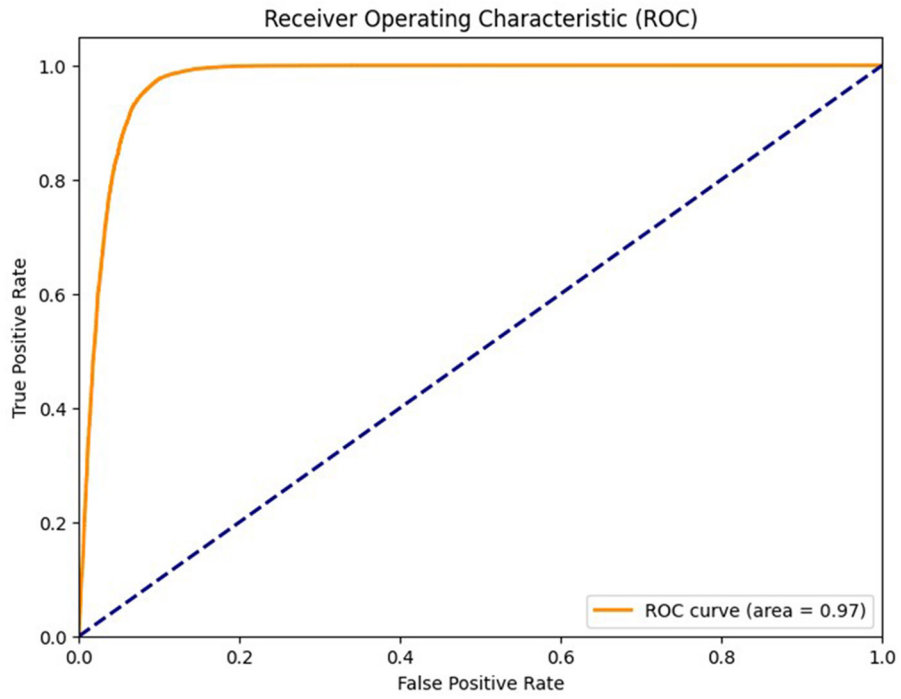


FIGURE 12
Receiver operating characteristic curve of the proposed method of epileptic seizure prediction.

preprocessing, feature extraction, and ensemble-based classification on a high-performance system equipped with an NVIDIA GeForce RTX 3090 and 64 GB of RAM. With GPU acceleration, the average processing time per 15-s EEG segment was approximately 0.12 s. This includes Butterworth filtering, wavelet and Fourier-based feature extraction, spatial filtering, and ensemble inference. The 1DCNN module benefited significantly from GPU parallelism using PyTorch, while classical models such as RF and SVM, as well as handcrafted feature operations, were efficiently handled on the CPU. All modules were implemented using optimized scientific computing libraries, including PyTorch, SciPy, and PyWavelets. The peak memory usage remained well within the hardware limits, ensuring that the proposed approach is suitable for real-time or near real-time deployment in high-throughput clinical environments.

An important aspect in real-time seizure prediction is the time taken to classify the test sample. EEG signals have high dimensionality due to the number of channels. It is extremely important to either reduce the number of channels by performing a channel selection method or by combining all channels to form a single surrogate channel. It was observed that the surrogate channel using an optimized spatial filter outperformed channel

selection. It is extremely important to extract a feature vector with high interclass variance and low intraclass variance. Therefore, we propose a customized architecture of 1DCNN that consists of five convolutional layers followed by batch normalization and max pooling. A Leaky ReLU with the value of 0.01 has been used to avoid the problem of vanishing gradients. In this research, a comprehensive feature vector is formed by concatenating the handcrafted, and features extracted using a customized 1DCNN. We also propose an ensemble classifier that uses MAML with three base classifiers, including SVM, RF, and LSTM. We used *k*-fold cross validation and were able to achieve an accuracy of 99.52% along with sensitivity of 99.22% and specificity of 99.62%, with standard deviation of 0.53, 0.61, and 0.59, respectively.

TABLE 5 Comparison of results achieved by proposed method with state-of-the-art existing methods.

Authors	Accuracy (%)	Sensitivity (%)	Specificity (%)
Birjandtalab et al. (3)	95	96.27	Not reported
Birjandtalab et al. (4)	Not reported	89.80	Not reported
Alotaiby et al. (5)	Not reported	89	37
Fei et al. (6)	89.67	89.50	89.75
Cogan et al. (7)	86	100	73
Cho et al. (8)	80.74	80.54	80.50
Jana et al. (9)	90.66	97	95.87
Daoud and Bayoumi (10)	99.60	99.72	99.6
Asharindavida et al. (11)	82.7	Not reported	Not reported
Borhade et al. (12)	96.54	96.52	97.53
Zhang et al. (13)	89.98	92.9	87.04
Usman et al. (14)	Not reported	92.7	90.8
Tamanna et al. (15)	96.38	76.73	83.16
Jana and Mukherjee (16)	99.47	97.83	92.35
Jemal et al. (17)	90.9	96.1	84.6
Koutsouvelis et al. (19)	97.32	99.31	95.34
Quadri et al. (34)	98.3	97.63	Not reported
Proposed method	99.47	97.83	92.35

To further validate the robustness of the proposed model, we computed the Matthews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUC-ROC). The ensemble classifier achieved an MCC score of 0.99, reflecting a strong correlation between predicted and actual class labels even in the presence of class imbalance. Furthermore, the AUC-ROC score of 0.997 confirms the high discriminative power of the proposed model in distinguishing between preictal and interictal states. Figure 11 shows the ROC curve of the proposed method. To evaluate the learning behavior and check for overfitting, we plotted the training and validation accuracy and loss curves, as shown in Figure 12. Table 5 compares the performance of our proposed method with recent state-of-the-art methods proposed by researchers on the same dataset, and it shows that the proposed method outperforms not only in terms of accuracy, sensitivity, and specificity but also uses less computational power due to reduced dimensionality. Although the proposed model achieves a low false positive rate during evaluation, its practical implications must be considered in continuous monitoring scenarios. Even a few false alarms per day can lead to alarm fatigue, reduced trust in the system, and clinical inefficiencies. In real-world deployment, such issues could be mitigated by incorporating post-processing techniques such as temporal smoothing, majority voting across time windows, or

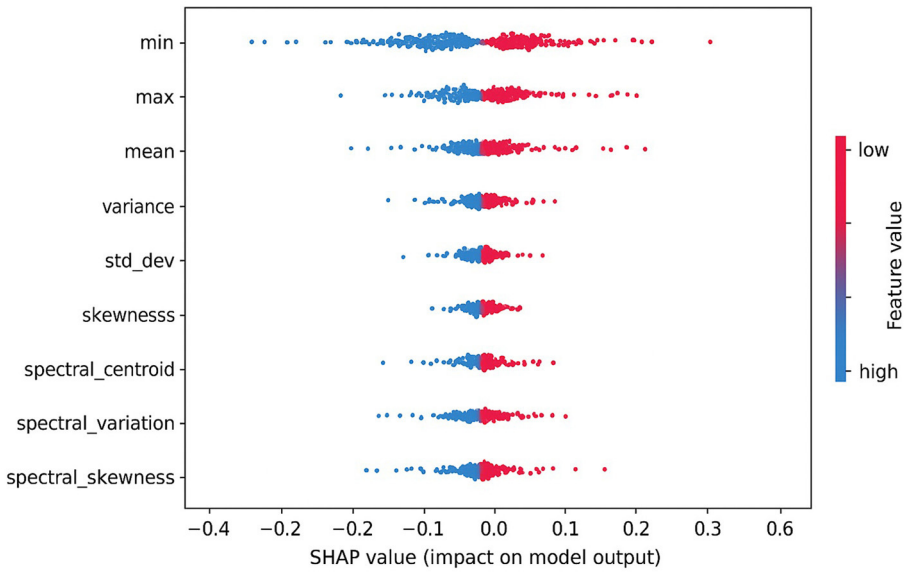


FIGURE 13
SHAP summary plot showing the impact of top handcrafted EEG features on the output of the proposed ensemble model.

hybrid decision systems that validate alerts through additional signals. These enhancements would further improve the practical viability of the proposed method in continuous, long-term monitoring contexts.

To ensure transparency in model decision-making, we applied Shapley additive explanations (SHAP) to interpret the influence of individual handcrafted features on the predicted seizure class. As shown in Figure 13, features like min, max, and mean had the most significant positive impact on the model's output. The direction and magnitude of each feature's contribution can be observed from the horizontal spread of SHAP values. For instance, high values of max and mean features (indicated in red) consistently push the model toward predicting the preictal state. This interpretability analysis enhances trust in the model's outputs and provides useful insights for potential clinical validation.

4 Conclusion and future directions

In this research, we propose a novel method for the prediction of epileptic seizures using scalp electroencephalographic (EEG) signals. The proposed method consists of three steps, including preprocessing, feature extraction, and classification. We propose a robust preprocessing method that involves conversion of 23 channels into a single surrogate channel using an optimized spatial pattern filter to reduce the dimensionality, followed by denoising using a Butterworth filter, wavelet, and Fourier transform. We also propose a customized architecture of a one-dimensional convolutional neural network (1DCNN), which is not only lightweight but also provides a feature vector with high interclass variance. Both handcrafted and 1DCNN features are concatenated to form a feature vector, which is then fed into three classifiers, including support vector machines, random forest, long short-term memory, and a model-agnostic meta learner ensemble classifier. The proposed method performs better compared to existing state-of-the-art methods in terms of accuracy, sensitivity, and specificity, and is also computationally less complex due to reduced dimensionality and a customized light-weight architecture. In the future, integrating other physiological signals, such as heart rate and blood oxygen levels, with EEG data could provide a more comprehensive understanding of seizures before onset. The proposed method can also be applied in real-time analysis of epileptic seizures. As part of future work, we plan to develop a lightweight graphical user interface to facilitate user interaction with the proposed model. This interface will enable real-time EEG data input, feature visualization, and display of model predictions and performance metrics, thereby enhancing the practical applicability of the system in clinical or research environments.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: Direct Data Link: https://physionet.org/content/chbmit/1.0.0/, Repository: PhysioNet, DOI: https://doi.org/10.13026/C2K01R.

Author contributions

YA: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. SK: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. SU: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. AJ: Investigation, Software, Validation, Visualization, Writing – original draft. MZ: Funding acquisition, Investigation, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft. HA: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. AA: Funding acquisition, Investigation, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing. SA: Funding acquisition, Investigation, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The authors extend their appreciation to the King Salman center For Disability Research for funding this work through Research Group no KSRG-2024-376.

Acknowledgments

The authors extend their appreciation to the King Salman center For Disability Research for funding this work through Research Group no KSRG-2024-376.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Mormann F, Andrzejak RG, Elger CE, Lehnertz K. Seizure prediction: the long and winding road. *Brain*. (2007) 130:314–33. doi: 10.1093/brain/awl241

2. Pitton Rissardo J, Fornari Caprara AL, Casares M, Skinner HJ, Hamid U. Antiseizure medication-induced alopecia: a literature review. *Medicines*. (2023) 10:35. doi: 10.3390/medicines10060035

3. Birjandtalab J, Heydarzadeh M, Nourani M. Automated EEG-based epileptic seizure detection using deep neural networks. In: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE (2017). p. 552–5. doi: 10.1109/ICHI.2017.55

4. Birjandtalab J, Pouyan MB, Cogan D, Nourani M, Harvey J. Automated seizure detection using limited-channel EEG and non-linear dimension reduction. *Comput Biol Med*. (2017) 82:49–58. doi: 10.1016/j.compbiomed.2017.01.011

5. Alotaiby TN, Alshebeili SA, Alotaibi FM, Alrshoud SR. Epileptic seizure prediction using CSP and LDA for scalp EEG signals. Comput Intell Neurosci. (2017) 2017:1240323. doi: 10.1155/2017/1240323

6. Fei K, Wang W, Yang Q, Tang S. Chaos feature study in fractional Fourier domain for preictal prediction of epileptic seizure. *Neurocomputing*. (2017) 249:290–8. doi: 10.1016/j.neucom.2017.04.019

7. Cogan D, Birjandtalab J, Nourani M, Harvey J, Nagaraddi V. Multi-biosignal analysis for epileptic seizure monitoring. *Int J Neural Syst*. (2017) 27:1650031. doi: 10.1142/S0129065716500313

8. Cho D, Min B, Kim J, Lee B. EEG-based prediction of epileptic seizures using phase synchronization elicited from noise-assisted multivariate empirical mode decomposition. *IEEE Trans Neural Syst Rehabil Eng*. (2016) 25:1309–18. doi: 10.1109/TNSRE.2016.2618937

9. Jana R, Bhattacharyya S, Das S. Epileptic seizure prediction from EEG signals using DenseNet. In: *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. Xiamen: IEEE (2019). p. 604–9. doi: 10.1109/SSCI44817.2019.9003059

10. Daoud H, Bayoumi MA. Efficient epileptic seizure prediction based on deep learning. *IEEE Trans Biomed Circuits Syst*. (2019) 13:804–13. doi: 10.1109/TBCAS.2019.2929053

11. Asharindavida F, Shamim Hossain M, Thacham A, Khammari H, Ahmed I, Alraddady F, et al. A forecasting tool for prediction of epileptic seizures using a machine learning approach. *Wiley Online Library*. (2020) 32:e5111. doi: 10.1002/cpe.5111

12. Borhade RR, Nagmode MS. Modified atom search optimization-based deep recurrent neural network for epileptic seizure prediction using electroencephalogram signals. *Biocybern Biomed Eng*. (2020) 40:1638–53. doi: 10.1016/j.bbe.2020.10.001

13. Zhang S, Chen D, Ranjan R, Ke H, Tang Y, Zomaya AY, et al. A lightweight solution to epileptic seizure prediction based on EEG synchronization measurement. *J Supercomput*. (2021) 77:3914–32. doi: 10.1007/s11227-020-03426-4

14. Usman SM, Khalid S, Aslam MH. Epileptic seizures prediction using deep learning techniques. *IEEE Access*. (2020) 8:39998–40007. doi: 10.1109/ACCESS.2020.2976866

15. Tamanna T, Rahman MA, Sultana S, Haque MH, Parvez MZ. Predicting seizure onset based on time-frequency analysis of EEG signals. *Chaos, Solitons Fractals*. (2021) 145:110796. doi: 10.1016/j.chaos.2021.110796

16. Jana R, Mukherjee I. Deep learning based efficient epileptic seizure prediction with EEG channel optimization. *Biomed Signal Process Control*. (2021) 68:102767. doi: 10.1016/j.bspc.2021.102767

17. Jemal I, Mezghani N, Abou-Abbas L, Mitiche A. An interpretable deep learning classifier for epileptic seizure prediction using EEG data. *IEEE Access*. (2022) 10:60141–50. doi: 10.1109/ACCESS.2022.3176367

18. Zarei A, Zhu B, Shoaran M. Enhancing epileptic seizure detection with eeg feature embeddings. In: *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. Toronto, ON: IEEE (2023). p. 1–5. doi: 10.1109/BioCAS58349.2023.10388670

19. Koutsouvelis P, Chybowski B, Gonzalez-Sulser A, Abdullateef S, Escudero J. Preictal period optimization for deep learning-based epileptic seizure prediction. *J Neural Eng*. (2024) 21:ad9ad0. doi: 10.1088/1741-2552/ad9ad0

20. Wang Y, Long X, Van Dijk H, Aarts R, Arends J. Adaptive EEG channel selection for nonconvulsive seizure analysis. In: *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*. Shanghai: IEEE (2018). p. 1–5. doi: 10.1109/ICDSP.2018.8631844

21. Arif S, Munawar S, Marie RR, Shah SA. Leveraging wavelets and deep CNN for sleep pattern recognition in road safety: an EEG study. In: *International Conference on Recent Trends in Image Processing and Pattern Recognition*. Springer: New York (2023). p. 227–41. doi: 10.1007/978-3-031-53082-1_19

22. Truong ND, Nguyen AD, Kuhlmann L, Bonyadi MR, Yang J, Ippolito S, et al. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Networks*. (2018) 105:104–11. doi: 10.1016/j.neunet.2018.04.018

23. Duun-Henriksen J, Kjaer TW, Madsen RE, Remvig LS, Thomsen CE, Sorensen HBD. Channel selection for automatic seizure detection. *Clin Neurophysiol*. (2012) 123:84–92. doi: 10.1016/j.clinph.2011.06.001

24. Parvez MZ, Paul M. EEG signal classification using frequency band analysis towards epileptic seizure prediction. In: *IEEE International conference on Computer and Information Technology*. IEEE: Bangladesh (2014). p. 126–130. doi: 10.1109/ICCITechn.2014.6997315

25. Tsiouris KM, Pezoulas VC, Zervakis M, Konitsiotis S, Koutsouris DD, Fotiadis DI. A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals. *Elsevier*. (2018) 99:24–37. doi: 10.1016/j.compbiomed.2018.05.019

26. Arif S, Munawar S, Ali H. Driving drowsiness detection using spectral signatures of EEG-based neurophysiology. *Front Physiol*. (2023) 14:1–23. doi: 10.3389/fphys.2023.1153268

27. Alickovic E, Kevric J, Subasi A. Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. *Biomed Signal Process Control*. (2018) 39:94–102. doi: 10.1016/j.bspc.2017.07.022

28. Dong Q, Zhang H, Xiao J, Sun J. Multi-scale spatio-temporal attention network for epileptic seizure prediction. *IEEE J Biomed Health Inform*. (2025) 29:4784–95. doi: 10.1109/JBHI.2025.3545265

29. Alasiry A, Sampedro GA, Almadhor A, Juanatas RA, Alsubai S, Karovic V. Epileptic seizures diagnosis and prognosis from EEG signals using heterogeneous graph neural network. *PeerJ Computer Science*. (2025) 11:e2765. doi: 10.7717/peerj-cs.2765

30. Cao X, Zheng S, Zhang J, Chen W, Du G. A hybrid CNN-Bi-LSTM model with feature fusion for accurate epilepsy seizure detection. *BMC Med Inform Decis Mak*. (2025) 25:6. doi: 10.1186/s12911-024-02845-0

31. Bajaj A, Sharma Vl. EEG-based epileptic seizure prediction using variants of the long short term memory algorithm. *Int J Comput Inf Syst Ind Manag Appl*. (2025) 17:13–13. doi: 10.70917/ijcisim-2025-0001

32. Meng Y, Liu Y, Wang G, Song H, Zhang Y, Lu J, et al. M-NIG: mobile network information gain for EEG-based epileptic seizure prediction. *Sci Rep*. (2025) 15:15181. doi: 10.1038/s41598-025-97696-8

33. PhysioNet. *CHB-MIT EEG Database*. (2024). Available online at: https://www.physionet.org/content/chbmit/1.0.0/ (Accessed August 29, 2024).

34. Quadri ZF, Akhoon MS, Loan SA. Epileptic seizure prediction using stacked CNN-BiLSTM: a novel approach. *IEEE Trans Artif Intellig*. (2024) 5:5553–60. doi: 10.1109/TAI.2024.3410928

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 22 August 2025
DOI 10.3389/fmed.2025.1600855

OPEN ACCESS

EDITED BY
Ateeq Ur Rehman,
Gachon University, Republic of Korea

REVIEWED BY
Mingfeng Jiang,
Zhejiang Sci-Tech University, China
Rao Asif,
Superior University, Pakistan

*CORRESPONDENCE
Sunnia Ikram
✉ sunnia.ikram@iub.edu.com

RECEIVED 27 March 2025
ACCEPTED 22 July 2025
PUBLISHED 22 August 2025

CITATION
Ikram S, Ikram A, Singh H, Ali Awan MD, Naveed S, De la Torre Díez I, Gongora HF and Candelaria Chio Montero T (2025) Transformer-based ECG classification for early detection of cardiac arrhythmias. *Front. Med.* 12:1600855. doi: 10.3389/fmed.2025.1600855

Transformer-based ECG classification for early detection of cardiac arrhythmias

Sunnia Ikram[1]*, Amna Ikram[2], Harvinder Singh[3], Malik Daler Ali Awan[1], Sajid Naveed[1], Isabel De la Torre Díez[4], Henry Fabian Gongora[5] and Thania Candelaria Chio Montero[6]

[1]Department of Software Engineering, The Islamia University of Bahawalpur, Bahawalpur, Pakistan, [2]Faculty of Computing, The Govt Sadiq College Women University, Bahawalpur, Pakistan, [3]Department of Mechanical Engineering, Chandigarh Group of Colleges, Landran, Mohali, Punjab, India, [4]Department of Signal Theory and Communications and Telematics Engineering, University of Valladolid, Valladolid, Spain, [5]Universidad Internacional Iberoamericana, Campeche, Mexico, [6]Universidad Internacional Iberoamericana, Arecibo, PR, United States

Electrocardiogram (ECG) classification plays a critical role in early detection and trocardiogram (ECG) classification plays a critical role in early detection and monitoring cardiovascular diseases. This study presents a Transformer-based deep learning framework for automated ECG classification, integrating advanced preprocessing, feature selection, and dimensionality reduction techniques to improve model performance. The pipeline begins with signal preprocessing, where raw ECG data are denoised, normalized, and relabeled for compatibility with attention-based architectures. Principal component analysis (PCA), correlation analysis, and feature engineering is applied to retain the most informative features. To assess the discriminative quality of the selected features, t-distributed stochastic neighbor embedding (t-SNE) is used for visualization, revealing clear class separability in the transformed feature space. The refined dataset is then input to a Transformer-based model trained with optimized loss functions, regularization strategies, and hyperparameter tuning. The proposed model demonstrates strong performance on the MIT-BIH benchmark dataset, showing results consistent with or exceeding prior studies. However, due to differences in datasets and evaluation protocols, these comparisons are indicative rather than conclusive. The model effectively classifies ECG signals into categories such as Normal, atrial premature contraction (APC), ventricular premature contraction (VPC), and Fusion beats. These results underscore the effectiveness of Transformer-based models in biomedical signal processing and suggest potential for scalable, automated ECG diagnostics. However, deployment in real-time or resource-constrained settings will require further optimization and validation.

KEYWORDS

cardiac monitoring, ECG classification, electrocardiogram analysis, PCA, t-SNE, Transformer-based model, VPC, feature engineering

1 Introduction

Electrocardiography is a primary and most used technique in cardiology that records electrical signals of the heart and analyzes the state of the heart. The increasing number of patients with CVDs, arrhythmia, myocardial infarction and heart failure proves that accurate and reliable diagnostic tools are needed (1). The initial stages of automated ECG classification were supported by convolutional models, which provided high accuracy and efficiency, although they typically relied on fixed-size kernels and local feature extraction (2). As such,

there is a growing need for automated ECG classification systems that can efficiently assist clinical decision-making and improve the quality of diagnostic results.

In recent years, the global incidents of CVDs has increased, making them one of the leading causes of death worldwide (3). ECG as a non-invasive technique is widely used for diagnosing cardiac arrhythmia and abnormalities. Prodromal signs of CVDs often manifest as irregular electrical patterns, detectable via ECG signals. For instance, cardiac arrhythmias can be fatal if not monitored properly, as they may indicate conditions leading to sudden cardiac arrest. Acharya et al. (4) employed CNN-based architectures to classify ECG signals and achieved 95% accuracy. Similarly, Liu et al. proposed the RNN-based approaches, demonstrating the ability of sequence-based models to capture temporal dependencies, achieving 95% accuracy in arrhythmia classification (5–7). These results indicate that deep learning models are well-suited for ECG classification.

The ability to differentiate between normal and arrhythmic ECG signals is critical for improving CVD diagnosis and identification (8). However, due to small amplitude variations and short- duration signals, ECG classification remains challenging. Additionally, inherent differences in ECG patterns across different CVDs, and difficulty in distinguishing similar features between patients make classification even more complex. As a result, deep learning-based automated diagnostic tools are crucial in complementing traditional ECG analysis to improve accuracy and efficiency in CVD detection. Chang and Limon (9) demonstrated that transformers could effectively classify ECG signal by focusing on the most relevant signal characteristics using the attention mechanism. Transformers can capture long-range dependencies in ECG measurements well-suited for complex classification tasks.

Building upon these advancements, this study proposes a novel Transformer-based model for multi-class ECG classification, specifically targeting five distinct classes: Normal, APC, VPC, Fusion beat and others. To enhance classification performance, a Transformer-based model is trained on refined ECG features rather than raw ECG signals, enabling better features extraction and reducing noise interference. The model is trained and tested on a publicly available ECG dataset, demonstrating its effectiveness in classifying various cardiac pathologies. To further evaluate the model's performance, various evaluation metrics are used, ensuring its reliability in real-world applications. Motivations behind this work are:

- Variability of ECG waveforms across individuals due to age, physical condition and emotional state, making it challenging to distinguish between normal and abnormal rhythms.
- Arrhythmic events often have low amplitude and short duration, making them difficult to identify amidst noise.
- Distinguishing between automatically and mechanically mediated arrhythmias remains ambiguous due to overlapping signal characteristic.
- Bio-noise, such as muscle contractions or improper electrode placement, increases signal distortion, affecting classification accuracy.
- Traditional convolutional methods used for noise reduction may also remove critical ECG features, impacting arrhythmia detection.

The analysis of electrocardiogram (ECG) data now generates better results for recognizing heart rhythm irregularities together with better classification of cardiac conditions. Modern approaches solve many problems of traditional techniques through direct ECG signal analysis which removes the requirement for human involvement (10). Recent systems, such as Transformer-based architectures, build upon CNN strengths by enabling long-range dependency modeling and adaptive attention, which enhances recognition of subtle and infrequent ECG patterns (11–13).

These approaches demonstrate strong capabilities in detecting relationships throughout long duration within ECG recordings. Their ability to detect irregular heartbeats that appear infrequently makes these methods highly effective (14, 15). The ensured reliable operation across different patient groups and improved diagnostic accuracy comes from this approach's capabilities. Real-world ECG measurements do not affect these systems because they demonstrate enhanced resistance to both interference and measurement distortions.

The ability to understand model prediction processes through these techniques increases the potential for medical practitioners to adopt the model. Transformers are particularly well-suited for capturing long-range temporal dependencies across ECG sequences, complementing the local feature extraction of CNNs.

This paper is organized:

- Section 1 presents the Literature Review.
- Section 2 describes Methodology, including data preprocessing, feature selection and model training.
- Section 3 presents the Results and Analysis, where classification outcomes are evaluated.
- Section 4 discusses Findings, Limitations and Future Research Directions.

2 Literature review

The identification and classification of cardiovascular disease (CVDs), particularly arrhythmia, remain critical areas of research due to the pivotal role of electrocardiography (ECG) in diagnosing heart disorder. Over the past few decades, various methodologies have been employed for ECG-based arrhythmia detection, ranging from classical machine learning techniques to advanced deep learning approaches, with the primary objective of enhancing accuracy, efficiency and robustness. Martis et al. (16) proposed an SVM-based classification method that relied on handcrafted features such as wavelet coefficients and heart rate variability, as discussed in Table 1. Similarly, Marinho et al. (17, 18) explored feature engineering techniques to improve arrhythmia classification. However, these models exhibit poor generalization on large databases due to their dependence on manual feature extraction, making them highly sensitive to noise and variation in patients.

To address the limitations of early rule-based and statistical ECG analysis methods, Hannun et al. (15) explored recurrent neural networks (RNNs) and LSTM architectures to preserve temporal information over longer durations. While LSTMs improved arrhythmia classification, they often struggled with vanishing gradient problems and incurred high computational costs—posing a challenge for real-time or resource-constrained deployment.

Transformer models, originally introduced for natural language processing, have recently gained traction in biomedical signal processing due to their ability to model long-range dependencies efficiently. In one of the earliest applications of Transformers to ECG

TABLE 1 State-of-the-art methods for ECG classification.

References	Techniques	Goals	Findings
Lee and Shin (30)	Hierarchical Transformer	Lead-aware ECG modeling	High-performance arrhythmia detection
Hannun et al. (31)	deep neural network (DNN)	improve the accuracy and scalability	reduce the rate of misdiagnosed
Rajpurkar et al. (32)	CNN	exceeds the performance	Exceed cardiologist performance
Arabi et al. (19)	MSW-Transformer	Multi-scale attention ECG classifier	Macro-F1: 77.85%
Ait Bourkha et al. (33)	DCETEN (1D-CNN + Transformer)	Efficient ECG classification	Accuracy: 99.84%
Kailan et al. (34)	PSO-based feature selection + SVM, KNN, RF, DT	Improve ECG classification accuracy & reduce dimensionality for IoT deployment	Accuracy: 98% (PSO-SVM) vs. 84% (non-PSO); Features reduced: 4000 → 888
Mavaddati (35)	ResNet-34 + Time–Frequency Scalogram + Transfer Learning	Classify 3 types of cardiovascular diseases (CVDs); compare with CNN, RNN, SNMF	ResNet-34 outperformed CNN, RNN, and SNMF in accuracy, sensitivity, and robustness for clinical use

signals, a 2021 study (19) demonstrated their effectiveness in arrhythmia classification. Li et al. (14) further extended this by integrating a Transformer with a 2D-UNet architecture to capture both spatial and temporal ECG features, improving classification accuracy and interpretability.

Despite their promise, Transformers also come with challenges. Training large-scale Transformer models demands significant computational resources and careful hyperparameter optimization. Additionally, their integration into clinical workflows requires further work on improving interpretability and operational efficiency. The contribution of our work is as follows:

- The proposed Transformer-based model was evaluated on five ECG arrhythmia classes: Normal, APC, VPC, Fusion Beat, and Others demonstrating its effectiveness in multi-class ECG classification tasks.
- The model exploits the attention mechanism to learn long-range temporal dependencies, offering improved performance over conventional CNN and RNN approaches.
- It addresses key challenges in ECG analysis, such as noise and signal variability, by focusing on clinically informative signal segments.
- While deployment in clinical settings remains a future goal, the model shows promise for scalable and automated ECG analysis, suitable for integration into health-monitoring systems.

Despite notable advancements in CNNs, LSTMs, and Transformer-based techniques, several key challenges persist. These include limited generalizability across datasets, vulnerability to signal artifacts, and the computational intensity required for model training and inference. Overcoming these obstacles is essential for creating robust, interpretable, and deployable ECG classification systems suitable for real-world clinical use.

3 Materials and methods

The proposed ECG classification framework is designed to detect and categorize cardiac arrhythmia using a Transformer-based deep learning model trained on preprocessed ECG signals. The system integrates data acquisition from a wearable device, such as a smartwatch, with a mobile application that transmits ECG data to a cloud server via Wi-Fi for further processing. Upon receipt, the raw ECG signals undergo a structured preprocessing pipeline that includes denoising to eliminate motion artifacts and baseline drift, normalization to standardize signal amplitude, and segmentation to extract uniform time windows for analysis.

Following preprocessing, feature extraction and selection are conducted using techniques such as principal component analysis (PCA) and correlation-based filtering to identify the most discriminative signal characteristics. These selected features serve as input to the Transformer-based architecture, which is trained in the cloud environment using supervised learning. The training phase incorporates hyperparameter tuning, loss function optimization, and regularization strategies to improve generalization and mitigate overfitting.

Once trained, the optimized model is intended for future deployment on mobile devices, where it can support real-time ECG classification. The mobile application will be able to receive ECG signals and output classification results, identifying patterns such as Normal, atrial premature contraction (APC), premature ventricular contraction (PVC), Fusion beat, and other arrhythmic events. While the system is structured for scalability and real-time analysis, on-device inference and hardware-level performance optimization remain areas of future work to ensure clinical reliability and deployment in resource-constrained settings (Figure 1).

Workflow of the proposed ECG classification system, illustrating the integration of hardware components (wearable smart watch, mobile application, and cloud server) and data processing stages including signal acquisition, preprocessing, feature extraction, Transformer-based classification, and result delivery. The framework is designed to improve the accessibility of cardiac monitoring and supports the goal of enabling earlier detection of arrhythmias, though deployment and validation on real-world hardware remain subjects for future work.

3.1 Dataset description

The dataset employed in this study comprises a collection of ECG recordings representing both normal rhythms and a range of arrhythmic conditions. All recordings are sampled at a consistent frequency, ensuring temporal uniformity across the dataset (20, 21). The dataset includes five clinically relevant classes: Normal, atrial premature contraction (APC), premature ventricular contraction (PVC), Fusion beat, and others, as illustrated in Figure 2. Although slightly imbalanced, it provides a diverse representation of common arrhythmic patterns. To ensure signal quality and reliability for downstream classification,

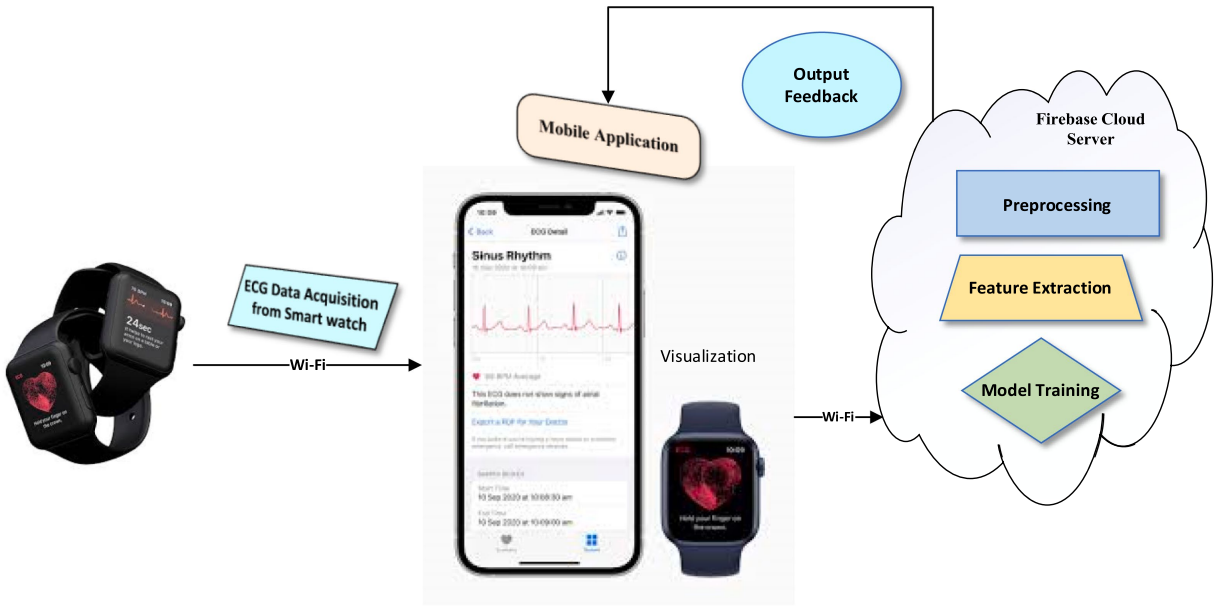


FIGURE 1
Hardware architecture of the Transformer-based ECG classification model.

preprocessing pipelines are applied to the raw ECG signals. This includes denoising, normalization, and segmentation steps, which help mitigate baseline drift, reduce motion artifacts, and standardize input lengths. These steps are essential to prepare the data for the attention-based Transformer model used in this study (6).

3.2 Data preprocessing

The preprocessing pipeline ensures the ECG signals are structured and standardized for input into the Transformer-based model. The key steps are as follows:

- Dataset loading and partitioning: The ECG dataset is first loaded and divided into training and testing subsets. Each row represents a single ECG sample, with the final column indicating the class label associated with the corresponding cardiac condition.
- Feature and label separation: The dataset is then split into feature matrices and target vectors. The features X_train, X_test → Contain Raw ECG features. While the Y_train, y_test → Contain corresponding class labels.
- Normalization: Given the variability in ECG signal amplitudes, normalization is applied to scale all feature values between 0 and 1. This mitigates amplitude-related noise, stabilizes the data distribution, and improves training convergence. The normalization is applied using the min-max scaling as shown in Equation 1:

$$x`_i = \frac{x_i - \mu}{\sigma} \quad (1)$$

Where,

o xi is the normalized signal value.
o x_i is the original signal value.
o μ is the meaning of the signal segment.
o σ is the standard deviation of the signal segment.

- Normalization not only stabilizes input ranges but also accelerates model convergence and enhances classification performance by minimizing bias introduced by amplitude variations across different recordings.
- To analyze how well the normalized features represent different heartbeat categories, the t-distributed stochastic neighbor embedding (t-SNE) technique is applied. This dimensionality reduction method maps high-dimensional ECG features into a 2D space, allowing visual assessment of class separability prior to training. This step is particularly valuable for evaluating whether the features preserve inter-class distinctions.
- Since the task involves multi-class classification, categorical labels are transformed into numerical representations using a label encoding technique. This conversion is essential for training the deep learning model, allowing loss functions and optimization routines to operate effectively on class indices.
- Transformer models require input in a sequence-based format. Thus, the ECG data is reshaped into a 3D tensor with the structure.

o Samples (batch size).
o Time steps (ECG sequence length).
o Feature (single ECG value per step).

The reshaping is illustrated in Equation 2:

$$X_{down}[i] = X[i.n^{-}] \quad (2)$$

Where,

o $X_{down}[i]$ is downsampled signal at index i.
o X[i] is an original signal.
o n is the down sampling factor.

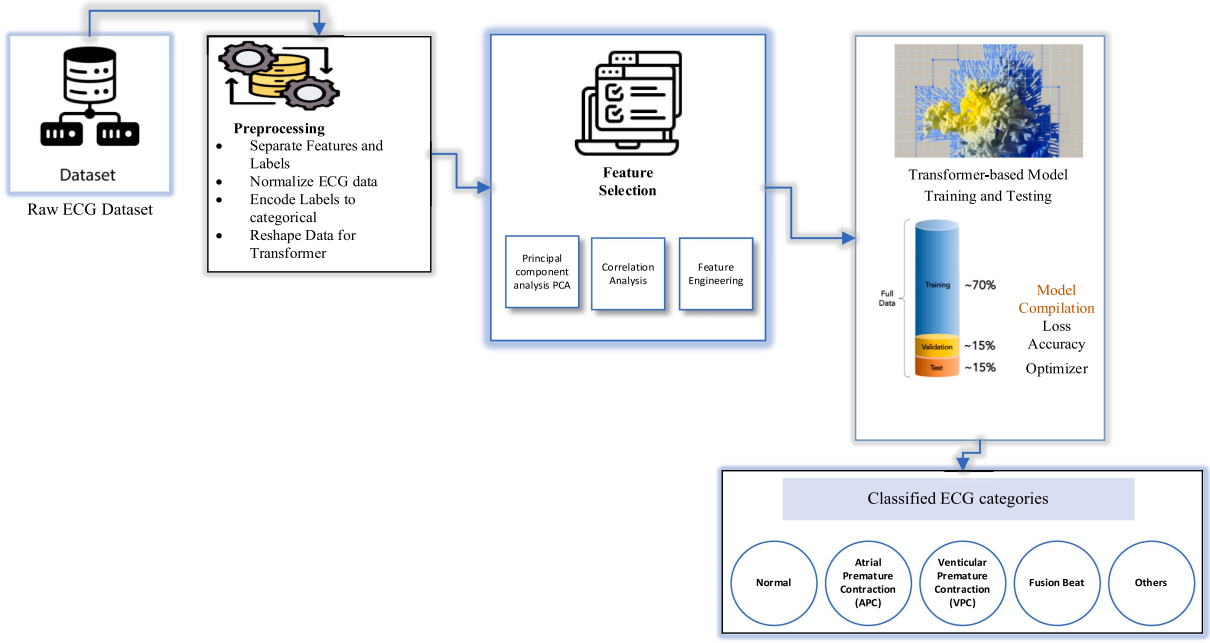


FIGURE 2
Architecture diagram of Transformer-based ECG classification model.

This reshaping enables the model to process ECG signals as temporal sequences, ensuring that temporal dependencies and waveform dynamics are preserved during training. It aligns the data structure with the self-attention mechanism used by Transformers, which excels in modeling long-range dependencies without relying on fixed kernel sizes.

3.2.1 ClassLabels

The dataset includes five distinct heart rhythm categories, each representing a specific type of arrhythmia or normal pattern:

- Normal: Represents a healthy, regular heart rhythm.
- Atrial premature contraction (APC): Premature beats originating from the atria, indicating irregular early electrical activity.
- Premature ventricular contraction (PVC): Extra systolic beats that originate in the ventricles, often associated with more serious cardiac conditions.
- Fusion beat: A waveform resulting from the combination of normal and abnormal heart contractions, leading to a hybrid signal.
- Other: Patterns that do not clearly fall into any of the above categories, encompassing miscellaneous or undefined anomalies.

Through the implementation of these preprocessing techniques, the ECG data is sanitized, segmented, and properly formatted before being fed into the Transformer-based classification model ensuring more accurate identification of a wide range of heart conditions.

3.3 Feature extraction techniques used in proposed model

Feature selection enhances model performance by identifying critical patterns and discarding irrelevant or less useful signal components (12, 13). After data preprocessing, multiple feature selection techniques are applied to ensure that only the most relevant features are retained for classification. The techniques used for feature extraction and selection include:

- Principal component analysis (PCA): A dimensionality reduction technique that transforms a set of potentially correlated variables into a smaller set of uncorrelated principal components, preserving the majority of the data's variance.
- t-distributed stochastic neighbor embedding (t-SNE): A nonlinear dimensionality reduction technique primarily used for visualizing high-dimensional data in 2D or 3D.
- Correlation analysis: Used to detect and eliminate redundant features that show strong inter-feature correlation but do not contribute independently to classification performance.
- Feature engineering: The process of generating new, domain-relevant features derived from existing data to improve model accuracy.

While PCA and correlation-based feature selection significantly improved classification performance, their clinical interpretability remains limited. The principal components produced by PCA are linear combinations of original ECG features and, while they effectively capture statistical variance, they do not directly correspond to established clinical indicators such as P-wave duration, QRS complex width, or T-wave inversion. This raises uncertainty about whether the most influential features in the model's predictions align with clinically accepted diagnostic markers used by cardiologists. This limitation underscores the need for future research that incorporates clinically annotated datasets and domain-informed feature selection strategies. Such efforts could bridge the gap between deep learning representations and clinically meaningful

interpretations, improving trust and applicability in real-world diagnostic settings.

To evaluate the discriminative quality of the extracted features before model training, we applied t-distributed Stochastic Neighbor Embedding (t-SNE) to the preprocessed dataset. As shown in Figure 3, the resulting 2D embedding reveals distinct clustering patterns for most arrhythmia types. This indicates that the features refined through PCA and correlation analysis retain sufficient discriminatory power for effective classification. The visual separation also validates the structure of the input space before learning begins, providing insight into class overlap and guiding model architecture decisions.

3.4 Transformer-based model training and testing

The Transformer-based model is trained on reshaped ECG input, where each sample represents a time-series sequence of cardiac electrical activity. The input data is formatted as a two-dimensional matrix, with dimensions corresponding to the sequence length and the feature dimension. The sequence length reflects the number of time steps (i.e., signal samples) in each ECG segment. The feature dimension represents the amplitude of the ECG signal at each time step, typically one-dimensional for raw ECG traces. This sequential structure is well-suited for Transformer architectures, which rely on self-attention mechanisms to capture long-range dependencies and temporal relationships in the input. Positional encodings are incorporated to retain temporal order information, as the Transformer lacks inherent recurrence or convolution. The model is trained using supervised learning, where ECG signals are paired with corresponding class labels (e.g., Normal, APC, VPC, Fusion Beat, Others). Training includes the use of optimized loss functions (e.g., sparse categorical cross-entropy), regularization techniques such as dropout, and hyperparameter tuning (e.g., number of attention heads, embedding dimensions, and learning rate) to improve generalization and prevent overfitting.

Tables 2, 3 illustrate the detailed architecture of the model, including the layer-wise parameters used in training. To assist with the initial level of feature extraction, the model incorporates an optional Dense Layer containing 64 neurons. This layer acts as a feature extractor, transforming the original input into a high-dimensional space (22). As a result, it highlights underlying steady-state patterns in ECG signals and enhances the model's ability to recognize complex patterns in subsequent layers. Notably, no activation is applied in this Dense Layer, ensuring that the transformation remains linear (23, 24). After passing through the Dense Layer, the data undergoes a crucial reshaping step. This step resizes the input dimensions to be compatible length and an embedding dimension of 64, optimizing it for processing within the core Transformer block.

The core component of the model is the Transformer Block, which is specifically designed to capture temporal dependencies in ECG signals. This block begins with a Multi-Head Attention mechanism consisting of four heads and an embedding size of 64. These attention heads allow the model to process multiple time segments

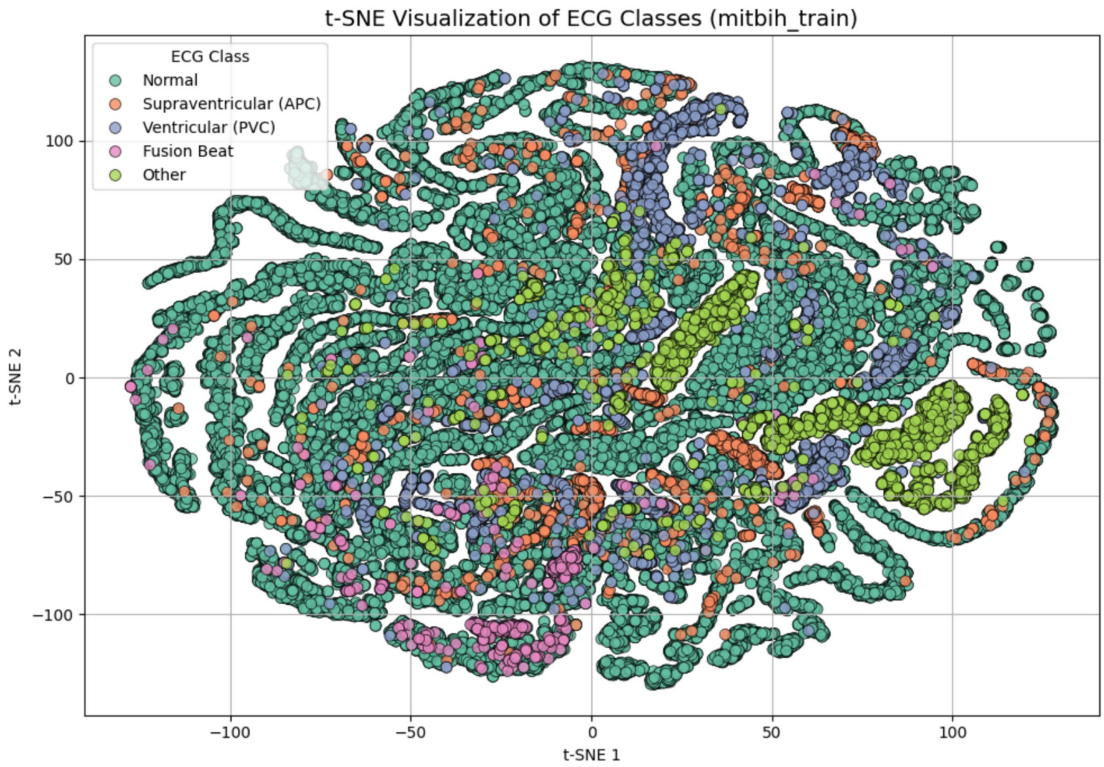


FIGURE 3
t-SNE visualization of ECG signal features after dimensionality reduction and preprocessing. Each point represents one ECG sample projected in a 2D space, colored by class label: Normal, supraventricular (APC), ventricular (PVC), fusion beat, and other. The visualization demonstrates that the extracted features possess natural class separation, indicating their suitability for classification.

simultaneously, capturing both local and global features within ECG signals. This capability is crucial for identifying arrhythmia, as different time steps may contribute to abnormal heart rhythms.

To further refine feature extraction and visualization, the model leverages t-SNE after training. T-SNE is applied to the high-dimensional feature representations extracted by the Transformer blocks, providing an interpretable 2D visualization of how ECG patterns are separated based on different heart conditions. This technique helps assess how well the model distinguishes between normal and abnormal heartbeat, enhancing its explainability in real-world applications.

The self-attention mechanism for each head is computed in Equation 3:

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^{T}}{\sqrt{dk}}\right)V \tag{3}$$

Where,

- $Q = W^{Q}$ X, $K = W^{k}X$ and $V = W^{v}X$ are the query, key and value matrices.
- d_k is the dimension of Key vector.
- W_Q, W_k, and W_v are learnable weights matrices.
- QK^T is Dot product of the query and key matrices.
- Softmax ensures attention weights sum to 1.
- Scaling by $\sqrt{d_k}$ helps with gradient stability.

For multi-head attention as shown in Equation 4:

$$MultiHead\left(Q,K,V\right)=Concat\left(head_1,\ldots\ldots,head_h\right)Wo \tag{4}$$

Where,

- h is the number of attention heads.
- $head_i$ is output of the i-th attention head.
- W^o is an output weight matrix and h is the number of heads.

For feed-forward network (FFN):
Each transformer layer includes a position-wise FFN:

$$FFN\left(x\right)=ReLU\left(xW1+b1\right)W2+b2 \tag{5}$$

Where,

- W1, W2 are weight matrices for 2 linear layers.
- b_1, b_2 are bias terms for each layer.
- ReLU activation function applied after the first linear transformation.

For layer normalization and dropout:
After each attention and FFN block, layer normalization and dropout are applied:

$$LayerNorm\left(x\right)=\frac{x-\mu}{\sqrt{\sigma^{2}+\dot{o}}}.\gamma+\beta \tag{6}$$

Where,

- μ is the meaning of x.
- σ^2 is the variance.
- ε is a small constant for numerical stability.
- Υ, β are learnable parameters.

Following the attention mechanism, the output proceeds through a Feed-Forward Neural Network (FFN) which comprises of two Dense Layers. The first Layer again makes the function non-linear by using the ReLU activation function thus enabling the model to detect higher order compounding in the data. The second layer scales the output back to the embedding size of 64 needed for attention computations. This is further added by layer Normalization that settles the training process as well as Dropout that discards some neurons at random to avoid overfitting. To address the issues of high dimensionality of the data in the model with important features preserved, the model uses Global Average Pooling Layer. This layer pools the learned features over the time steps making it easy to work on an informed representation of the entire sequence.

The output from the transformer encoder is passed to a fully connected layer for classification, where softmax activation is used to assign probabilities to each ECG class as shown in Equation 7:

$$Y=softmax\left(ZWc+bc\right) \tag{7}$$

Where,

- Z is the output from the encoder.

TABLE 2 Layer structure and parameters used in proposed model.

Layer type	Layer name	Parameters	Description
Input layer	Input	Input_shape = (X-train. Shape [1], 1)	Accepts input data reshaped to have one channel.
Flatten layer	Flatten	None	Flattens the input into a 1D array for initial processing.
Dense layer	Dense	Units = 64	Fully connected layer for initial feature extraction
Reshape layer	Reshape	Target-Shape = (−1, 64)	Reshapes the output to prepare it for the Transformer block.
Transformer block	TransformerBlock	Embed_dim = 64, num_heads = 4, ff_dim = 64	Custom layer implementing multi-head self-attention and feed-forward networks.
Global average pooling	GlobalAveragePooling1D	None	Reduce the output sequence to a single vector by averaging.
Output layer	Dense	Units = num_classes, activation = 'softmax'	Final layer for classification, providing class probability.

TABLE 3 Transformer block breakdown of the proposed model.

Component	Parameters	Description
Multi-head attention	Num_heads = 4, key_dim = 64	Computers attention scores for different subspaces of the input.
Feed-forward network	Dense_layers: [64, 64]	It consists of two dense layers with a ReLU activation in between.
Layer normalization	Epsilon = 1e-6	Normalize the output for better training stability.
Dropout	Rate = 0.1	Regularization to prevent overfitting, applied after attention and feed-forward layers.

- Wc is the weight matrix.
- b_c bias terms for the classifier.

The output of the layer is feed to the Dense layer and a Softmax activation function is used. This last step computes probability for each of the five ECG classes which makes it possible for the model to perform multi-class classification. The model's prediction is based on the maximum probability, which shows to which category the ECG signals belong, thus helping to diagnose arrhythmia correctly.

4 Results and evaluation metrics

The Transformer-based model's performance was measured using various metrics to provide a comprehensive evaluation of its classification capabilities. The model achieved a final validation accuracy of 97% after 10 epochs, reflecting strong generalization on unseen data.

The correlation heatmap in Figure 4 depicts relationships among ECG features. Strong correlations (values close to 1 or −1) suggest redundancy, which guided the feature selection process using PCA. Features with low correlation were preserved to retain signal diversity. These insights helped reduce dimensionality while maintaining important clinical features. In this heatmap, every cell indicates the correlation between the two features of the bioinformatics dataset based on a coefficient varying between −1 to 1. Here, a value close to 1 reveals positive correlation, which makes one feature dependent on the other, whereas if one rises the other is also likely to rise. On the other hand, the value will be near −1, if the features are negative, thus suggesting that one of the features increases the other is likely to decrease (25). The heatmap uses a color gradient where darker colors signify higher positive correlation, lighter color signify low or negative correlation and black areas signify low correlation. Since each feature is compared to itself on the diagonal of the heatmap, it is obvious that the correlation between features would be 1. Some blocks in the heatmap contain areas with a clearly higher correlation, that can be attributed to groups of features that likely possess similar characteristics or possibly act in concert to manifest certain patterns in the ECG signals (26). Some blocks in the heatmap contain areas with a clearly higher correlation, that can be attributed to groups of features that likely possess similar characteristics or possibly act in concert to manifest certain patterns in the ECG signal. These observations indicated the possible redundancy or relevance of feature groups and might be helpful for the feature selection or dimensionality reduction (27). The lighter-colored areas or the areas with correlations near zero show that the features of these regions are least dependent on each other. Such features may be valuable for capturing some of the temporal qualities of the ECG signals that may be essential for the classification of arrhythmias. The heatmap analysis may show how various features related to arrhythmia are related to each other based on the pattern analysis. For instance, some attributes might appear to be more effective in identifying sorts of cardiac pathologies, knowledge of which can help to determine the model's architecture.

Figure 5 visualizes class imbalance in the dataset. The "Normal" class dominates with 18,000 samples, compared to 560 for APC and 1,400 for VPC. This imbalance motivated the use of augmentation and class-weighted training to prevent overfitting toward the majority class and improve minority class detection. The above figure provides the visual representation of the class distribution in the dataset, offering a clear view of the count of samples in each category (28). By using a heatmap, it emphasizes the significant class imbalance where the Normal class has a much larger sample size compared to other classes like APC, VPC, Fusion Beat and others. This disparity may impact the model's performance, potentially leading to bias toward the majority class during training.

Figure 6 shows the progression of training and validation accuracy/loss over 10 epochs. Accuracy steadily increased while loss decreased, with both curves converging by the 10th epoch. This indicates minimal overfitting and efficient learning. This trend suggests that the model is learning effectively and improving its predictions over time (29). The closeness of the training and validation accuracy curves indicates minimal overfitting, as the validation accuracy closely follows the training accuracy. The right curve, loss curve, augments downward with the training time, showing less error of prediction. The training and validation losses converge closely by the final epoch, indicating stable performance, which is additional evidence of model performance on unseen data. But the early epoch oscillates a bit, and this could mean the model is making changes to the learning rate or complexities in some classes. All these plots show that the model performed very well and with little overfitting which implies that there was good or sufficient balancing between the training and the validation accuracy models. This model appears well-optimized for this dataset, though further comparisons with baseline models are required to confirm its superiority, since both the accuracy and the loss rate converge quite steadily.

4.1 Quantifying the impact of PCA

While both principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) were used in the study, their individual contributions were distinctly different. PCA was applied as a dimensionality reduction technique prior to training, aiming to eliminate redundancy and retain the most informative features. To evaluate its effectiveness, an ablation experiment was conducted where the Transformer model was trained once with PCA and once without PCA, using the same training configuration (Table 4).

These results confirm that PCA significantly improved model performance by reducing feature noise and enhancing separability in the feature space. In contrast, t-SNE was used exclusively for visualization to illustrate class-wise separability and decision

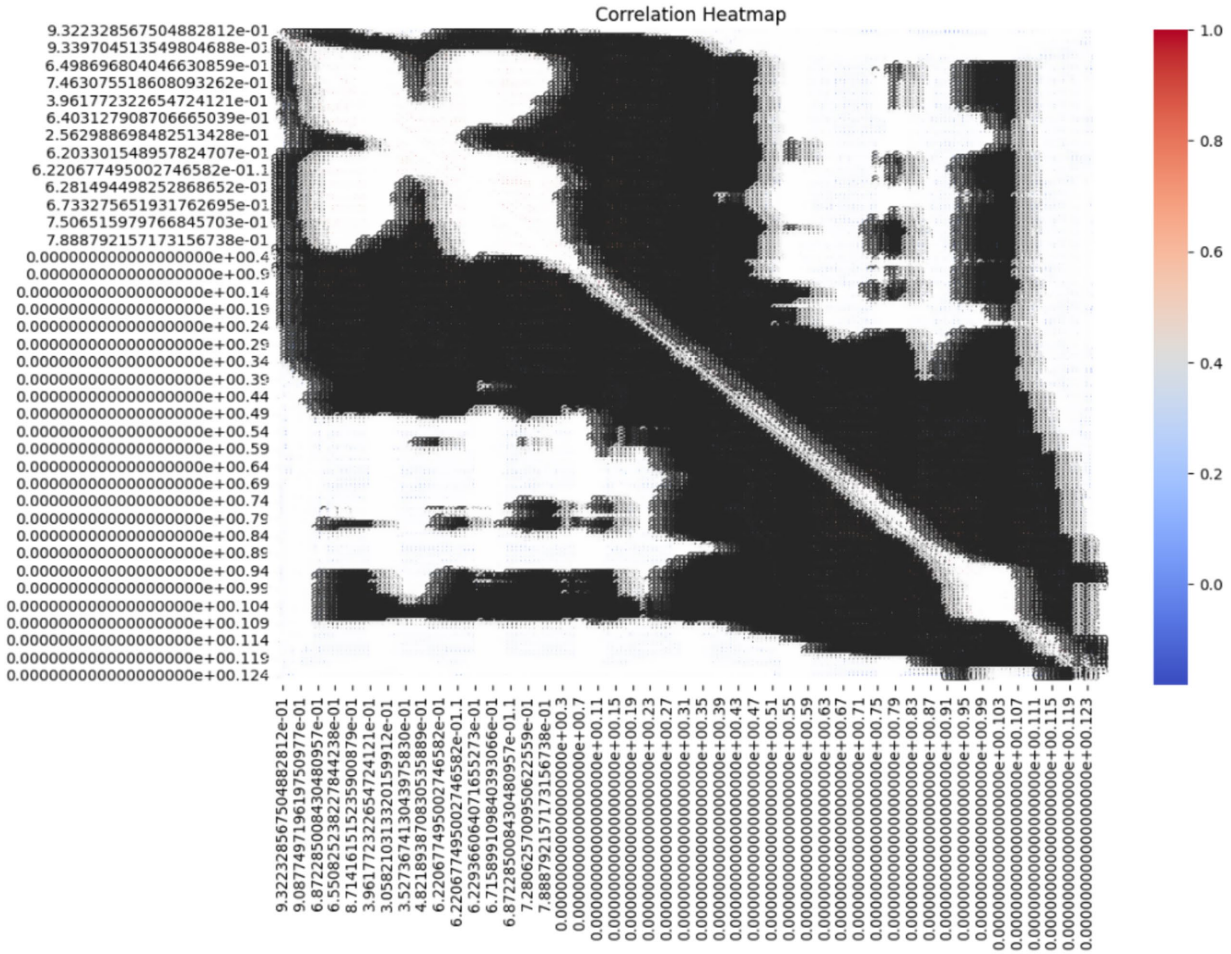


FIGURE 4
Correlation heatmap of the ECG dataset.

boundaries in a reduced feature space. It was not used during training and did not influence model accuracy directly.

To interpret the model's behavior after training, we applied t-distributed Stochastic Neighbor Embedding (t-SNE) to the learned feature embeddings and visualized the decision boundaries for each ECG class. As shown in Figure 7, the background color represents the class regions predicted by the trained Transformer-based model, while the overlaid dots indicate actual test samples projected into the 2D t-SNE space. The clear separation in some regions particularly for classes like "Normal" and "Fusion Beat" indicates strong class-specific learning. However, overlapping regions involving "APC" and "VPC" reflect residual class confusion, consistent with class imbalance and similar signal morphology. This visualization confirms that the model has successfully learned a meaningful embedding space for ECG classification, while also highlighting opportunities for further refinement.

Figure 8 illustrates the precision, recall, and F1 score for each ECG class, reflecting the model's classification performance across different arrhythmia types. The results show that the model achieves near-perfect precision and F1 scores for the "Normal," "Fusion Beat," and "Other" categories, indicating excellent classification for these classes. For the *Atrial Premature Contraction* (APC) class, the model demonstrates strong recall (100%), suggesting it detects nearly all APC instances. However, the precision is relatively low, resulting in an F1 score above 85%. This implies the model over-predicts APC, likely due to its confusion with similar classes such as Normal. The *Ventricular Premature Contraction* (VPC) class exhibits the weakest performance, with noticeably lower recall and F1 score. This may be due to class imbalance and the morphological similarity of VPC to APC and Fusion Beat in ECG waveforms particularly within the QRS complex, where overlapping features can confuse the classifier.

Interestingly, the VPC class shows a perfect AUC (1.00), indicating that the model is capable of ranking VPC instances correctly. However, the low recall suggests that classification thresholds or insufficient representation in the training data may limit actual detection. This highlights the need for possible threshold adjustment or targeted data augmentation.

Figure 9 displays the ROC curves for each ECG class in the classification model, showing the trade-off between the true positive rate (TPR) and false positive rate (FPR) at various classification thresholds. The ROC curve is a standard diagnostic tool to evaluate the model's ability to distinguish between different classes. The area under the ROC curve (AUC) provides a scalar measure of this discriminative ability. AUC values closer to 1.0 indicate excellent class separability, while values near 0.5 suggest random guessing. In this

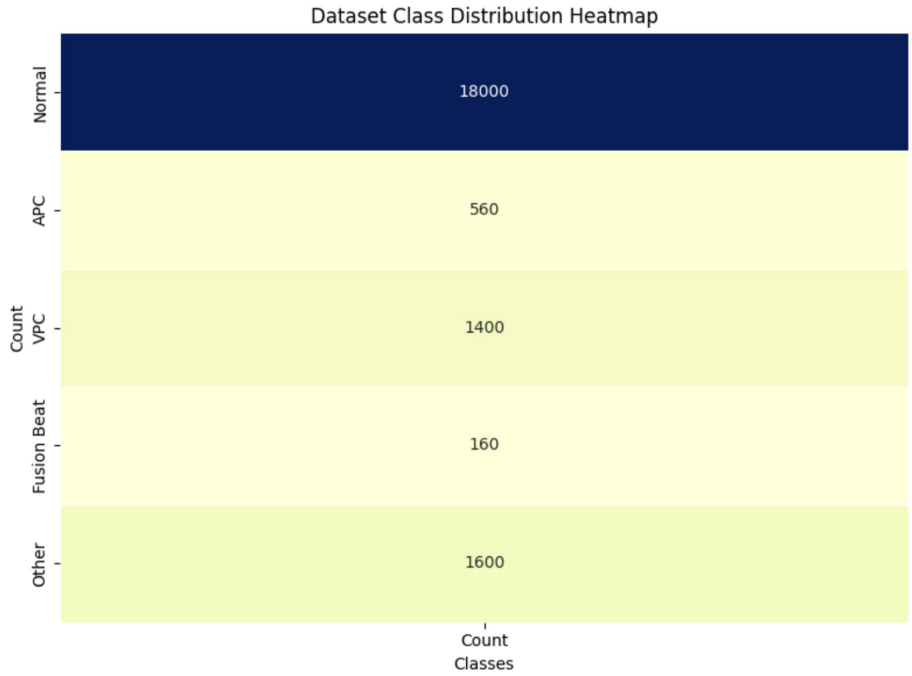


FIGURE 5
Class distribution heatmap of ECG dataset.

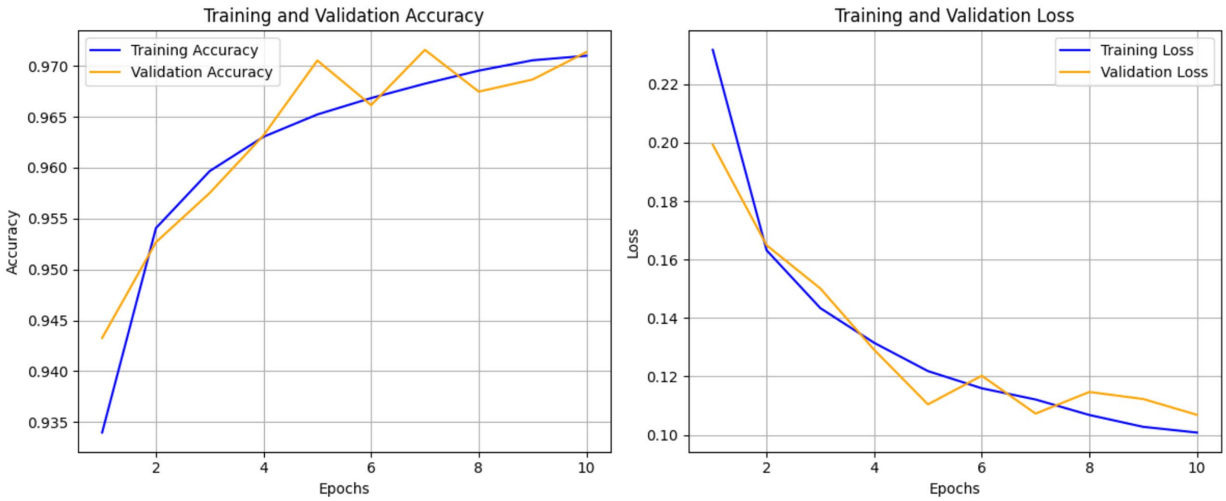


FIGURE 6
Training and validation accuracy and loss over 10 epochs of ECG classification model.

TABLE 4 Impact of PCA on the model performance.

Model setup	Accuracy	AUC
Without PCA	92.3	0.91
With PCA	97.1	0.96

model, all ECG classes achieved high AUC scores, reflecting strong performance:

- Normal: 0.98.
- APC: 0.94.
- VPC: 1.00.
- Fusion Beat: 0.98.
- Other: 1.00.

These results indicate that the model is highly capable of distinguishing between the different rhythm types, even for more challenging arrhythmias like APC and VPC. Despite some misclassifications seen in the confusion matrix and F1 scores (particularly for VPC), the high AUC values suggest that the model's ranking ability is robust. This discrepancy implies that classification

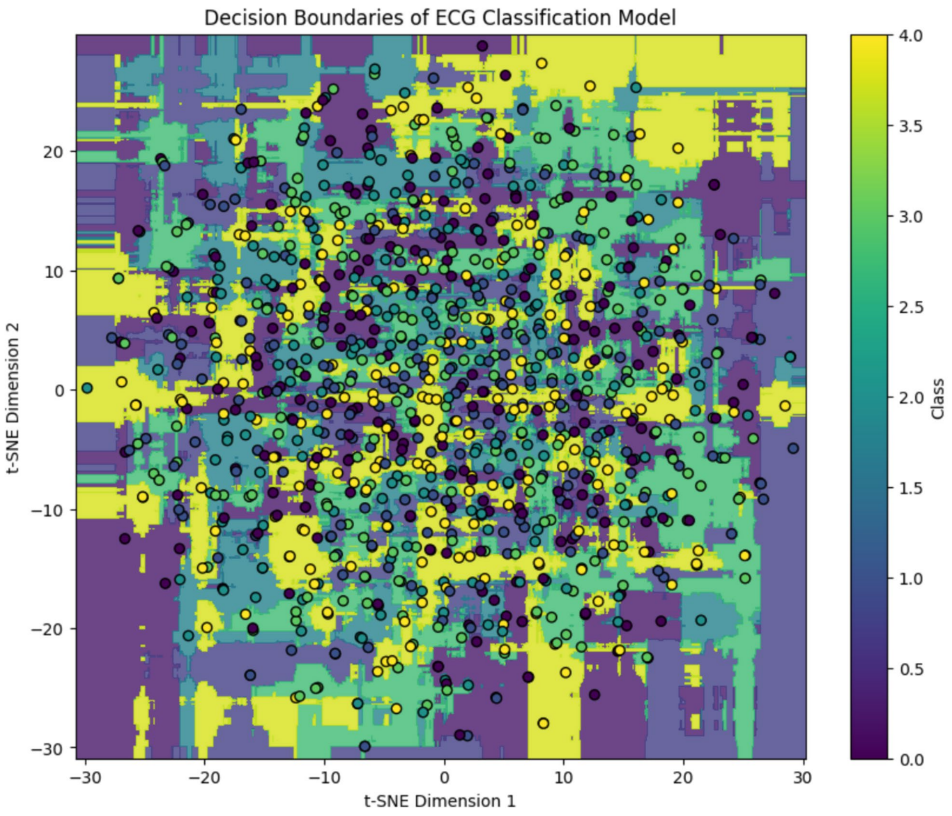


FIGURE 7
Post-training t-SNE decision boundary visualization of the ECG classification model. Background regions indicate model-predicted class clusters, and colored circles represent projected ECG samples. While distinct clusters emerge for dominant classes, class overlap remains in minority arrhythmias.

thresholds, class imbalance, or feature overlap might be affecting precision and recall, rather than the model's core ability to separate classes. Therefore, further improvements could be made through threshold tuning, class-specific loss weighting, or augmentation strategies, rather than architecture changes:

$$\mathrm{TPR}(t) = \mathrm{TP} / \mathrm{TP} + \mathrm{FN}, \mathrm{FPR}(t) = \mathrm{FP} / \mathrm{FP} + \mathrm{TN} \quad (8)$$

The final AUC score is computed by integrating the area under the ROC curve.

Figure 10 presents the normalized confusion matrix, providing a detailed view of the model's classification performance across ECG rhythm categories: Normal (0), APC (1), VPC (2), Fusion Beat (3), and Other (4). Each cell indicates the percentage of instances from a true class (rows) predicted as a certain class (columns). Diagonal values represent correct classifications, while off-diagonal values indicate misclassifications.

The matrix shows excellent performance on the Normal class, with 99.5% of samples correctly classified, reflecting the model's high sensitivity and specificity for detecting normal heartbeats. The "Other" category also shows strong results, with over 96% correctly identified.

However, some confusion is evident among arrhythmic classes:

- APC is often misclassified as normal (33.1%), despite a high recall.
- Fusion Beat is frequently predicted as normal (43.8%), suggesting difficulty in distinguishing Fusion morphology from typical ECG rhythms.
- VPC shows good accuracy (86.7%), but a small portion is misclassified as normal (10.2%) or fusion (2.4%).

These misclassifications likely arise from morphological similarities in the QRS complexes and overlapping waveform features across arrhythmia types. In particular, the confusion between APC and Normal, and Fusion and Normal, may stem from subtle variations in signal patterns that challenge the model's feature extractor.

To enhance class separability, especially for VPC and Fusion, future work could focus on improving the feature extraction pipeline, incorporating class-specific augmentation, or using contrastive learning techniques to better differentiate similar waveform classes in the learned embedding space.

The model was trained to minimize sparse categorical cross-entropy loss, which quantifies the difference between the predicted probability distribution ppp and the true distribution qqq. The loss function is defined in Equation 9:

$$Loss = -\Sigma N\, i = 1\, qi \log(pi) \quad (9)$$

Where,

- N is the number of classes.
- q_i is 1 for the correct class and 0 otherwise.
- p_i is predicted for class I.

FIGURE 8
Highlighting model performance across various arrhythmia types.

Overall accuracy, which is the ratio of correctly predicted instances to the total number of instances, is defined as:

$$Accuracy = True\ Positive + True\ Negatives / \text{Total Instances} \quad (10)$$

These metrics help to evaluate the model's performance in each class:

Precision measures the accuracy of positive predictions:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (11)$$

Recall measures the model's ability to capture all relevant instances:

$$Recall = True\ Positives / True\ Positives + False\ Negatives$$

F1-score is the harmonic mean of precision and recall, balancing the two metrics:

$$F1 - Score = 2 * Precision * Recall / Precision + Recall \quad (12)$$

When evaluating the model with respect to precision, recall and F1-score as well as the analysis of the confusion matrix, the model would be strong in predicting classes that make the majority such as 29 K, 44 K thus indicating the areas that would require improvement in the minority classes such as APC and VPC. The model architecture could also be improved further and overspecification of hyperparameters could be done to achieve a better balance among all classes.

5 Comparative evaluation of transformer variants

To demonstrate the effectiveness of our proposed model, we compared its performance with other state-of-the-art Transformer-based ECG classifiers, including ECG-BERT, time series transformer (TST), and Informer. These models were selected based on their recent use in biomedical signal processing and sequential data tasks.

Table 5 summarizes the comparative performance of various state-of-the-art Transformer-based models applied in biomedical signal classification. Among them, ECG-BERT, Informer, and time series transformer (TST) demonstrate strong performance on arrhythmia detection tasks, with AUC scores ranging from 0.94 to 0.95. These models leverage attention mechanisms to effectively model temporal dependencies within ECG signals. MN-STDT model proposes a brand-new multimodal framework, where chest X-ray spatial features and EHRs temporal features are combined, with an AUC of 0.8620 in in-hospital mortality prediction of heart failure. Despite not being directly applicable to ECG classification, MN-STDT demonstrates the increased nexus of multimodal Transformer models in clinical research and their ability to perform more context-aware predictions. In their turn, the suggested Transformer model of the present research, based on the use of the PCA-based feature selection, engineered representations as well as t-SNE visualization, attains higher performance, with an accuracy ratio of 97.1, F1-score rate of 0.95 and the value of AUC equals to 0.96. It suggests that, besides the overall success of the Transformer backbone at modeling ECG sequences, the well-optimized preprocessing, dimensionality reduction, and hyperparameters tuning play a central role. As opposed to other models, the proposed one has a high degree of interpretability and generalization to different classes of ECG, indicating its strong potential to be broadly integrated into the clinical routine in automated pipelines of ECG analysis.

6 Ablation study of hyperparameter settings

An ablation study was undertaken to assess the effectiveness of parameter ablation by varying the number of attention heads, the size

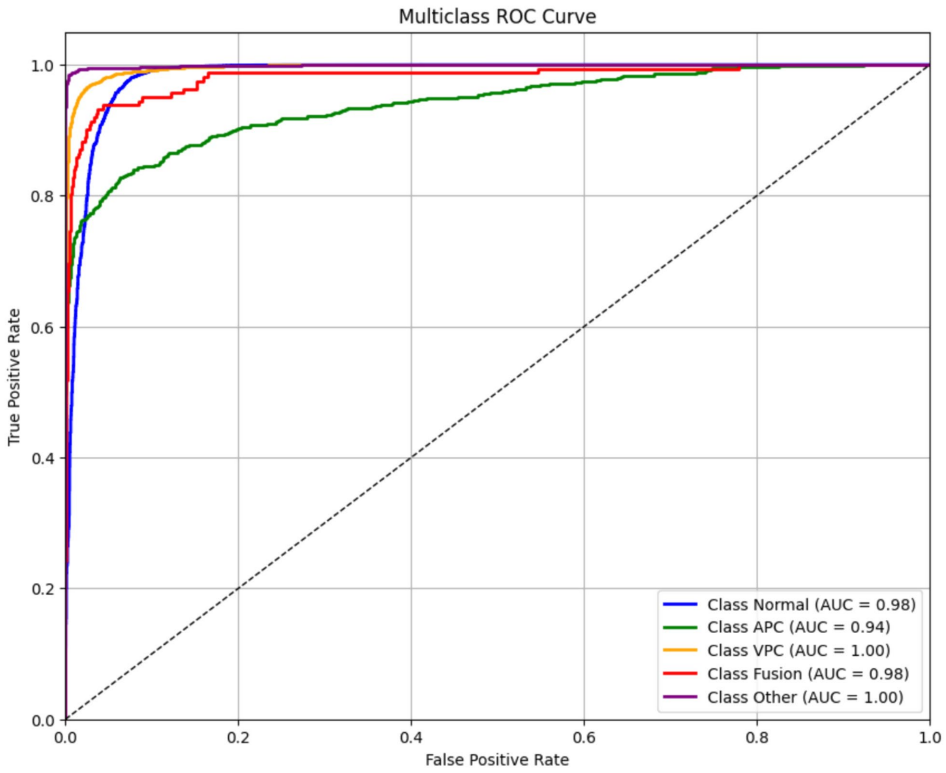


FIGURE 9
Receiver operating characteristic (ROC) curves for ECG classes.

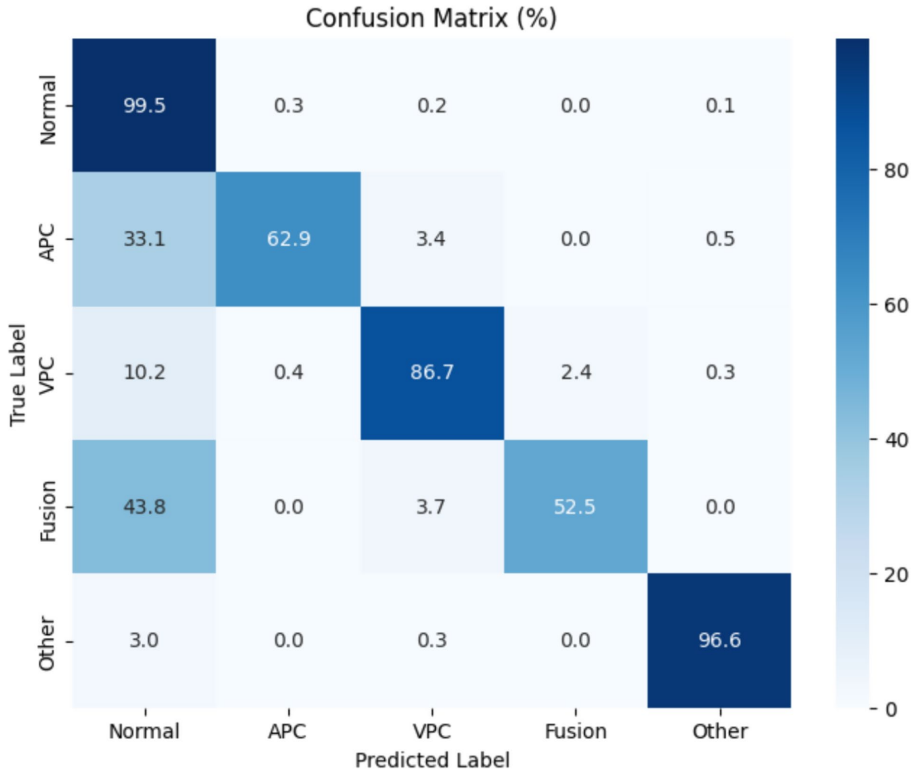


FIGURE 10
Confusion matrix for Transformer-based ECG classification model across classes.

TABLE 5 Comparison with Transformer-based and SOTA ECG models.

Model	Architecture	Accuracy	F1-score	AUC	Reference
ECG-BERT	Pre-trained transformer (BERT-based)	94.6	0.92	0.94	(36)
Time series transformer (TST)	Encoder-only transformer with positional encoding	95.3	0.93	0.95	(37)
DRL-ECG-HF	DRL + Multi-instance learning + PER + SHAP	–	0.58	9.90	(38)
MN-STDT	Spatially and temporally decoupled transformer with multimodal fusion (CXRs + EHR)	–	–	0.86	(37)
Proposed transformer model	Transformer + PCA + Feature engineering	97.1	0.95	0.96	Current study

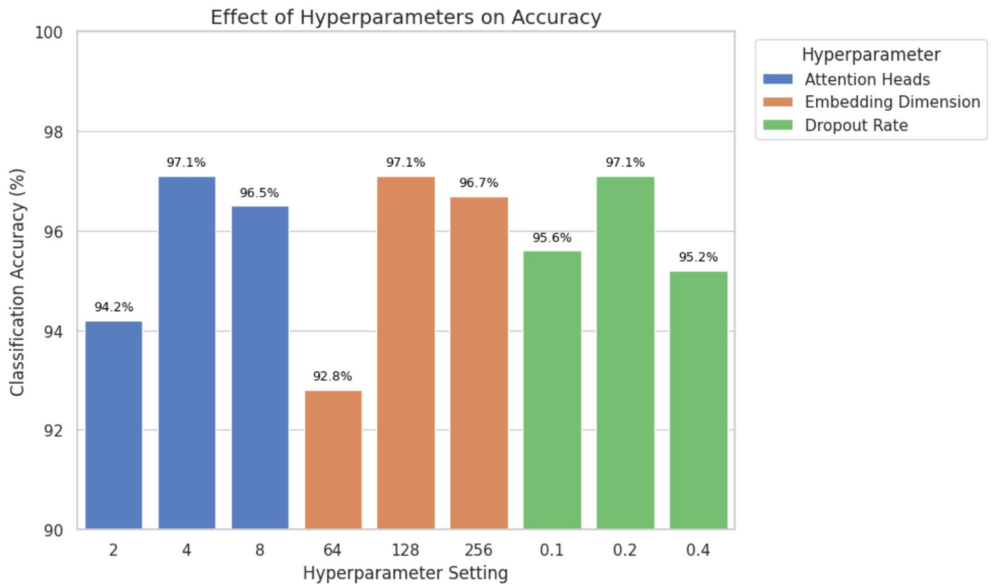


FIGURE 11
Ablation study showing the effect of attention heads, embedding dimension, and dropout rate on classification accuracy. Optimal performance was achieved with 4 attention heads, 128-dimensional embeddings, and a dropout rate of 0.2.

of embedding dimension and dropout rate independently. This analysis aimed at finding the most optimal values that would offer classification accuracy and model complexity. Table 5 presents the classification accuracy and AUC values obtained by modifying one hyperparameter at a time while keeping the others constant (Figure 11).

The results demonstrate that using four attention heads and an embedding dimension of 128 achieved the highest classification accuracy and AUC without significantly increasing the computational cost. A dropout rate of 0.2 provided effective regularization, reducing the risk of overfitting while preserving performance. Higher dropout values (e.g., 0.4) led to underfitting, while lower values (e.g., 0.1) increased variance during training. These findings support the final hyperparameter configuration used in the proposed model and confirm that the selected values contribute meaningfully to improved classification outcomes, particularly for clinically relevant ECG classes.

7 Discussion

The transformer model as applied to the ECG has high classification accuracy across various classes of arrhythmias which implies that the model can handle temporal variability and complex morphologies of the ECG signals. By using self-attention, the model learns dependencies that are long-range without constraints to fixed-size temporal windows and recurrent architecture. This is because it can accommodate ECG sequences with different sequences and dynamics; this is a common feature in clinical data. Consequently, the sensitivity to the slight variation of the waveforms which is important in identifying the classification of arrhythmia is better enhanced on the model. Although CNN-based models have shown strong results in ECG analysis and remain widely used in clinical and research settings, their reliance on local receptive fields limits their capacity to capture long-range dependencies. Transformers overcome this by using self-attention mechanisms that dynamically model relationships across the entire signal length. Conversely, Transformer global attention mechanism better captures temporal dependencies to yield better classification results. Although model performance is one of the priorities, explaining the model still is a big challenge. Transformer-based models can be regarded as black boxes and even the presented techniques such as visualization of attention weights may provide some insight into the models, but this paper does not envisage an analysis of interpretability. Future research is advised to include implementations of explainability algorithms like attention mapping or SHAP analysis, seeking to make the inclusion of such systems more clinically acceptable and easy to adapt to.

Additionally, although architecture holds potential for integration into edge devices and wearable technologies, this study does not

evaluate inference latency, computational resource requirements, or hardware deployment feasibility. As such, claims regarding real-time or mobile deployment are beyond the scope of this work. Future research may explore model simplification, quantization, or pruning strategies to enable deployment in resource-constrained environments, such as wearable health monitoring systems.

Overall, this study underscores the applicability of Transformer architectures to biomedical signal classification tasks, particularly ECG interpretation, and provides a foundation for future research focused on explainability, deployment, and clinical validation.

8 Limitations

Although the current Transformer-based ECG classification model shows promising results, several limitations must be acknowledged.

First, the dataset used for training and evaluation lacked significant diversity and exhibited class imbalance. While the model performed well on majority classes such as "Normal" and "Fusion Beats," it underperformed on minority classes like "Ventricular Premature Contractions (VPC)," which had relatively few samples. This imbalance likely affected the model's ability to accurately classify rare arrhythmias and limits its generalizability to diverse or unseen clinical scenarios.

Second, the transformer architecture is computationally intensive, both during training and inference. Memory and processing demand pose challenges for deployment in resource-constrained environments, such as mobile or wearable healthcare devices. This limitation impacts the model's scalability and increases the cost and complexity of real-world implementation.

Third, interpretability remains a significant concern. Despite the theoretical advantages of attention mechanisms in revealing important features, Transformer-based models continue to function largely as black boxes. Current attention visualization techniques provide limited insight into the model's reasoning, which hinders clinical trust and diagnostic transparency. Clinicians require explainable models to validate predictions and make informed decisions, and the lack of interpretability restricts practical adoption in healthcare settings. Finally, direct comparison with prior studies is constrained by inconsistencies in datasets, preprocessing pipelines, and evaluation metrics. Although Table 5 summarizes performance metrics and limitations of previous approaches, such comparisons should be interpreted cautiously due to differing experimental setups.

In summary, these limitations underscore key areas for future improvement, including addressing class imbalance, optimizing model efficiency for deployment, and enhancing model transparency. Addressing these challenges is essential to advance the clinical applicability of deep learning-based ECG analysis systems (Table 6).

Lastly, generalizability remains a fundamental concern due to the homogeneity of the dataset, which was collected from a specific demographic using a single device type. ECG signals can vary across different populations, age groups, and acquisition devices, potentially affecting the model's performance in diverse clinical settings. As a result, the effectiveness of the proposed model may be limited when applied outside the specific context in which it was trained.

To enhance generalizability and clinical robustness, future studies should aim to validate the model on datasets collected from multiple sources, encompassing both homogeneous and heterogeneous subject groups. This includes variations in age, ethnicity, health conditions, and recording hardware. Such external validation would provide a stronger basis for assessing the model's adaptability and reliability in real-world clinical environments.

9 Future work

In subsequent studies, efforts will focus on enhancing the robustness, clinical reliability, and deployment readiness of Transformer-based models for ECG classification.

TABLE 6 Limitations of various approaches used in ECG classification.

References	Model	Accuracies	Limitations of previous work	Limitations of current transformer model
Smith et al. (39)	Transformer-based model for ECG diagnosis	Evaluated by sensitivity, PPV, and detection of major abnormalities	Lower accuracy in detecting major abnormalities; higher false positives/negatives leading to reduced diagnostic reliability	Sensitive to ECG noise; misclassification of subtle abnormalities; requires large, annotated datasets; trade-off between sensitivity and specificity
Zhao et al. (40)	CNN-RNN (Deep Convolutional Neural Network – Recurrent Neural Network)	97.6% (for 2-s ECG segments)	Lacked real-time inference; limited performance in heart failure staging; complex feature extraction pipeline	Requires intensive preprocessing (segmentation, augmentation); limited capacity to capture long-range dependencies
Chithra et al. (41)	ANN-based Decision Tree	93.4%	Poor integration of clinical and ECG features; low model interpretability	High feature engineering cost; poor scalability to multilead/multiclass ECGs
Arabi et al. (19)	MSW-Transformer	Macro-F1 up to 77.85%	CNN only captures local patterns	Complex architecture, data-hungry sliding windows
Uğraş et al. (42)	CardioPatternFormer	Interpretable, multi-pathology	Opaque black-box models	May overfit attention map, needs clinical validation
Luo et al. (43)	Hierarchical Transformer	-	Single-scale Transformers	Multi-stage model is resource-intensive
Alghieth (44)	DCETEN	99.84% acc. (MIT-BIH)	Heavyweight transformer models	Still GPU-reliant despite pruning
Current study	Transformer Model	97%	N/A	High Computational demand, requiring advanced GPUs or TPUs, limited interpretability, challenging clinical transparency.

One key direction is addressing class imbalance, particularly for underrepresented arrhythmia types such as Ventricular Premature Contractions (VPC), which currently contribute to lower classification accuracy. Sensitivity to rare classes may be improved by applying techniques such as class-specific data augmentation, oversampling, and class-weighted loss functions.

Another priority is improving the diversity and representativeness of the training data. Incorporating ECG signals from a broader population encompassing different demographics, acquisition devices, and arrhythmia types can increase the model's generalizability and reduce bias toward specific data sources or conditions.

To further improve diagnostic accuracy, future work may explore multimodal learning by integrating additional physiological signals such as heart rate variability, blood oxygen saturation, and blood pressure. These complementary modalities could enhance the feature space and provide more context for ECG interpretation.

Optimizing the model for deployment in resource-constrained environments, such as mobile or wearable devices, is also a critical focus. While Transformers offer high accuracy, their computational demands limit feasibility on low-power platforms. Future research will investigate lightweight Transformer variants, as well as model compression techniques such as pruning and quantization, to reduce inference costs while preserving clinical performance.

Finally, improving model interpretability remains a central challenge. Future studies will incorporate explainability techniques such as attention weight visualization, relevance mapping, and lead-wise contribution analysis. These tools can help clinicians better understand the basis for automated predictions, thereby increasing trust and promoting adoption in real-world healthcare settings.

10 Conclusion

The proposed Transformer-based ECG classification model demonstrates strong potential in accurately diagnosing multiple cardiac arrhythmias from raw ECG signals. Leveraging the self-attention mechanism inherent in Transformer architecture, the model effectively captures the temporal dependencies of ECG sequences and achieves high classification accuracy across several classes, including Normal, APC, VPC, Fusion Beats, and Others. These results confirm the suitability of attention-based models for analyzing the complex and sequential nature of biomedical time-series data.

A key contribution of this work is the demonstration that transformer models can serve as effective tools for ECG signal classification, providing clinically relevant outputs with high precision, recall, and F1-scores particularly for classes with ample training data. This suggests that such models can complement existing machine learning techniques in automated ECG interpretation.

In addition to performance, the model offers potential for integration into future clinical workflows, where automated ECG analysis can support healthcare professionals by reducing manual diagnostic load and improving consistency. However, several challenges remain before deployment in real-world settings. These include improving classification for underrepresented arrhythmia classes, validating the model across more diverse populations and device types, and enhancing model interpretability and computational efficiency.

Future work should focus on optimizing the model for broader generalization, incorporating multimodal physiological data, and adapting the architecture for deployment in resource-constrained environments such as wearable healthcare devices. With further development and clinical validation, Transformer-based models may play an important role in advancing automated, scalable, and accessible cardiac diagnostics.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article because, dataset was publicly available on Kaggle.

Author contributions

SI: Data curation, Writing – review & editing, Conceptualization, Methodology, Writing – original draft. AI: Validation, Conceptualization, Data curation, Writing – review & editing, Writing – original draft. HS: Visualization, Writing – original draft, Writing – review & editing. MA: Formal analysis, Writing – review & editing, Writing – original draft. SN: Conceptualization, Writing – original draft, Writing – review & editing, Software. IT: Visualization, Funding acquisition, Writing – original draft, Writing – review & editing. HG: Data curation, Project administration, Writing – review & editing, Writing – original draft. TC: Writing – original draft, Conceptualization, Investigation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was funded by the European University of Atlantic.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol*. (2021) 18:465–78. doi: 10.1038/s41569-020-00503-2

2. Labati RD, et al. Deep-ECG: convolutional neural networks for ECG biometric recognition. *Pattern Recogn Lett*. (2019) 126:78–85. doi: 10.1016/j.patrec.2018.03.028

3. Bhatnagar P, Wickramasinghe K, Williams J, Rayner M, Townsend N. The epidemiology of cardiovascular disease in the UK 2014. *Heart*. (2015) 101:1182–9. doi: 10.1136/heartjnl-2015-307516

4. Acharya UR, Fujita H, Lih OS, Hagiwara Y, Tan JH, Adam M. Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. *Inf Sci*. (2017) 405:81–90. doi: 10.1016/j.ins.2017.04.012

5. Singh S, Pandey SK, Pawar U, Janghel RR. Classification of ECG arrhythmia using recurrent neural networks. *Proc Comput Sci*. (2018) 132:1290–7. doi: 10.1016/j.procs.2018.05.045

6. Ikram A, Aslam W. Enhancing intercropping yield predictability using optimally driven feedback neural network and loss functions. *IEEE Access*. (2024) 12:162769–87. doi: 10.1109/ACCESS.2024.3486101

7. Ikram A, Ikram S, el-kenawy ESM, Hussain A, Alharbi AH, Eid MM. A fuzzy-optimized hybrid ensemble model for yield prediction in maize-soybean intercropping system. *Front Plant Sci*. (2025) 16:1567679. doi: 10.3389/fpls.2025.1567679

8. Yang H, Wei Z. Arrhythmia recognition and classification using combined parametric and visual pattern features of ECG morphology. *IEEE Access*. (2020) 8:47103–17. doi: 10.1109/ACCESS.2020.2979256

9. Chang AC, Limon A. Introduction to artificial intelligence for cardiovascular clinicians In: Intelligence-based cardiology and cardiac surgery: Elsevier (2024). 3–120.

10. Murat F, Yildirim O, Talo M, Demir Y, Tan RS, Ciaccio EJ, et al. Exploring deep features and ECG attributes to detect cardiac rhythm classes. *Knowl-Based Syst*. (2021) 232:107473. doi: 10.1016/j.knosys.2021.107473

11. Xia Y, Xiong Y, Wang K. A transformer model blended with CNN and denoising autoencoder for inter-patient ECG arrhythmia classification. *Biomed Signal Proc Control*. (2023) 86:105271. doi: 10.1016/j.bspc.2023.105271

12. Ikram S, Bajwa IS, Ikram A, Abdullah-al-Wadud M, Pk H. A transformer-based multimodal object detection system for real-world applications. *IEEE Access*. (2025) 13:29162–76. doi: 10.1109/ACCESS.2025.3539569

13. Ikram S, Bajwa IS, Ikram A, Díez IT, Ríos CEU, Castilla ÁK. Obstacle detection and warning system for visually impaired using IoT sensors. *IEEE Access*. (2025) 13:35309–21. doi: 10.1109/ACCESS.2025.3543299

14. Din S, Qaraqe M, Mourad O, Qaraqe K, Serpedin E. ECG-based cardiac arrhythmias detection through ensemble learning and fusion of deep spatial–temporal and long-range dependency features. *Artif Intell Med*. (2024) 150:102818. doi: 10.1016/j.artmed.2024.102818

15. Ikram S, Sarwar Bajwa I, Gyawali S, Ikram A, Alsubaie N. Enhancing object detection in assistive Technology for the Visually Impaired: a DETR-based approach. *IEEE Access*. (2025) 13:71647–61. doi: 10.1109/ACCESS.2025.3558370

16. Martis RJ, Chakraborty C, Ray AK. Wavelet-based machine learning techniques for ECG signal analysis. *Machine Learn Healthc Informat*. (2014):25–45. doi: 10.1007/978-3-642-40017-9_2

17. Marinho LB, Nascimento NMM, Souza JWM, Gurgel MV, Rebouças Filho PP, de Albuquerque VHC. A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification. *Futur Gener Comput Syst*. (2019) 97:564–77. doi: 10.1016/j.future.2019.03.025

18. Martis RJ, Acharya UR, Lim CM, Suri JS. Characterization of ECG beats from cardiac arrhythmia using discrete cosine transform in PCA framework. *Knowl-Based Syst*. (2013) 45:76–82. doi: 10.1016/j.knosys.2013.02.007

19. Arabi Z., Pourbahrami S., Abedi A., Hybrid CNN-transformer architecture with adaptive positional embedding for interpretable ECG arrhythmia classification. (2025).

20. Ikram A, et al. Crop yield maximization using an IoT-based smart decision. *J Sens*. (2022) 2022:2022923.

21. Latif MS, et al. Pest prediction in rice using IoT and feed forward neural network. *KSII Trans Internet Informat Syst*. (2022) 16:133–52.

22. Dong Y, Zhang M, Qiu L, Wang L, Yu Y. An arrhythmia classification model based on vision transformer with deformable attention. *Micromachines*. (2023) 14:1155. doi: 10.3390/mi14061155

23. Meng L, Tan W, Ma J, Wang R, Yin X, Zhang Y. Enhancing dynamic ECG heartbeat classification with lightweight transformer model. *Artif Intell Med*. (2022) 124:102236. doi: 10.1016/j.artmed.2022.102236

24. Natarajan A, et al. A wide and deep transformer neural network for 12-lead ECG classification In: 2020 computing in cardiology: IEEE (2020)

25. Li H, Han J, Zhang H, Zhang X, Si Y, Zhang Y, et al. Clinical knowledge-based ECG abnormalities detection using dual-view CNN-transformer and external attention mechanism. *Comput Biol Med*. (2024) 178:108751. doi: 10.1016/j.compbiomed.2024.108751

26. Varghese A, Kamal S, Kurian J. Transformer-based temporal sequence learners for arrhythmia classification. *Med Biol Eng Comput*. (2023) 61:1993–2000. doi: 10.1007/s11517-023-02858-3

27. Zhang X, et al. MSFT: a multi-scale feature-based transformer model for arrhythmia classification. *Biomed Signal Proce Control*. (2025) 100:106968

28. Natarajan A, et al. Convolution-free waveform transformers for multi-lead ECG classification In: 2021 Computing in Cardiology (CinC): IEEE (2021)

29. Le MD, et al. Multi-module recurrent convolutional neural network with transformer encoder for ECG arrhythmia classification In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI): IEEE (2021)

30. Lee J, Shin M. Cross-database learning framework for electrocardiogram arrhythmia classification using two-dimensional beat-score-map representation. *Appl Sci*. (2025) 15:5535. doi: 10.3390/app15105535

31. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. (2019) 25:65–9. doi: 10.1038/s41591-018-0268-3

32. Rajpurkar P, et al. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv*. (2017)

33. Ait Bourkha ME, et al. Optimized wavelet scattering network and cnn for ecg heartbeat classification from mit–bih arrhythmia database. *Int J Adv Comput Sci Appl*. (2025) 16

34. Kailan SL, et al. Efficient ECG classification based on machine learning and feature selection algorithm for IoT-5G enabled health monitoring systems. *Int J Intell Eng Syst*. (2025) 18

35. Mavaddati S. ECG arrhythmias classification based on deep learning methods and transfer learning technique. *Biomed Signal Proc Control*. (2025) 101:107236. doi: 10.1016/j.bspc.2024.107236

36. Tao T., Advances in artificial intelligence for electrocardiogram analysis: a comprehensive review of architectures, clinical applications, and future directions. (2025).

37. Li D, et al. Multimodal deep learning for predicting in-hospital mortality in heart failure patients using longitudinal chest X-rays and electronic health records. *Int J Cardiovasc Imaging*. (2025):1–14.

38. Zhao B, Gao Z, Liu X, Zhang Z, Xiao W, Zhang S. DRL-ECG-HF: deep reinforcement learning for enhanced automated diagnosis of heart failure with imbalanced ECG data. *Biomed Signal Proce Control*. (2025) 107:107680. doi: 10.1016/j.bspc.2025.107680

39. Smith SW, Walsh B, Grauer K, Wang K, Rapin J, Li J, et al. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. *J Electrocardiol*. (2019) 52:88–95. doi: 10.1016/j.jelectrocard.2018.11.013

40. Li D, Li X, Zhao J, Bai X. Automatic staging model of heart failure based on deep learning. *Biomed Signal Proc Control*. (2019) 52:77–83. doi: 10.1016/j.bspc.2019.03.009

41. Chithra V, Shashank H, Patel DS. Heart disease detection using ensemble machine learning technique based on various risk factors In: 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON): IEEE (2024)

42. Uğraş BK, Gerek ÖN, Saygı İT. CardioPatternFormer: pattern-guided attention for interpretable ECG classification with transformer architecture. *arXiv*. (2025)

43. Luo Y, et al. Detection of atrial fibrillation with a hybrid deep learning model and time-frequency representations. *medRxiv*. (2025)

44. Alghieth M. DeepECG-net: a hybrid transformer-based deep learning model for real-time ECG anomaly detection. *Sci Rep*. (2025) 15:20714. doi: 10.1038/s41598-025-07781-1

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 01 September 2025
DOI 10.3389/fmed.2025.1619706

OPEN ACCESS

EDITED BY
Ateeq Ur Rehman,
Gachon University, Republic of Korea

REVIEWED BY
Junaid Zafar,
Government College University, Lahore, Pakistan
Noman Shabbir,
Tallinn University of Technology, Estonia

*CORRESPONDENCE
Doaa Shehab
✉ damenshehab@stu.kau.edu.sa

RECEIVED 28 April 2025
ACCEPTED 31 July 2025
PUBLISHED 01 September 2025

CITATION
Shehab D and Alhaddad M (2025) Image steganalysis using LSTM fused convolutional neural networks for secure telemedicine. *Front. Med.* 12:1619706. doi: 10.3389/fmed.2025.1619706

Image steganalysis using LSTM fused convolutional neural networks for secure telemedicine

Doaa Shehab* and Mohmmed Alhaddad

Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Deep learning-based image steganalysis has progressed in recent times, with efforts more concerted toward prioritizing detection accuracy over lightweight frameworks. In the context of AI-driven health solutions, ensuring the security and integrity of medical images is imperative. This study introduces a novel approach that leverages the correlation between local image features using a CNN fused Long Short-Term Memory (LSTM) model for enhanced feature extraction. By replacing the fully connected layers of conventional CNN architectures with LSTM, our proposed method prioritizes high-relevance features, making it a viable choice for detecting hidden data within medical and sensitive imaging datasets. The LSTM layers in our hybrid model demonstrate better sensitivity characteristics for ensuring privacy in AI-driven diagnostics and telemedicine. Experiments were conducted on Break Our Steganographic System (BOSS Base 1.01) and Break Our Watermarking System (BOWS) datasets, followed by validation on the ALASKA2 Image Steganalysis dataset. The results confirm that our approach generalizes effectively and would serve as impetus to ensure security and privacy for digital healthcare solutions.

KEYWORDS

steganalysis, steganography, data hiding, healthcare security, LSTM, lightweight

1 Introduction

AI-based digital healthcare solutions require security and data privacy while handling sensitive medical images; therefore, robust techniques are essential to maintain data integrity (1, 2). Particularly, the medical images contain embedded metadata and annotations that may compromise patient privacy (3). Image steganalysis helps in preserving sensitive medical records (4) and by leveraging artificial intelligence (AI) techniques, healthcare professionals can identify potential threats posed by steganographic attacks (5, 6). Beyond privacy concerns, the integrity of medical data is another essential dimension for AI diagnostic systems (7, 8). Malicious actors could use steganography to manipulate images, alter tumor regions, or embed misleading data without detection (1). Advanced steganalysis techniques and emerging telemedicine issues necessitate the integration of robust AI-driven steganalysis tools to improve the security of sensitive health data (2).

Recent image steganalysis techniques exploited the traditional machine learning to extract meaningful features, but human dependencies limited their scope in image steganalysis (9). Low embedding capacity and poor image retrieval rates necessitated the deployment of deep learning assisted steganalysis algorithms. Detailed reviews regarding the recent deep learning strategies and network developments are included elsewhere (10, 11). In this connection, numerous deep learning algorithms were reported for rapid detection of steganographic payloads with reasonable accuracies (12–15). Key modifications include enhancing filters and different activation operators (16), high-order

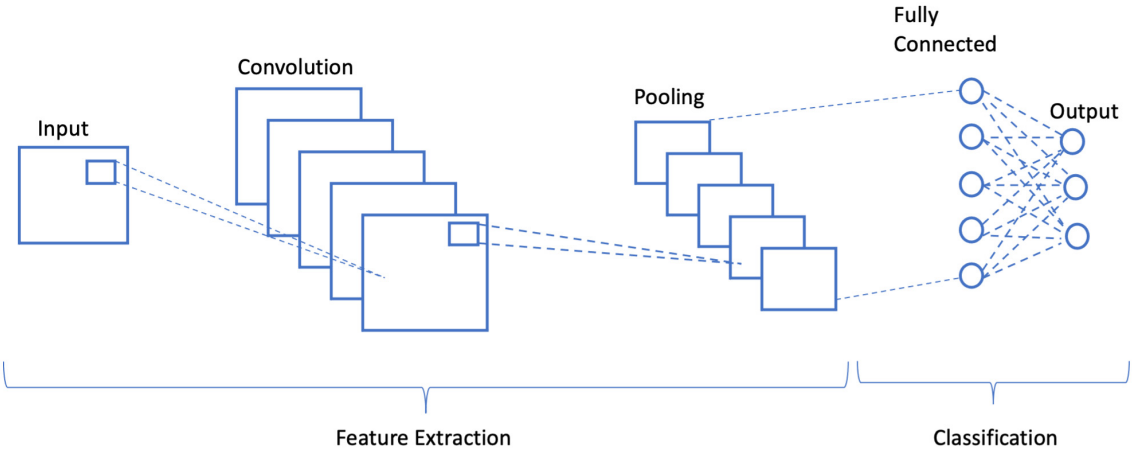


FIGURE 1
Schematic illustration of a generalized convolution neural network.

co-occurrence matrices to capture sensitivity (17, 18), periodic weight capture (19), dimensionality reduction schemes (20), and covariance pooling techniques (16, 21–24).

Moreover, various DL-based models such as Qian et al. (25), Yedroudj et al. (18), Boroumand et al. (19), Deng et al. (16), Zhang et al. (26), Reinel et al. (22), Öztürk Ş and Özkaya (27), and Ozdemir et al. (28) tried to improvise on the stego image feature extraction. In this regard, You et al. (29) exploited EfficientNet, MixNet, and ResNet by removing pooling and stride operations in the first layers. Similarly, (24) applied floating-point quantization to XuNet (24). Recently, LSTM was reported to capture data correlation for image classification tasks (21, 30–32).

In this study, we propose a CNN architecture fused with LSTM by replacing the fully connected layers of the CNN. Our proposed model leverages LSTM to optimize weight matrices and bias vector parameters, ensuring effective training at each time step. In addition, LSTM nodes extract essential contextual features, which is vital for detecting hidden threats within medical images. This research contributes to the field by demonstrating the effectiveness of LSTM fused CNNs in medical image steganalysis by offering a robust security framework to protect sensitive patient data. Furthermore, we compare our proposed architecture with state-of-the-art deep learning models in terms of computational efficiency. By significantly reducing the number of trainable parameters, our model offers a resource efficient and scalable solution for secure medical image transmission and integrity in telemedicine.

The remaining of this work is organized as follows: Explain the Architecture of CNN and LSTM in Section 2. The materials and methods are presented in Section 3. The results discussion is detailed in Section 4. Section 5 concludes the study.

2 A brief on CNN and LSTM architecture

The encoder in any CNN-based steganography scheme employs binary inputs: one for the cover image and the other for secret image to foster a stego image. It includes pre-processing, feature extraction, and classification stage as illustrated in Figure 1. In the feature extraction phase, convolution is performed multiple times to ameliorate the signal-to-noise ratio of the image and to characterize local features, whereas in classification, the extracted local features are average-pooled and concatenated to yield final feature maps. These feature maps were then classified in terms of class probabilities using SoftMax function.

Though LSTM networks improve the functioning of recurrent neural networks (RNNs) in terms of vanishing gradient, LSTM contains three gates which are an input gate, a forget gate, and an output gate, where x_t, C_t, and C_{t-1} represent the current input, new, and previous cell states, respectively. h_t and h_{t-1} refer to the current and previous outputs, respectively. A non-linear function is used to activate these three gates, which makes LSTM a dynamic model with changing contexts (33). The internal architecture of an LSTM cell is shown in Figure 2.

Within an LSTM cell, forget gate controls the contribution of the previous state C_{t-1} to the current state by using sigmoid function σ and is responsible for LSTM cell memory as given by the expression in Equation 1.

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

where f_t is the forget vector, and x_t and h_{t-1} are the current input and previous output. As given in Equation 1, x_t and h_{t-1} are multiplied by the trained weights matrix W_f with offset b_f. Due to sigmoid function, the input vector ranges between 0 and 1, indicating the degree to which values are to be remembered or forgotten. h_{t-1} and x_t are passed via input updated gate to append the relevant information and is governed by Equation 2. Thereafter, new information is obtained as $\widetilde{C}_t$ from Equation 3 after passing h_{t-1} and x_t via tanh function. Finally, the candidate of the cell state C_t for the next time step is generated by combining current moment information $\widetilde{C}_t$ and long-term memory information C_{t-1}

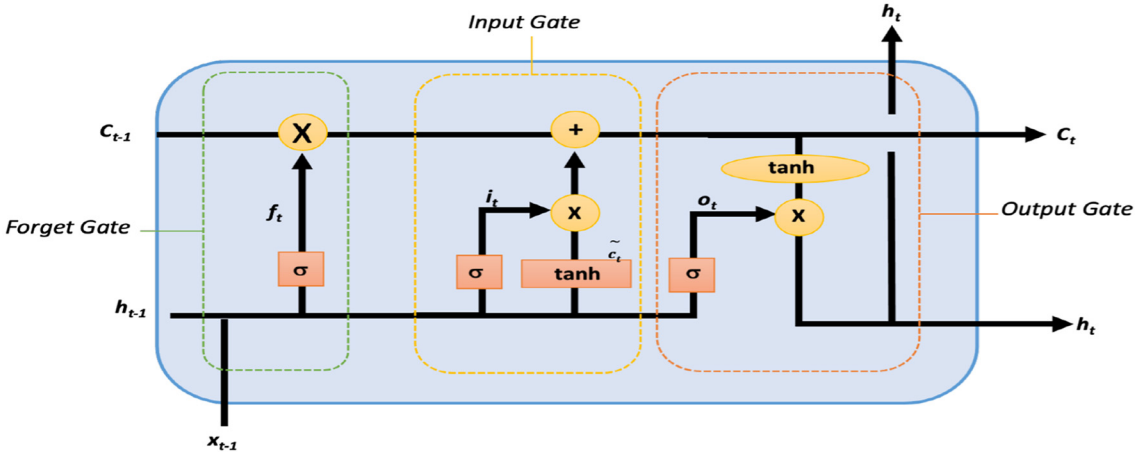


FIGURE 2
Internal architecture of a single LSTM cell.

as shown in Equation 4.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \quad (4)$$

Here, W_i denotes weight matrices that are produced from sigmoid function, and b_i denotes the input gate bias. The output gate controls the require output O_t using the expression in Equations 5, 6.

$$h_t = O_t \tanh(C_t) \quad (5)$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

Where W_o and b_o are the weighted matrices of the output gate and LSTM bias, respectively.

3 Materials and methods

With the rapid adoption of remote healthcare services, the risk of cyberattacks and data tampering has increased significantly. The main endeavor of this research is to detect and analyze hidden embeddings in medical images for secure medical data transmission. By continuously analyzing incoming medical images using AI-driven image steganalysis, data security and privacy risks can be minimized. In our proposed architecture, LSTMs were fused within the CNN by replacing the fully connected layers. The idea was to capture and rank the correlation between different stego-noises and to reduce the number of trainable parameters for time efficient classification.

3.1 Pre-processing BOSSBase 1.01 and BOWS 2 databases

For the experiments, Break Our Steganographic System (BOSSBase 1.01) (34) and Break Our Watermarking System (BOWS 2) (35) databases were used. Each database has 10,000 cover images in a Portable Gray Map (PGM) format. The data were prepared by resizing all images to 256 × 256 pixels (36). Then, a corresponding steganographic image for each cover image was generated using with payloads of 0.4 bits per pixel (bpp). In the next stage, the data were partitioned to training, validation, and testing sets. 4,000 images were used pairs for training, 1,000 for validation, and 5,000 for testing purposes. Both datasets were merged to generate a database of 20,000 images in which split 14,000 images were used for training (10,000 BOWS 2 + 4,000 BOSSBase 1.01), 1,000 pairs for validation (BOSSBase 1.01), and 5,000 for testing (BOSSBase 1.01).

3.2 Pre-processing ALASKA2 image steganalysis database

ALASKA2 dataset was chosen due to its massive size and heterogeneous nature for an in-depth validation of our proposed steganalysis algorithm. In this dataset, steganography algorithms transform data with an unknown payload. All the images were resized to 256×256 pixels and compressed with JPEG quality factors of 95, 90, and 75. This database is available on Kaggle platform (37). ALASKA2 database includes 7,500 pairs of images in JPEG format (cover and stego) which were randomly shuffled before partition. We prepared the ALASKA2 database by portioning split 6,000 pairs for training, 1500 pairs for validation, and 7,500 pairs were randomly chosen testing purposes. Furthermore, we prepared another ALASKA2 dataset by using all images via three steganographic algorithms. This database was partitioned in which 9,000 pairs were used for training, 2,250 pairs for validation, and 11,250 pairs for validations.

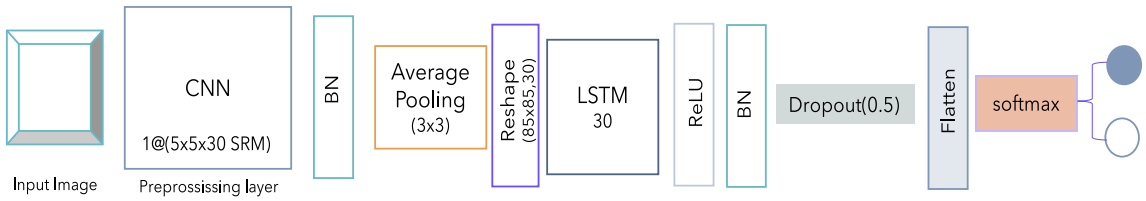


FIGURE 3
Schematic illustration of LSTM for feature representations and classification.

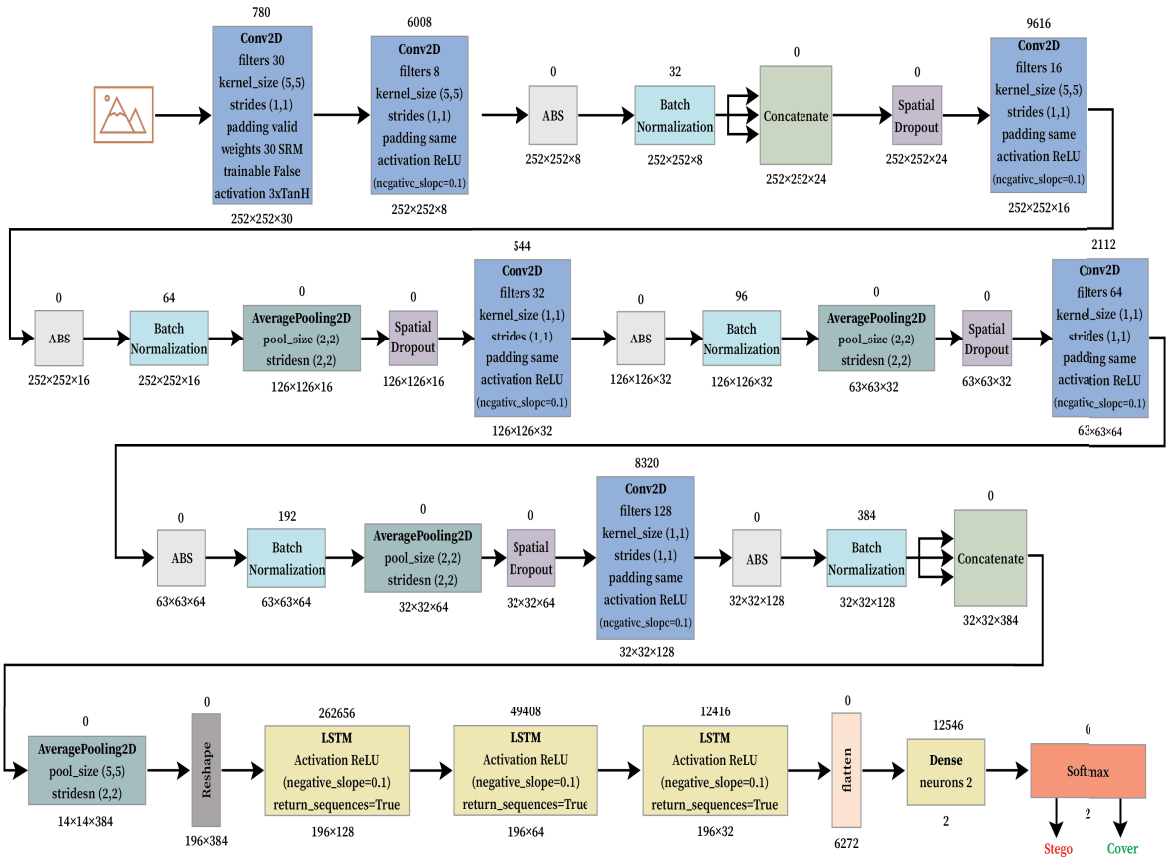


FIGURE 4
Proposed LSTM fused Xu-Net neural network architecture for secured telemedicine.

3.3 Proposed LSTM fused CNN architecture

Initially, we establish the effectiveness of LSTM for steganalysis in securing telemedicine communications and then integrate it into a CNN architecture to enhance both detection accuracy and processing efficiency. Given the critical need for real-time threat detection in remote healthcare, we provide a detailed analysis and comparison with state-of-the-art architectures to assess our model's capability. To simulate real-world security threats in telemedicine, we embedded noise in cover images using five steganographic algorithms. Two of them are spatial steganographic algorithms: S-UNIWARD (38) and WOW (39) with 0.4 bpp payloads. The other three are transform steganographic algorithms: JMiPOD (40), JUNIWARD (38), and UERD (41). Our implementation ensures robust steganalysis for secure medical image transmission.

Our initial approach investigates the applicability of LSTM in image steganalysis and is presented in Figure 3. It starts with an input image, which is first passed through a preprocessing layer using a convolutional neural network (CNN) filter of dimensions ($5 \times 5 \times 30$), indicating the use of 30 SRM (Spatial Rich Model) filters for extracting high-frequency residuals. This is followed by

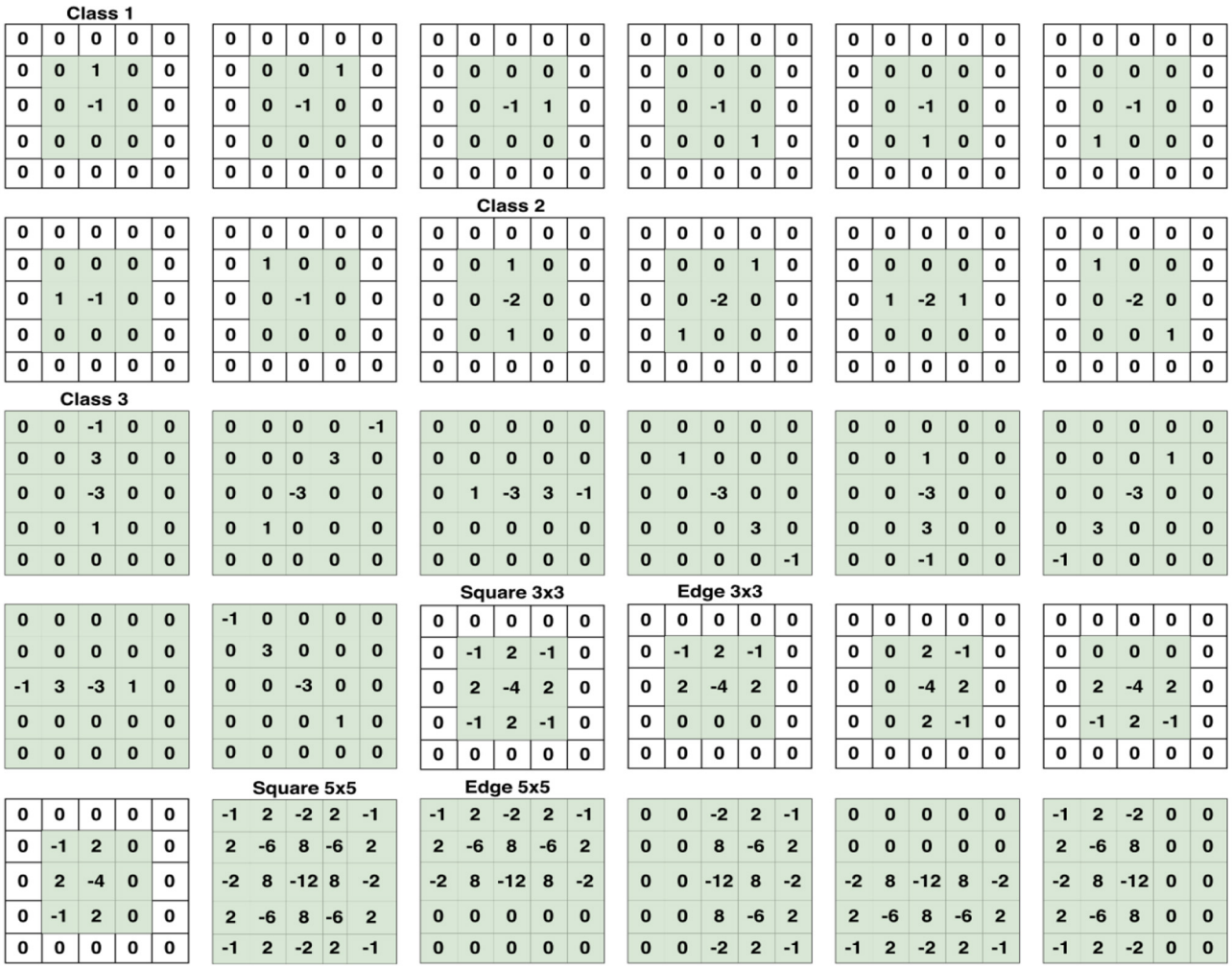


FIGURE 5
Set of 30 SRM Filters per category which are used in the first convolution, or preprocessing stage. Taken from Reinel et al. (22).

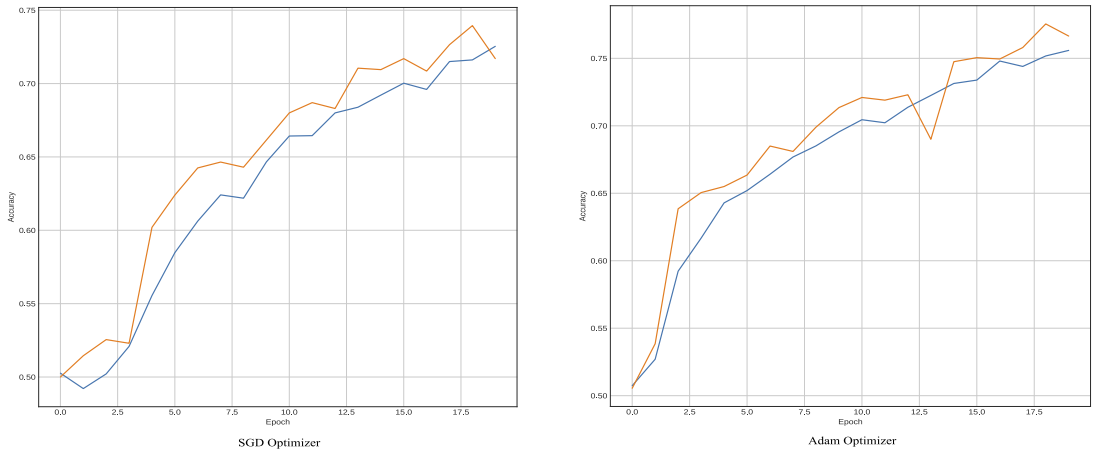


FIGURE 6
Training plots in terms of accuracy for Yedroudj-Net Model using LSTM as a classifier with BOSSBase 1.01 WOW 0.4 bpp.

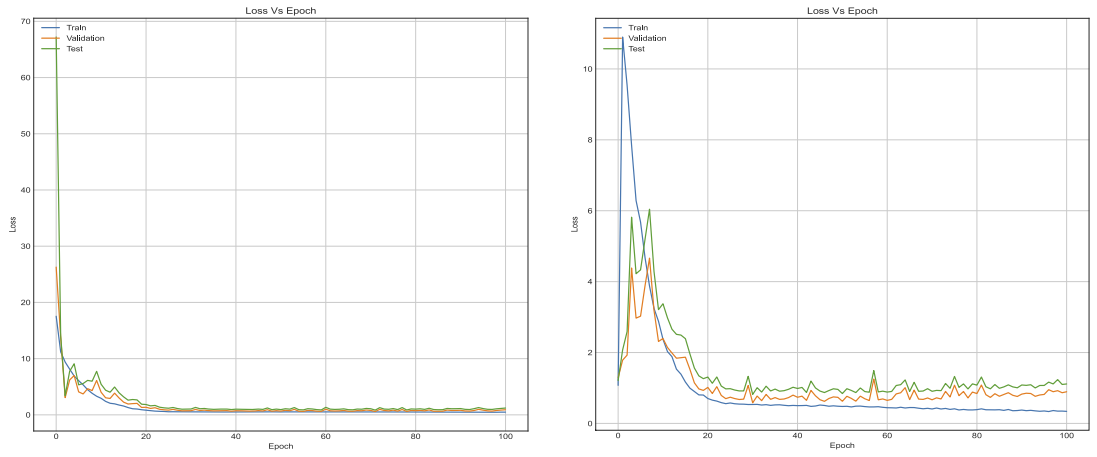


FIGURE 7
Training curves in terms of learning loss for the first method, when using stacked LSTM and FC layers in the classification stage, respectively. With BOSSBase 1.01 S-UNWARD 0.4 bpp.

TABLE 1 Accuracy percentage and number of trainable parameters of the fist method model, when using FC layer and LSTM layer for the S-UNWARD steganographic algorithm with payload 0.4 bpp using BOSSBase 1.01 database.

Scenario	with LSTM	with FC
Training Acc.	75%	**85%**
Validation Acc.	**76%**	75%
Test Acc.	**67%**	**67%**
# Trainable parameters	**433,592**	434,522

The best performances are shown in bold for each scenario.

TABLE 2 Accuracy percentage and loss value of the fist method model, when using FC layer and LSTM layer for ALASKA2 database.

Scenario	with LSTM		with FC	
Database	Acc.	Loss	Acc.	loss
JMiPOD	62%	**.99**	**65%**	1.45
JUNIWARD	60%	**1.00**	**62%**	1.00
UERD	61%	**0.90**	**63%**	0.94
ALASKA2_All	**49%**	**1.00**	46%	1.7

The best performances are shown in bold for each scenario.

batch normalization (BN) to stabilize and accelerate training. Next, average pooling with a 3×3 kernel is applied to reduce spatial dimensions while preserving critical features. This is then reshaped into a sequence format (65×30), which is suitable for temporal modeling via LSTM. After reshaping, the feature map is fed into an LSTM layer with 30 units as illustrated in Figure 3. The output of LSTM is passed through a ReLU activation to introduce non-linearity, followed by another batch normalization to standardize feature distributions. A dropout layer with a rate of 0.5 is included to prevent overfitting by randomly deactivating neurons during the training. The resulting features are flattened into a one-dimensional vector and are further passed through a Softmax classifier. This architecture combines the spatial feature extraction capability of CNNs with the sequential modeling strength of LSTMs, making it particularly robust for detecting subtle patterns in stego and manipulated images.

After the initial proof of concept regarding LSTM architecture for steganalysis, we fused LSTM as a classifier into the CNN architecture by replacing its three fully connected layers which is presented in Figure 4. The model begins with a convolutional preprocessing layer using fixed SRM filters, which are effective in extracting the noise residuals from the images. These initial outputs are passed through several convolutional blocks, each containing Conv2D layers, batch normalization, and spatial dropout. It is further followed by average pooling to reduce spatial dimensions while maintaining the important feature structures. The model uses concatenation operations to merge different channels for a multi-level residual learning. After the hierarchical CNN feature extraction, the architecture transitions into a temporal modeling phase using LSTM layers. Before entering the LSTM block, features are reshaped and passed through an average pooling 2D layer. The sequence of two LSTM layers allows the model to capture long-range dependencies across spatially transformed image features. The final output from the LSTM is flattened and passed into a dense layer with two neurons, corresponding to a binary classification: Stego and Cover. A softmax layer provides probabilistic outputs for the final decision. This hybrid CNN-LSTM design, coupled with residual modeling, makes the architecture well-suited for subtle signal detection tasks.

For this experiment, four famous and recent CNNs for image steganalysis were used, which include Xu-Net (24), Ye-Net (15), Yedroudj-Net (18), and Zhu-Net (26). SRM filters were used to improve the ratio of stego- to image-noise signal. Since the stego signal is always embedded in the high-frequency part of an image, we utilized these filters to initialize the kernels of a convolutional

TABLE 3 Accuracy percentage of the second method models for the S-UNWARD steganographic algorithm with payload 0.4 bpp.

Dataset	BOSSBase 1.01			BOSSBase 1.01+ BOWS		
results	Original	Strategy	With LSTM	Original	Strategy	With LSTM
Xu-Net	73%	78%	76%	–	82%	81%
Ye-Net	68%	81%	80%	–	83%	81%
Yedroudj-Net	77%	79%	79%	–	84%	82%
Zhu-Net	84.5%	78.6%	80.7%	–	86%	81.3%

TABLE 4 Accuracy percentage of the second method models for the WOW steganographic algorithm with payload 0.4 bpp.

Dataset	BOSSBase 1.01			BOSSBase 1.01+ BOWS		
Results	Original	Strategy	With LSTM	Original	Strategy	With LSTM
Xu-Net	79%	82%	81%	–	85%	83%
Ye-Net	75%	84%	83%	–	86%	85%
Yedroudj-Net	84%	85%	83%	–	86%	85%
Zhu-Net	88.1%	82.9%	83.5%	–	75%	83.5%

layer. A bulk of 30 high-pass filters from the SRM are used in the pre-processing block prior to feature extraction phase as indicated in Figure 5.

Experimental implementations used Python 3.8.1 and TensorFlow 2.2.0. In our model using LSTM only, network was trained for 100 epochs using S-UNWARD steganography with payload 0.4 bpp (BOSSBase 1.01 dataset). The LSTM fused CNN implementations presented in Figure 4 used the Google Colaboratory platform on Tesla P100 PCIe (16 GB) having CUDA Version 10.1 with 32 GB RAM to speed up simulations.

4 Results and discussion

4.1 Validation of LSTM classifier on BOSSBase 1.01, BOWS 2, and ALASKA2 dataset

To ensure reliable telemedicine, the LSTM classifier was trained for 100 epochs on the BOSSBase 1.01 and BOWS 2 databases and 50 epochs on the ALASKA2 database. A batch size of 64 images was used, with the Stochastic Gradient Descent (SGD) optimizer set at a momentum of 0.95 and an initial learning rate of 0.005. The training curves, illustrating accuracy and learning loss, are presented in Figure 6. Our model incorporates gating mechanisms to regulate gradients, enabling the architecture to retain critical information necessary for detecting hidden threats in transmitted medical images. This ability to learn and preserve information over extended sequences enhances the reliability of telemedicine via secure data transmission.

Figure 7 reflects the loss function which is binary cross entropy. The results indicate that LSTM model reaches saturation in a time-efficient manner very as the training data hyperparameters were tuned quickly. The gap between validation loss and the training loss using LSTM model is indicative of the fact that LSTM have the ability to adapt to diverse datasets and can generalize to new data. Moreover, the loss value of LSTM model is small and less than that of FC model. The classification accuracy and number of trainable parameters are reported in Table 1 with a fully connected layer and hybrid LSTM for S-UNWARD steganographic algorithm. As presented in Table 1, the fully connected model achieves higher training accuracy (85%) as compared to the LSTM-based model (75%), which suggests that the FC model is better at fitting the training data. However, the similarity in test accuracy between both models indicates that the FC model suffers from overfitting. This is due to specific patterns in the training set that do not generalize well to the unseen data. In contrast, the LSTM model with its inherent regularization via likely promotes better generalization despite its lower training accuracy. This behavior is consistent with the hypothesis that the FC model's capacity to memorize leads to overfitting, while the LSTM model trades some training performance for improved robustness to the unseen data.

Table 2 provides the accuracy and loss results of the CNNs when using either of fully connected (FC) layer or LSTM layer for ALASKA2 databases. Similarly, LSTM classifier outperforms FC on ALASKA2 dataset.

4.2 Validation of LSTM fused CNN architecture against BOSSBase 1.01, BOWS 2, and ALASKA2 dataset

In our proposed model for secure telemedicine, the training batch size was set to 64 images for Xu-Net, Ye-Net, and Yedroudj-Net, while Zhu-Net utilized a batch size of 32. These mini-batches optimize computational efficiency, ensuring rapid and scalable analysis of medical images in remote healthcare environments. To enhance model stability and accuracy in detecting hidden threats in transmitted medical data, we trained Xu-Net, Ye-Net, and

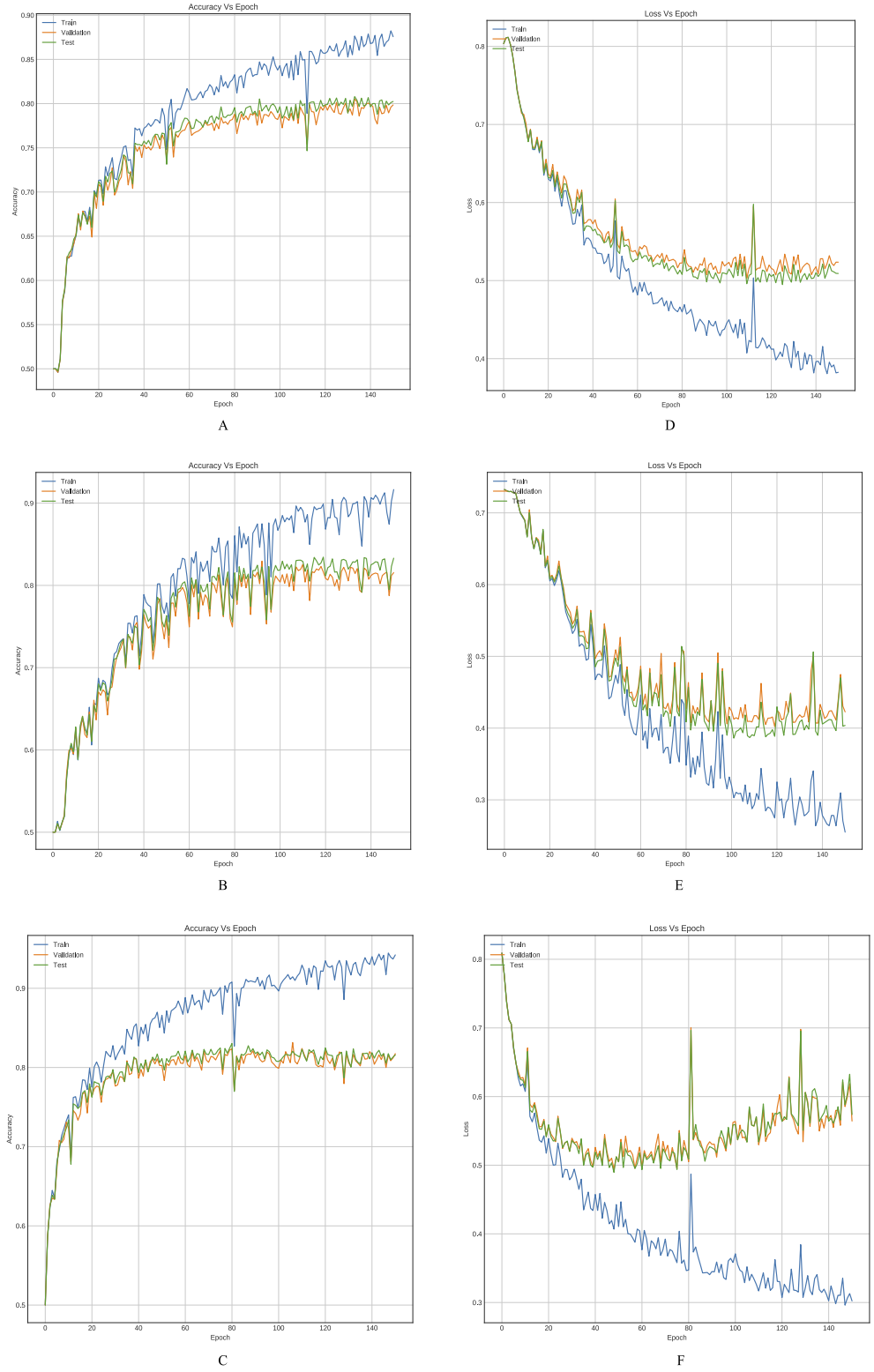


FIGURE 8
Training curves, **(A–C)** reflect the accuracy, and **(D–F)** reflect the learning loss for Xu-Net based on LSTM, Ye-Net based on LSTM, and Yedroudj-Net based on LSTM, respectively, with BOSSBase 1.01 WOW 0.4 bpp.

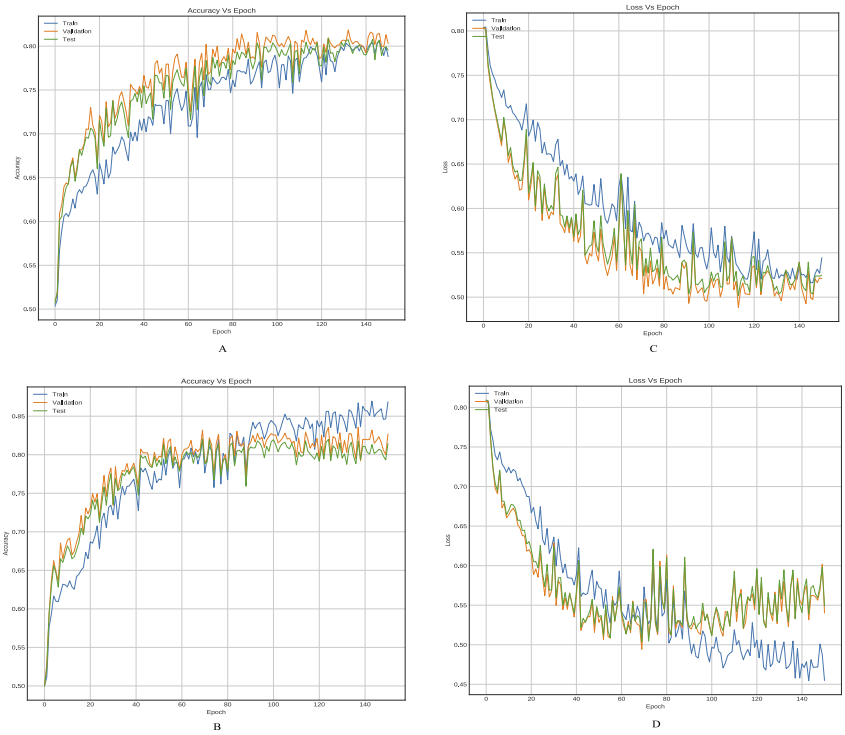


FIGURE 9
Training curves, **(A, B)** reflect the accuracy, and **(C, D)** reflect the learning loss for Xu-Net based on LSTM, and Yedroudj-Net based on LSTM, respectively, with BOSSBase 1.01 + BOWS 2 S-UNWARD 0.4 bpp.

Yedroudj-Net for 150 epochs, while Zhu-Net was trained for 70 epochs. A spatial dropout rate of 0.1 was applied across all layers to prevent overfitting, and batch normalization was configured with a momentum of 0.2, epsilon of 0.001, and renorm momentum of 0.4. The Adam optimizer, with a learning rate of 0.001, beta 1 of 0.9, beta 2 of 0.999, and an epsilon value of $1e-08$, was employed to ensure efficient convergence. To reinforce security in telemedicine image transmission, all layers were regularized for weights and bias, enabling the model to detect anomalies and steganographic threats in real-time. The accuracy results for both the S-UNWARD and WOW steganographic algorithms, which assess the model's ability to identify hidden data in medical images, are presented in Tables 3, 4.

Tables 3, 4 provide an inter-comparison between the accuracy of our proposed LSTM fused CNN architecture with the reported results (36). We achieved a high agreement between strategy and our model in terms of accuracy. The results highlighted in Tables 3, 4 are extracted from Figures 8, 9.

Trainable parameters refer to those parameters which can be learned and updated during the training cycle and has direct relationship with the computation time. Table 5 presents the number of trainable parameters for each model when applying the strategy reported in Tabares-Soto et al. (36) and when we used our proposed hybrid LSTM model.

The results presented in Table 5 confirm that our proposed model significantly decreased the number of trainable parameters as compared to leading available models and hence the computational effort required.

TABLE 5 Number of trainable parameters for state-of-the-arts architectures.

Results	Based on FC		Based on LSTM	
#Trainable parameters	Total	Classification stage	Total	Classification stage
Xu-Net	86,554	59,616	**39,418**	**0**
Ye-Net	**87,562**	22,752	118,570	**0**
Yedroudj-Net	251,110	59,616	**203,974**	**0**
Zhu-Net	275,684	59,616	**265,156**	**0**

The best results are shown in bold for each scenario.

5 Conclusion

Our proposed architecture proves to be highly effective in capturing complex interrelations among different features, making it a viable choice for steganalysis in telemedicine. Experiments conducted on BOSSBase 1.01, BOWS, and ALASKA2 datasets validate that our model demonstrates strong adaptability and generalization capabilities, which are essential for detecting hidden manipulations in telemedicine imaging systems. The achieved validation loss characteristics further reinforce the robustness of our approach in identifying steganographic threats in medical data transmission. A comparative analysis with leading architectures highlights that our model achieves significant dimensionality reduction in terms of training parameters, making it more efficient

without compromising accuracy. This efficiency is critical for real-time telemedicine applications.

However, we acknowledge that the current study does not include validation on real-world clinical datasets or standard medical image formats such as DICOM. Addressing this limitation forms a key part of our future work, where we aim to evaluate the model's performance on actual clinical imaging data to strengthen its practical applicability in telemedicine settings. By continuing to refine and expand our approach, we can contribute to a more secure and reliable telemedicine ecosystem.

Data availability statement

Publicly available datasets were analyzed in this study. The code is available on GitHub: https://github.com/DrDoaaSh/phd-code.git. The data set used to reproduce the results can be downloaded from this link: 10.5281/zenodo.4884116 or from this link https://drive.google.com/drive/folders/18KaJnn432D89WJarNY5NCTAxZB2Z3nw7?usp=drive_link.

Author contributions

DS: Project administration, Data curation, Formal analysis, Methodology, Investigation, Validation, Conceptualization, Writing – original draft. MA: Project administration, Supervision, Conceptualization, Resources, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by the Cybersecurity Research and Innovation Pioneers Grants Initiative, National Program for Research, Development, and Innovation (RDI) in Cybersecurity, Kingdom of Saudi Arabia, under Grant Number CRPG-25-2030.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Magdy M, Hosny KM, Ghali NI, Ghoniemy S. Security of medical images for telemedicine: a systematic review. *Multimed Tools Appl*. (2022) 81:25101–45. doi: 10.1007/s11042-022-11956-7

2. Hameed MA, Hassaballah M, Bekhet S, Kenk MA, et al. A high quality secure medical image steganography method. In: *2023 3rd International Conference on Computing and Information Technology (ICCIT)*. Tabuk: IEEE (2023). p. 465–70. doi: 10.1109/ICCIT58132.2023.10273950

3. Saidi H, Tibermacine O, Elhadad A. High-capacity data hiding for medical images based on the mask-RCNN model. *Sci Rep*. (2024) 14:7166. doi: 10.1038/s41598-024-55639-9

4. Abdulla AA. Digital image steganography: challenges, investigation, and recommendation for the future direction. *Soft Comput*. (2024) 28:8963–76. doi: 10.1007/s00500-023-09130-8

5. Sirisha BL, Ahamed SF, Aruna V. Patient data hiding and transmitting during COVID-19 for telemedicine application using image steganography. *Curr Med Imaging*. (2024) 20:e15734056276785. doi: 10.2174/0115734056276785240229073917

6. Mansour RF, Girgis MR. Steganography-based transmission of medical images over unsecure network for telemedicine applications. *Comput Mater Contin*. (2021) 68:4069–85. doi: 10.32604/cmc.2021.017064

7. Mansour RF, Abdelrahim EM. An evolutionary computing enriched RS attack resilient medical image steganography model for telemedicine applications. *Multidimens Syst Signal Process*. (2019) 30:791–814. doi: 10.1007/s11045-018-0575-3

8. Le Guern N. Les développements récents de la photographie: de la menace des smartphones au potentiel démesuré de l'IA. *Marché Organ*. (2025) 52:217–47. doi: 10.3917/maorg.pr1.0117

9. Shehab DA, Alhaddad MJ. Comprehensive survey of multimedia steganalysis: techniques, evaluations, and trends in future research. *Symmetry*. (2022) 14:117. doi: 10.3390/sym14010117

10. Himthani V, Dhaka VS, Kaur M, Rani G, Oza M, Lee HN. Comparative performance assessment of deep learning based image steganography techniques. *Sci Rep*. (2022) 12:16895. doi: 10.1038/s41598-022-17362-1

11. Jahromi ZT, Hasheminejad SMH, Shojaedini SV. Deep learning semantic image synthesis: a novel method for unlimited capacity, high noise resistance coverless video steganography. *Multimed Tools Appl*. (2024) 83:17047–65. doi: 10.1007/s11042-023-16278-w

12. Telli M, Othmani M, Ltifi H. A new approach to video steganography models with 3D deep CNN autoencoders. *Multimed Tools Appl*. (2024) 83:51423–39. doi: 10.1007/s11042-023-17358-7

13. Ding K, Hu T, Niu W, Liu X, He J, Yin M, et al. A novel steganography method for character-level text image based on adversarial attacks. *Sensors*. (2022) 22:6497. doi: 10.3390/s22176497

14. Zhuo P, Yan D, Ying K, Wang R, Dong L. Audio steganography cover enhancement via reinforcement learning. *Signal Image Video Process*. (2024) 18:1007–13. doi: 10.1007/s11760-023-02819-1

15. Ye J, Ni J, Yi Y. Deep learning hierarchical representations for image steganalysis. *IEEE Trans Inf Forensics Secur*. (2017) 12:2545–57. doi: 10.1109/TIFS.2017.2710946

16. Deng X, Chen B, Luo W, Luo D. Fast and effective global covariance pooling network for image steganalysis. In: *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*. New York, NY: ACM (2019). p. 230–4. doi: 10.1145/3335203.3335739

17. Wu S, Zhong S, Liu Y. Deep residual learning for image steganalysis. *Multimed Tools Appl*. (2018) 77:10437–53. doi: 10.1007/s11042-017-4440-4

18. Yedroudj M, Comby F, Chaumont M. Yedroudj-net: an efficient CNN for spatial steganalysis. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB: IEEE (2018). p. 2092–6. doi: 10.1109/ICASSP.2018.8461438

19. Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images. *IEEE Trans Inf Forensics Secur*. (2018) 14:1181–93. doi: 10.1109/TIFS.2018.2871749

20. Zhang X, Kong X, Wang P, Wang B. Cover-source mismatch in deep spatial steganalysis. In: *International Workshop on Digital Watermarking*. Cham: Springer (2019). p. 71–83. doi: 10.1007/978-3-030-43575-2_6

21. Wang L, Xu X, Gui R, Yang R, Pu F. Learning rotation domain deep mutual information using convolutional LSTM for unsupervised PolSAR image classification. *Remote Sens*. (2020) 12:4075. doi: 10.3390/rs12244075

22. Reinel TS, Brayan AAH, Alejandro BOM, Alejandro MR, Daniel AG, Alejandro AGJ, et al. GBRAS-Net: a convolutional neural network architecture for spatial image steganalysis. *IEEE Access*. (2021) 9:14340–50. doi: 10.1109/ACCESS.2021.3052494

23. Zhu Y, Wang X, Chen HS, Salloum R, Kuo CCJ. Green steganalyzer: a green learning approach to image steganalysis. *APSIPA Trans Signal Inf Process*. (2023) 12:e41. doi: 10.1561/116.00000136

24. Xu G, Wu HZ, Shi YQ. Ensemble of CNNs for steganalysis: an empirical study. In: *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*. New York, NY: ACM (2016). p. 103–7. doi: 10.1145/2909827.2930798

25. Qian Y, Dong J, Wang W, Tan T. Deep learning for steganalysis via convolutional neural networks. In: *Media Watermarking, Security, and Forensics 2015, Vol. 9409*. SPIE (2015). p. 171–80. doi: 10.1117/12.2083479

26. Zhang R, Zhu F, Liu J, Liu G. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis. *IEEE Trans Inf Forensics Secur*. (2019) 15:1138–50. doi: 10.1109/TIFS.2019.2936913

27. Öztürk Ş, Özkaya U. Gastrointestinal tract classification using improved LSTM based CNN. *Multimed Tools Appl*. (2020) 79:28825–40. doi: 10.1007/s11042-020-09468-3

28. Ozdemir T, Taher F, Ayinde BO, Zurada JM, Tuzun Ozmen O. Comparison of feedforward perceptron network with LSTM for solar cell radiation prediction. *Appl Sci*. (2022) 12:4463. doi: 10.3390/app12094463

29. You W, Zhang H, Zhao X. A siamese CNN for image steganalysis. *IEEE Trans Inf Forensics Secur*. (2020) 16:291–306. doi: 10.1109/TIFS.2020.3013204

30. Islam MZ, Islam MM, Asraf A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Inform Med Unlocked*. (2020) 20:100412. doi: 10.1016/j.imu.2020.100412

31. Wang L, Xu X, Dong H, Gui R, Yang R, Pu F. Exploring convolutional LSTM for PolSAR image classification. In: *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. Valencia: IEEE (2018). p. 8452–5. doi: 10.1109/IGARSS.2018.8518517

32. Li P, Tang H, Yu J, Song W. LSTM and multiple CNNs based event image classification. *Multimed Tools Appl*. (2021) 80:30743–60. doi: 10.1007/s11042-020-10165-4

33. Zhuang N, Qi GJ, Kieu TD, Hua KA. Differential recurrent neural network and its application for human activity recognition. *arXiv*. (2019) [Preprint] arXiv:1905.04293. doi: 10.48550/arXiv.1905.04293

34. Bas P, Filler T, Pevnỳ T. "Break our steganographic system": the ins and outs of organizing BOSS. In: *International Workshop on Information Hiding*. Cham: Springer (2011). p. 59–70. doi: 10.1007/978-3-642-24178-9_5

35. Piva A, Barni M. The first BOWS contest (break our watermarking system). In: *Security, Steganography, and Watermarking of Multimedia Contents IX, Vol. 6505*. SPIE (2007). p. 425–34. doi: 10.1117/12.704969

36. Tabares-Soto R, Arteaga-Arteaga HB, Mora-Rubio A. Bravo-Ortíz MA, Arias-Garzón D, Grisales JAA, et al. Strategy to improve the accuracy of convolutional neural network architectures applied to digital image steganalysis in the spatial domain. *PeerJ Comput Sci*. (2021) 7:e451. doi: 10.7717/peerj-cs.451

37. Kaggle. *ALASKA2 Image Steganalysis*. (2020). Available online at: https://www.kaggle.com/c/alaska2-image-steganalysis (Accessed February 20, 2023).

38. Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain. *EURASIP J Inf Secur*. (2014) 2014:1–13. doi: 10.1186/1687-417X-2014-1

39. Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. *IEEE Trans Inf Forensics Secur*. (2012) 7:868–82. doi: 10.1109/TIFS.2012.2190402

40. Cogranne R, Giboulot Q, Bas P. Steganography by minimizing statistical detectability: The cases of JPEG and color images. In: *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*. New York, NY: ACM (2020). p. 161–7. doi: 10.1145/3369412.3395075

41. Guo L, Ni J, Su W, Tang C, Shi YQ. Using statistical image model for JPEG steganography: uniform embedding revisited. *IEEE Trans Inf Forensics Secur*. (2015) 10:2669–80. doi: 10.1109/TIFS.2015.2473815

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 02 September 2025
DOI 10.3389/fmed.2025.1619228

OPEN ACCESS

EDITED BY
Habib Hamam,
Université de Moncton, Canada

REVIEWED BY
M. Rajesh Khanna,
Vel Tech Multi Tech Dr Rangarajan Dr Sakunthala Engineering College, India
Noman Sohail,
Universitetssjukhuset i Linkoping Paramedicinska enheten, Sweden

*CORRESPONDENCE
Fehaid Salem Alshammari
✉ falshammari@imamu.edu.sa

RECEIVED 27 April 2025
ACCEPTED 21 July 2025
PUBLISHED 02 September 2025

CITATION
Kaur A, Alshammari FS, Rehman AU and Bharany S (2025) Intelligent Alzheimer's diagnosis and disability assessment: robust medical imaging analysis using ensemble learning with ResNet-50 and EfficientNet-B3. *Front. Med.* 12:1619228. doi: 10.3389/fmed.2025.1619228

Intelligent Alzheimer's diagnosis and disability assessment: robust medical imaging analysis using ensemble learning with ResNet-50 and EfficientNet-B3

Arpanpreet Kaur[1], Fehaid Salem Alshammari[2,3]*, Ateeq Ur Rehman[4,5] and Salil Bharany[1]

[1]Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India, [2]Department of Mathematics and Statistics, College of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia, [3]King Salman Center for Disability Research, Riyadh, Saudi Arabia, [4]Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India, [5]Applied Science Research Center, Applied Science Private University, Amman, Jordan

Neurodegenerative disorder Alzheimer's disease (AD) has progressive characteristics and leads to severe cognitive impairment that reduces life quality. Disease management along with effective intervention depends on the detailed diagnosis conducted early. The proposed framework builds an ensemble system from ResNet-50 and EfficientNet-B3 to conduct automated AD diagnostics by processing high-resolution Magnetic Resonance Imaging (MRI) images. The proposed model uses ResNet-50 to extract features coupled with EfficientNet-B3 as its robust classifier which achieves high accuracy alongside generalization performance. A large, high-quality dataset comprising 33,984 MRI images was used, ensuring diverse representation of different disease stages: the study included participants with four dementia stages organized as Mild, Moderate, Non-demented, and Very Mild Demented. The research applied several comprehensive data preprocessing methods combining normalization steps with rescaling algorithms alongside noise elimination techniques to achieve enhanced performance. Performance tests on the model required examination of accuracy along with precision and recall metrics and F1-score and ROC curve area measurements. The ensemble model delivered remarkable overall accuracy reaching 99.32% while surpassing separate deep learning architectures. The confusion matrix evaluation results showed superb classification results for Mild and Moderate stages along with non-dementia cases while maintaining minimal Wrong choices in Very Mild Demented cases. Experimental findings demonstrate the strength of deep learning algorithms to detect AD disease stages accurately. The robust and accurate performance of the proposed model indicates it has potential for use in medical environments to support radiologists in their work of early-stage AD screening and treatment development. Additional research in diverse clinical environments will strive to optimize and validate the model so it can meet real-world diagnostic requirements for medical use.

KEYWORDS

Alzheimer's disease, neurodegenerative disorder, deep learning, MRI analysis, ResNet-50, EfficientNet-B3, ensemble model, feature extraction

1 Introduction

Alzheimer's disease (AD) is a primary neurodegenerative disease that is responsible for 60%−70% of all dementia cases across the globe, it results in progressive impairment of cognitive and memory function, and overall physical disability mainly in old age. The disease is defined clinically by the deposit of amyloid plaques and neurofibrillary tangles in the brains, leading to the gradual decline in brain volume, and resulting in confusion, poor judgement, language disorder, personality changes, and the inability to carry out activities of daily living (1). To date, aging is still the biggest risk factor for developing AD, but there are also genetic factors, unhealthy life styles, cardiovascular diseases and physical environments that affect the development as well as the progress of AD (2). To date, there is no known cure for Alzheimer's disease but major advancements in medical research have provided methods of managing the disease, these include; cholinesterase inhibitors, memantine, health and safety promotion through changes in diets and coming up with strict exercise regimes that can reduce deterioration of the patient's condition (3). Prior to the publication of DSM IV-Tre quantitative diagnosis of Alzheimer's disease primarily depended on clinical assessment, patient history, and neuropsychological assessment that even though still today have their utility, were reported to provide low sensitivity in early diagnosis of Alzheimer's disease as well as being time consuming and labor intensive. Also, Magnetic Resonance Imaging (MRI) and PET scans have been used to detect abnormalities in the brains of mentally ill patients, although these approaches lack high accuracy when no computational tools are applied (4). Over the last few years, the incorporation of deep learning methods in medical imaging has definitely advanced diagnosis, particularly for Alzheimer's disease as a more precise, fast, and less error-prone approach (5).

Among these, Convolutional Neural Networks (CNNs) have shown exceptional performance in efforts to diagnose MRI patterns that point toward AD, all while surpassing conventional machine learning models by learning features from raw image data. In the context of Alzheimer's disease, the required diagnostic tools are significantly more diverse and refined; this is why ensemble deep learning models have recently become popular as they unite the results of several architectures in one model (6). As for the CNN model selection, two advanced structures including ResNet-50 and EfficientNet-B3 have become the most popular pro forma architectures in recent years due to the higher image classification performance. The vanishing gradient problem is solved through using the ResNet architecture of a deep residual network of 50 layers; deeper networks converge well while capturing details of the images at the same time (7, 8). On the other hand, EfficientNet-B3 uses compound scaling method to control the network depth, width, and so on, making it highly efficient and accurate to extract features with little computational need. Thus, the ensemble of ResNet-50 and EfficientNet-B3 models, where the weaknesses of each of them are masked, and the strengths are combined, contributes to increasing the efficiency of diagnostics compared to using only such architectures and increases the robustness when detecting subtle abnormalities in MRI scans. The main goal of this research is to enhance a deep learning model for distinguishing between the Alzheimer's disease and the Normal Cognitive status by integrating ResNet-50 and EfficientNet-B3 models for MRI data. This approach operates in an attempt to overcome the recognized deficiencies of conventional diagnostic check techniques for AD through the development of an efficient diagnosis system that would be automated, accurate, and fairly easy to implement in the different human populations at the various stages of the disease development (9). In addition, the problem statement focuses on the requirement of an accurate diagnostic tool to differentiate between distinct phases of Alzheimer's disease with robust performance, despite data imbalance, MRI scan noise, and variation (10). Therefore, the major contributions of this study are the development of an ensemble model that comprises ResNet-50 and EfficientNet-B3, an assessment of the performance of the proposed ensemble model against existing deep learning architectures, and a proof of the usefulness of the suggested model in enhancing the diagnostic accuracy of Alzheimer's disease classification. Several works have been extensively conducted on AD detection using standalone CNNs, CNNs with Attention Mechanisms, Ensemble of CNNs and the hybrid of them; their performance is sometimes constrained by a limited number of available diagnostic samples, non-normative database information, and high computational costs (4, 8). For example, Ajagbe et al. (3) and Shirbandi et al. (6) pointed out that applying CNN-based models in MRI-based classification is promising; however, that architectures should be improved to learn deeper and abstract features. Finally, the studies by Sorour et al. (8) and Mujahid et al. (7) showed that the setup based on the ensemble learning is extremely valuable for the detection of AD, as the results of multiple models enhanced positive prediction and diminished the numbers of false-positives. Thus, basing on these achievements, the development of our proposed model is intended to fill the gap in the identified scientific studies and integrate the advantages of ResNet-50 and EfficientNet-B3, including their residual learning ability and computational efficiency. Furthermore, the given work uses techniques like data augmentation and employs adaptive learning to deal with issues that are hard to solve for, including overfitting and imbalance, in order to have a high model accuracy on various MRI datasets (7). The reason as to which ResNet-50 and EfficientNet-B3 were selected for the experiment is because these two architectures have demonstrated good performance across multiple tasks and are robust combinations of feature extraction and classification (8). Based on its deep residual connections which allow the model to learn complex features, ResNet-50 is well-suited to this task, whereas EfficientNet-B3 which incorporates optimized scaling for efficient computations is equally efficient and accurate for the task at hand. This combination is specifically advantageous for medical imaging applications where the minor differences have to be identified between the structures of normal brains and that of the AD patients (6). Moreover, ensemble learning is beneficial in increasing the generalizability of the model, since the combination of more predictions means decreasing the model bias and variance and thus, increasing the diagnostic reliability (7). Finally, this paper intends to make a positive contribution to the available body of knowledge on Alzheimer's disease by proposing a new, yet highly effective, deep learning structure that encompasses the best facets of the ensemble learning technique to deliver the highest possible

diagnostic accuracy. Of critical value and practical applicability, the proposed model can help clinicians make quick and precise diagnosis decisions, which will lead to earlier diagnosis, target treatment plans, and enhanced patient care (2, 8).

In this context, this study contributes to fill the gap of the current diagnostic techniques in Alzheimer's disease and to establish the base for future studies that will promote the creation of new, available and reliable tools with deep learning for Alzheimer's disease diagnosis in magnetic resonance images. This work couples two important elements for the construction of an effective diagnostic test for Alzheimer's disease based on high classification accuracy and explainability. Section 2 gives an extensive literature review of the existing diagnostic conventional approaches, deep learning in neuroimaging. Next, in Section 3, the method is described, more specifically, details about the dataset, the preprocessing of MRI scans, the architecture of the proposed ensemble model based on ResNet-50 and Efficient Net B3. In Section 4, the authors report the findings analyzing the effectiveness of the ensemble model and taking them up against the other classification models. Section 5 contains a discussion of the study's results and their potential, possible clinical uses of the proposed model, its weaknesses, and potential improvements for future work. Also in Section 6, the conclusion of the paper points to the contributions of the study and the implication of applying the proposed approach to timely diagnosis of Alzheimer's disease.

2 Literature review

Alzheimer's disease (AD) classification has received a considerable amount of focus in the medical research sector mainly due to the development of new approaches such as deep learning, which have indicated that they can outperform conventional diagnostic approaches. The two best performing deep learners in this study are the Convolutional Neural Networks (CNNs), specifically ResNet-50 and EfficientNet-B3 reveal promising features for efficient AD diagnosis from brain MRI scans. The subjects of Raza et al.'s (11) study involved segmentation and classification of MRI images of Alzheimer's disease employing transfer learning (TL) and proposed particular CNNs. The approach works on images that segment objects as divided by the brain's Gray Matter. Rather than training from the ground up, there existed a pre-trained deep learning model, to which the process proceeded as transfer learning. The model was compared at 10, 25, and 50 epochs and the mean accuracy was found to be 97.84%. Ironically, transfer learning and segmentation techniques stand as prominent methodologies in a comprehensive framework of medical imaging analysis in diagnosing Alzheimer's disease this study shows the enhancement of accuracy (11). Sharma et al. presents a machine learning model based on transfer learning (TL) and permutation-based voting classifiers for Alzheimer's detection from MRI images. DenseNet-121 and DenseNet-201 extract features in phase one and phase two has classifiers such as support vector machine, Naïve Bayes and XGBoost to classify. Therefore, in the voting mechanism the final predictions are improved with accuracy of 91.75%, specificity of 96.5% and F1-score of 90.25. The model was trained from scratch using a Kaggle data set consisting of 6,200 images in four dementia classes. Mentioned results are completely compatible with the statements and show the higher effectiveness of the offered model compared with state-of-the-art methods; thus, there is perspective to consider the proposed model for usage in clinician applications for Alzheimer's disease identification (12). The authors Zhang, Zhang, Du, and Wang (13) in their study proposed an enhanced neural network known as ADNet from the VGG-16 model for detection of Alzheimer's diseases applying 2D MRI slices. Those modifications consist of depthwise separable convolution to decrease the number of parameters; however, the model uses ELU activation to avoid the problem of exploding gradients; the model also incorporated an SE module for effective feature recalibration. Similarly, training is combined with auxiliary tasks: regression of clinical dementia and mental state score. Experimental results proved that the proposed approach gives 4.18% higher accuracy of AD compared with cognitively normal (CN) and 6% of MCI accuracy compared with CN than the VGG16 model. These outcomes indicate that multitask learning solutions and better architecture for the neural network may help ADNet to support early Alzheimer's detection. Solano et al. (14) uses a three dimensional DenseNet model for the detection of Alzheimer's disease using Magnetic Resonance Imaging (MRI). Using the proposed deep neural network classifier, an overall accuracy of 0.86, sensitivity of 0.86, specificity of 0.85, and the area under the ROC curve (micro-average) of 0.91 for five disease stages. Focusing on the ability to produce replicable results, the approach uses only the tools available freely online, which means it should be more easily implemented in poorer countries as well. This approach helps to show that deep learning is useful in medical diagnosis and the equitable distribution of technology for installation and use. Carcagnì et al. (15) investigate the performance of CNNs and the adaptive self-attention mechanism for identifying Alzheimer's using brain MRI data. In particular, the study utilizes deep learning methods in improving the detection accuracy and speed of Alzheimer's disease, through exploiting the features of CNN, through a feature extraction step and exploiting self-attention to learn the long-range dependencies. In addition, proofs reveal a vast scope for the use of some automated diagnostic tools to have a high sensitivity and specificity compared with conventional practices. The work focuses on the implementation of the new AI models in the early diagnosis and effective individualized approach to the disease, providing a solid base for non-invasive and horizontally scalable dementia diagnostics (16). In recent years, deep learning proved to be a valuable approach in analyzing genomes, responding to the large and dependent features' patterns and correlations. The recent innovations include variation in model structures, paradigms of model establishment, and techniques of model decoding all focused on the prophetic models of genetic variants and their influence on the disease causation. In such context, this review addresses how genomic deep learning techniques remain rather flexible for disease-oriented investigations with reference to neurodegenerative disorders including Alzheimer. It uses primarily the articles on Alzheimer's disease and considers more general methods, explaining the potential value of these approaches. To the best of our knowledge, the review conducted

by Jo et al. aimed at reviewing future research directions at the crossroads of neurodegeneration, genomics, and deep learning (16). Deep learning has emerged as an essential element of genomic analysis because of its capability to handle large genomic data by identifying the diverse relationships between them. Progress includes the following new trends in models: model architecture, model development philosophies, and model interpretation techniques for estimating the effects of genetic variants on disease progression. This review shows how to incorporate genomic deep learning methods into disease-specific models with an emphasis on neurodegenerative diseases such as Alzheimer's. It focuses on Alzheimer's literature and where it identifies more general methodological approaches, it explores their suitability. In addition, Qui et al. have discussed directions for future work involving neurodegeneration genomics, and deep learning (17). Hazarika et al. compares different deep learning (DL) models in AD classification using brain Magnetic Resonance (MR) images collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. However, the DenseNet-121 model showed the highest accuracy of 88.78%, a bit slower than the others because of the extensive convolutions. Thus, to overcome this kind of limitation, the authors suggested a new DenseNet-121 structure, where instead of the conventional convolutional layers, the depth-wise convolutional layers should be used. These optimizations improved computational and accuracy rates making the average accuracy to be 90.22%. The results discussed above imply future possibilities of depth-wise convolution in enhancing the DL-based AD classification models (18). In their paper, Helaly et al. describes a system for early detection of Alzheimer's disease (AD) and multi-stage classification with the help of convolutional neural networks (CNNs). Two methods are explored: specifically, the use of 2D and 3D CNNs for structural images, and apply transfer learning with VGG19 to improve the classification performance. Therefore, based on the ADNI dataset, the highest precision rate established was 93.61% in 2D; 95.17% in 3D, and 97% in VGG19. A web application helps in diagnosing and staging AD remotely, and improving health care access during COVID-19. The approach is simple and less computationally demanding, and the method's performance is stable and suitable for medical applications based on its evaluation on nine criteria (19). Jo et al. employed the 3D convolutional neural networks (CNN) and layer-wise relevance propagation designed to diagnose AD using tau PET scans. MCI using the proposed model he has come up with a result of 90.8% accuracy by using AD and cognitively normal (CN) subjects. Using information from voxel-wise analysis the key regions identified were hippocampus, thalamus, and entorhinal cortex. Probability of AD, calculated from cognitive measures, was associated with medial temporal tau deposition in MCI, proving useful in detection at this stage (20). Table 1 below shows the state of art comparison.

3 Proposed methodology

On the same note the proposed methodology outlines a comprehensive framework of Alzheimer's disease diagnosis. First, a clear overview of the dataset is provided, including its characteristics, which is diverse, clean and has high quality ground truth labels to enable accurate training and testing. Normalization, rescaling, center cropping, and elimination of noisy regions also prepares the data to be in the right standard. Each of these transformations enhances model robustness and, at the same time, can help increase its ability to generalize. The diagnostic framework involves an ensemble model of ResNet-50 and EfficientNet-B3 networks which are the best for the feature extraction and the classification, respectively. Moreover, evaluation criteria by accuracy, precision, recall, F1-score, and area under the ROC curve are used to provide more detailed analysis of the performance of the model. A general idea of the proposed methodology flowchart is presented in Figure 1 below.

3.1 Dataset description

The dataset used in this study is a publicly available MRI dataset sourced from Kaggle, titled the "Augmented Alzheimer MRI Dataset" (22). It comprises a total of 33,984 2D T1-weighted MRI slice images, not full 3D volumes, evenly divided among four diagnostic categories: Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented as shown in the Figure 2, the images are saved in JPEG format and have undergone data augmentation and applied solely to the training set to enhance diversity and prevent overfitting. The validation and test sets were left unaltered to ensure unbiased evaluation and preprocessing by the original dataset providers and represent 2D slices extracted from volumetric MRI scans. The dataset does not contain subject-level metadata such as age, gender, imaging protocol, or acquisition parameters. Due to the absence of subject identifiers, the dataset was split at the image level rather than the patient level. As a result, adjacent slices from the same volume may exist across training, validation, and test sets, potentially introducing correlation-based bias. The images were divided into training (80%), validation (10%), and testing (10%) subsets, corresponding to 27,188, 3,397, and 3,399 images, respectively. Due to the absence of patient identifiers, the split was performed at the image level, and this limitation is acknowledged as a potential source of correlation bias (23). It is important to note that this dataset includes images that were augmented by the dataset provider prior to release. Therefore, it is most appropriate for use in training and internal evaluation. The lack of access to original, non-augmented scans limits the dataset's suitability for external validation or generalization studies.

The original dataset does not include metadata regarding MRI acquisition protocols, sequence parameters, scanner types, or image reconstruction software, and thus, such details could not be reported in this study. It is important to note that this dataset includes images that were augmented by the dataset provider prior to release. Therefore, it is most appropriate for use in training and internal evaluation. The lack of access to original, non-augmented scans limits the dataset's suitability for external validation or generalization studies. Further, no documentation regarding ethics approval, patient consent, or institutional data sourcing is available for this dataset, and its origin cannot be independently verified. The distribution of the classes is tabulated as follows in Table 2.

TABLE 1 Comparison with state of art.

Reference	Technique used	Advantages	Disadvantages
Sharma et al. (2022) (12)	Hybrid artificial system (HTLML) for Alzheimer's disease diagnosis	Finally, the use of multiple Artificial Intelligence techniques for a better result	They proposed that complexity manifested in hybrid models could result in longer time taken during training and high computational costs
Qiu et al. (2020) (17)	A clear and understandable deep learning structure	Used for explaining model adult human decision making	
Jo et al. (2020) (20)	Using residual deep learning on tau PET imaging	Concentrates in the identification of tau protein images in Alzheimer's	May be tuned to small fluctuations in MRI data
Solano-Rojas and Villalón-Fonseca (2021) (14)	A DenseNet neural network for early identification of Alzheimer's disease	A less expensive method with reasonable efficiency for early detection	Lacks capability of real time and high processing speed for 3D data
Jo et al. (2022) (16)	Application of deep learning for the analysis of genetic variants	It allows the analysis of massive genetic data to classify Alzheimer's	Is highly dependent on the availability of large high quality genotype data for use in training
Hazarika et al. (2022) (21)	Different Deep Learning Architectures for Alzheimer's Classification	Compared and contrasted several models, toward the decision-making process of selecting the right approach	Some of these techniques may compromise the model's accuracy or, sometimes, make it less complex
Helaly et al. (2022) (18)	AI based early diagnosis of Alzheimer's disease	Another stamina is early identification abilities since the program detects omissions at the beginning	Mixed evidence provided by models; models need to be chosen more carefully
Raza et al. (2023) (11)	Preprocessing and feature selection in Alzheimer's disease identification	Utilizes pre-trained models that mostly help to decrease the time and amount of training data needed	Some of native to the domain features might not be recognized by the pre-trained models
Carcagnì et al. (2023) (15)	CNN and self-attention learners	Proper to extract features from the brain MRI images	Self-attention mechanism may be costly
Zhang et al. (2024) (13)	This proposal addresses multi-task learning with an enhanced or modified version of a neural network	Multi-talented and able to work on a number of projects at once, hence increasing productivity	Complexity in models often leads to over fitting and these models will need large data sets

3.2 Preprocessing

In this paper, data preprocessing is found to be a fundamental step in enhancing the machine learning outcomes especially in classifying Alzheimer diseases using MRI scans. Because of the variations witnessed in the quality of images and the small differences in the brain boundaries some preprocessing techniques are very essential to improve the input images (24). First, a process of image normalization is conducted so that the pixel values range from 0 to 1 to reduce possible deviations due to image sizes. Although no explicit denoising or contrast enhancement was applied, several data augmentation techniques were used to enhance the training data and improve model robustness. These included random rotations, zooming, flipping, and brightness variation. All images were resized to 224 × 224 pixels and normalized to a pixel intensity range of [0, 1] before being fed into the models. This makes the model generalized better and also relieves it from overfitting (25). All these preprocessing steps serve to enhance the quality of data put into the ensemble model for the correct identification of Alzheimer's stages (26).

a.) **Normalization:** normalization is the task of adjusting the range of pixel intensities of an image to a standard range, often the interval [0:1]. The most common method is min-max normalization, which can be expressed mathematically as given in the Equation 1 below:

$$x_{norm} = x - x_{min}\frac{x - x_{min}}{x_{-}\ x_{min}} \tag{1}$$

where x_{norm} is the normalized portion of the pixel value, x is the actual pixel value, and x_{min} is the minimum value of the pixel in the picture, x_{max} is the maximum value of the pixel in the picture.

b.) **Resizing:** resizing is the process of moving each pixel of an image to a new location in relation to desired width and height of the targeted image. If (W_{in}, H_{in}) is the width and height of the original image and (W_{out}, H_{out}) is the width and height of the resized image. While maintaining the spatial relationships. If (W_{in}, H_{in}) be the width and height of the original image and (W_{out}, H_{out}) be the width and height of the resized image. To standardize input dimensions for model training, each 2D MRI slice was resized to 224 × 224 pixels using bilinear interpolation. This resizing adjusted the number of image pixels but did not account for physical voxel dimensions, which could not be preserved due to the absence of spatial resolution metadata in the JPEG-formatted dataset. So, the scaling factors for width and height are computed as in Equation 2 below:

$$s_w = \frac{W_{out}}{W_{in}},\ s_h = \frac{H_{out}}{H_{in}} \tag{2}$$

c.) **Data augmentation:** data augmentation involves applying various operations on the existing dataset in order to create an enlarged and diversified set in order to improve generalization. It features different augmentations like rotation, scaling, shifting, flipping among others as shown in Figure 3. Rotating an image by angle θ is given by the formula as shown in the

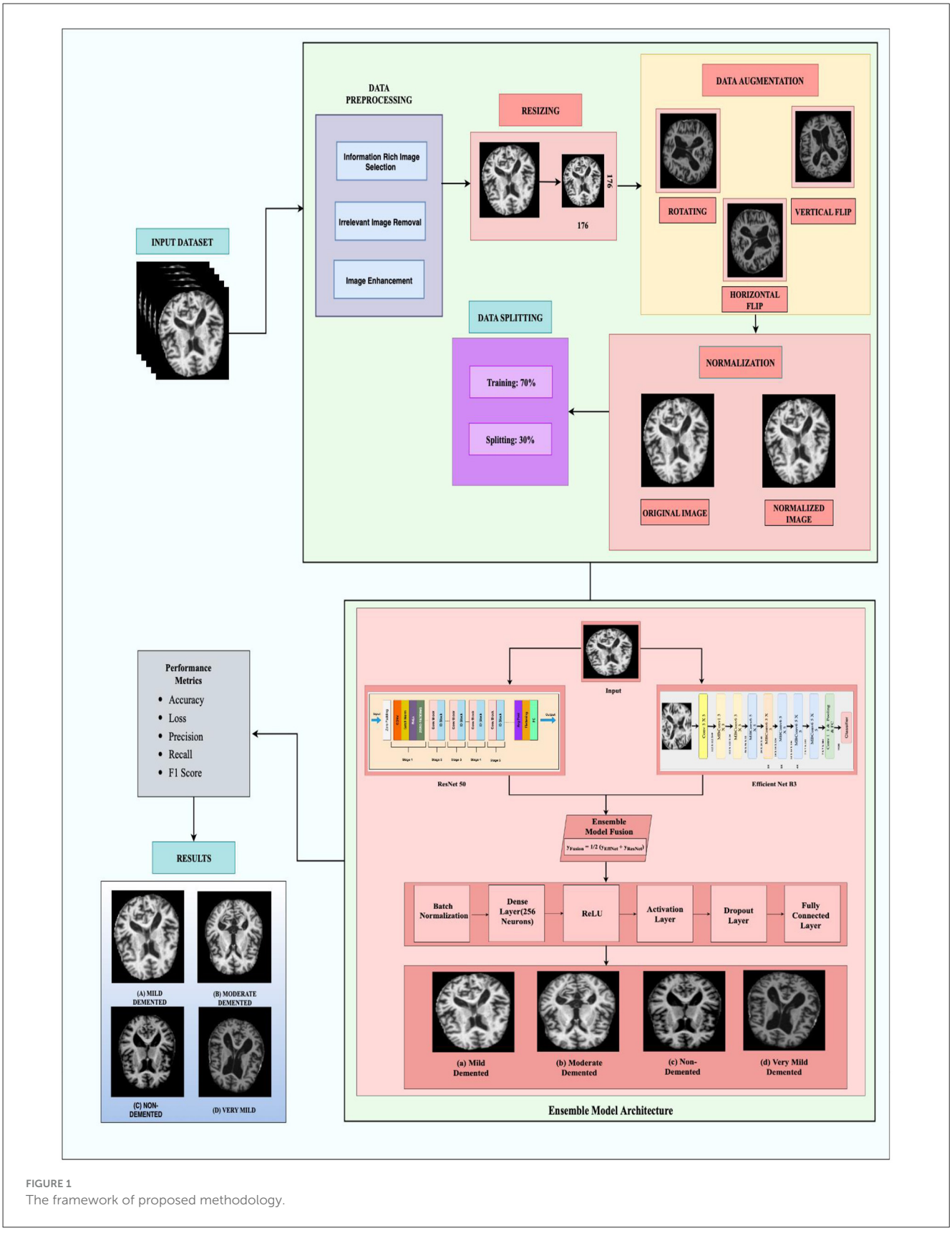


FIGURE 1
The framework of proposed methodology.

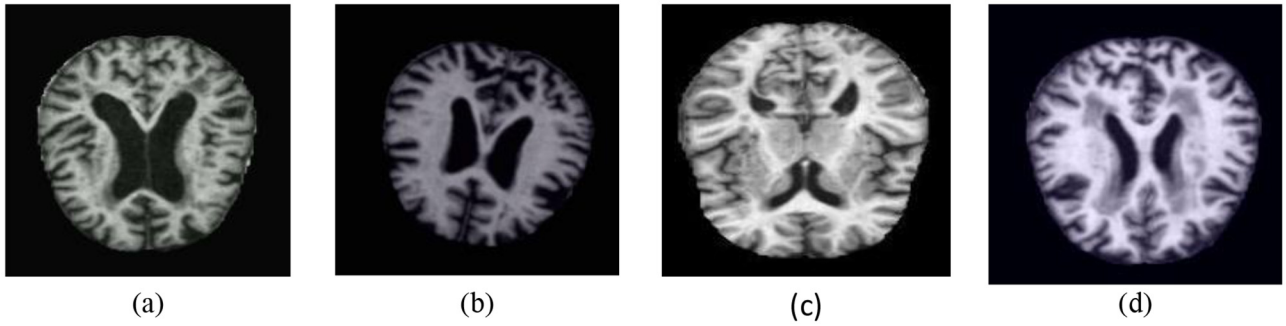


FIGURE 2
Dataset classes: **(a)** Mild Demented **(b)** Moderate Demented **(c)** Non-Demented **(d)** Very Mild Demented.

TABLE 2 Class wise dataset distribution.

Dataset	No. of images in 'Mild Demented' class	No. of images in 'Moderate Demented' class	No. of images in 'Non-Demented' class	No. of images in 'Very Mild' class	Total images
Training	6,797	6,797	6,797	6,797	27,188
Validation	850	850	850	850	3,400
Testing	850	850	850	850	3,400
Total	8,497	8,497	8,497	8,497	33,988

Equation 3 below.

$$\left[x'\ y'\right] = \left[cos\theta - sin\theta\ sin\theta cos\theta\right]\left[x\ y\right] \tag{3}$$

where, the coordinate position of the original raster image pixel is designated by (x, y) and that of the new position is by (x', y') and the angle of rotation is θ in radians. Horizontal flipping reflects an image across the vertical axis. This transformation can be mathematically represented by reversing the x-coordinate of each pixel as in Equation 4:

$$x' = -x,\ \ y' = y \tag{4}$$

This augmentation is particularly useful in medical imaging to introduce left–right symmetry, thereby improving the model's robustness to orientation variance.

Vertical flipping reflects the image across the horizontal axis and is represented as in Equation 5 below:

$$x' = x,\ \ y' = -y \tag{5}$$

This operation helps simulate top–bottom inversion, further enhancing the model's ability to learn invariant spatial features, especially when orientation does not impact diagnostic relevance.

3.3 Model building

Two architectures of deep learning models, the ResNet50 and EfficientNet-B7 that form the basis of the ensemble model are generated by this method. Each model is established meticulously to construct components of MRI images essential for satisfying classification exclusively.

3.3.1 ResNet-50

ResNet-50 consists of 50 layers, including convolutional layers, pooling layers, batch normalization (BN), and fully connected layers, as illustrated in Figure 4a. ResNet' s principal invention is the residual block; this essential function is a "shortcut" or direct pathway that sends the input to the layer through to the output. This allows the model they base to skip certain layers and decrease the gradient disappearance problem in very deep networks (27). These shortcut connections help the network retain accuracies of deeper models possible without crossing the degradation issue by "jumping" other layers. The recognized blocks of architecture include the pooling layers, batch normalization, ReLU activation functions, and convolutional layers in sequence is given mathematically by Equation 6.

$$y = F\left(x, \{w_i\}\right) + x \tag{6}$$

Here x is the input to the residual block, y is the output, $F(x\{w_i\})$ is the function that is applied on the input x. The last layer of classification produces output z_{resnet} as described below in Equation 7 after passing through the network.

$$y_{resnet} = softmax\left(W_{resnet} \times Y_{global} + b_{resnet}\right) \tag{7}$$

Here, W_{resnet} and b_{resnet} are the weights and biases of the dense layer, and Y_{global} is the output from the global average pooling layer.

The convolutional block from ResNet-50 as illustrated in the Figure 4b, is a deep convolutional neural network that aids in the vanishing gradient problem through the element of residual learning. This block was implemented with the intent of being used to extract features while still allowing deeper networks to learn. The convolutional block includes three types of convolutional layers

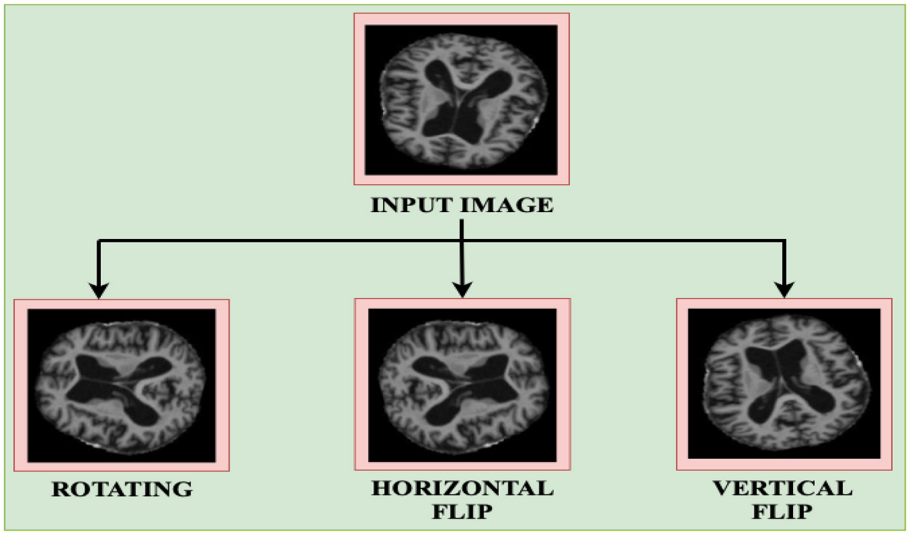


FIGURE 3
Data augmentation techniques.

implemented in a sequence. The first layer is a 1 × 1 convolution that decreases the dimension of the input feature maps in order to lessen computational cost Stage 3: technology 3. The second one is another convolution layer with size 3 × 3 to cover spatial connections and explicit features. The third layer is another 1 × 1 convolution to get back to the original dimensions of the feature maps. After each convolution there is normalization to make the training process faster and more stable, as well as using activation function (ReLU). The feature that is unique to the convolutional block is the projection shortcut connection, which uses 1 × 1 convolution to bring the dimensions of the input to match that of the processed features. This makes some sense as it actually establishes compatibility for the element-wise addition on the shortcut and the convoluted feature maps. Then a feedback layer addition is applied, and finally has the activation function to get the output. This design makes it possible for ResNet-50 to learn initially both low level and high-level features in deep networks.

In addition, an identity block in ResNet-50 as depicted in Figure 4c is an essential building block aimed at transferring features well through deep architectures. As it will be seen, the identity block retains the input dimensions since it uses a skip connection that feeds the input directly to the output without any change of dimension. This helps in making the model fast and stable while processing in the later stage of the training. The identity block contains three layers of convolution. The first is a 1 × 11 times one convolution layer that is aimed at the dimensionality of the input feature maps. This is succeeded by a 3 × 33 times three convolution which extracts spatial features and patterns, and one more 1 × 11 times one convolution which brings back dimensionality. Each convolutional layer is associated with batch normalization to update the activation for acceleration of convergence as well as activation function like ReLU. The key feature of the identity block is that the input directly connects to the output without passing through the convolutional layers by adding the input feature maps with the corresponding feature maps after passing through the network. After this addition there is an activation function to produce the output. The added identity block makes ResNet-50 deepen this network while allowing it to maintain the hoisting of features and avoid the vanishing gradient issue, making it a great architecture for acquiring features.

3.3.2 Efficient Net

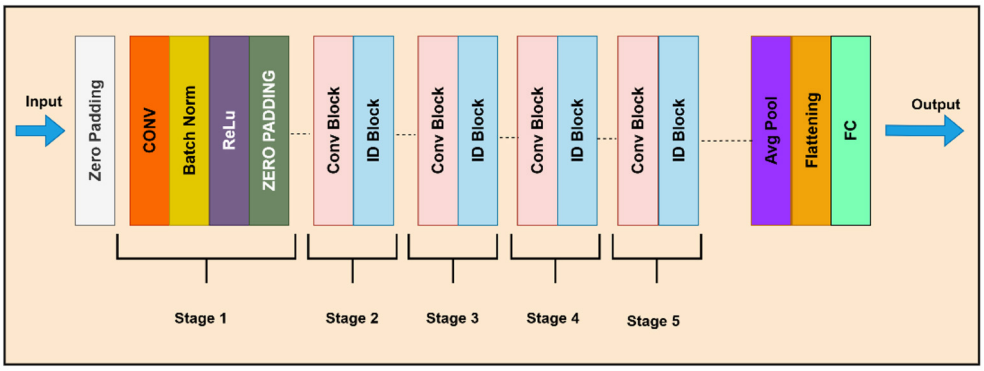
Based on a compound scaling coefficient, Efficient Net aims to optimize at the same time depth, width and the resolution according to a parameter Ø that represents a family of models. EfficientNet-B3 is one particular network in the Efficient Net series of models and, as with all models in this series, this network enforces a balance between these three aspects to yield decent compromise between model complexity, model accuracy, and compute requirements (26). The scaling is governed by Equation 8:

$$d = \alpha^{\phi}, w = \beta^{\phi}, r = \gamma^{\phi} \tag{8}$$

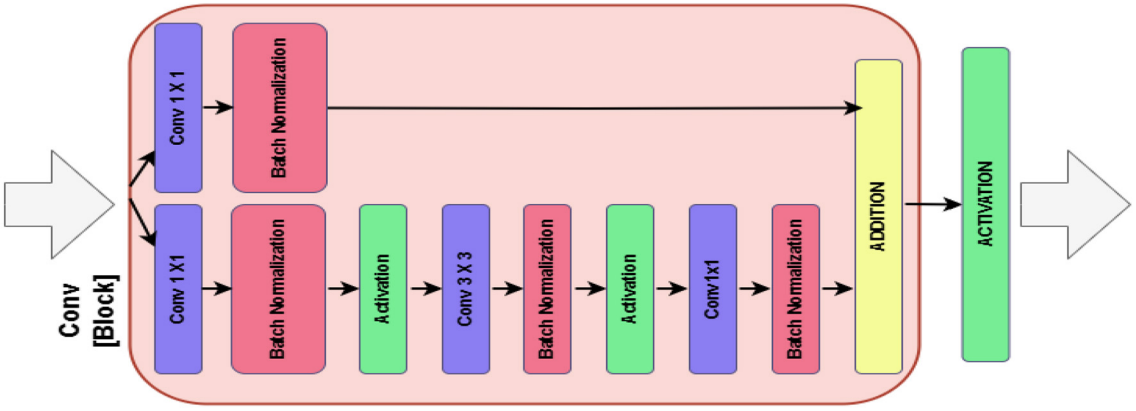
where d, w, and r are the network's depth, width, and resolution, respectively and where α, β, and γ are parameters. The output of EfficientNet-B7, after global average pooling, is shown in Equation 9:

$$y_{efficientnet} = softmax\left(W_{efficientnet} \times f_{global} + b_{efficientnet}\right) \tag{9}$$

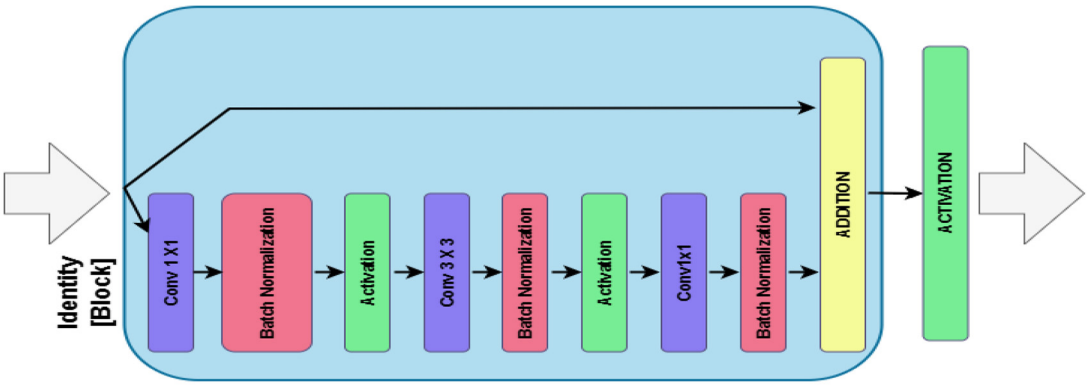
where f_{global} is the feature vector, and $W_{efficientnet}$, $b_{efficientnet}$ are the weights and the biases of the dense layer. The architecture of EfficientNet-B7 demands for many important components: from original input, features are extracted by convolutional layers to improve gradient flow and achieve batch normalization and the Swish activation function. The Figure 5 shows the architecture of Efficient Net B3.



(a) Resnet 50 Architecture



(b) Convolutional Block



(c) Identity Block

FIGURE 4
Resnet 50 **(a)** Resnet 50 architecture **(b)** Convolutional block **(c)** Identity block.

3.3.3 Ensemble model architecture

In the proposed ensemble model, ResNet-50 and EfficientNet-B3 were trained independently using the same training dataset to classify MRI slices into four Alzheimer's disease stages. During inference, both models generate probability scores for each class through softmax layers, and these outputs are combined using a soft voting approach by simply averaging the predictions. This fusion allows the ensemble to benefit from the complementary strengths of both networks: EfficientNet-B3 offers high efficiency with fewer parameters, while ResNet-50 contributes deep hierarchical feature extraction through residual learning. To stabilize training and reduce internal covariate shift, batch normalization is applied to the fused features, followed by a dense layer with 256 neurons and ReLU activation for non-linearity. Regularization techniques,

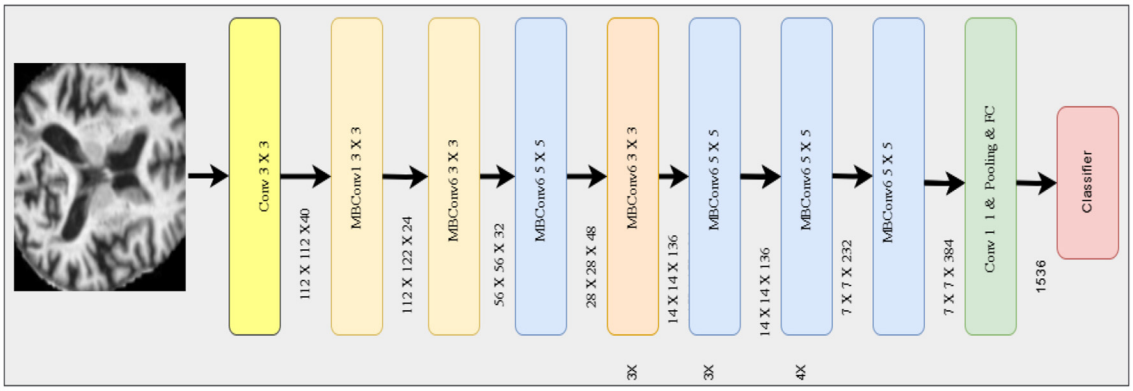


FIGURE 5
Efficient Net B3 architecture.

including both L1 and L2 penalties, are applied to prevent overfitting, and a dropout layer is used to further improve generalization. The final classification is performed through a fully connected layer that maps the processed features to class probabilities. The model is trained using categorical cross-entropy loss, which evaluates the difference between predicted and true class labels. Overall, this ensemble design enhances diagnostic performance by combining the robustness of two diverse deep learning architectures as in the Figure 6. Rather than assigning weighted average or performing any other operation, the outputs from both the models are then simply averaged as they have been observed to complement each other. EfficientNet-B3 gives state of the art efficient feature representation using fewer number of parameters compared to ResNet-50 which offers strong hierarchical feature representation due to its residual learning (11). The combined output fusion is computed as shown in Equation 10 where y_{EffNet} is the final prediction of Efficient B3 and y_{ResNet} is the final prediction of Resnet 50.

$$y_{Fusion} = \frac{1}{2}.(y_{EffNet} + y_{ResNet}) \tag{10}$$

This particular fusion strategy also ensures that both models contribute equally enough to the ensemble so that generalization over the various patterns across images will be well-captured.

Batch normalization (BN) is then employed on the fused features to stabilize and enhance the speed of the whole training process by normalizing the outcomes. The normalized feature vector $\hat{y}$ is computed as in Equation 11:

$$\hat{y} = \frac{y_{Fusion} - \mu}{\sqrt{\sigma^2} + \epsilon} \tag{11}$$

where μ and σ^2 are the estimate of average of the batch, and variance of the batch respectively and ϵ is a small constant value so as to avoid division by zero. Trainable scaling (γ) and shifting (β) parameters further refine the normalized features by using the Equation 12:

$$y' = \gamma.\hat{y} + \beta \tag{12}$$

This step reduces the covariate shift problem within the organization's internal environment, meaning that there is a more stable distribution of the particular features through the layers. The features being batch normalized are then fed through a dense layer with 256 output neurons. This layer applies a linear transformation followed by a ReLU activation for non-linearity as shown in Equation 13:

$$z = ReLU(W.y' + b) \tag{13}$$

where, W is Weight matrix, b is Biase vector and ReLU(a) = max (0, a) To prevent overfitting, L1 and L2 regularization terms are added to the loss function, penalizing large weights as in Equation 14:

$$Regularization\ Loss = \lambda_1 . ||\ W||_1 + \lambda_2 . ||\ W||_2^2 \tag{14}$$

Also, Dropout layer which drops out neurons with the probability p is implemented to increase the ability of generalization of the model. The last fully connected layer adopts the SoftMax function in order to convert the distilled features to probabilistic outcomes reflecting the number of categories of the output. For each class k, the output probability y_k is given by Equation 15:

$$y_k = \frac{exp\ exp\ (z_k)}{\sum_{j=1}^{C} = exp(z_j)} \tag{15}$$

where C represents the number of classes, while z_k is the logit for class k. The model is trained using categorical cross-entropy loss, minimizing the divergence between true labels $y_{i,k}$ and predicted probabilities $y_{(i,k)}$ as in Equation 16

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{C} y_{i,k}\ log(\hat{y}_{i,k}\ ) \tag{16}$$

All the enhancement methods used in the proposed ensemble model, namely, feature fusion, normalization, dense layers, and regularization, make it highly capable to perform well in the classification of Alzheimer's disease. Using EfficientNet-B3 and ResNet-50, this approach offers significant capabilities for the early diagnosis, which further outperform the outcomes of separate models with higher accuracy and their generality. A dropout layer is applied after the ReLU-activated dense layer and before the final classification layer to reduce overfitting and improve generalization.

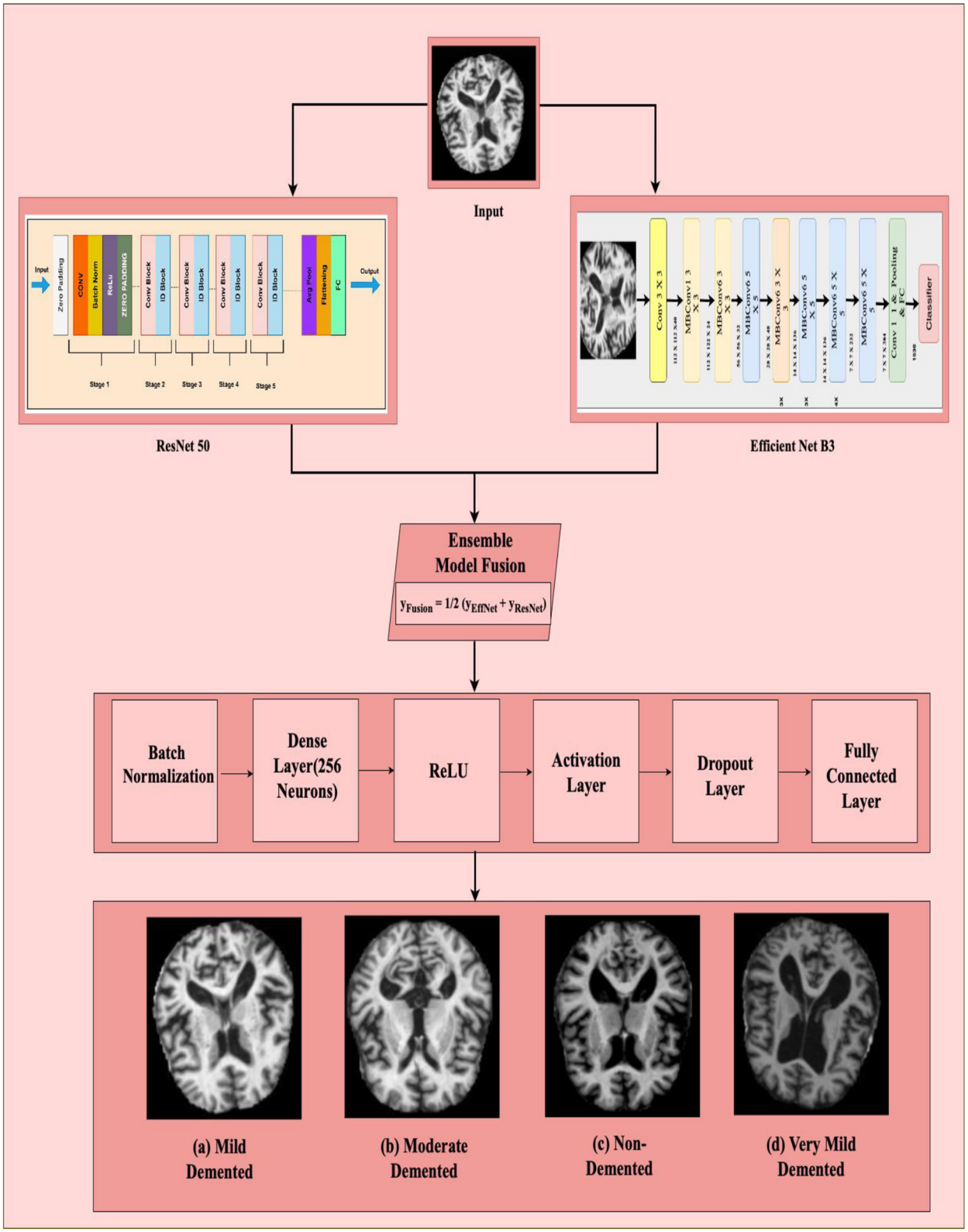


FIGURE 6
Ensemble model architecture. The framework consists of: (1) Input brain MRI image; (2) Feature extraction using ResNet-50 and EfficientNet-B3; (3) Ensemble model fusion, where outputs of ResNet-50 and EfficientNet-B3 are combined; (4) Classification head composed of Batch Normalization, Dense Layer (256 neurons), ReLU, Activation Layer, Dropout Layer, and Fully Connected Layer; and (5) Final classification into four categories: **(a)** Mild Demented, **(b)** Moderate Demented, **(c)** Non-Demented, and **(d)** Very Mild Demented.

TABLE 3 Training hyperparameters.

Hyperparameter details	Value/description
Optimizer	Adam
Learning rate	0.0001
Loss function	Categorical cross entropy
Batch size	32
Number of epochs	10
Input image size	224 × 224 × 3
Dropout rate	0.5
Data split ratio	80% Training 10% Validation 10% Testing
Data augmentation	Rotation, Zooming
Framework used	Python 3.8, TensorFlow 2.9, Keras, OpenCV, NumPy, Matplotlib

3.3.4 Hyperparameter details

To ensure optimal model performance and training stability, a carefully selected and tuned range of hyperparameters for both ResNet-50 and EfficientNet-B3 models used in the ensemble (11). These parameters were chosen based on preliminary experimentation and established best practices in deep learning for medical imaging. Key hyperparameters include the choice of optimizer, learning rate, batch size, number of training epochs. A detailed summary of the hyperparameters used in this study is provided in Table 3. These settings were consistent across both models to ensure fairness and effective ensemble integration. The models were developed using Python 3.8 with the TensorFlow 2.9 and Keras libraries. Additional preprocessing and evaluation were performed using NumPy, OpenCV, scikit-learn, and Matplotlib.

4 Results

This section presents the experimental results obtained from evaluating the proposed ensemble model comprising ResNet-50 and EfficientNet-B3 on the Alzheimer's MRI classification task. The model's performance was assessed using standard evaluation metrics, including accuracy, precision, recall, and F1-score across four Alzheimer's disease stages: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The results demonstrate that the ensemble approach outperforms individual models in terms of both classification accuracy and generalization capability. Detailed comparisons, confusion matrices, and performance tables are provided to illustrate the effectiveness of the proposed method and support its potential for clinical deployment in diagnostic workflows.

4.1 Evaluation parameters

An evaluation parameter is a measure by which the performance, efficiency or effectiveness of a model, process, or system can be judged. Such parameters are commonly applied in different areas including machine learning, statistics, finance and engineering.

a) Accuracy: accuracy in multi-class classification is defined as the ratio of correctly predicted samples to the total number of samples across all classes. It measures the overall effectiveness of the model in assigning the correct label to each input as in Equation 17 below:

$$Accuracy = \frac{No.\ of\ correct\ predictions}{Total\ No.\ of\ predictions} = \frac{\sum i = 1CTPi}{N} \quad (17)$$

Where TPi = True Positives for class i, C = Total number of classes, N = Total number of samples, where i can be any class out of four classes of Alzheimer.

b) Precision: precision measures the proportion of correct positive predictions for each class out of all predictions made for that class. It indicates how many of the predicted instances for a specific class are actually correct. Precision is presented by the formula of precision expressed in Equation 18 below:

$$Precision_i = \frac{TPi}{TPi + FPi} \quad (18)$$

c) Recall: recall, also known as sensitivity, measures the proportion of actual positives that were correctly identified for each class. It shows how well the model captures the true instances of each class. The formula of precision is expressed below in Equation 19 below:

$$Recall_i = \frac{TPi}{TPi + FNi} \quad (19)$$

Where FN_i is false negative for class i.

d) F1-Score: the F1-score is the harmonic mean of precision and recall for each class. It balances the trade-off between precision and recall, especially useful when classes are imbalanced. The F1-score is calculated as shown in Equation 20:

$$F1_i - Score = 2 \times \frac{(Precision_i \times Recall_i)}{(Precision_i + Recall_i)} \quad (20)$$

4.2 Training and validation results

Comparative analysis of performance was conducted between ResNet-50 and EfficientNet-B3 during their training and validation stages. Two different computational frameworks trained against a predefined dataset to evaluate their performance by calculating their accuracy and precision during validation with recall and F1-score metrics achieved alongside AUC-ROC value evaluations. The feature extraction abilities of ResNet-50 were excellent but required precision adjustments through fine-tuning to reach its best levels of operation. The efficient scaling of EfficientNet-B3 produced superior accuracy results while maintaining better generalization capabilities. The validation results showed that EfficientNet-B3 demonstrated better performance than ResNet-50 models primarily because of its superior structural design. Background inference speed

retained similarity between ResNet-50 and other comparison models. A decision between the two systems depends on whether applications prioritize accuracy or computational speed. The model was evaluated using multi-class performance metrics, including overall accuracy, precision, recall, and F1-score. These metrics were calculated for each of the four classes individually and macro-averaged to provide an overall assessment.

4.2.1 Training and validation results of efficient net B3

Performance trends from the EfficientNetB3 based Alzheimer's disease detection model can be found in the depicted accuracy and loss data plots. The deployment of 10 epochs throughout training yielded positive results which appeared in both training and validation metrics. Both training and validation data show continuous performance improvements throughout the epochs according to the accuracy plot displayed on the left. The initial training accuracy level was ∼65% before reaching near 95% stability. The generalization capacity becomes evident through the validation accuracy which shows a start value higher than training accuracy and converges to 95%. The models training and validation accuracy graphs remain close together which means the model avoids major overfitting problems. Training along with validation loss shows continuous reduction throughout the overall training process according to the loss plot. Training losses initiate at 0.7 but continuously decrease and settle near 0.1 by the end of training (28). The validation loss chain shows a downward movement which starts underneath the training loss mark then reaches similar value terminals at epoch completion. The model's robust structure receives additional confirmation through the parallel changes observed in validation and training loss metrics. Effective learning and generalization abilities stand out in the EfficientNetB3 architecture when used for Alzheimer's disease detection based on its metric convergence performance. The balanced performance of training and validation curves demonstrates that the model effectively extracts significant data features while avoiding overfitting which demonstrates its practical utility in clinical diagnostics settings. All performance metrics are displayed through the graphs presented in Figure 7.

4.2.2 Training and validation results of ResNet 50

Multiple plots show the performance metrics between training data accuracy and validation data accuracy alongside training data loss and validation data loss when using ResNet-50 for Alzheimer's disease prediction. The training process required 10 epochs toward model evolution yet the performance metrics showed some separateness between training and validation results. The accuracy graph (left) demonstrates that model training accuracy gradually improved from 60% to a nearly 95% level throughout ten epochs. Initially the validation accuracy started at ∼70% then climbed to reach nearly 87% values. Beyond the fifth epoch the validation accuracy demonstrates unstable patterns which could be explained by overfitting and changes found within the validation dataset. The decreasing trend on loss data demonstrates successful learning between training data along with validation data. Training loss begins at 0.9 before reaching 0.2 only after completing the training period. From its starting point at 0.8 the validation loss gradually lowers until reaching a minimum of 0.4 at epoch five. Beyond epoch 5 the validation loss exhibits a tiny upward trend because the model effectively performs on training data however, it misses essential patterns needed for unseen input recognition (29). Throughout the later part of training the separation between validation and training performance metrics demonstrates that ResNet-50 successfully grasps patterns from the data although it needs further development for generalized results. Early stopping alongside data augmentation and standard techniques for regularization offer potential solutions to reduce overfitting. The ResNet-50 model shows promise for Alzheimer's disease detection capabilities through its excellent training accuracy results and fair validation performance potential that creates opportunities for future clinical diagnostic applications. All performance metrics have their graphical representations displayed in Figure 8.

4.2.3 Training and validation results of proposed ensemble model

These graphic displays show how an ensemble with ResNet-50 and EfficientNetB3 models detects Alzheimer's disease throughout 10 training cycles. The left graph shows accuracy performance which demonstrates exceptional model behavior through rapid improvement of training and validation accuracy toward perfect scores. The model establishes an initial training accuracy baseline at 70% which evolves into 100% accuracy during the fourth epoch then maintains peak performance for the remaining epochs. The baseline validation accuracy sits at 85% during the initial stage after which it establishes perfect synchronization with training accuracy throughout subsequent epochs. The coaches' curves align perfectly which demonstrates the model will generalize successfully and avoids excessive overfitting behavior. A loss plot analysis reveals that both training and validation loss decrease sharply in initial epochs to stabilize at low levels. Training loss displays initial values of about 3.5 that diminish rapidly to less than one unit during epoch 5 then settles down at that minimum value point. Validation loss displays a parallel reduction pattern which starts near 2.5 before decreasing under 0.5 during epoch 4 while training loss tracks closely in subsequent epochs (30). The parallel development of accurate results and low loss data points demonstrates the sturdy characteristics of the ensemble model system. The ensemble methodology uses ResNet-50 and EfficientNetB0 to extract complementary functionality which delivers outstanding results for Alzheimer's disease diagnosis. The model demonstrates accurate pattern recognition in the data through quick criterion alignment and data metric convergence without producing overfitting issues. The ensemble approach demonstrates potential utility as a dependable medical diagnostic instrument since it delivers accurate results alongside sharp dataset generalization abilities. All performance metrics are displayed graphically in Figure 9.

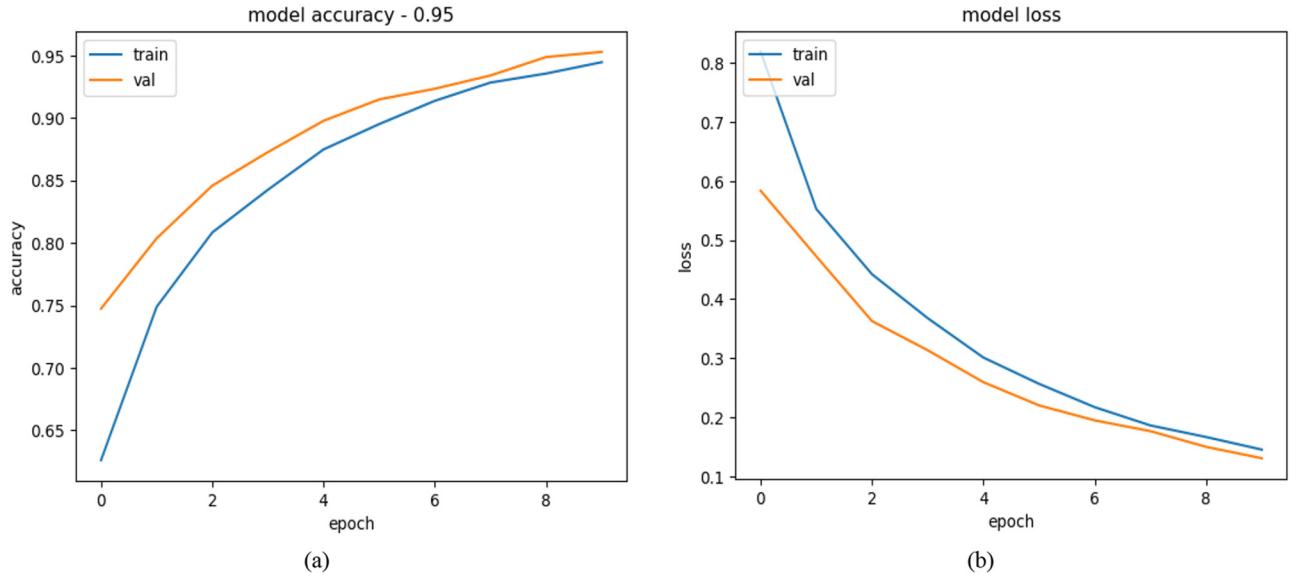


FIGURE 7
(a) Model accuracy **(b)** model loss of efficient net B3.

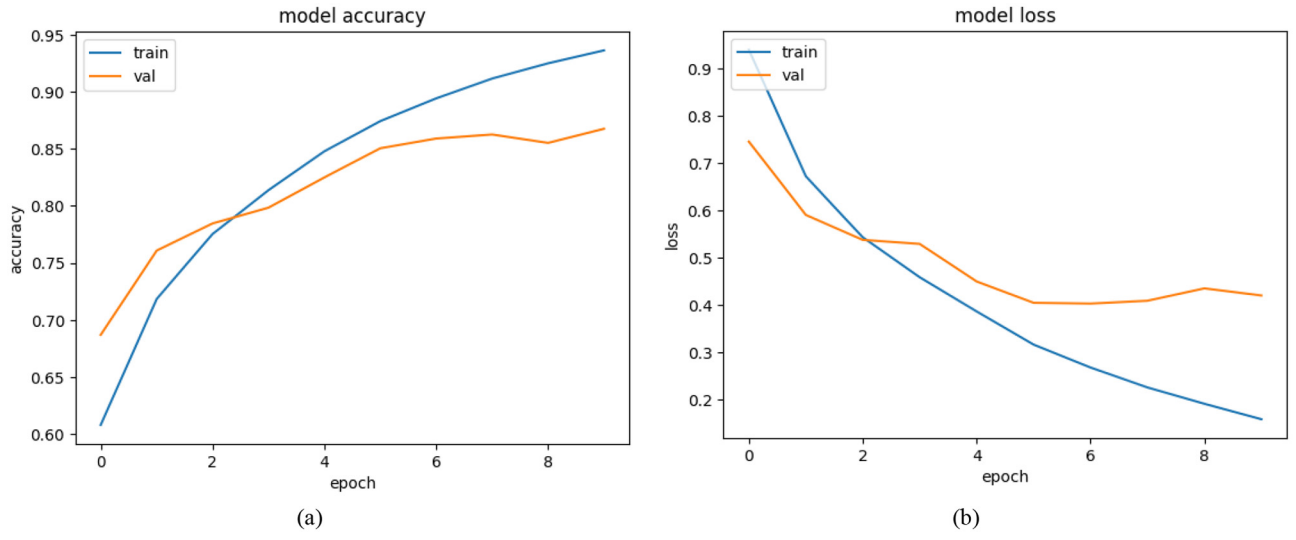


FIGURE 8
(a) Model accuracy **(b)** model loss of Resnet 50.

4.2.4 Comparison results of ensemble model, EfficientNet-B3 and ResNet50

The performance metrics for multiple deep learning models across ten epochs are shown in Table 4 where training accuracy and validation accuracy and validation F1-score are evaluated. Scientists apply equivalent deep learning technologies from this domain to detect Alzheimer's disease through MRI medical imaging. The progressive neurodegenerative psychiatric condition Alzheimer's disease leads to cognitive decline so it requires early diagnosis to deliver effective therapeutic measures. The diagnostic systems built with CAD capabilities utilize EfficientNet-B3 along with ResNet50 and ensemble models as they demonstrate exceptional accuracy in image recognition tasks. The training and validation accuracy of both EfficientNet-B3 and ResNet50 increase through epochs and the ensemble model exceeds the performance of each model individually. All performance metrics, including accuracy, precision, recall, and F1-score, were calculated in a multi-class setting across four classes. Per-class metrics were computed and macro-averaged to summarize overall model performance. Ensemble learning proves beneficial because diverse model combinations increase generalization ability which then produces superior diagnostic results. Deep learning models trained with Alzheimer's Disease Neuroimaging Initiative (ADNI) medical images demonstrate potential for Alzheimer's disease

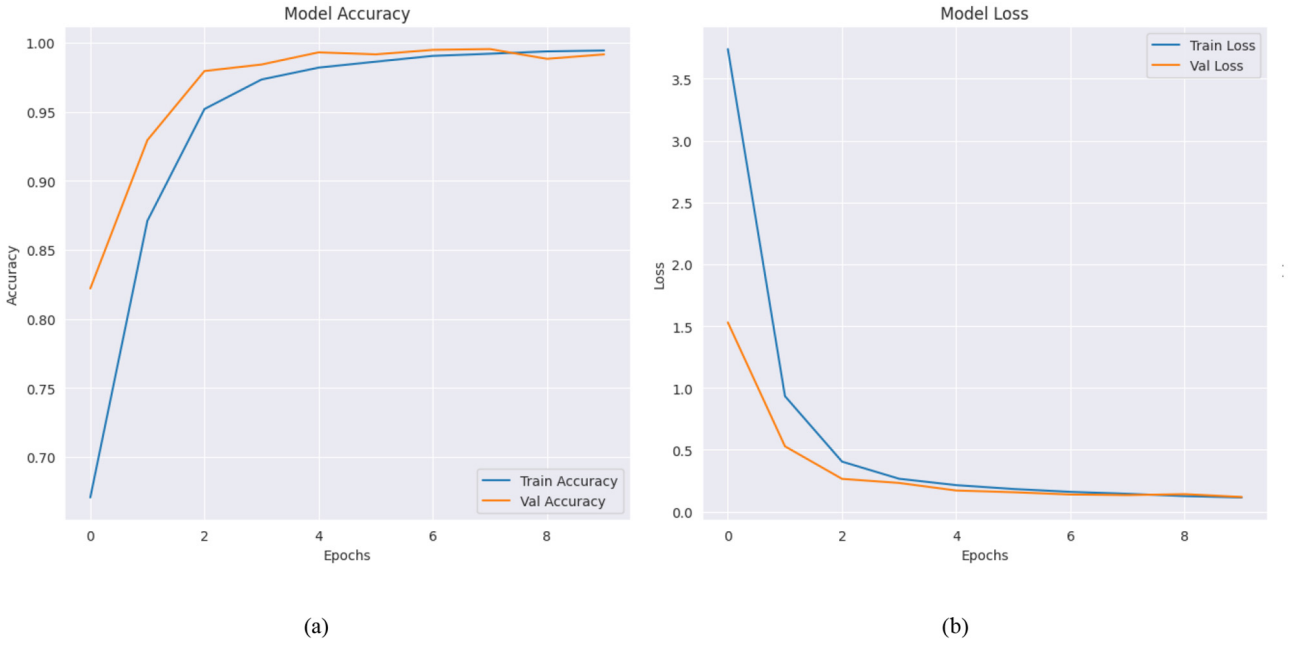


FIGURE 9
(a) Model accuracy **(b)** model loss of ensemble model.

detection applications. The EfficientNet-B3 model demonstrates top capability in extracting MRI scan features followed by ResNet50 which automatically adjusts training depths to overcome vanishing-gradient difficulties by using its residual learning method. The ensemble model's high performing results indicate that using multiple architectures enhances detection accuracy for early-stage Alzheimer's disease. The F1-score acts as a vital tool for medical researchers because it evaluates model performance specifically during assessment of diagnosis systems which operate on imbalanced datasets primarily featuring underrepresented early-stage and mild Alzheimer's cases. Analysis of the F1-score values shows that the ensemble model maintains its superior performance throughout all epochs while achieving optimal precision and recall ratings. Morocco's scientific research benefits from F1-score accuracy which strives to improve disease detection at both non-diseased and diseased case levels thereby supporting clinical tools development. Model learning effectiveness and generalization ability increase concurrently with validation accuracy across epochs which proves fundamental when applying medical approaches to real-world situations. Deep learning algorithms with similar models from the table enable researchers to create dependable CAD systems which benefit neurologists through improved Alzheimer's disease diagnosis accuracy. The diagnostic accuracy can be improved by two techniques: domain-specific transfer learning fine-tuning and additional multimodal data analysis. Deep learning demonstrates its critical role in disease detection through the data trends presented in the table. Researchers implementing these technologies in Alzheimer's detection will achieve early diagnosis while enabling faster interventions that ultimately lead to better patient results. The Table 4 below shows the comparison of Resnet 50, Efficient Net B3, and ensemble model.

4.3 Testing results

Real-world testing of ResNet-50 and EfficientNet-B3 produced evaluation results. The superior generalization capabilities of EfficientNet-B3 became evident through improved accuracy and precision together with enhanced recall. The model was superior to ResNet-50 in recognizing minimal patterns while producing fewer mistakes. The real-time applications could benefit from the ResNet-50 model because it delivers inference operations at a faster pace. The scoring system emphasized EfficientNet-B3 as the best model in discrimination capability assessment. The efficiency of ResNet-50 did not reduce its competitive strength unless optimum hyperparameters were used. Two efficient network choices exist: EfficientNet-B3 provides enhanced accuracy while ResNet-50 delivers crucial speed performance for applications. Additional adjustments to model parameters combined with better data preparation will help increase test results from both systems.

4.3.1 Classification results of EcientNet-B3, ResNet50, and ensemble model

The classification report in Table 5 provides a comprehensive breakdown on testing models across four categories by showing accuracy data as well as recall metrics alongside F1-score percentages and class support counts. Our results show the ensemble model based on ResNet50 plus EfficientNet-B3 delivers advanced detection of Alzheimer's disease across all four disease classification levels. The ensemble model executed with ResNet50 and EfficientNet-B3 demonstrated absolute classification precision and recall and F1-score values of 1.00 for detecting Mild Demented, Moderate Demented and Non-Demented cases. The

TABLE 4 Comparison of ResNet-50, EfficientNet-B3, and Ensemble model.

Epoch	Model	Training accuracy	Validation accuracy	Validation F1-score
1	EfficientNet-B3	0.6261	0.7473	0.5482
	ResNet50	0.608	0.687	0.686
	Ensemble model	0.6707	0.822	0.8365
2	EfficientNet-B3	0.7489	0.8037	0.6947
	ResNet50	0.7184	0.7608	0.758
	Ensemble model	0.8709	0.9294	0.9355
3	EfficientNet-B3	0.8083	0.8458	0.7797
	ResNet50	0.7754	0.7846	0.784
	Ensemble model	0.9519	0.9794	0.9809
4	EfficientNet-B3	0.8425	0.8726	0.8349
	ResNet50	0.8138	0.7985	0.798
	Ensemble model	0.9733	0.9841	0.9854
5	EfficientNet-B3	0.8748	0.8977	0.8889
	ResNet50	0.8479	0.8249	0.824
	Ensemble model	0.9819	0.9929	0.9935
6	EfficientNet-B3	0.8951	0.9148	0.9124
	ResNet50	0.8744	0.8505	0.85
	Ensemble model	0.9862	0.9915	0.9922
7	EfficientNet-B3	0.9137	0.9233	0.9201
	ResNet50	0.8942	0.8591	0.859
	Ensemble model	0.9904	0.9947	0.9951
8	EfficientNet-B3	0.9283	0.934	0.9311
	ResNet50	0.9116	0.8626	0.862
	Ensemble model	0.9919	0.9953	0.9957
9	EfficientNet-B3	0.9355	0.9487	0.946
	ResNet50	0.925	0.8553	0.855
	Ensemble model	0.9936	0.9882	0.9891
10	EfficientNet-B3	0.9446	0.9528	0.9504
	ResNet50	0.9363	0.8676	0.868
	Ensemble model	0.9943	0.9915	0.9922

model maintains a precision rate of 0.98 and recall rate of 1.00 when classifying Very Mild Demented images. This produces an F1-score of 0.99. Evaluation shows that when measuring performance separately, the EfficientNet-B3 model produces superior results than ResNet50 because it achieves 0.95 precision compared to 0.87 precision together with 0.95 recall compared to 0.87 recall which generates a superior overall F1-score. The F1-score of EfficientNet-B3 achieves 1.00 in detecting Moderate Demented cases in particular together with strong performance in all present classes. ResNet50 demonstrates reduced performance in identifying Very Mild Demented cases and achieves recall levels of 0.76 thereby affecting its overall classification precision. The coordinating method capitalizes on the individual capabilities of both systems thereby enhancing overall classification performance. The ensemble model demonstrates reliable performance with an overall accuracy rating of 0.9932 which confirms its potential use for automated Alzheimer's disease detection.

4.3.2 Confusion matrix of EfficientNet-B3

A confusion matrix serves as a performance evaluation tool which enables researchers to evaluate how machine learning models classify different data points. A basic mathematical unit that displays the real classification output with the model prediction output during model analysis. The rows display real-world labeling and the columns deliver model prediction classes. The research invests in studying the confusion matrices obtained from the Ensemble Model alongside ResNet50 and

TABLE 5 Comparison of various parameters under different models.

Class	Model	Precision	Recall	F1-score	Support
Mild Demented	Ensemble model	1	1	1	896
Moderate Demented		1	1	1	647
Non-Demented		1	1	0.99	960
Very Mild Demented		0.98	1	0.99	896
Mild Demented	ResNet50	0.82	0.92	0.87	896
Moderate Demented		0.99	0.98	0.98	927
Non-Demented		0.84	0.85	0.84	927
Very Mild Demented		0.86	0.76	0.81	907
Mild Demented	Efficient net B3	0.96	0.98	0.97	932
Moderate Demented		0.99	1	1	602
Non-Demented		0.93	0.94	0.94	979
Very Mild Demented		0.94	0.91	0.93	886
Overall accuracy	Ensemble model	0.99	0.99	0.99	3,399
	ResNet50	0.87	0.87	0.87	3,399
	Efficient Net B3	0.95	0.95	0.95	3,399

EfficientNet-B3. The confusion matrix in Figure 10 evaluates the EfficientNet-B3 model's performance in classifying Alzheimer's disease stages: Mild, Moderate, Non, and Very. The model demonstrates impressive accuracy by accurately identifying Mild (915 correct) and Moderate (602 correct) cases paired with sparse misdiagnosis occurrences. The identification of non-Alzheimer's international cases proves reliable at 928 while showing some wrong assignments of very severity. Severe cases (805 correct) show occasional confusion with Non-cases (56 misclassified). The successful early and moderate stage differentiation by EfficientNet-B3 needs improvements for better discrimination between severe disease presentations and non-diseased conditions to create accurate tools for clinical diagnosis.

4.3.3 Confusion matrix of ResNet 50

The Resnet 50 model delivers excellent diagnostic accuracy when distinguishing between Mild Demented and Non-Demented groups since it makes 823 and 784 correct determinations at once. The evaluation shows certain classification errors occur most frequently between Very Mild Demented and Non-Demented categories. Habitat Resnet 50 demonstrates accurate performance detecting Moderate Demented stages because it delivers 653 precise identification results while minimally misclassifying any samples. A significant number of Very Mild Demented cases get assigned to the Mild Demented group in addition to the 111 diagnoses which the classifier labels as non-demented based on Figure 11. Distinguishing dementia at early stages from healthy individuals remains a challenge for early intervention because both cases present similar symptoms.

4.3.4 Confusion matrix of ensemble model

Each category shows robust performance in classification based on the ensemble model where most instances fall within correct interpretations. Our analysis showed the model correctly identified 896 cases of Mild Demented and 647 cases of Moderate Demented along with 938 non-demented cases and 895 Very Mild Demented patients. The classification method shows minimal mistakes because occasional Very Mild Demented cases accidentally overlapped with non-demented cases (19 images) while other classification results were unaffected (31). The integrated ResNet50 and EfficientNet-B3 model successfully identifies different dementia stages because of its powerful feature extraction strengths. Both ResNet50 and EfficientNet-B3 contribute remarkable capabilities to classification accuracy by demonstrating strong combinations of deep learning methodology and parameter optimization capabilities. The ensemble model proves highly suitable for early-stage Alzheimer's detection through its minimal misidentification errors in identifying groups of Moderate Demented patients along with Mild Demented patients as shown in Figure 12. The ensemble model demonstrates high diagnostic accuracy which makes it suitable for automated Alzheimer's disease detection systems that would help doctors intervene early and make better medical choices. The ensemble model demonstrates superior performance by attaining maximum accuracy while making the fewest classification errors especially in subjects with Mild and Moderate Demented diagnosis. The EfficientNet-B3 performs exceptionally well in mild and moderate case identification although it displays challenges when trying to identify severe cases. The ResNet50 Model demonstrates successful operation however, its efficiency decreases when attempting to distinguish very mild Dementia from persons who do not have dementia.

5 External validation

To evaluate the generalization ability of the proposed ensemble model, an external validation was performed using

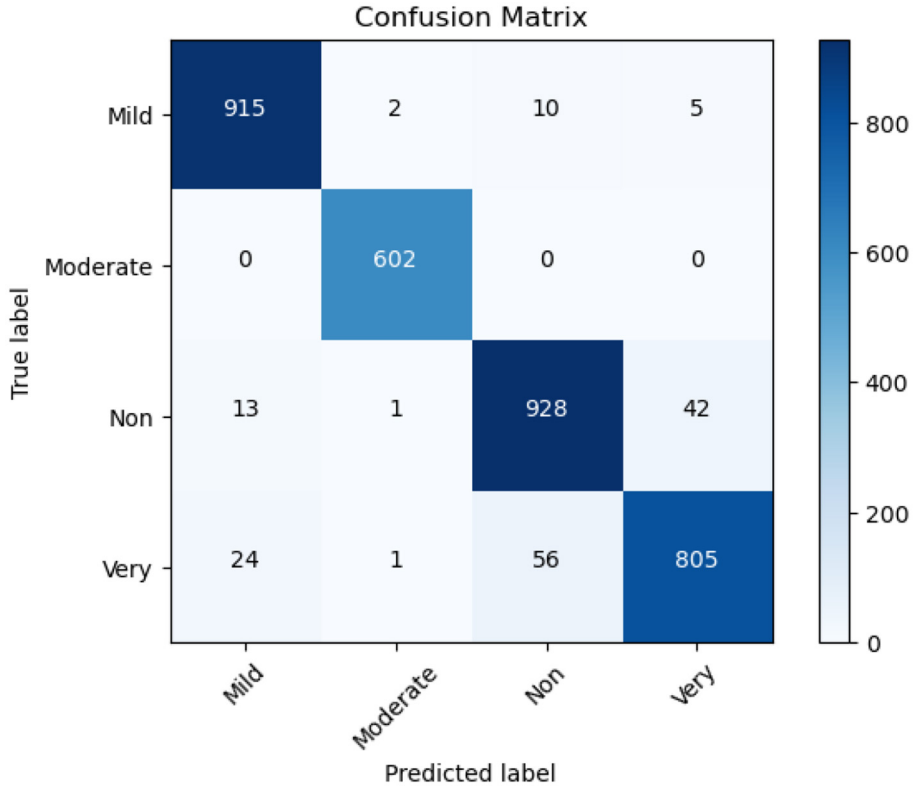


FIGURE 10
Confusion matrix of Efficient Net B3.

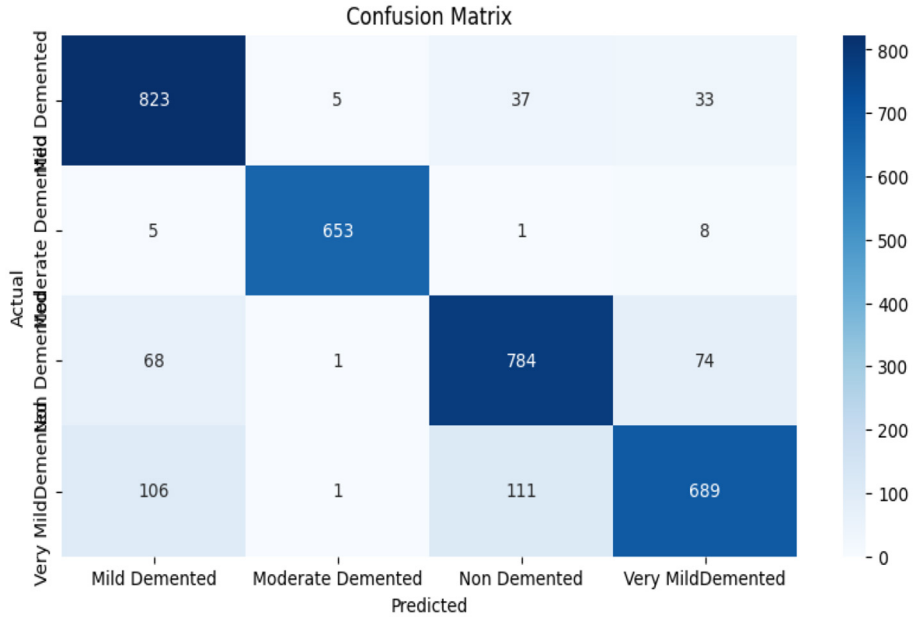


FIGURE 11
Confusion matrix of ResNet 50.

a separate dataset comprising 6,400 MRI images representing four stages of Alzheimer's disease: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The results confirm the robustness and accuracy of the model beyond the training data, demonstrating its potential for real-world clinical application (32).

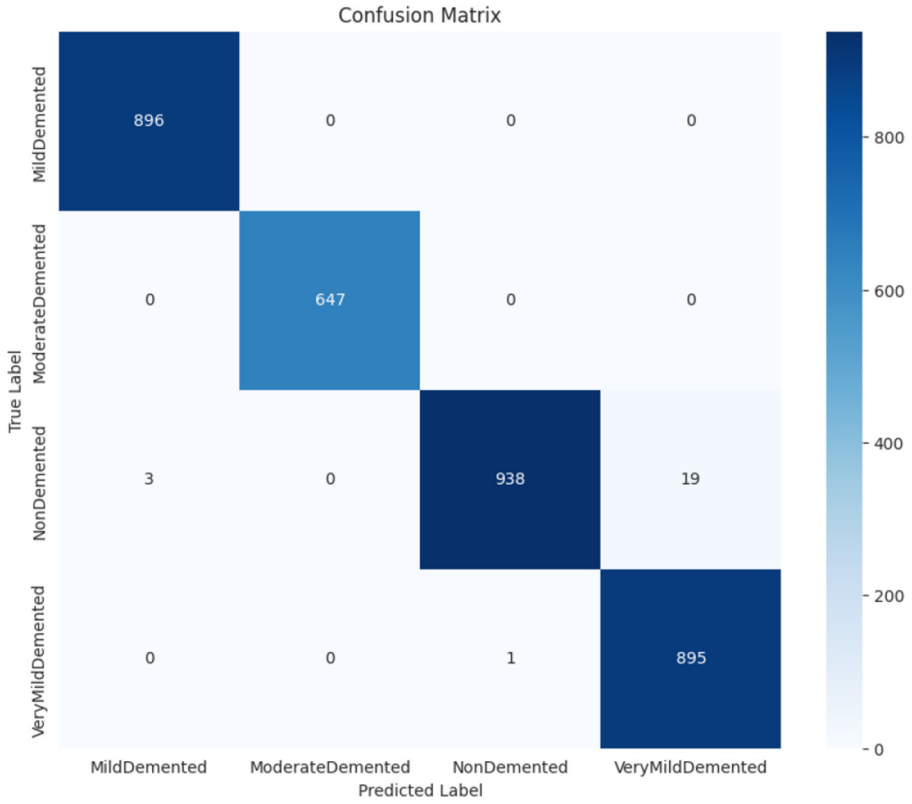


FIGURE 12
Confusion matrix of ensemble model.

The model achieved an overall accuracy of 97%, with consistently high precision, recall, and F1-scores across all classes. Specifically, the Non-Demented class yielded a precision of 0.96 and a recall of 0.94, resulting in an F1-score of 0.95. The Very Mild Demented class, which represents early-stage Alzheimer's detection, achieved perfect scores—precision, recall, and F1-score all at 1.00—though this result should be interpreted with caution due to the relatively small sample size ($n = 10$). The model also performed well on the Mild Demented and Moderate Demented categories, achieving F1-scores of 0.97 and 0.96, respectively as depicted in the Table 6 below.

Macro and weighted averages for all metrics were uniformly 0.97, indicating that the model maintains consistent performance across both balanced and imbalanced class distributions. These results suggest that the ensemble model, which combines ResNet-50 and EfficientNet-B3, is capable of accurately distinguishing between Alzheimer's disease stages even when evaluated on data not seen during training.

The results are promising, but the limited number of samples in some classes—especially Very Mild Demented—warrants further validation using larger, clinically diverse datasets. Future work will focus on subject-level validation using datasets with patient identifiers, clinical metadata, and imaging protocols to assess the model's robustness in practical diagnostic environments.

TABLE 6 Performance metrics on external validation dataset.

Class	Precision	Recall	F1-score	Support
Mild Demented	0.96	0.94	0.95	145
Moderate Demented	1.00	1.00	1.00	10
Non-Demented	0.97	0.98	0.97	513
Very Mild Demented	0.96	0.96	0.96	356
Accuracy			0.97	1,024
Macro Avg	0.97	0.97	0.97	1,024
Weighted Avg	0.97	0.97	0.97	1,024

6 Comparison with state-of-the-art

This research demonstrates how recent developments improve disease detection models and dataset capabilities and classification metrics when compared to current field-leading detection approaches. Research using deep learning algorithms ResNet50, EfficientNet, VGG16, and DenseNet has evaluated Alzheimer's disease classification from MRI scans with different degrees of achievement. The application of CAM-CNN on MRI scans with VGG19 and ResNet101 network models produced a 98.85% accuracy outcome where ResNet101 provided better performance

TABLE 7 Comparison on the basis of aspects.

Ref No	Year	Technique used	Number of classes	Name of classes	Accuracy
(4)	2024	VGG19 and RESNET 101 with CAM-CNN	4	• Non-Dementia • Without Dementia • Very Mild Dementia • Mild Dementia • Moderate Dementia	98.85%
(7)	2023	Ensemble of EfficientNet-B2 and VGG-16	4	• Mild Demented • Moderate Demented • Non-Demented • Very Mild Demented	97.35%
(9)	2024	Using various architectures like VGG 16, VGG 19, Dense Net 121	5	• Binswanger Dementia • Hemorrhagic Dementia • Multi-infarct dementia • Strategical dementia subcortical dementia	84.67%
(10)	2024	Using deep learning techniques	4	• Mild Demented • Moderate Demented • Non-Demented • Very Mild Demented	80.14%
(15)	2024	Using ResNet, Dense Net, and Efficient Net	4	• Mild Demented • Moderate Demented • Non-Demented • Very Mild Demented	75.06%
Proposed model		Ensemble Model of Resnet 50 and Efficient Net-B3	4	• Mild Demented • Moderate Demented • Non-Demented • Very Mild Demented	99.32%

than VGG19. The combination of EfficientNet-B2 with VGG16 allowed researchers to produce a model that reached 97.35% accuracy through transfer learning applications. Individual use of ResNet50 in previous research reached an accuracy of 80.14% yet displayed spaces where its classification accuracy might be enhanced. Research results using multiple models including VGG16 and DenseNet121 with ResNet50 demonstrated an accuracy level of 84.67 percent which indicates the requirement for better ensemble strategies. The research introduces an ensemble model that joins ResNet50 with EfficientNet-B3 to improve classification outcomes in a major way. The proposed model delivers 99% overall performance accuracy because Mild Demented, Moderate Demented, and Non-Demented classes achieve precision, recall and F1-score values of 1.00. Feature extraction capabilities of EfficientNet-B3 reveal its superiority over ResNet50 since individual assessments show precision at 0.95 vs. 0.87 and an F1-score of 0.99. To surpass benchmarked models this research generated an ensemble method that brings together beneficial characteristics from EfficientNet-B3 and ResNet50 including their optimized architecture and deep feature learning ability. Its high classification accuracy makes this approach a promising option for automated Alzheimer's detection while enabling better medical decision support particularly during early diagnosis. A summary of these two methods appears in Table 7.

Several recent studies have contributed valuable insights into the development of intelligent diagnostic systems, which support the objective of this research. For instance, Zhang et al. (33) demonstrated the clinical benefits of precision imaging techniques in neurosurgical applications, highlighting the importance of targeted image-guided interventions in neurological disorders, a concept that aligns with the need for accurate neuroimaging analysis in Alzheimer's disease. Yin et al. (34) proposed an EEG-based emotion recognition system using autoencoder feature fusion and MSC-TimesNet, which exemplifies the utility of deep learning in neurocognitive data interpretation. Similarly, Tian et al. (35) introduced a novel self-supervised learning model for binocular disparity estimation, indicating the growing potential of self-supervised frameworks that could be extended to medical imaging applications such as Alzheimer's classification. Furthermore, Xiao et al. (36) presented a large-scale machine learning-based dementia risk model tailored to elderly populations with depression, providing a strong clinical basis for integrating predictive analytics in Alzheimer's risk assessment. Zhu (37) explored memory impairment detection through computational intelligence in substance abuse patients, reinforcing the relevance of machine learning in cognitive disorder diagnostics. Zhan et al. (38) investigated brain strain analysis using *in-vivo* and simulation data, underlining the value of biomechanical modeling in neurodegenerative research. Li et al. (39) applied machine learning to diagnose sarcopenia using sEMG signals, showing the adaptability of ML in aging-related disease detection. Lastly, Xiang et al. (40) employed a systems biology approach to explore potential therapeutic mechanisms in Alzheimer's, offering complementary biological insights that support a multimodal understanding of the disease. Together, these works underscore the feasibility and importance of leveraging advanced machine learning, neuroimaging, and multimodal integration strategies—paralleling the aims of our ensemble learning-based framework using ResNet-50 and EfficientNet-B3 for Alzheimer's diagnosis and disability assessment.

7 Discussion

Research development centers on building an ensemble model for Alzheimer's disease detection while showcasing its value for clinical assessments. The proposed model extends clinical abilities of neurologists and radiologists through its accuracy enhancement and robustness while facilitating timely precise diagnostic procedures that minimize human error and enhance early treatment strategies. The absence of patient-level demographic data, including age and gender, limits the model's ability to analyze performance variations across different population subgroups. Future work will utilize clinically annotated datasets to enhance interpretability and fairness and use datasets that allow patient-wise splitting to ensure proper generalization. The lack of patient identifiers prevented subject-level data splitting. Consequently, the model may have been exposed to highly correlated adjacent slices across training and test sets, increasing the risk of overfitting and overestimating performance. Although augmentation and splitting were carefully performed, the absence of subject identifiers may result in correlated slices from the same subject appearing in different data subsets, potentially impacting generalization. Through implementation in hospital imaging platforms the ensemble model functions as a medical decision tool which enables specialists to detect Alzheimer's disease manifestations at different stages confidently. Due to the absence of raw volumetric MRI files and acquisition metadata, advanced corrections such as N4 bias field correction could not be applied, which may affect intensity uniformity across slices. Since the dataset was pre-augmented and lacks original raw scans, it may not be suitable for standalone testing or external benchmarking. This restricts our ability to fully assess generalization and may introduce bias if augmentation artifacts influenced the model. Deep learning methods showcase their potential to outperform conventional diagnostic methods through the successful ensemble architecture which unites ResNet50 and EfficientNet-B3 networks. A key limitation of this work is the absence of imaging acquisition metadata, such as sequence types and scanner specifications, as the dataset was sourced from a publicly available platform (Kaggle) that did not include these details. This limits our ability to assess the model's robustness across different clinical imaging conditions. The enhanced accuracy of combined model identifications results in increased abilities to distinguish dementia's early stages from standard brain abnormalities thereby enabling prompt medical care. The improved diagnosis system reliability comes from better misclassification control which decreases false-positive and false-negative outcomes leading to incorrect diagnosis. Medical imaging is undergoing significant change through artificial intelligence as studies demonstrate the practical benefits of automatic Alzheimer's disease detection on a wide scale basis. Due to the lack of publicly available documentation the possibility of synthetic or unverified image generation cannot be ruled out, and this represents a significant limitation in terms of compliance and reproducibility. To ensure broader applicability and robustness, future work will involve validating the model on external datasets Deep learning-based models demonstrate clinically appropriate applications in patient workflows for early detection and personalized treatment development which leads to better neurodegenerative disease outcomes. Further, the proposed ensemble model can serve as an assistive tool for radiologists by providing automated classification of Alzheimer's disease stages from MRI scans. This can help flag early-stage or high-risk patients for further investigation. However, it should not replace expert interpretation. The model may produce false positives or false negatives, especially in very mild or atypical cases. Therefore, recommendation in its integration with standard clinical workflows, cognitive scoring systems, and physician review to ensure accurate diagnosis and decision-making.

8 Conclusion

Using MRI high-resolution scans, the research team developed an ensemble deep learning diagnostic system which performed with 99% accuracy in detecting Alzheimer's disease. The model utilized ResNet-50 to extract efficient features and EfficientNet-B3 to classify robustly while remaining effective against challenges in medical imaging applications. Precise model training and evaluation became possible through the reliable annotations and diverse high-quality image dataset which contained 33,984 images. Preprocessing methods performed through normalization, rescaling, and noise removal improved the model quality for enhanced robustness. The model demonstrated superior performance as shown through precision and recall scores together with F1-score and area under the ROC curve metrics during comprehensive evaluations across all stages of Alzheimer's disease. Our model achieved consistent training and validation accuracy improvements which converged at 99.32% with minimal overfitting observed in loss plots thus, proving its strong generalization potential. Analysis of the confusion matrix demonstrated that the model produced accurate results for both Mild and Moderate cases along with non-demented cases and achieved commendable accuracy when identifying Very Mild Demented cases. The research data shows that the ensemble model delivers strong diagnostic capabilities for Alzheimer's detection across severe disease manifestations. High-quality data alongside deep learning produces better diagnostic accuracy according to the research findings. Its performance quality makes the model suitable for clinical use because it provides essential medical decisions to doctors for early disease detection and ongoing care regulation. Further studies must evaluate both model optimization and implementation across multiple clinical settings as part of broader application validation.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

AK: Conceptualization, Funding acquisition, Formal analysis, Writing – review & editing, Writing – original draft, Investigation,

Methodology. FA: Funding acquisition, Resources, Writing – original draft, Project administration, Visualization, Methodology, Investigation, Validation, Conceptualization. AR: Writing – original draft, Formal analysis, Conceptualization, Project administration, Investigation, Methodology, Data curation. SB: Data curation, Methodology, Conceptualization, Writing – review & editing, Investigation, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The authors extend their appreciation to the King Salman Center for Disability Research for funding this work through Research Group no KSRG-2024-408.

Acknowledgments

The authors extend their appreciation to the King Salman Center for Disability Research for supporting this publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Zammataro L, Rovetta S, Greco D. MEDIGUI-ConvNet–Interactive Architecture Combining the Power of Convolutional Neural Networks and Medical Imaging. In *INI-DH 2024: Workshop on Innovative Interfaces in Digital Healthcare, in conjunction with International Conference on Advanced Visual Interfaces.* (2024). pp. 3–7.

2. Assmi A, Elhabyb K, Benba A, Jilbab A. Alzheimer's disease classification: a comprehensive study. *Multimed Tools Appl.* (2024) 83:1–24. doi: 10.1007/s11042-024-18306-9

3. Ajagbe SA, Amuda KA, Oladipupo MA, Oluwaseyi FA, Okesola KI. Multi-classification of Alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (DCNN) approaches. *Int J Adv Comput Res.* (2021) 11:51. doi: 10.19101/IJACR.2021.1152001

4. Khosravi M, Parsaei H, Rezaee K. Novel classification scheme for early Alzheimer's disease (AD) severity diagnosis using deep features of the hybrid cascade attention architecture: early detection of AD on MRI Scans. *Tsinghua Sci Technol.* (2024) 30:2572–91. doi: 10.26599/TST.2024.9010080

5. Li Q, Yang MQ. Comparison of machine learning approaches for enhancing Alzheimer's disease classification. *PeerJ.* (2021) 9:e10549. doi: 10.7717/peerj.10549

6. Shirbandi K, Khalafi M, Mirza-Aghazadeh-Attari M, Tahmasbi M, Shahvandi HK, Javanmardi P, et al. Accuracy of deep learning model-assisted amyloid positron emission tomography scan in predicting Alzheimer's disease: a systematic review and meta-analysis. *Inform Med Unlocked.* (2021) 25:100710. doi: 10.1016/j.imu.2021.100710

7. Mujahid M, Rehman A, Alam T, Alamri FS, Fati SM, Saba T. An efficient ensemble approach for Alzheimer's disease detection using an adaptive synthetic technique and deep learning. *Diagnostics.* (2023) 13:2489. doi: 10.3390/diagnostics13152489

8. Sorour SE, Abd El-Mageed AA, Albarrak KM, Alnaim AK, Wafa AA, El-Shafeiy E. Classification of Alzheimer's disease using MRI data based on Deep Learning Techniques. *J King Saud Univ Comput Inform Sci.* (2024) 36:101940. doi: 10.1016/j.jksuci.2024.101940

9. Tufail H, Ahad A, Naqvi MH, Maqsood R, Pires IM. Classification of vascular dementia on magnetic resonance imaging using deep learning architectures. *Intell Syst Appl.* (2024) 22:200388. doi: 10.1016/j.iswa.2024.200388

10. Goyal P, Rani R, Singh K. A multilayered framework for diagnosis and classification of Alzheimer's disease using transfer learned Alexnet and LSTM. *Neural Comput Appl.* (2024) 36:3777–801. doi: 10.1007/s00521-023-09301-6

11. Raza N, Naseer A, Tamoor M, Zafar K. Alzheimer disease classification through transfer learning approach. *Diagnostics.* (2023) 13:801. doi: 10.3390/diagnostics13040801

12. Sharma S, Gupta S, Gupta D, Altameem A, Saudagar AKJ, Poonia RC, et al. HTLML: Hybrid AI based model for detection of Alzheimer's disease. *Diagnostics.* (2022) 12:1833. doi: 10.3390/diagnostics12081833

13. Zhang X, Gao L, Wang Z, Yu Y, Zhang Y, Hong J. Improved neural network with multi-task learning for Alzheimer's disease classification. *Heliyon.* (2024) 10:e26405. doi: 10.1016/j.heliyon.2024.e26405

14. Solano-Rojas B, Villalón-Fonseca R. A low-cost three-dimensional DenseNet neural network for Alzheimer's disease early discovery. *Sensors.* (2021) 21:1302. doi: 10.3390/s21041302

15. Carcagnì P, Leo M, Del Coco M, Distante C, De Salve A. Convolution neural networks and self-attention learners for Alzheimer dementia diagnosis from brain MRI. *Sensors.* (2023) 23:1694. doi: 10.3390/s23031694

16. Jo T, Nho K, Bice P, Saykin AJ, Alzheimer's Disease Neuroimaging Initiative. Deep learning-based identification of genetic variants: application to Alzheimer's disease classification. *Brief Bioinform.* (2022) 23:bbac022. doi: 10.1093/bib/bbac022

17. Qiu S, Joshi PS, Miller MI, Xue C, Zhou X, Karjadi C, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain.* (2020) 143:1920–33. doi: 10.1093/brain/awaa137

18. Helaly HA, Badawy M, Haikal AY. Deep learning approach for early detection of Alzheimer's disease. *Cognit Comput.* (2022) 14:1711–27. doi: 10.1007/s12559-021-09946-2

19. Jo T, Nho K, Risacher SL, Saykin AJ. Alzheimer's Neuroimaging Initiative. Deep learning detection of informative features in tau PET for Alzheimer's disease classification. *BMC Bioinform.* (2020) 21:1–13. doi: 10.1186/s12859-020-03848-0

20. Awarayi NS, Twum F, Hayfron-Acquah JB, Owusu-Agyemang K. A bilateral filtering-based image enhancement for Alzheimer disease classification using CNN. *PLoS ONE.* (2024) 19:e0302358. doi: 10.1371/journal.pone.0302358

21. Hazarika RA, Kandar D, Maji AK. An experimental analysis of different deep learning based models for Alzheimer's disease classification using brain magnetic resonance images. *J King Saud Univ Comput Inform Sci.* (2022) 34:8576–98. doi: 10.1016/j.jksuci.2021.09.003

22. Kaggle.com. [Online]. Available online at: https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset/data (Accessed February 28, 2025).

23. Wong PC, Abdullah SS, Shapiai MI. Exceptional performance with minimal data using a generative adversarial network for Alzheimer's disease classification. *Sci Rep.* (2024) 14:17037. doi: 10.1038/s41598-024-66874-5

24. Angkoso CV, Agustin Tjahyaningtijas HP, Purnama I, Purnomo MH. Multiplane Convolutional Neural Network (Mp-CNN) for Alzheimer's Disease Classification. *Int J Intell Eng Syst.* (2022) 15:329–40. doi: 10.22266/ijies2022.0228.30

25. Wu Y, Zhou Y, Zeng W, Qian Q, Song M. An attention-based 3D CNN with multi-scale integration block for Alzheimer's disease classification. *IEEE J Biomed Health Inform.* (2022) 26:5665–73. doi: 10.1109/JBHI.2022.3197331

26. Al-Adhaileh MH. Diagnosis and classification of Alzheimer's disease by using a convolution neural network algorithm. *Soft Computing.* (2022) 26:7751–62. doi: 10.1007/s00500-022-06762-0

27. Khan R, Qaisar ZH, Mehmood A, Ali G, Alkhalifah T, Alturise F, et al. A practical multiclass classification network for the diagnosis of Alzheimer's disease. *Appl Sci.* (2022) 12:6507. doi: 10.3390/app12136507

28. Dao Q, El-Yacoubi MA, Rigaud AS. Detection of Alzheimer disease on online handwriting using 1D convolutional neural network. *IEEE Access.* (2022) 11:2148–55. doi: 10.1109/ACCESS.2022.3232396

29. Tripathi T, Kumar R. Speech-based detection of multi-class Alzheimer's disease classification using machine learning. *Int J Data Sci Analyt.* (2024) 18:83–96. doi: 10.1007/s41060-023-00475-9

30. Ujilast NA, Firdausita NS, Aditya CSK, Azhar Y, MRI. Image based Alzheimer's disease classification using convolutional neural network: EfficientNet architecture. *J RESTI.* (2024) 8:18–25. doi: 10.29207/resti.v8i1.5457

31. Pandey PK, Pruthi J, Alzahrani S, Verma A, Zohra B. Enhancing healthcare recommendation: transfer learning in deep convolutional neural networks for Alzheimer disease detection. *Front Med.* (2024) 11:1445325. doi: 10.3389/fmed.2024.1445325

32. Kaggle Link. Available online at: https://www.kaggle.com/datasets/borhanitrash/alzheimer-mri-disease-classification-dataset (Accessed February 28, 2025).

33. Zhang C, Ge H, Zhang S, Liu D, Jiang Z, Lan C et al. Hematoma evacuation via image-guided para-corticospinal tract approach in patients with spontaneous intracerebral hemorrhage. *Neurol Therapy.* (2021) 10:1001–13. doi: 10.1007/s40120-021-00279-8

34. Yin J, Qiao Z, Han L, Zhang X. EEG-based emotion recognition with autoencoder feature fusion and MSC-TimesNet model. *Comput Methods Biomech Biomed Eng.* (2025) 1–18. doi: 10.1080/10255842.2025.2477801

35. Tian J, Zhou Y, Chen X, AlQahtani SA, Chen H, Yang B et al. A novel self-supervised learning network for binocular disparity estimation. *CMES Compu Model Eng Sci.* (2024) 142:209–29. doi: 10.32604/cmes.2024.057032

36. Xiao X, Li Y, Wu Q, Liu X, Cao X, Li M, et al. Development and validation of a novel predictive model for dementia risk in middle-aged and elderly depression individuals: a large and longitudinal machine learning cohort study. *Alzheimers Res Ther.* (2025) 17:103. doi: 10.1186/s13195-025-01750-6

37. Zhu C. Computational intelligence-based classification system for the diagnosis of memory impairment in psychoactive substance users. *J Cloud Comput.* (2024) 13:119. doi: 10.1186/s13677-024-00675-z

38. Zhan X, Zhou Z, Liu Y, Cecchi NJ, Hajiahamemar M, Zeineh MM et al. Differences between two maximal principal strain rate calculation schemes in traumatic brain analysis with *in-vivo* and *in-silico* datasets. *J Biomech.* (2025) 179:112456. doi: 10.1016/j.jbiomech.2024.112456

39. Li N, Ou J, He H, He J, Zhang L, Peng Z et al. Exploration of a machine learning approach for diagnosing sarcopenia among Chinese community-dwelling older adults using sEMG-based data. *J Neuroeng Rehabil.* (2024) 21:69. doi: 10.1186/s12984-024-01369-y

40. Xiang Q, Xiang Y, Liu Y, Chen Y, He Q, Chen T, et al. Revealing the potential therapeutic mechanism of *Lonicerae japonicae* Flos in Alzheimer's disease: a computational biology approach. *Front Med.* (2024) 11:1468561. doi: 10.3389/fmed.2024.1468561

Frontiers in Medicine

Translating medical research and innovation into improved patient care

A multidisciplinary journal which advances our medical knowledge. It supports the translation of scientific advances into new therapies and diagnostic tools that will improve patient care.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact